Tests for separability in nonparametric covariance operators of random surfaces

J. A. D. Aston[*], D. Pigoli and S. Tavakoli[†]

Statistical Laboratory
Department of Pure Mathematics and Mathematical Statistics
University of Cambridge

Abstract

. The assumption of separability of the covariance operator for a random image or hypersurface can be of substantial use in applications, especially in situations where the accurate estimation of the full covariance structure is unfeasible, either for computational reasons, or due to a small sample size. However, inferential tools to verify this assumption are somewhat lacking in high-dimensional or functional data analysis settings, where this assumption is most relevant. We propose here to test separability by focusing on K-dimensional projections of the difference between the covariance operator and a nonparametric separable approximation. The subspace we project onto is one generated by the eigenfunctions of the covariance operator estimated under the separability hypothesis, negating the need to ever estimate the full non-separable covariance. We show that the rescaled difference of the sample covariance operator with its separable approximation is asymptotically Gaussian. As a by-product of this result, we derive asymptotically pivotal tests under Gaussian assumptions, and propose bootstrap methods for approximating the distribution of the test statistics. We probe the finite sample performance through simulations studies, and present an application to log-spectrogram images from a phonetic linguistics dataset.

Keywords: Acoustic Phonetic Data, Bootstrap, Dimensional Reduction, Functional Data, Partial Trace, Sparsity.

1 Introduction

Many applications involve hypersurface data, data that is both functional (as in functional data analysis, see e.g. Ramsay & Silverman 2005, Ferraty & Vieu 2006, Horváth & Kokoszka 2012*a*, Wang et al. 2015) and multidimensional. Examples abound and include images from medical devices such as MRI (Lindquist 2008) or PET (Worsley et al. 1996), spectrograms

[*]Research Supported by EPSRC grant EP/K021672/2.

[†]Address for correspondence: Shahin Tavakoli, Statistical laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WB, United Kingdom. Email: s.tavakoli@statslab.cam.ac.uk

derived from audio signals (Rabiner & Schafer 1978, and as in the application we consider in Section 4) or geolocalized data (see, e.g., Secchi et al. 2015). In these kinds of problem, the number of available observations (hypersurfaces) is often small relative to the high-dimensional nature of the individual observation, and not usually large enough to estimate a full multivariate covariance function.

It is usually, therefore, necessary to make some simplifying assumptions about the data or their covariance structure. If the covariance structure is of interest, such as for PCA or network modeling, for instance, it is usually assumed to have some kind of lower dimensional structure. Traditionally, this translates into a *sparsity* assumption: one assumes that most entries of the covariance matrix or function are zero. Though being relevant for a number of applications (Tibshirani 2014), this traditional definition of sparsity may not be appropriate in some cases, such as in imaging, as this can give rise to artefacts in the analysis (for example, holes in an image). In such problems, where the data is multidimensional, a natural assumption that can be made is that the covariance is *separable*. This assumption greatly simplifies both the estimation and the computational cost in dealing with multivariate covariance functions, while still allowing for a positive definite covariance to be specified. In the context of space-time data $X(s,t)$, for instance, where $s \in [-S,S]^d$, $S > 0$, denotes the location in space, and $t \in [0,T]$, $T > 0$, is the time index, the assumption of separability translates into

$$c(s,t,s',t') = c_1(s,s')c_2(t,t'), \quad s,s' \in [-S,S]^d; t,t' \in [0,T], \tag{1.1}$$

where c, c_1, and c_2, are respectively the full covariance function, the space covariance function and the time covariance function. In words, this means that the full covariance function factorises as a product of the spatial covariance function with the time covariance function.

The separability assumption (see e.g. Gneiting et al. 2007, Genton 2007) simplifies the covariance structure of the process and makes it far easier to estimate; in some sense, the separability assumption results in a estimator of the covariance which has less variance, at the expense of a possible bias. As an illustrative example, consider that we observe a discretized version of the process through measurements on a two dimensional grid (without loss of generality, as the same arguments apply for any dimension greater than 2) being a $q \times p$ matrix (of course, the functional data analysis approach taken here does *not* assume that the replications of the process are observed on same grid, nor that they are observed on a grid). Since we are not assuming a parametric form for the covariance, the degrees of freedom in the full covariance are $qp(qp+1)/2$, while the separability assumption reduces them to $q(q+1)/2 + p(p+1)/2$. This reflects a dramatic reduction in the dimension of the problem even for moderate value of q, p, and overcomes both computational and estimation problems due to the relatively small sample sizes available in applications. For example, for $q = p = 10$, we have $qp(qp+1)/2 = 5050$ degrees of freedom, however, if the separability holds, then we have only $q(q+1) + p(p+1) = 110$ degrees of freedom. Of course, this is only one example, and our approach is not restricted to data on a grid, but this illustrates the computational savings that such assumptions can possess.

Three related computational classes of problem can be identified. In the first case, the full covariance structure can be computed and stored. In the second one, it is still possible, although burdensome, to compute the full covariance matrix but it can not be stored, while the last class includes problems where even computation of the full covariance is infeasible. The values of q, p that set the boundaries for these classes depend of course on the available hardware and they are rapidly changing. At the present time however, for widely available systems, storage is feasible up to $q, p \approx 100$ while computation becomes unfeasible when

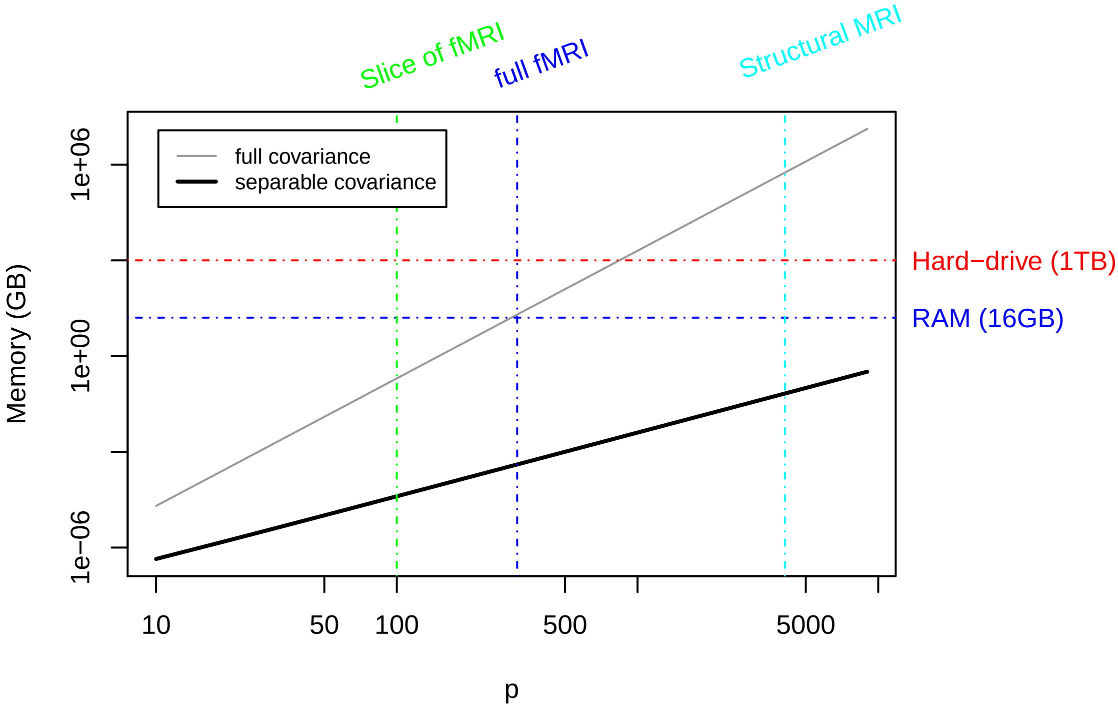


Figure 1: Memory required to store the full covariance and the separable covariance of $p \times p$ matrix data, as a function of p. Several types of data related to Neuroimaging (structural and functional Magnetic Resonance Imaging) are used as exemplars of data sizes, as they naturally have multidimensional structure.

q, p get close to 1000 (see Figure 1). On the contrary, a separable covariance structure can be usually both computed and stored without effort even for these sizes of problem. We would like to stress however that the constraints coming from the need for statistical accuracy are usually tighter. The estimation of the full covariance structure even for $q, p = 100$ presents about 5×10^7 unknown parameters, when typical sample sizes are in the order of hundreds at most. If we are able to assume separability, we can rely on far more accurate estimates.

While the separability assumption can be very useful, and is indeed often implicitly made in many higher dimensional applications when using isotropic smoothing (Worsley et al. 1996, Lindquist 2008), very little has been done to develop tools to assess its validity on a case by case basis. In the classical multivariate setup, some tests for the separability assumption are available. These have been mainly developed in the field of spatial statistics (see Lu & Zimmerman 2005, Fuentes 2006, and references therein), where the discussion of separable covariance functions is well-established, or for applications involving repeated measures (Mitchell et al. 2005). These methods, however, rely on the estimation of the full multidimensional covariance structure, which can be troublesome. It is sometimes possible to circumvent this problem by considering a parametric model for the full covariance structure (Simpson 2010, Simpson et al. 2014, Liu et al. 2014). On the contrary, when the covariance is being non-parametrically specified, as will be the case in this paper, estimation of the full covariance is at best computationally complex with large estimation errors, and in many cases simply computationally infeasible. Indeed, we highlight that, while the focus of this paper is on checking the viability of a separable structure for the covariance, this is done without any parametric assumption on the form of $c_1(s, s')$ and $c_2(t, t')$, thus allowing for the maximum flexibility. This is opposed to assuming a parametric separable form with only few unknown parameters, which is usually too restrictive in many applications, something that has led to separability being rightly criticised and viewed with suspicion in the spatio-temporal statistics literature (Gneiting 2002, Gneiting et al. 2007). Moreover, the methods we develop here are aimed to applications typical of functional data, where replicates from the underlying random process are available. This is different from the spatio-temporal setting, where usually only one realization of the process is observed. See also Constantinou et al. (2015) for another approach to test for separability in functional data.

It is important to notice that a separable covariance structure (or equivalently, a separable correlation structure) is not necessarily connected with the original data being separable. Furthermore, sums or differences of separable hypersurfaces are not necessarily separable. On the other hand, the error structure may be separable even if the mean is not. Given that in many applications of functional data analysis, the estimation of the covariance is the first step in the analysis, we concentrate on covariance separability. Indeed, covariance separability is an extremely useful assumption as it implies separability of the eigenfunctions, allowing computationally efficient estimation of the eigenfunctions (and principal components). Even if separability is misspecified, separable eigenfunctions can still form a basis representation for the data, they simply no longer carry optimal efficiency guarantees in this case (Aston & Kirch 2012), but can often have near-optimality under the appropriate assumptions (Chen et al. 2015) .

In this paper, we propose a test to verify if the data at hand are in agreement with a separability assumption. Our test does not require the estimation of the full covariance structure, but only the estimation of the separable structure (1.1), thus avoiding both the computational issues and the diminished accuracy involved in the former. To do this, we rely on a strategy from Functional Data Analysis (Ramsay & Silverman 2002, 2005, Ferraty & Vieu 2006, Ramsay et al. 2009, Horváth & Kokoszka 2012*b*), which consists in projecting the observations onto a carefully chosen low-dimensional subspace. The key fact for the success of our approach is that, under the null hypothesis, it is possible to determine this subspace using only the marginal covariance functions. While the optimal choice for the dimension of this subspace is a non-trivial problem, some insight can be obtained through our extensive simulation studies (Section 4.1). Ultimately, the proposed test checks the separability in the chosen subspace, which will often be the focus of following analyses.

The paper proceeds as follows. In Section 2, we examine the ideas behind separability, propose a separable approximation of a covariance operator, and study the asymptotics of the difference between the sample covariance operator and its separable approximation. This difference will be the building block of the testing procedures introduced in Section 3, and whose distribution we propose to approximate by bootstrap techniques. In Section 4, we investigate by means of simulation studies the finite sample behaviour of our testing procedures and apply our methods to acoustic phonetic data. A conclusion, given in Section 5, summarizes the main contributions of this paper. Proofs are collected in appendices A, B, and C, while implementation details, theoretical background and additional figures can be found in the appendices E, D and F. All the tests introduced in the paper are available as an R package `covsep` (Tavakoli 2016), available on the Comprehensive R Archive Network

(CRAN).

For notational simplicity, the proposed method will be described for two dimensional functional data (e.g. random surfaces), hence a four dimensional covariance structure (i.e. the covariance of a random surface), but the generalization to higher dimensional cases is straightforward. The methodology is developed in general for data that take values in a Hilbert space, but the case of square integrable surfaces—being relevant for the case of acoustic phonetic data—is used throughout the paper as a demonstration. We recall that the proposed approach is not restricted to data observed on a regular grid, although for simplicity of exposition we consider here the case where data are observed densely and a pre-processing smoothing step allows to consider the smooth surfaces as our observations, as happens for example the case of the acoustic phonetic data described in Section 4. If data are observed sparsely, the proposed approach can still be applied but there may be the need to use more appropriate estimators for the marginal covariance functions (see, e.g. Yao et al. n.d.) and these need to satisfy the properties described in Section 2.

2 Separable Covariances: definitions, estimators and asymptotic results

While the general idea of the factorization of a multi-dimensional covariance structure as the product of lower dimensional covariances is easy to describe, the development of a testing procedure asks for a rigorous mathematical definition and the introduction of some technical results. In this section we propose a definition of separability for covariance operators, show how it is possible to estimate a separable version of a covariance operator and evaluate the difference between the empirical covariance operator and its separable version. Moreover, we derive some asymptotic results for these estimators. To do this, we first set the problem in the framework of random elements in Hilbert spaces and their covariance operators. The benefit in doing this is twofold. First, our results become applicable in more general settings (e.g. multidimensional functional data, data on multidimensional grids, fixed size rectangular random matrices) and do not depend on a specific choice of smoothness of the data (which is implicitly assumed when modeling the data as e.g. square integrable surfaces). They only rely on the Hilbert space structure of the space in which the data lie. Second, it highlights the importance of the *partial trace* operator in the estimation of the separable covariance structure, and how the properties of the partial trace (Appendix C) play a crucial role in the asymptotic behavior of the proposed test statistics. However, to ease explanation, we use the case of the Hilbert space of square integrable surfaces (which shall be used in our linguistic application, see Section 4) as an illustration of our testing procedure.

2.1 Notation

Let us first introduce some definitions and notation about operators in a Hilbert space (see e.g. Gohberg et al. 1990, Kadison & Ringrose 1997*a*, Ringrose 1971). Let H be a real separable Hilbert space (that is, a Hilbert space with a countable orthonormal basis), whose inner product and norm are denoted by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$, respectively. The space of bounded (linear) operators on H is denoted by $\mathcal{S}_\infty(H)$, and its norm is $|||T|||_\infty = \sup_{x \neq 0} \|Tx\|/\|x\|$. The space of Hilbert–Schmidt operators on H is denoted by $\mathcal{S}_2(H)$, and is a Hilbert space with the inner-product $\langle S, T \rangle_{\mathcal{S}_2} = \sum_{i \geq 1} \langle Se_i, Te_i \rangle$ and induced norm $|||\cdot|||_2$, where $(e_i)_{i \geq 1} \subset H$ is an orthonormal basis of H. The space of trace-class operator on H is denoted by $\mathcal{S}_1(H)$, and consists of all compact operators T with finite trace-norm, i.e. $|||T|||_1 = \sum_{n\geq 1} s_n(T) < \infty$, where $s_n(T) \geq 0$ denotes the n-th singular value of T. For any trace-class operator $T \in \mathcal{S}_1(H)$, we define its trace by $\mathrm{Tr}(T) = \sum_{i\geq 1} \langle Te_i, e_i \rangle$, where $(e_i)_{i\geq 1} \subset H$ is an orthonormal basis, and the sum is independent of the choice of the orthonormal basis.

If H_1, H_2 are real separable Hilbert spaces, we denote by $H = H_1 \otimes H_2$ their tensor product Hilbert space, which is obtained by the completion of all finite sums $\sum_{i,j=1}^{N} u_i \otimes v_j$, $u_i \in H_1, v_j \in H_2$, under the inner-product $\langle u \otimes v, z \otimes w \rangle = \langle u, z \rangle \langle v, w \rangle$, $u, z \in H_1, z, w \in H_2$ (see e.g. Kadison & Ringrose 1997*a*). If $C_1 \in \mathcal{S}_\infty(H_1)$, $C_2 \in \mathcal{S}_\infty(H_2)$, we denote by $C_1 \tilde{\otimes} C_2$ the unique linear operator on $H_1 \otimes H_2$ satisfying

$$\left(C_1 \tilde{\otimes} C_2\right)(u \otimes v) = C_1 u \otimes C_2 v, \quad \text{for all } u \in H_1, v \in H_2. \tag{2.1}$$

It is a bounded operator on H, with $\left|\left|\left|C_1 \tilde{\otimes} C_2\right|\right|\right|_\infty = |||C_1|||_\infty |||C_2|||_\infty$. Furthermore, if $C_1 \in \mathcal{S}_1(H_1)$ and $C_2 \in \mathcal{S}_1(H_2)$, then $C_1 \tilde{\otimes} C_2 \in \mathcal{S}_1(H_1 \otimes H_2)$ and $\left|\left|\left|C_1 \tilde{\otimes} C_2\right|\right|\right|_1 = |||C_1|||_1 |||C_2|||_1$. We denote by $\mathrm{Tr}_1 : \mathcal{S}_1(H_1 \otimes H_2) \to \mathcal{S}_1(H_2)$ the *partial trace with respect to* H_1. It is the unique bounded linear operator satisfying $\mathrm{Tr}_1(A \tilde{\otimes} B) = \mathrm{Tr}(A)B$, for all $A \in \mathcal{S}_1(H_1), B \in \mathcal{S}_1(H_2)$. $\mathrm{Tr}_2 : \mathcal{S}_1(H_1 \otimes H_2) \to \mathcal{S}_1(H_1)$ is defined symmetrically (see Appendix C for more details).

If $X \in H$ is a random element with $\mathbb{E}\,\|X\| < \infty$, then $\mu = \mathbb{E}\,X \in H$, the mean of X, is well defined. Furthermore, if $\mathbb{E}\,\|X\|^2 < \infty$, then $C = \mathbb{E}\,[(X - \mu) \otimes_2 (X - \mu)]$ defines the *covariance operator* of X, where $f \otimes_2 g$ is the operator on H defined by $(f \otimes_2 g)h = \langle h, g \rangle f$, for $f, g, h \in H$. The covariance operator C is a trace-class hermitian operator on H, and encodes all the second-order fluctuations of X around its mean.

Using this nomenclature, we are going to deal with random variables belonging to a tensor product Hilbert space. This framework encompasses the situation where X is a random surface, for example a space-time indexed data, i.e. $X = X(s,t)$, $s \in [-S,S]^d, t \in [0,T]$, $S, T > 0$, by setting $H = L^2\left([-S,S]^d \times [0,T], \mathbb{R}\right)$, for instance (notice however that additional smoothness assumptions on X would lead to assume that X belongs to some other Hilbert space). In this case, the covariance operator of the random element $X \in L^2\left([-S,S]^d \times [0,T], \mathbb{R}\right)$ satisfies

$$Cf(s,t) = \int_{[-S,S]^d} \int_0^T c(s,t,s',t') f(s',t') ds' dt', \quad s \in [-S,S]^d, t \in [0,T],$$

$f \in L^2\left([-S,S]^d \times [0,T], \mathbb{R}\right)$, where $c(s,t,s',t') = \mathrm{cov}\,[X(s,t), X(s',t')]$ is the *covariance function* of X. The space of square integrable surfaces,

$$L^2\left([-S,S]^d \times [0,T], \mathbb{R}\right),$$

is a tensor product Hilbert space because it can can be identified with

$$L^2\left([-S,S]^d, \mathbb{R}\right) \otimes L^2\left([0,T], \mathbb{R}\right).$$

2.2 Separability

We recall now that we want to define separability so that the covariance function can be written as $c(s,t,s',t') = c_1(s,s')c_2(t,t')$, for some $c_1 \in L^2\left([-S,S]^d \times [-S,S]^d, \mathbb{R}\right)$ and $c_2 \in L^2\left([0,T] \times [0,T], \mathbb{R}\right)$. This can be extended to the covariance operator of a random elements $X \in H = H_1 \otimes H_2$, where H_1, H_2 are arbitrary separable real Hilbert spaces. We call its covariance operator C *separable* if

$$C = C_1 \tilde{\otimes} C_2, \tag{2.2}$$

where C_1, respectively C_2, are trace-class operators on H_1, respectively on H_2, and $C_1 \mathbin{\tilde{\otimes}} C_2$ is defined in (2.1). Notice that though the decomposition (2.2) is not unique, since $C_1 \mathbin{\tilde{\otimes}} C_2 = (\alpha C_1) \mathbin{\tilde{\otimes}} (\alpha^{-1} C_2)$ for any $\alpha \neq 0$, this will not cause any problem at a later stage since we will ultimately be dealing with the product $C_1 \mathbin{\tilde{\otimes}} C_2$, which is identifiable.

In practice, neither C nor $C_1 \mathbin{\tilde{\otimes}} C_2$ are known. If $X_1, \ldots, X_N \overset{\text{i.i.d.}}{\sim} X$ and (2.2) holds, the sample covariance operator $\widehat{C}_N$ is not necessarily separable in finite samples. However, we can estimate a separable approximation of it by

$$\widehat{C}_{1,N} \mathbin{\tilde{\otimes}} \widehat{C}_{2,N}, \tag{2.3}$$

where $\widehat{C}_{1,N} = \mathrm{Tr}_2(\widehat{C}_N)/\sqrt{\mathrm{Tr}(\widehat{C}_N)}$, $\widehat{C}_{2,N} = \mathrm{Tr}_1(\widehat{C}_N)/\sqrt{\mathrm{Tr}(\widehat{C}_N)}$. The intuition behind (2.3) is that

$$\mathrm{Tr}(T)T = \mathrm{Tr}_2(T) \mathbin{\tilde{\otimes}} \mathrm{Tr}_1(T),$$

for all $T \in \mathcal{S}_1(H_1 \otimes H_2)$ of the form $T = A \mathbin{\tilde{\otimes}} B$, $A \in \mathcal{S}_1(H_1), B \in \mathcal{S}_1(H_2)$, with $\mathrm{Tr}(T) \neq 0$.

Let us consider again what this means when X is a random element of $L^2\left([-S,S]^d \times [0,T], \mathbb{R}\right)$—i.e. the realization of a space-time process—of which we observe N i.i.d. replications $X_1, \ldots, X_N \sim X$. In this case, Proposition C.2 tells us that if the covariance function is continuous, the operators $\widehat{C}_{1,N}$ and $\widehat{C}_{2,N}$ are defined by

$$\begin{aligned}
\widehat{C}_{1,N} f(s) &= \int_{[-S,S]^d} \widehat{c}_{1,N}(s,s') f(s) ds, \quad f \in L^2\left([-S,S]^d, \mathbb{R}\right), \\
\widehat{C}_{2,N} g(t) &= \int_0^T \widehat{c}_{2,N}(t,t') g(t) dt, \quad g \in L^2\left([0,T], \mathbb{R}\right),
\end{aligned}$$

where

$$\widehat{c}_{1,N}(s,s') = \frac{\tilde{c}_{1,N}(s,s')}{\sqrt{\int_{[-S,S]^d} \tilde{c}_{1,N}(s,s) ds}}, \qquad \widehat{c}_{2,N}(t,t') = \frac{\tilde{c}_{2,N}(t,t')}{\sqrt{\int_0^T \tilde{c}_{2,N}(t,t) dt}},$$

and

$$\begin{aligned}
\tilde{c}_{1,N}(s,s') &= \frac{1}{N} \sum_{i=1}^N \int_0^T \left(X_i(s,t) - \overline{X}(s,t)\right)\left(X_i(s',t) - \overline{X}(s',t)\right) dt = \int_0^T c_N(s,t,s',t) dt, \\
\tilde{c}_{2,N}(t,t') &= \frac{1}{N} \sum_{i=1}^N \int_{[-S,S]^d} \left(X_i(s,t) - \overline{X}(s,t)\right)\left(X_i(s,t') - \overline{X}(s,t')\right) ds = \int_{[-S,S]^d} c_N(s,t,s,t') ds, \\
\overline{X}(s,t) &= \frac{1}{N} \sum_{i=1}^N X_i(s,t), \widehat{c}_N(s,t,s',t') = \frac{1}{N} \sum_{i=1}^N \left(X_i(s,t) - \overline{X}(s,t)\right)\left(X_i(s',t') - \overline{X}(s',t')\right),
\end{aligned}$$

for all $s, s' \in [-S,S]^d, t, t' \in [0,T]$. The assumption of separability here means that the estimated covariance is written as a product of a purely spatial component and a purely temporal component, thus making both modeling and estimation easier in many practical applications.

We stress again that we aim to develop a test statistic that solely relies on the estimation of the separable components C_1 and C_2, and does not require the estimation of the full covariance C. We can expect that under the null hypothesis $H_0 : C = C_1 \mathbin{\tilde{\otimes}} C_2$, the difference $D_N = \widehat{C}_N - \widehat{C}_{1,N} \mathbin{\tilde{\otimes}} \widehat{C}_{2,N}$ between the sample covariance operator and its separable approximation should take small values. We propose therefore to construct our test statistic by projecting D_N onto the first eigenfunctions of C, since these encode the directions along which X has the most variability. If we denote by $C_1 = \sum_{i\geq 1} \lambda_i u_i \otimes_2 u_i$ and $C_2 = \sum_{j\geq 1} \gamma_j v_j \otimes_2 v_j$ the Mercer decompositions of C_1 and C_2, we have

$$C = C_1 \;\widetilde{\otimes}\; C_2 = \sum_{i,j\geq 1} \lambda_i \gamma_j (u_i \otimes v_j) \bigotimes\nolimits_2 (u_i \otimes v_j),$$

where we have used results from Appendix D.1. The eigenfunctions of C are therefore of the form $u_r \otimes v_s$, where $u_r \in H_1$ is the r-th eigenfunction of C_1 and $v_s \in H_2$ is the s-th eigenfunction of C_2. We define a test statistic based on the projection

$$T_N(r,s) = \sqrt{N} \left\langle D_N(\hat{u}_r \otimes \hat{v}_s), \hat{u}_r \otimes \hat{v}_s \right\rangle, \quad r,s \geq 1 \text{ fixed}, \tag{2.4}$$

where we have replaced the eigenfunctions of C_1 and C_2 by their empirical counterpart, i.e. the Mercer decompositions of $\widehat{C}_{1,N}$, respectively $\widehat{C}_{2,N}$, are given by $\widehat{C}_{1,N} = \sum_{i\geq 1} \hat{\lambda}_i \hat{u}_i \otimes \hat{u}_i$, respectively $\widehat{C}_{2,N} = \sum_{j\geq 1} \hat{\gamma}_j \hat{v}_j \otimes \hat{v}_j$. Notice that though the eigenfunctions of $\widehat{C}_{1,N}$ and $\widehat{C}_{2,N}$ are defined up to a multiplicative constant $\alpha = \pm 1$, our test statistic is well defined. The key fact for the practical implementation of the method is that $T_N(r,s)$ can be computed without the need to estimate (and store in memory) the operator D_N, since $T_N(r,s) = \sqrt{N}\left(\frac{1}{N}\sum_{k=1}^{N} \left\langle X_k - \overline{X}_N, \widehat{v}_i \otimes \widehat{u}_j \right\rangle^2 - \hat{\lambda}_r \hat{\gamma}_s\right)$. In particular, the computation of $T_N(r,s)$ does *not* require an estimation of the full covariance operator C, but only the estimation of the marginal covariance operators C_1 and C_2, and their eigenstructure.

2.3 Asymptotics

The theoretical justification for using a projection of D_N to define a test procedure is that, under the null hypothesis $H_0 : C = C_1 \;\widetilde{\otimes}\; C_2$, we have $|||D_N|||_1 \xrightarrow{p} 0$ as $N \to \infty$, i.e. D_N convergences in probability to zero with respect to the trace norm. In fact, we will show in Theorem 2.3 that $\sqrt{N} D_N$ is asymptotically Gaussian under the following regularity conditions:

Condition 2.1. *X is a random element of the real Hilbert space H satisfying*

$$\sum_{j=1}^{\infty} \left(\mathbb{E}\left[\langle X, e_j\rangle^4\right]\right)^{1/4} < \infty, \tag{2.5}$$

for some orthonormal basis $(e_j)_{j\geq 1}$ of H.

The implications of this condition can be better understood in light of the following remark.

remark 2.2 (Mas (2006))**.**

1. *Condition 2.1 implies that $\mathbb{E}\|X\|^4 < \infty$.*
2. *If $\mathbb{E}\|X\|^4 < \infty$, then $\sqrt{N}(C_N - C)$ converges in distribution to a Gaussian random element of $\mathcal{S}_2(H)$ for $N \to \infty$, with respect to the Hilbert–Schmidt topology. Under Condition 2.1, a stronger form of convergence holds: $\sqrt{N}(C_N - C)$ converges in distribution to a random element of $\mathcal{S}_1(H)$ for $N \to \infty$, with respect to the trace-norm topology.*

3. *If X is Gaussian and $(\lambda_j)_{j\geq 1}$ is the sequence of eigenvalues of its covariance operator, a sufficient condition for* (2.5) *is* $\sum_{j\geq 1}\sqrt{\lambda_j} < \infty$.

Condition 2.1 requires fourth order moments rather than the usual second order moments often assumed in functional data, as in this case we are interested in investigating the variation of the second moment, and hence require assumptions on the fourth order structure. Recall that $\widehat{C}_N = \frac{1}{N}\sum_{j=1}^{N}(X_i - \overline{X})\otimes_2(X_i-\overline{X})$, where $\overline{X} = N^{-1}\sum_{k=1}^{N} X_k$. The following result establishes the asymptotic distribution of $D_N = \widehat{C}_N - \frac{\mathrm{Tr}_2(\widehat{C}_N)\ \tilde{\otimes}\ \mathrm{Tr}_1(\widehat{C}_N)}{\mathrm{Tr}(\widehat{C}_N)}$:

theorem 2.3. *Let H_1, H_2 be separable real Hilbert spaces, $X_1,\ldots,X_N \sim X$ be i.i.d. random elements on $H_1\otimes H_2$ with covariance operator C, and $\mathrm{Tr}\, C \neq 0$.*

If X satisfies Condition 2.1 (with $H = H_1\otimes H_2$), then, under the null hypothesis

$$H_0 : C = C_1\ \tilde{\otimes}\ C_2, \qquad C_1\in\mathcal{S}_1(H_1), C_2\in\mathcal{S}_1(H_2),$$

we have

$$\sqrt{N}\left(\widehat{C}_N - \frac{\mathrm{Tr}_2(\widehat{C}_N)\ \tilde{\otimes}\ \mathrm{Tr}_1(\widehat{C}_N)}{\mathrm{Tr}(\widehat{C}_N)}\right) \xrightarrow{d} Z, \quad \text{as } N\to\infty, \tag{2.6}$$

where Z is a Gaussian random element of $\mathcal{S}_1(H_1\otimes H_2)$ with mean zero, whose covariance structure is given in Lemma A.1.

Condition 2.1 is used here because we need $\sqrt{N}(\widehat{C}_N - C)$ to converge in distribution in the topology of the space $\mathcal{S}_1(H_1\otimes H_2)$; it could be replaced by any (weaker) condition ensuring such convergence. The assumption $\mathrm{Tr}\, C\neq 0$ is equivalent to assuming that X is not almost surely constant.

Proof of Theorem 2.3. First, notice that $C = C_1\ \tilde{\otimes}\ C_2 = \frac{\mathrm{Tr}_2(C)\ \tilde{\otimes}\ \mathrm{Tr}_1(C)}{\mathrm{Tr}(C)}$ under H_0. Therefore, using the linearity of the partial trace, we get

$$\begin{aligned}
\sqrt{N}\left(\widehat{C}_N - \frac{\mathrm{Tr}_2(\widehat{C}_N)\ \tilde{\otimes}\ \mathrm{Tr}_1(\widehat{C}_N)}{\mathrm{Tr}(\widehat{C}_N)}\right) &= \sqrt{N}(\widehat{C}_N - C) \\
&\quad + \sqrt{N}\left(\frac{\mathrm{Tr}_2(C)\ \tilde{\otimes}\ \mathrm{Tr}_1(C)}{\mathrm{Tr}(C)} + \frac{\mathrm{Tr}_2(\widehat{C}_N)\ \tilde{\otimes}\ \mathrm{Tr}_1(\widehat{C}_N)}{\mathrm{Tr}(\widehat{C}_N)}\right) \\
&= \sqrt{N}(\widehat{C}_N - C) + \frac{\mathrm{Tr}\left(\sqrt{N}(\widehat{C}_N - C)\right)C}{\mathrm{Tr}(\widehat{C}_N)} \\
&\quad - \frac{\mathrm{Tr}_2\left(\sqrt{N}(\widehat{C}_N - C)\right)\ \tilde{\otimes}\ \mathrm{Tr}_1(C)}{\mathrm{Tr}(\widehat{C}_N)} \\
&\quad - \frac{\mathrm{Tr}_2(\widehat{C}_N)\ \tilde{\otimes}\ \mathrm{Tr}_1\left(\sqrt{N}(\widehat{C}_N - C)\right)}{\mathrm{Tr}(\widehat{C}_N)}. \\
&= \Psi\left(\sqrt{N}(\widehat{C}_N - C), \widehat{C}_N\right),
\end{aligned}$$

where

$$\Psi(T,S) = T + \frac{\mathrm{Tr}(T)C}{\mathrm{Tr}(S)} - \frac{\mathrm{Tr}_2(T)\ \tilde{\otimes}\ \mathrm{Tr}_1(C)}{\mathrm{Tr}(S)} - \frac{\mathrm{Tr}_2(S)\ \tilde{\otimes}\ \mathrm{Tr}_1(T)}{\mathrm{Tr}(S)}; \quad T, S\in\mathcal{S}_1(H_1\otimes H_2).$$

Notice that the function $\Psi : \mathcal{S}_1(H_1 \otimes H_2) \times \mathcal{S}_1(H_1 \otimes H_2) \to \mathcal{S}_1(H_1 \otimes H_2)$ is continuous at $(T, S) \in \mathcal{S}_1(H_1 \otimes H_2) \times \mathcal{S}_1(H_1 \otimes H_2)$ in each coordinate, with respect to the trace norm, provided $\mathrm{Tr}(S) \neq 0$. Since $\sqrt{N}(\widehat{C}_N - C)$ converges in distribution—under Condition 2.1—to a Gaussian random element $Y \in \mathcal{S}_1(H_1 \otimes H_2)$, with respect to the trace norm $|||\cdot|||_1$ (see Mas 2006, Proposition 5), $\Psi\left(\sqrt{N}(\widehat{C}_N - C), \widehat{C}_N\right)$ converges in distribution to

$$\Psi(Y, C) = Y + \frac{\mathrm{Tr}(Y)C}{\mathrm{Tr}(C)} - \frac{\mathrm{Tr}_2(Y) \,\widetilde{\otimes}\, \mathrm{Tr}_1(C)}{\mathrm{Tr}(C)} - \frac{\mathrm{Tr}_2(C) \,\widetilde{\otimes}\, \mathrm{Tr}_1(Y)}{\mathrm{Tr}(C)} \tag{2.7}$$

by the continuous mapping theorem in metric spaces (Billingsley 1999). $\Psi(Y, C)$ is Gaussian because each of the summands of (2.7) are Gaussian. Indeed, the first and second summands are obviously Gaussian, and the last two summands are Gaussian by Proposition C.3, and Proposition D.2. □

We can now give the asymptotic distribution of $T_N(r, s)$, defined in (2.4) as the (scaled) projection of D_N in a direction given by the tensor product of the empirical eigenfunctions $\hat{u}_r$ and $\hat{v}_s$. The proof of the following result is given in Appendix B.

corollary 2.4. *Under the conditions of Theorem 2.3, if $\mathcal{I} \subset \{(i, j) : i, j \geq 1\}$ is a* finite *set of indices such that $\lambda_r \gamma_s > 0$ for each $(r, s) \in \mathcal{I}$, then*

$$(T_N(r, s))_{(r,s)\in\mathcal{I}} \xrightarrow{d} N(0, \Sigma), \quad \textit{as } N \to \infty.$$

This means that the vector $(T_N(r, s))_{(r,s)\in\mathcal{I}}$ is asymptotically multivariate Gaussian, with asymptotic variance-covariance matrix $\Sigma = \left(\Sigma_{(r,s),(r',s')}\right)_{(r,s),(r',s')\in\mathcal{I}}$ is given by

$$\begin{aligned}
\Sigma_{(r,s),(r',s')} = \tilde{\beta}_{rsr's'} &+ \frac{\alpha_{rs}\tilde{\beta}_{r's'\cdot\cdot} + \alpha_{r's}\tilde{\beta}_{r\cdot\cdot s'} + \alpha_{rs'}\tilde{\beta}_{r'\cdot\cdot s} + \alpha_{r's'}\tilde{\beta}_{rs\cdot\cdot}}{\mathrm{Tr}(C)} \\
&+ \frac{\alpha_{rs}\alpha_{r's'}\tilde{\beta}_{\cdot\cdot\cdot\cdot}}{\mathrm{Tr}(C)^2} + \frac{\lambda_r\lambda_{r'}\tilde{\beta}_{\cdot s\cdot s'}}{\mathrm{Tr}(C_1)^2} + \frac{\gamma_s\gamma_{s'}\tilde{\beta}_{r\cdot r'\cdot}}{\mathrm{Tr}(C_2)^2} \\
&- \frac{\lambda_r\tilde{\beta}_{r's'\cdot s} + \lambda_{r'}\tilde{\beta}_{rs\cdot s'}}{\mathrm{Tr}(C_1)} - \frac{\gamma_s\tilde{\beta}_{r's'r\cdot} + \gamma_{s'}\tilde{\beta}_{rsr'\cdot}}{\mathrm{Tr}(C_2)} \\
&- \frac{\alpha_{rs}}{\mathrm{Tr}(C)}\left(\frac{\gamma_{s'}\tilde{\beta}_{r'\cdot\cdot\cdot}}{\mathrm{Tr}(C_2)} + \frac{\lambda_{r'}\tilde{\beta}_{\cdot s'\cdot\cdot}}{\mathrm{Tr}(C_1)}\right) \\
&- \frac{\alpha_{r's'}}{\mathrm{Tr}(C)}\left(\frac{\gamma_s\tilde{\beta}_{r\cdot\cdot\cdot}}{\mathrm{Tr}(C_2)} + \frac{\lambda_r\tilde{\beta}_{\cdot s\cdot\cdot}}{\mathrm{Tr}(C_1)}\right)
\end{aligned}$$

where $\mu = \mathbb{E}[X]$, $\alpha_{rs} = \lambda_r\gamma_s$,

$$\tilde{\beta}_{ijkl} = \mathbb{E}\left[\langle X - \mu, u_i \otimes v_j\rangle^2 \langle X - \mu, u_k \otimes v_l\rangle^2\right],$$

and '$\cdot$' denotes summation over the corresponding index, i.e. $\tilde{\beta}_{r\cdot jk} = \sum_{i\geq 1}\tilde{\beta}_{rijk}$.

We note that the asymptotic variance-covariance of $(T_N(r, s))_{(r,s)\in\mathcal{I}}$ depends on the second and fourth order moments of X, which is not surprising since it is based on estimators of the covariance of X. Under the additional assumption that X is Gaussian, the asymptotic variance-covariance of $(T_N(r, s))_{(r,s)\in\mathcal{I}}$ can be entirely expressed in terms of the covariance operator C. The proof of the following result is given in Appendix B.

corollary 2.5. *Assume the conditions of Theorem 2.3 hold, and that X is Gaussian. If $\mathcal{I} \subset \{(i,j) : i,j \geq 1\}$ is a* finite *set of indices such that $\lambda_r \gamma_s > 0$ for each $(r,s) \in \mathcal{I}$, then*

$$(T_N(r,s))_{(r,s)\in\mathcal{I}} \xrightarrow{d} N(0,\Sigma), \quad \textit{as } N \to \infty.$$

where

$$\Sigma_{(r,s),(r',s')} = \frac{2\lambda_r \lambda_{r'} \gamma_s \gamma_{s'}}{\mathrm{Tr}(C)^2} \Big(\delta_{rr'} \mathrm{Tr}(C_1)^2 + |||C_1|||_2^2 - (\lambda_r + \lambda_{r'}) \mathrm{Tr}(C_1) \Big) \times \Big(\delta_{ss'} \mathrm{Tr}(C_2)^2 + |||C_2|||_2^2 - (\gamma_s + \gamma_{s'}) \mathrm{Tr}(C_2) \Big),$$

and $\delta_{ij} = 1$ if $i = j$, and zero otherwise. In particular, notice that Σ itself is separable.

It will be seen in the next section that even in the case where we use a bootstrap test, knowledge of the asymptotic distribution can be very useful to establish a pivotal bootstrap test, which will be seen to have very good performance in simulation.

3 Separability Tests and Bootstrap Approximations

In this section we use the estimation procedures and the theoretical results presented in Section 2 to develop a test for $H_0 : C = C_1 \widetilde{\otimes} C_2$, against the alternative that C cannot be written as a tensor product.

First, it is straightforward to define a testing procedure when X is Gaussian. Indeed, if we let

$$G_N(r,s) = T_N^2(r,s) = N \left(\frac{1}{N} \sum_{k=1}^{N} \left\langle X_k - \overline{X}, \widehat{u}_r \otimes \widehat{v}_s \right\rangle^2 - \hat{\lambda}_r \hat{\gamma}_s \right)^2, \tag{3.1}$$

and

$$\hat{\sigma}^2(r,s) = \Big(\mathrm{Tr}(\widehat{C}_{1,N})^2 \, \mathrm{Tr}(\widehat{C}_{2,N})^2 \Big)^{-1} 2\hat{\lambda}_r^2 \hat{\gamma}_s^2 \times \left(\mathrm{Tr}(\widehat{C}_{1,N})^2 + \left|\left|\left| \widehat{C}_{1,N} \right|\right|\right|_2^2 - 2\hat{\lambda}_r \mathrm{Tr}(\widehat{C}_{1,N}) \right) \times \left(\mathrm{Tr}(\widehat{C}_{2,N})^2 + \left|\left|\left| \widehat{C}_{2,N} \right|\right|\right|_2^2 - 2\hat{\gamma}_s \mathrm{Tr}(\widehat{C}_{2,N}) \right), \tag{3.2}$$

then $\hat{\sigma}^{-2}(r,s) G_N(r,s)$ is asymptotically χ_1^2 distributed, and $\{G_N^2(r,s) > \hat{\sigma}^2(r,s)\chi_1^2(1-\alpha)\}$, where $\chi_1^2(1-\alpha)$ is the $1-\alpha$ quantile of the χ_1^2 distribution, would be a rejection region of level approximately α, for $\alpha \in [0,1]$ and N large.

Apart for the distributional assumption for X to be Gaussian, this approach suffers also the important limitation that it only tests the separability assumption along *one* eigendirection. It is possible to extend this approach to take into account several eigendirections. For simplicity, let us consider the case $\mathcal{I} = \{1,\ldots,p\} \times \{1,\ldots,q\}$. Denote by $\mathbf{T}_N(\mathcal{I})$ the $p \times q$ matrix with entries $(\mathbf{T}_N(\mathcal{I}))_{ij} = T_N(i,j)$, and let

$$\widetilde{G}_N(\mathcal{I}) = \left| \hat{\Sigma}_{L,\mathcal{I}}^{-1/2} \mathbf{T}_N(\mathcal{I}) \hat{\Sigma}_{R,\mathcal{I}}^{-\mathsf{T}/2} \right|^2, \tag{3.3}$$

where $|A|^2$ denotes the sum of squared entries of a matrix A, $A^{-1/2}$ denotes the inverse of (any) square root of the matrix A, $A^{-\mathsf{T}/2} = (A^{-1/2})^{\mathsf{T}}$, and the matrices $\hat{\Sigma}_{L,\mathcal{I}}$, respectively

$\hat{\Sigma}_{R,\mathcal{I}}$, which are estimators of the row, resp. column, asymptotic covariances of $\mathbf{T}_N(\mathcal{I})$, are defined in Appendix E. Then $\widetilde{G}_N(\mathcal{I})$ is asymptotically χ^2_{pq} distributed. In the simulation studies (Section 4.1), we consider also an approximate version of this Studentized test statistics, $\widetilde{G}^a_N(\mathcal{I}) = \sum_{(r,s)\in\mathcal{I}} T^2_N(r,s)/\hat{\sigma}^2(r,s)$, which are obtained simply by standardizing marginally each entry $T^2_N(r,s)$, thus ignoring the dependence between the test statistics associated with different directions. In order to assess the advantage of Studentization, we also consider the non-Studentized test statistic

$$G_N(\mathcal{I}) = \sum_{(r,s)\in\mathcal{I}} T^2_N(r,s).$$

The computation details for $\widetilde{G}_N$, T_N, $\hat{\sigma}^2(r,s)$, $\hat{\Sigma}_{L,\mathcal{I}}$ and $\hat{\Sigma}_{R,\mathcal{I}}$ are described in Appendix E.

remark 3.1. *Notice that the only test whose asymptotic distribution is parameter free is $\widetilde{G}_N(\mathcal{I})$, under Gaussian assumptions. It would in principle be possible to construct an analogous test without the Gaussian assumptions (using Corollary 2.4). However, due to the large number of parameters that would need to be estimated in this case, we expect the asymptotics to come into force only for very large sample sizes (this is actually the case under Gaussian assumptions, specially if the set of projections $\mathcal{I}$ is large, as can be seen in Figure 10). For these reasons, we shall investigate bootstrap approximations to the test statistics.*

The choice of the number of eigenfunctions K (the number of elements in $\mathcal{I}$) onto which one should project is not trivial. The popular choice of including enough eigenfunctions to explain a fixed percentage of the variability in the dataset may seem inappropriate in this context, because under the alternative hypothesis there is no guarantee that the separable eigenfunctions explain that percentage of variation.

For fixed K, notice that the test at least guarantees the separability in the subspace of the respective K eigenfunctions, which is where the following analysis will be often focused. On the other hand, since our test statistic looks at an estimator of the non-separable component

$$D = C - \frac{\mathrm{Tr}_2(C)\ \widetilde{\otimes}\ \mathrm{Tr}_1(C)}{\mathrm{Tr}(C)},$$

restricted to the subspace spanned by the eigenfunctions $u_r \otimes v_s$, the test takes small values (and thus lacks power) when

$$\langle D(u_r \otimes v_s), u_r \otimes v_s\rangle = \left\langle D, (u_r \otimes_2 u_r)\ \widetilde{\otimes}\ (v_s \otimes_2 v_s)\right\rangle_{\mathcal{S}_2} = 0,$$

that is when the non-separable component D is orthogonal to

$$(u_r \otimes_2 u_r)\ \widetilde{\otimes}\ (v_s \otimes_2 v_s)$$

with respect to the Hilbert–Schmidt inner product. Thus the proposed test statistic $G_N(\mathcal{I})$ is powerful when D is not orthogonal to the subspace

$$V_{\mathcal{I}} = \mathrm{span}\{(u_i \otimes_2 u_i)\ \widetilde{\otimes}\ (v_j \otimes_2 v_j), (i,j) \in \mathcal{I}\},$$

and in general the power of the test for finite sample size depends on the properly rescaled norm of the projection of D onto $V_{\mathcal{I}}$.

In practice, it seems reasonable to use the subset of eigenfunctions that it is possible to estimate accurately given the available sample sizes. The accuracy of the estimates for the eigendirections can be in turn evaluated with bootstrap methods, see e.g. Hall & Hosseini-Nasab (2006) for the case of functional data. A good strategy may also be to consider more than one subset of eigenfunctions and then summarize the response obtained from the different tests using a Bonferroni correction.

As an alternative to these test statistics (based on projections of $D_N = C_N - C_{1,N} \widetilde{\otimes} C_{2,N}$), we consider also a test based on the squared Hilbert–Schmidt norm of D_N, i.e. $|||D_N|||_2^2$, whose null distribution will be approximated by a bootstrap procedure (this test will be referred to as *Hilbert–Schmidt test* hereafter). Though it seems that such tests would require one to store the full sample covariance of the data (which could be infeasible), we describe in Appendix E a way of circumventing such problem, although the computation of each entry of the full covariance is still needed. Therefore this could be used only for applications in which the dimension of the discretized covariance matrix is not too large.

In the following, we propose also a bootstrap approach to approximate the distribution of the test statistics $\widetilde{G}_N(\mathcal{I})$, $\widetilde{G}_N^a(\mathcal{I})$ and $G_N(\mathcal{I})$, with the aim to improve the finite sample properties of the procedure and to relax the distributional assumption on X.

3.1 Parametric Bootstrap

If we assume we know the distribution of X up to its mean μ and its covariance operator C, i.e. $X \sim F(\mu; C)$, we can approximate the distribution of $\widetilde{G}_N(\mathcal{I})$, $\widetilde{G}_N^a(\mathcal{I})$, $G_N(\mathcal{I})$ and $|||D_N|||_2^2$ under the separability hypothesis via a parametric bootstrap procedure. Since $C_{1,N} \widetilde{\otimes} C_{2,N}$, respectively $\overline{X}$, is an estimate of C, respectively μ, we simulate B bootstrap samples $X_1^b, \ldots, X_N^b \overset{\text{i.i.d.}}{\sim} F\left(\overline{X}, C_{1,N} \widetilde{\otimes} C_{2,N}\right)$, for $b = 1, \ldots, B$. For each sample, we compute $H_N^b = H_N(X_1^b, \ldots, X_N^b)$, where $H_N = G_N(\mathcal{I})$, $H_N = \widetilde{G}_N(\mathcal{I})$, $H_N = \widetilde{G}_N^a(\mathcal{I})$ respectively $H_N = |||D_N|||_2^2$, if we wish to use the non-Studentized projection test, the Studentized projection test, the approximated Studentized version or the Hilbert–Schmidt test, respectively. A formal description of the algorithm for obtaining the p-value of the test based on the statistic $H_N = H_N(X_1, \ldots, X_N)$ with the parametric bootstrap can be found in Appendix E, along with the details for the computation of H_N. We highlight that this procedure does not ask for the estimation of the full covariance structure, but only of its separable approximation, with the exception of the Hilbert–Schmidt test (and even in this case, it is possible to avoid the storage of the full covariance).

3.2 Empirical Bootstrap

In many applications it is not possible to assume a distribution for the random element X, and a non-parametric approach is therefore needed. In this setting, we can use the empirical bootstrap to estimate the distribution of the test statistic $G_N(\mathcal{I})$, $\widetilde{G}_N(\mathcal{I})$ or $|||D_N|||_2^2$ under the null hypothesis $H_0 : C = C_1 \widetilde{\otimes} C_2$. Let H_N denote the test statistic whose distribution is of interest. Based on an i.i.d. sample $X_1, \ldots, X_N \sim X$, we wish to approximate the distribution of H_N with the distribution of some test statistic $\Delta_N^* = \Delta_N(X_1^*, \ldots, X_N^*)$, where $X_1^*, \ldots, X_N^*$ is obtained by drawing with replacement from the set $\{X_1, \ldots, X_N\}$. Though it is tempting to use $\Delta_N^* = H_N(X_1^*, \ldots, X_N^*)$, this is not an appropriate choice. Indeed, let us look at the case $H_N = G_N(i, j)$. Notice that the true covariance of X is

$$C = \frac{\mathrm{Tr}_2(C) \widetilde{\otimes} \mathrm{Tr}_1(C)}{\mathrm{Tr}(C)} + D, \tag{3.4}$$

where D is a possibly non-zero operator, and that

$$H_N^* = G_N(i,j|X_1^*,\ldots,X_N^*) = N\left\langle (C_N^* - C_{1,N}^* \mathbin{\widetilde{\otimes}} C_{2,N}^*)(\hat{u}_i \otimes \hat{v}_j), \hat{u}_i \otimes \hat{v}_j \right\rangle^2,$$

where $C_N^* = C_N(X_1^*,\ldots,X_N^*)$, $C_{1,N}^* = C_{1,N}(X_1^*,\ldots,X_N^*)$, and $C_{2,N}^* = C_{2,N}(X_1^*,\ldots,X_N^*)$. Since $(C_N^* - C_{1,N}^* \mathbin{\widetilde{\otimes}} C_{2,N}^*) \approx (C_N - C_{1,N} \mathbin{\widetilde{\otimes}} C_{2,N}) \approx D$, the statistic H_N^* would approximate the distribution of H_N under the hypothesis (3.4), which is not what we want. We therefore propose the following choices of $\Delta_N^* = \Delta_n(X_1^*,\ldots,X_N^*;X_1,\ldots,X_N)$, depending on the choice of H_N:

1. $H_N = G_N(\mathcal{I})$, $\Delta_N^* = \sum_{(i,j)\in\mathcal{I}} (T_N^*(i,j) - T_N(i,j))^2$.
2. $H_N = \widetilde{G}_N(\mathcal{I})$, $\Delta_N^* = \left|\left(\hat{\Sigma}_{L,\mathcal{I}}^*\right)^{-1/2} (\mathbf{T}_N^*(\mathcal{I}) - \mathbf{T}_N(\mathcal{I})) \left(\hat{\Sigma}_{R,\mathcal{I}}^*\right)^{-\mathsf{T}/2}\right|^2$, where $\hat{\Sigma}_{L,\mathcal{I}}^* = \hat{\Sigma}_{L,\mathcal{I}}(X_1^*,\ldots,X_N^*)$, and $\hat{\Sigma}_{R,\mathcal{I}}^* = \hat{\Sigma}_{R,\mathcal{I}}(X_1^*,\ldots,X_N^*)$. are the row, resp. column, covariances estimated from the bootstrap sample.
3. $H_N = \widetilde{G}_N^a(\mathcal{I})$, $\Delta_N^* = \sum_{(i,j)\in\mathcal{I}} (T_N^*(i,j) - T_N(i,j))^2 / \hat{\sigma}_*^2(i,j)$, where $\hat{\sigma}_*^2(i,j) = \hat{\sigma}^2(i,j|X_1^*,\ldots,X_N^*)$.
4. $H_N = |||D_N|||_2^2$, $\Delta_N^* = |||D_N^* - D_N|||_2^2$, where $D_N^* = D_N(X_1^*,\ldots,X_N^*)$.

The algorithm to approximate the p-value of H_N by the empirical bootstrap is described in detail. The basic idea consists of generating B bootstrap samples, computing Δ_N^* for each bootstrap sample and looking at the proportion of bootstrap samples for which Δ_N^* is larger than the test statistic H_N computed from the original sample.

4 Empirical demonstrations of the method

4.1 Simulation studies

We investigated the finite sample behavior of our testing procedures through an intensive reproducible simulation study (its running time is equivalent to approximately 401 days on a single CPU computer). We compared the test based on the asymptotic distribution of (3.1), as well as the tests based on $G_N(\mathcal{I})$, $\widetilde{G}_N(\mathcal{I})$, $\widetilde{G}_N^a(\mathcal{I})$, and $|||D_N|||_2^2$, with the p-values obtained via the parametric bootstrap or the empirical bootstrap.

We generated discretized functional data $X_1,\ldots,X_N \in \mathbb{R}^{32\times 7}$ under two scenarios. In the first scenario (Gaussian scenario), the data were generated from a multivariate Gaussian distribution $\mathcal{N}(0,\mathbf{C})$. In the second scenario (Non-Gaussian scenario), the data were generated from a multivariate t distribution with 6 degrees of freedom and non centrality parameter equal to zero. In the Gaussian scenario, we set $\mathbf{C} = \mathbf{C}^{(\gamma)}$, where

$$\mathbf{C}^{(\gamma)}(i_1,j_1,i_2,j_2) = (1-\gamma)c_1(i_1,i_2)c_2(j_1,j_2) + \gamma\frac{1}{(j_1-j_2)^2+1}\exp\left\{-\frac{(i_1-i_2)^2}{(j_1-j_2)^2+1}\right\}, \quad (4.1)$$

$\gamma \in [0,1]; i_1,i_2 = 1,\ldots,32; j_1,j_2 = 1,\ldots,7$. The covariances c_1 and c_2 used in the simulations can be seen in Figure 2. For the Non-Gaussian scenario, we chose a multivariate t distribution with the correlation structure implied by $\mathbf{C}^{(\gamma)}$, $\gamma \in [0,1]; i_1,i_2 = 1,\ldots,32; j_1,j_2 = 1,\ldots,7$. The parameter $\gamma \in [0,1]$ controls the departure from the separability of the covariance $\mathbf{C}^{(\gamma)}$: $\gamma = 0$ yields a separable covariance, whereas $\gamma = 1$ yields a complete non-separable covariance structure (Cressie & Huang 1999). All the simulations have been performed using the R package `covsep` (Tavakoli 2016), available on CRAN, which implements the tests presented in the paper.

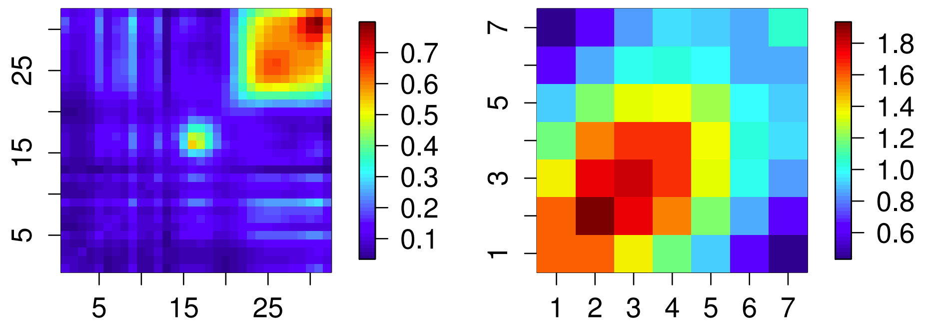
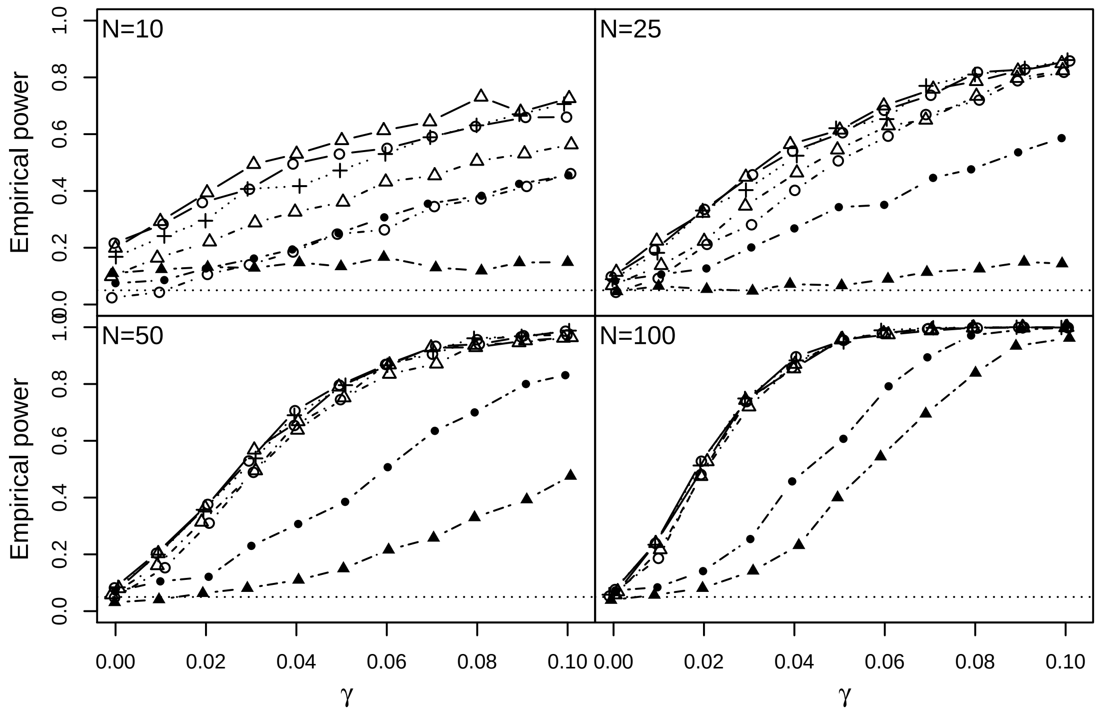


Figure 2: Covariance functions c_1 (left) and c_2 (right) used in the simulation study.

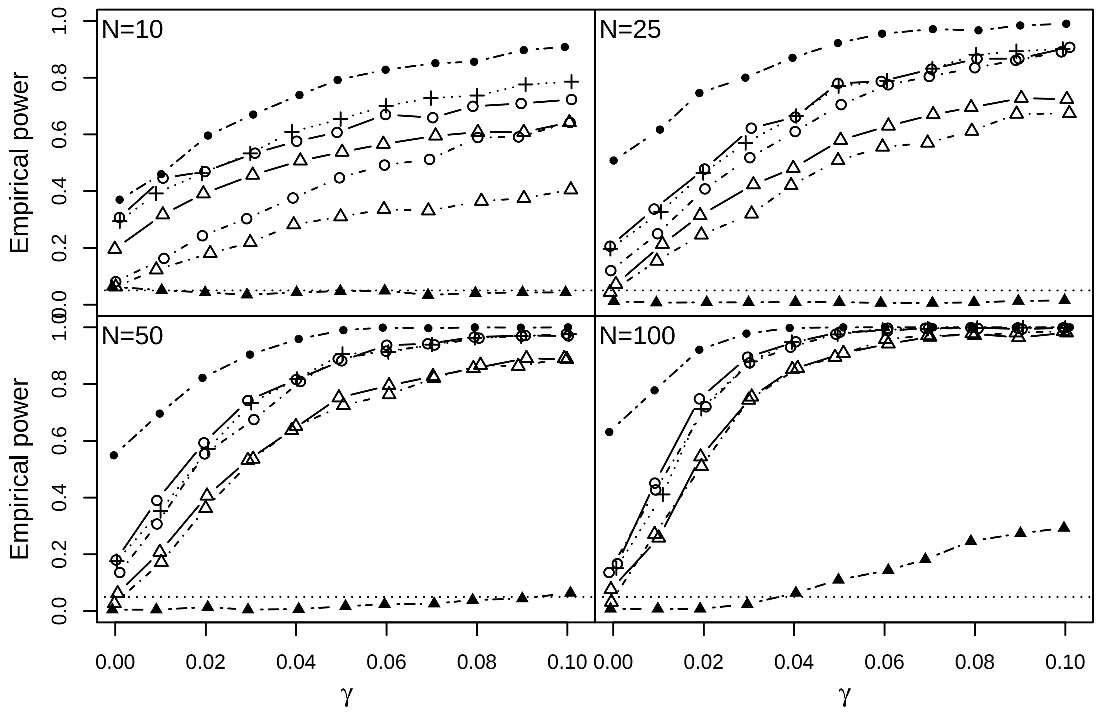
For each value of $\gamma \in \{0, 0.01, 0.02, \dots, 0.1\}$ and $N \in \{10, 25, 50, 100\}$, we performed 1000 replications for each of the above simulations, and estimated the power of the tests based on the asymptotic distribution of (3.1).

We first also estimated the power of the tests $\widetilde{G}_N(1,1)$, $G_N(1,1)$, and $|||D_N|||_2$, with distributions approximated by a Gaussian parametric bootstrap, and the empirical bootstrap, with $B = 1000$. The results are shown in Figure 3. In the Gaussian scenario (Figure 3, panel (a)), the empirical size of all the proposed tests gets closer to the nominal level (5%) as N increases (see also Table 2). Nevertheless, the non-Studentized tests $G_N(1,1)$, for both parametric and empirical bootstrap, seem to have a slower convergence with respect to the Studentized version, and even for $N = 100$ the level of these tests appear still higher than the nominal one (and a CLT-based 95% confidence interval for the true level does not contain the nominal level in both cases). The empirical bootstrap version of the Hilbert–Schmidt test also fails to respect the nominal level at $N = 100$, but its parametric bootstrap counterpart respects the level, even for $N = 25$. For $N = 25, 50, 100$, the most powerful tests (amongst those who respect the nominal level) are the parametric and empirical bootstrap versions of $\widetilde{G}_N(1,1)$, and they seem to have equal power. The power of the Hilbert–Schmidt test based on the parametric bootstrap seems to be competitive only for $N = 100$ and $\gamma = 0.1$, and is much lower for other values of the parameters. The test based on the asymptotic distribution does not respect the nominal level for small N but it does when N increases. Indeed, the convergence to the nominal level seems remarkably fast and its power is comparable with those of the parametric and empirical bootstrap tests based on $\widetilde{G}_N(1,1)$. Despite being based on an asymptotic result, its performance is quite good also in finite samples, and it is less computationally demanding than the bootstrap tests.

In the non-Gaussian scenario (Figure 3, panel (b)), only the empirical bootstrap version of $\widetilde{G}_N(1,1)$ and of the Hilbert–Schmidt test seem to respect the level for $N = 10$ (see also Table 2). Amongst these tests, the most powerful one is clearly the empirical bootstrap test based on $\widetilde{G}_N(1,1)$. Although the Gaussian parametric bootstrap test has higher empirical power, it does not have the correct level (as expected) and thus cannot be used in a non-Gaussian scenario. Notice also that the test based on the asymptotic distribution of $\widetilde{G}_N(1,1)$ (under Gaussian assumptions) does not respects the level of the test even for $N = 100$. The same holds for the Gaussian bootstrap version of the Hilbert–Schmidt test. Finally,



(a) Gaussian scenario



(b) Non Gaussian scenario

Figure 3: Empirical power of the testing procedures in the *Gaussian* scenario (panel (a)) and *non-Gaussian* scenario (panel (b)), for $N = 10, 25, 50, 100$ and $\mathcal{I} = \mathcal{I}_1$. The results shown correspond to the test (3.1) based on its asymptotic distribution (····+····), the Gaussian parametric bootstrap test (solid line with empty circles) and its studentized version (dash-dotted line with empty circles), the empirical parametric bootstrap test (—△—) and its Studentized version (– –△– –), the Gaussian parametric Hilbert–Schmidt test (dash-dotted line with filled circles) and the empirical Hilbert–Schmidt test (dash-dotted line with filled triangles). The horizontal dotted line indicates the nominal level (5%) of the test. Note that the points have been horizontally jittered for better visibility.

though the empirical bootstrap version of the Hilbert–Schmidt test respects the level for $N = 10, 25, 50, 100$, it has virtually no power for $N = 10, 25, 50$, and has very low power for $N = 100$ (at most 0.3 for $\gamma = 0.1$).

As mentioned previously, there is no guarantee that a violation in the separability of C is mostly reflected in the first separable eigensubspace. Therefore, we consider also a larger subspace for the test. Figure 9 shows the empirical power for the asymptotic test, the parametric and empirical bootstrap tests based on the test statistic $\widetilde{G}_N(\mathcal{I}_2)$, as well as parametric and bootstrap tests based on the test statistics $G_N(\mathcal{I})$, $\widetilde{G}^a_N(\mathcal{I}_2)$ where $\mathcal{I}_2 = \{(i,j) : i,j = 1,2\}$. In the Gaussian scenario, the asymptotic test is much slower in converging to the correct level compared to its univariate version based on $\widetilde{G}_N(1,1)$. For larger N its power is comparable to that of the parametric and empirical bootstrap based on the Studentized test statistics $\widetilde{G}_N(\mathcal{I}_2)$, which in addition respects the nominal level, even for $N = 10$. It is interesting to note that the approximated Studentized bootstrap tests $\widetilde{G}^a_N(\mathcal{I}_2)$ have a performance which is better than the non Studentized bootstrap tests $G_N(\mathcal{I}_2)$ but far worse than that of the Studentized tests $\widetilde{G}_N(\mathcal{I}_2)$. The Hilbert–Schmidt test is again outperformed by all the other tests, with the exception of the non-Studentized bootstrap test when $N = 10, 25$. The results are similar for the non-Gaussian scenario, apart for the fact that the asymptotic test does not respect the nominal level (as expected, since it asks for X to be Gaussian).

To investigate the difference between projecting on one or several eigensubspaces, we also compare the power of the empirical bootstrap version of the tests $\widetilde{G}_N(\mathcal{I})$ for increasing projection subspaces, i.e. for $\mathcal{I} = \mathcal{I}_l, l = 1,2,3$, where $\mathcal{I}_1 = \{(1,1)\}, \mathcal{I}_2 = \{(i,j) : i,j = 1,2\}$ and $\mathcal{I}_3 = \{(i,j) : i = 1,\ldots,4; j = 1,\ldots,10\}$. The results are shown in Figure 4 for the Gaussian scenario and Figure 6 for the non-Gaussian scenario. In the Gaussian scenario, for $N = 10$, the most powerful test is $\widetilde{G}_N(\mathcal{I}_2)$. In this case, projecting onto a larger eigensubspace decreases the power of the test dramatically. However, for $N \geq 25$ the power of the test is the largest for $\widetilde{G}_N(\mathcal{I}_3)$, albeit only significantly larger than that of $\widetilde{G}_N(\mathcal{I}_2)$ when $\gamma = 0.01$. Our interpretation is that when the sample size is too small, including too many eigendirection is bound to add only noise that degrades the performance of the test. However, as long as the separable eigenfunctions are estimated accurately, projecting in a larger eigenspace improves the performance of test. See also Figure 10 for the complete simulation results of the projection set $\mathcal{I}_3$.

This prompts us to investigate how the power of the test varies across all projection subsets

$$\mathcal{I}_{r,s} = \{(i,j) : 1 \leq i \leq r, 1 \leq j \leq s\},$$

$r = 1,\ldots,32, s = 1,\ldots,7$,=. The test used is $\widetilde{G}_N(\mathcal{I})$, with distribution approximated by the empirical bootstrap with $B = 1000$. Figure 5 shows the empirical size and power of the separability test in the Gaussian scenario for sample size $N = 25$, and Figure 7, respectively Figure 8, shows the power for different sample sizes in the Gaussian scenario, respectively the non-Gaussian scenario.

4.1.1 Discussion of simulation studies

The simulation studies above illustrate how the empirical bootstrap test based on the test statistics $\widetilde{G}_N(\mathcal{I})$ usually outperforms its competitors, albeit it is also much more computationally expensive than the asymptotic test, whose performance are comparable in the Gaussian scenario for large enough number of observations.

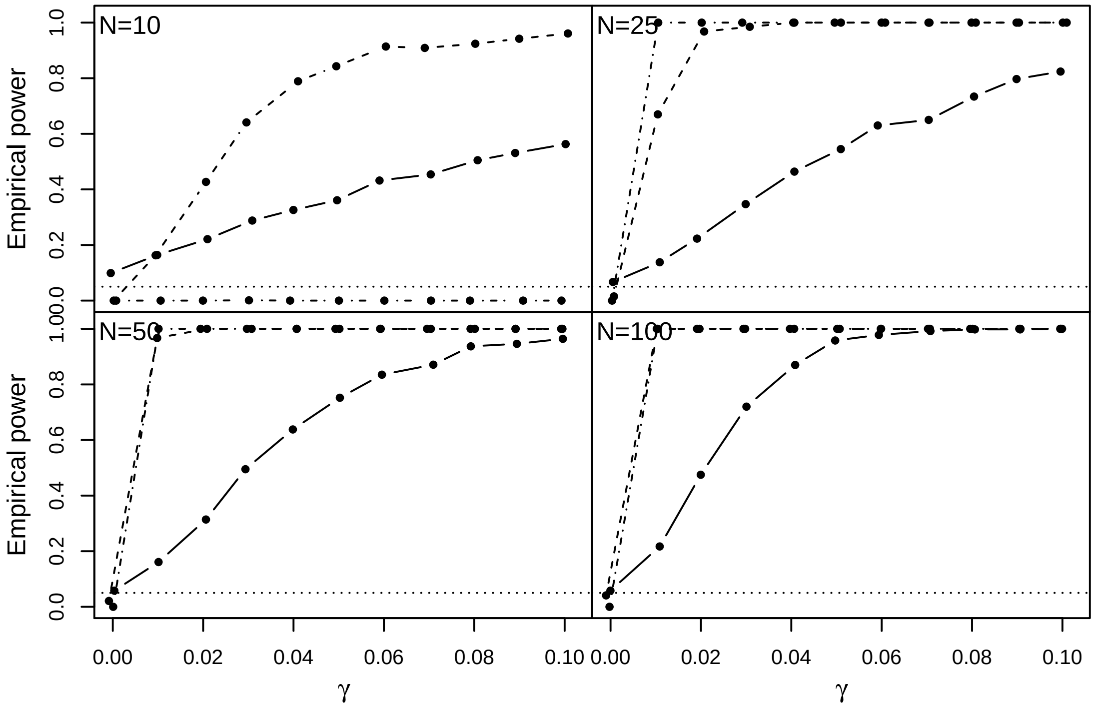


Figure 4: Empirical power of the empirical bootstrap version of $\widetilde{G}_N(\mathcal{I}_l)$, for $l = 1$ (solid line), $l = 2$ (dashed line) and $l = 3$ (dash-dotted line), in the *Gaussian* scenario. The horizontal dotted line indicates the nominal level (5%) of the test. Note that the points have been horizontally jittered for better visibility.

The choice of the best set of eigendirections to use in the definition of the test statistics is difficult. It seems that K should be ideally chosen to be increasing with N. This is reasonable, because larger values of N increase the accuracy of the estimation of the eigenfunctions and therefore we will be able to detect departures from the separability in more eigendirections, including ones not only associated with the largest eigenvalues. However, the optimal rate at which K should increase with N is still an open problem, and will certainly depend in a complex way on the eigenstructure of the true underlying covariance operator C.

This is confirmed by the results reported in Figure 5 and Figures 7 and 8. These indeed show that taking into account too few eigendirections can result in smaller power, while including too many of them can also decrease the power.

As an alternative to tests based on projections of D_N, the tests based on the squared Hilbert–Schmidt norm of D_N, i.e. $|||D_N|||_2^2$, could potentially detect any departure from the separability hypothesis—as opposed to the tests $\widetilde{G}_N(\mathcal{I})$. But as the simulation study illustrates, they might be far less powerful in practice, particularly in situations where the departure from separability is reflected in only in a few eigendirections. Moreover, this approach still requires the computation of the full covariance operator (although not its storage) and is therefore not feasible for all applications.

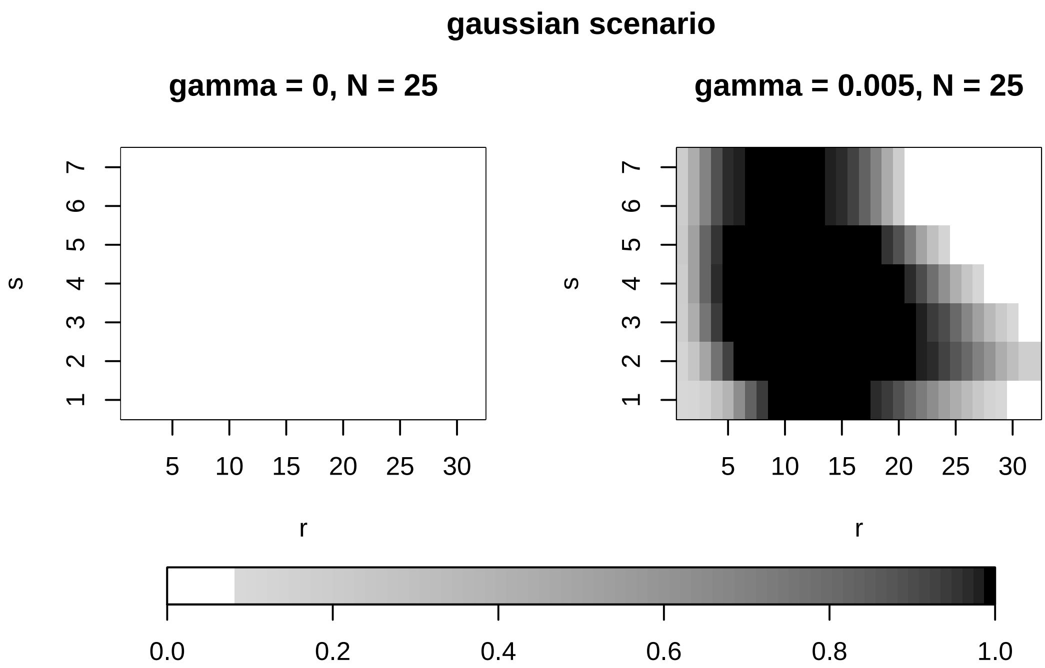


Figure 5: Empirical size (left) and power (right) of the separability test as functions of the projection set $\mathcal{I}$. The test used is $\widetilde{G}_N(\mathcal{I})$, with distribution approximated by the empirical bootstrap with $B = 1000$. The left plot, respectively the right plot, was simulated from the Gaussian scenario with $\gamma = 0$, respectively $\gamma = 0.005$, and $N = 25$. Each (r, s) rectangle represents the level/power of the test based on the projection set $\mathcal{I} = \{(i, j) : 1 \leq i \leq r, 1 \leq j \leq s\}$.

4.2 Application to acoustic phonetic data

An interesting case where the proposed methods can be useful are phonetic spectrograms. These data arise in the analysis of speech records, since relevant features of recorded sounds can be better explored in a two dimensional time-frequency domain.

In particular, we consider here the dataset of 23 speakers from five different Romance languages that has been first described in Pigoli et al. (2014). The speakers were recorded while pronouncing the words corresponding to the numbers from one to ten in their language and the recordings are converted to a sampling rate of 16000 samples per second. Since not all these words are available for all the speakers, we have a total of 219 speech records. We focus on the spectrum that speakers produce in each speech recording $x_{ik}^L(t)$, where L is the language, $i = 1, \dots, 10$ the pronounced word and $k = 1, \dots, n_L$ the speaker, n_L being the number of speakers available for language L. We then use a short-time Fourier transform to obtain a two dimensional log-spectrogram: we use a Gaussian window function $w(\cdot)$ with a window size of 10 milliseconds and we compute the short-time Fourier transform as

$$X_{ik}^L(\omega, t) = \int_{-\infty}^{+\infty} x_{ik}^L(\tau) w(\tau - t) e^{-j\omega\tau} \mathrm{d}\tau.$$

The spectrogram is defined as the magnitude of the Fourier transform and the log-spectrogram (in decibel) is therefore

$$\mathfrak{S}_{ik}^{L}(\omega, t) = 10 \log_{10}(|X_{ik}^{L}(\omega, t)|^2).$$

The raw log-spectrograms are then smoothed (with the robust spline smoothing method proposed in Garcia 2010) and aligned in time using an adaptation to 2-D of the procedure in Tang & Müller (2008). The alignment is needed because a phase distortion can be present in acoustic signals, due to difference in speech velocity between speakers. Since the different words of each language have different mean log-spectrograms, the focus of the linguistic analysis—which is the study cross-linguistics changes—is on the residual log-spectrograms

$$R_{ik}^{L}(\omega, t) = S_{ik}^{L}(\omega, t) - (1/n_i) \sum_{k=1}^{n_i} S_{ik}^{L}(\omega, t).$$

Assuming that all the words within the language have the same covariance structure, we disregard hereafter the information about the pronounced words that generated the residual log-spectrogram, and use the surface data $R_j^L(\omega, t), j = 1, \dots, N_L$, i.e. the set of observations for the language L including all speakers and words, for the separability test. These observations are measured on an equispaced grid with 81 points in the frequency direction and 100 points in the time direction. This translate on a full covariance structure with about 33×10^6 degrees of freedom. Thus, although the discretized covariance matrix is in principle computable, its storage is a problem. More importantly, the accuracy of its estimate is poor, since we have at most 50 observations within each language. For these reasons, we would like to investigate if a separable approximation of each covariance is appropriate.

We thus apply the Studentized version of the empirical bootstrap test for separability to the residual log-spectrograms for each language individually. Here, we take into consideration different choices for set of eigendirections to be used in the definition of the test statistic $\widetilde{G}_N(\mathcal{I})$, namely $\mathcal{I} = \mathcal{I}_1 = \{(1, 1)\}$, $\mathcal{I} = \mathcal{I}_2 = \{(r, s) : 1 \leq r \leq 2, 1 \leq s \leq 3\}$, $\mathcal{I} = \mathcal{I}_3 = \{(r, s) : 1 \leq r \leq 8, 1 \leq s \leq 10\}$. For all cases we use $B = 1000$ bootstrap replicates.

The resulting p-values for each language and for each set of indices can be found in Table 1. Taking into account the multiple testings with a Bonferroni correction, we can conclude that the separability assumption does not appear to hold. We can also see that the departure from separability is caught mainly on the first component for the two Spanish varieties. In conclusion, a separable covariance structure is not a good fit for these languages and thus, when practitioners use this approximation for computational or modeling reasons, they should bear in mind that relevant aspects of the covariance structure may be missed in the analysis.

5 Discussion and conclusions

We presented tests to verify the separability assumption for the covariance operators of random surfaces (or hypersurfaces) through hypothesis testing. These tests are based on the difference between the sample covariance operator and its separable approximation—which we have shown to be asymptotically Gaussian—projected onto subspaces spanned by the eigenfunctions of the covariance of the data. While the optimal choice for this subspace is still an open problem and it may depend on the eigenstructure of the full covariance operator, it is however possible to give some advice on how to choose $\mathcal{I}$ in practice:

Table 1: P-values for the test for the separability of the covariance operators of the residual log-spectrograms of the five Romance languages, using the Studentized version of the empirical bootstrap.

$\mathcal{I}$	French	Italian	Portuguese	American Spanish	Iberian Spanish
$\mathcal{I}_1$	0.65	< 0.001	< 0.001	< 0.001	< 0.001
$\mathcal{I}_2$	0.078	0.197	0.022	0.36	0.013
$\mathcal{I}_3$	0.001	0.002	0.001	0.001	< 0.001

- in many cases, a dimensional reduction based on the separable eigenfunctions is needed also for the follow up analysis. Then, it is recommended to use the same subspace for the test procedure as well, so that we are guaranteed at least that the projection of the covariance structure in the subspace that will be used for the analysis is separable, as shown in Section 3.
- As mentioned in Section 3, it is usually better to focus on the subset of eigenfunctions that it is possible to estimate accurately with the available data. These can be again identified with bootstrap methods such as the one described in Hall & Hosseini-Nasab (2006) or considering the dimension of the sample size. As highlighted by the results of the simulation studies in Figure 5 and in Figures 7 and 8, the empirical power of the test starts to decline when eigendirections that cannot be reasonably estimated with the available sample size are included.
- When in doubt, it is also possible to apply the test to more than one subset of eigenfunctions and then summarize the response using a Bonferroni correction. We follow this approach in the data application described in Section 4.2.

Though an asymptotic distribution is available in some cases, we also propose to approximate the distribution of our test statistics using either a parametric bootstrap (in case the distribution of the data is known) or an empirical bootstrap. A simulation study suggests that the Studentized version of the empirical bootstrap test gives the highest power in non-Gaussian settings, and has power comparable to its parametric bootstrap counterpart and to the asymptotic test in the Gaussian setting. We therefore use the Studentized empirical bootstrap for the application to linguistic data, since it is not easy to assess the distribution of the data generating process. The bootstrap test leads to the conclusion that the covariance structure is indeed not separable.

Our present approach implicitly assumed that the functional observations (e.g. the hypersurfaces) were densely observed. Though this approach is not restricted to data observed on a grid, it leaves aside the important class of functional data that are sparsely observed (e.g. Yao et al. n.d.). However, the extension of our methodology to the case of sparsely observed functional data is also possible, as long as the estimator used for the full covariance is consistent and satisfies a central limit theorem. Indeed, while we have only detailed the methods for 2-dimensional surfaces, the extension to higher-order multidimensional functions (such as 3-dimensional volumetric images from applications such as magnetic resonance imaging) is straightforward.

A The Asymptotic Covariance Structure

lemma A.1. *The covariance operator of the random operator Z, defined in Theorem 2.3, is characterized by the following equality, in which* $\Gamma = \mathbb{E}\left[(X \otimes X - C)\ \widetilde{\bigotimes}\ (X \otimes X - C)\right]$:

$$
\begin{aligned}
&\mathbb{E}\left[\operatorname{Tr}\left[(A_1 \,\widetilde{\otimes}\, A_2)Z\right]\operatorname{Tr}\left[(B_1 \,\widetilde{\otimes}\, B_2)Z\right]\right] = \qquad\qquad \text{(A.1)}\\
&\operatorname{Tr}\left[(A \,\widetilde{\bigotimes}\, B)\Gamma\right] + \frac{\operatorname{Tr}[BC]}{\operatorname{Tr}(C)}\operatorname{Tr}\left[(A \,\widetilde{\bigotimes}\, \mathrm{Id}_H)\Gamma\right] - \frac{\operatorname{Tr}[B_2C_2]}{\operatorname{Tr}[C_2]}\operatorname{Tr}\left[\left(A \,\widetilde{\bigotimes}\, (B_1 \,\widetilde{\otimes}\, \mathrm{Id}_{H_2})\right)\Gamma\right]\\
&-\frac{\operatorname{Tr}[B_1C_1]}{\operatorname{Tr}[C_1]}\operatorname{Tr}\left[\left(A \,\widetilde{\bigotimes}\, (\mathrm{Id}_{H_1} \,\widetilde{\otimes}\, B_2)\right)\Gamma\right] + \frac{\operatorname{Tr}[AC]}{\operatorname{Tr}[C]}\left\{\operatorname{Tr}\left[(\mathrm{Id}_H \,\widetilde{\bigotimes}\, B)\Gamma\right] + \frac{\operatorname{Tr}[BC]}{\operatorname{Tr}[C]}\operatorname{Tr}[\Gamma]\right.\\
&\left.-\frac{\operatorname{Tr}[B_2C_2]}{\operatorname{Tr}[C_2]}\operatorname{Tr}\left[(\mathrm{Id}_H \,\widetilde{\bigotimes}\, (B_1 \,\widetilde{\otimes}\, \mathrm{Id}_{H_2}))\Gamma\right] - \frac{\operatorname{Tr}[B_1C_1]}{\operatorname{Tr}[C_1]}\operatorname{Tr}\left[(\mathrm{Id}_H \,\widetilde{\bigotimes}\, (\mathrm{Id}_{H_1} \,\widetilde{\otimes}\, B_2))\Gamma\right]\right\}\\
&-\frac{\operatorname{Tr}[A_2C_2]}{\operatorname{Tr}[C_2]}\left\{\operatorname{Tr}\left[\left((A_1 \,\widetilde{\otimes}\, \mathrm{Id}_{H_2}) \,\widetilde{\bigotimes}\, B\right)\Gamma\right] + \frac{\operatorname{Tr}[BC]}{\operatorname{Tr}[C]}\operatorname{Tr}\left[\left((A_1 \,\widetilde{\otimes}\, \mathrm{Id}_{H_2}) \,\widetilde{\bigotimes}\, \mathrm{Id}_H\right)\Gamma\right]\right.\\
&-\frac{\operatorname{Tr}[B_2C_2]}{\operatorname{Tr}[C_2]}\operatorname{Tr}\left[\left((A_1 \,\widetilde{\otimes}\, \mathrm{Id}_{H_2}) \,\widetilde{\bigotimes}\, (B_1 \,\widetilde{\otimes}\, \mathrm{Id}_{H_2})\right)\Gamma\right]\\
&\left.-\frac{\operatorname{Tr}[B_1C_1]}{\operatorname{Tr}[C_1]}\operatorname{Tr}\left[\left((A_1 \,\widetilde{\otimes}\, \mathrm{Id}_{H_2}) \,\widetilde{\bigotimes}\, (\mathrm{Id}_{H_1} \,\widetilde{\otimes}\, B_2)\right)\Gamma\right]\right\}\\
&-\frac{\operatorname{Tr}[A_1C_1]}{\operatorname{Tr}[C_1]}\left\{\operatorname{Tr}\left[\left((\mathrm{Id}_{H_1} \,\widetilde{\otimes}\, A_2) \,\widetilde{\bigotimes}\, B\right)\Gamma\right] + \frac{\operatorname{Tr}[BC]}{\operatorname{Tr}[C]}\operatorname{Tr}\left[\left((\mathrm{Id}_{H_1} \,\widetilde{\otimes}\, A_2) \,\widetilde{\bigotimes}\, \mathrm{Id}_H\right)\Gamma\right]\right.\\
&-\frac{\operatorname{Tr}[B_2C_2]}{\operatorname{Tr}[C_2]}\operatorname{Tr}\left[\left((\mathrm{Id}_{H_1} \,\widetilde{\otimes}\, A_2) \,\widetilde{\bigotimes}\, (B_1 \,\widetilde{\otimes}\, \mathrm{Id}_{H_2})\right)\Gamma\right]\\
&\left.-\frac{\operatorname{Tr}[B_1C_1]}{\operatorname{Tr}[C_1]}\operatorname{Tr}\left[\left((\mathrm{Id}_{H_1} \,\widetilde{\otimes}\, A_2) \,\widetilde{\bigotimes}\, (\mathrm{Id}_{H_1} \,\widetilde{\otimes}\, B_2)\right)\Gamma\right]\right\},
\end{aligned}
$$

where $A_1, B_1 \in \mathcal{S}_\infty(H_1), A_2, B_2 \in \mathcal{S}_\infty(H_2)$*, and* $A = A_1 \,\widetilde{\otimes}\, A_2$, $B = B_1 \,\widetilde{\otimes}\, B_2$, $H = H_1 \otimes H_2$*, and* Id_H *denotes the identity operator on the Hilbert space* H.

Proof. By the linearity of the expectation and the trace, and by the properties of the partial trace, the computation of (A.1) boils down to the computation of expressions of the form

$$\mathbb{E}\left[\operatorname{Tr}\left[(A_1' \,\widetilde{\otimes}\, A_2')Y\right]\operatorname{Tr}\left[(B_1' \,\widetilde{\otimes}\, B_2')Y\right]\right],$$

for general $A_1', B_1' \in \mathcal{S}_\infty(H_1), A_2', B_2' \in \mathcal{S}_\infty(H_2)$. Since $\mathbb{E}\,|||Y|||_1^2 < \infty$, we have

$$
\begin{aligned}
&\mathbb{E}\left(\operatorname{Tr}\left[(A_1' \,\widetilde{\otimes}\, A_2')Y\right]\operatorname{Tr}\left[(B_1' \,\widetilde{\otimes}\, B_2')Y\right]\right) =\\
&= \operatorname{Tr}\left[\left((A_1' \,\widetilde{\otimes}\, A_2') \,\widetilde{\bigotimes}\, (B_1' \,\widetilde{\otimes}\, B_2')\right)\mathbb{E}\left(Y \,\widetilde{\bigotimes}\, Y\right)\right]\\
&= \operatorname{Tr}\left[\left((A_1' \,\widetilde{\otimes}\, A_2') \,\widetilde{\bigotimes}\, (B_1' \,\widetilde{\otimes}\, B_2')\right)\Gamma\right],
\end{aligned}
$$

where $\Gamma = \mathbb{E}\left[(X \otimes X - C)\ \widetilde{\bigotimes}\ (X \otimes X - C)\right]$. The computation of (A.1) follows directly. □

B Proofs

Proof of Corollary 2.4. To alleviate the notation, we shall assume without loss of generality that $\mu = \mathbb{E} X = 0$. Using the properties of the tensor product (see Appendix D.1, we get that $T_N(r,s) = \mathrm{Tr}\left[(\hat{A}_r \,\widetilde{\otimes}\, \hat{B}_s)\sqrt{N} D_N\right]$, where $\hat{A}_r = (\hat{u}_r \otimes_2 \hat{u}_r)$, $\hat{B}_s = (\hat{v}_s \otimes_2 \hat{v}_s)$. Now notice that though $A_r = u_r \otimes_2 u_r$ and $B_s = v_s \otimes_2 v_s$ are not estimable separately (since C_1 and C_2 are not identifiable), their $\widetilde{\otimes}$-product is identifiable, and is consistently estimated by $\hat{A}_r \,\widetilde{\otimes}\, \hat{B}_s$ (in trace norm). Slutsky's Lemma, Theorem 2.3 and the continuous mapping theorem imply therefore that $(T_N(r,s))_{(r,s)\in\mathcal{I}}$ has the same asymptotic distribution of $\left(\widetilde{T}_N(r,s)\right)_{(r,s)\in\mathcal{I}}$, where $\widetilde{T}_N(r,s) = \mathrm{Tr}\left[(A_r \,\widetilde{\otimes}\, B_s)\sqrt{N} D_N\right]$. This implies that

$$(T_N(r,s))_{(r,s)\in\mathcal{I}} \xrightarrow{d} Z' = \left(\mathrm{Tr}\left[(A_r \,\widetilde{\otimes}\, B_s)Z\right]\right)_{(r,s)\in\mathcal{I}}, \quad \text{as } N \to \infty,$$

where Z is a mean zero Gaussian random element of $\mathcal{S}_1(H_1 \otimes H_2)$ whose covariance structure is given by Lemma A.1. Z' is therefore also Gaussian random element, with mean zero and covariances

$$\Sigma_{(r,s),(r',s')} = \mathrm{cov}(Z'_{(r,s)}, Z'_{(r',s')}) = \mathbb{E}\left[\mathrm{Tr}\left[(A_r \,\widetilde{\otimes}\, B_s)Z\right]\mathrm{Tr}\left[A_{r'} \,\widetilde{\otimes}\, B_{s'})Z\right]\right].$$

Using Lemma A.1, we see that the computation of $\Sigma_{(r,s),(r',s')}$ depends on the terms $\mathrm{Tr}\left[(A_r \,\widetilde{\otimes}\, B_s)C\right] = \lambda_r\gamma_s$, $\mathrm{Tr}\left[A_r C_1\right] = \lambda_r$, $\mathrm{Tr}\left[B_s C_2\right] = \gamma_s$, as well as on the value of

$$\mathrm{Tr}\left[\left((A'_1 \,\widetilde{\otimes}\, B'_1) \,\widetilde{\bigotimes}\, (A'_2 \,\widetilde{\otimes}\, B'_2)\right)\Gamma\right]$$

for general $A'_1, A'_2 \in \mathcal{S}_\infty(H_1), B'_1, B'_2 \in \mathcal{S}_\infty(H_2)$. Using the Karhunen–Loève expansion $X = \sum_{i,i'\geq 1} \xi_{ii'} u_i \otimes v_{i'}$, where $\xi_{ii'} = \langle X, u_i \otimes v_{i'}\rangle$, we get

$$\begin{aligned}\Gamma &= \mathbb{E}\left((X \otimes_2 X - C) \,\widetilde{\otimes}\, (X \otimes_2 X - C)\right) \\ &= \sum_{i,i',j,j',k,k',l,l'\geq 1} \beta_{ii'jj'kk'll'} \left(u_{ij} \,\widetilde{\otimes}\, v_{i'j'}\right) \,\widetilde{\bigotimes}\, \left(u_{kl} \,\widetilde{\otimes}\, v_{k'l'}\right) \\ &\quad - \sum_{i,i',j,j'} \alpha_{ii'}\alpha_{jj'} \left(u_{ii} \,\widetilde{\otimes}\, v_{i'i'}\right) \,\widetilde{\bigotimes}\, \left(u_{jj} \,\widetilde{\otimes}\, v_{j'j'}\right)\end{aligned}$$

where we have written $u_{ij} = u_i \otimes_2 u_j \in \mathcal{S}_1(H_1), v_{ij} = v_i \otimes_2 v_j \in \mathcal{S}_1(H_2)$, $\beta_{ii'jj'kk'll'} = \mathbb{E}\left[\xi_{ii'}\xi_{jj'}\xi_{kk'}\xi_{ll'}\right]$, $\alpha_{ij} = \lambda_i\gamma_j$ and used the identity $u_{ij} \,\widetilde{\otimes}\, v_{i'j'} = (u_i \otimes v_{i'}) \otimes_2 (u_j \otimes v_{j'})$. Therefore,

$$\begin{aligned}&\mathrm{Tr}\left[\left((A'_1 \,\widetilde{\otimes}\, A'_2) \,\widetilde{\bigotimes}\, (B'_1 \,\widetilde{\otimes}\, B'_2)\right)\Gamma\right] = \\ &= \sum_{i,i',j,j',k,k',l,l'\geq 1} \beta_{ii'jj'kk'll'} \,\mathrm{Tr}[A'_1 u_{ij}]\,\mathrm{Tr}[A'_2 v_{i'j'}]\,\mathrm{Tr}[B'_1 u_{kl}]\,\mathrm{Tr}[B'_2 v_{k'l'}] \\ &\quad - \sum_{i,i',j,j'} \alpha_{ii'}\alpha_{jj'} \,\mathrm{Tr}[A'_1 u_{ii}]\,\mathrm{Tr}[B'_1 u_{jj}]\,\mathrm{Tr}[A'_2 v_{i'i'}]\,\mathrm{Tr}[B'_2 v_{j'j'}],\end{aligned}$$

and the computation of the variance $\Sigma_{(r,s),(r',s')}$ follows from a straightforward (though tedious) calculation. □

Proof of Corollary 2.5. We only need to compute and substitute the values of the fourth order moments terms $\tilde{\beta}_{ijkl}$ in the expression given by Corollary 2.4. Since $\tilde{\beta}_{ijkl} = \mathbb{E}\left[\xi_{ij}^2 \xi_{kl}^2\right] = 3\alpha_{kl}^2$ if $(i,j) = (k,l)$, and $\tilde{\beta}_{ijkl} = \alpha_{ij}\alpha_{kl}$ if $(i,j) \neq (k,l)$, straightforward calculations give

$$\tilde{\beta}_{rs\cdot\cdot} = 2\alpha_{rs}^2 + \alpha_{rs}\operatorname{Tr}(C) = \tilde{\beta}_{r\cdot\cdot s}, \qquad \tilde{\beta}_{\cdots\cdot} = \operatorname{Tr}(C)^2 + 2|||C|||_2^2,$$
$$\tilde{\beta}_{\cdot s\cdot s'} = \gamma_s\gamma_{s'}(\operatorname{Tr}(C_1)^2 + 2\delta_{ss'}|||C_1|||_2^2), \qquad \tilde{\beta}_{r\cdot r'\cdot} = \lambda_r\lambda_{r'}(\operatorname{Tr}(C_2)^2 + 2\delta_{rr'}|||C_2|||_2^2),$$
$$\tilde{\beta}_{rs\cdot s'} = 2\delta_{ss'}\alpha_{rs}^2 + \alpha_{rs}\gamma_{s'}\operatorname{Tr}(C_1), \qquad \tilde{\beta}_{rsr'\cdot} = 2\delta_{rr'}\alpha_{rs}^2 + \alpha_{rs}\lambda_{r'}\operatorname{Tr}(C_2),$$
$$\tilde{\beta}_{\cdots s} = 2\gamma_s^2|||C_1|||_2^2 + \gamma_s\operatorname{Tr}(C_1)^2\operatorname{Tr}(C_2), \qquad \tilde{\beta}_{r\cdots} = 2\lambda_r^2|||C_2|||_2^2 + \lambda_r\operatorname{Tr}(C_1)\operatorname{Tr}(C_2)^2,$$

where $\delta_{ij} = 1$ if $i = j$, and zero otherwise. The proof is finished by direct calculations. □

C Partial Traces

Letting $\mathcal{S}_1(H_1 \otimes H_2)$ denote the space of trace-class operators on $H_1 \otimes H_2$, we define the partial trace with respect to H_1 as the unique linear operator $\operatorname{Tr}_1 : \mathcal{S}_1(H_1 \otimes H_2) \to \mathcal{S}_1(H_2)$ satisfying $\operatorname{Tr}_1(A \,\tilde{\otimes}\, B) = \operatorname{Tr}(A)B$ for all $A \in \mathcal{S}_1(H_1)$, $B \in \mathcal{S}_1(H_2)$.

proposition C.1. *The operator* Tr_1 *is well-defined, linear, continuous, and satisfies*

$$|||\operatorname{Tr}_1(A)|||_1 \leq |||A|||_1, \quad A \in \mathcal{S}_1(H_1 \otimes H_2) \tag{C.1}$$

Furthermore,

$$\operatorname{Tr}(S\operatorname{Tr}_1(T)) = \operatorname{Tr}((\mathrm{Id}_1 \,\tilde{\otimes}\, S)T), \qquad T \in \mathcal{S}_1(H_1 \otimes H_2), S \in \mathcal{S}_\infty(H_2), \tag{C.2}$$

where Id_1 *is the identity operator on* H_1.

Proof. Let us start by proving that the operator $\operatorname{Tr}_1(\cdot)$ is well defined. By Lemma D.6, the space

$$B_0 = \{\sum_{i=1}^n A_i \,\tilde{\otimes}\, B_i : A_i \in \mathcal{S}_1(H_1), B_i \in \mathcal{S}_1(H_2), n = 1, 2, \ldots\}$$

is a dense subset of $\mathcal{S}_1(H_1 \otimes H_2)$. We therefore only need to show that $\operatorname{Tr}_1(\cdot)$ is continuous on B_0. Let $T \in B_0$, $T = \sum_{i=1}^n A_i \,\tilde{\otimes}\, B_i$. Then, for any $S \in \mathcal{S}_\infty(H_2)$, we have

$$\operatorname{Tr}(S\operatorname{Tr}_1(T)) = \sum_{i=1}^n \operatorname{Tr}(A_i)\operatorname{Tr}(SB_i) = \sum_{i=1}^n \operatorname{Tr}\left((\mathrm{Id}_1 \,\tilde{\otimes}\, S)(A_i \,\tilde{\otimes}\, B_i)\right) =$$
$$= \operatorname{Tr}\left((\mathrm{Id}_1 \,\tilde{\otimes}\, S)\left[\sum_{i=1}^n A_i \,\tilde{\otimes}\, B_i\right]\right) = \operatorname{Tr}\left((\mathrm{Id}_1 \,\tilde{\otimes}\, S)T\right).$$

Hence, using the following formula for the trace norm,

$$|||T|||_1 = \sup\{|\operatorname{Tr}(ST)| : |||S|||_\infty = 1\},$$

we get $|||\operatorname{Tr}_1(T)|||_1 \leq |||T|||_1$ for all $T \in B_0$. Thus $\operatorname{Tr}_1(\cdot)$ can be extended by continuity to $H_1 \otimes H_2$, and (C.1) (of the paper) holds.

Let us now show (C.2) (of the paper). Fix $S \in \mathcal{S}_\infty(H_2)$, and define the linear functionals $g_S, h_S : \mathcal{S}_1(H_1 \otimes H_2) \to \mathbb{R}$ by $g_S(T) = \operatorname{Tr}\left((S \,\tilde{\otimes}\, \mathrm{Id}_2)T\right)$ and $h_S(T) = \operatorname{Tr}(S\operatorname{Tr}_1(T))$, for $T \in \mathcal{S}_1(H_1 \otimes H_2)$. By Hölder's inequality and (C.1) (of the paper), g_S and h_S are both continuous. Since they are equal on the dense subset B_0, they are in fact equal everywhere, and (C.2) (of the paper) follows. □

We can also define $\mathrm{Tr}_2 : \mathcal{S}_1(H_1 \otimes H_2) \to \mathcal{S}_1(H_1)$ analogously. The following result gives an explicit formula for the partial traces of integral operators with continuous kernels.

proposition C.2. *Let $D_s \subset \mathbb{R}^p, D_t \subset \mathbb{R}^q$ be compact subsets, $H_1 = L^2(D_s, \mathbb{R}), H_2 = L^2(D_t, \mathbb{R})$, and $H = L^2(D_s \times D_t, \mathbb{R}) = H_1 \otimes H_2$. If $C \in \mathcal{S}_1(L^2(D_s \times D_t, \mathbb{R}))$ is a positive definite operator with symmetric continuous kernel $c = c(s,t,s',t')$, i.e. $c(s,t,s',t') = c(s',t',s,t)$ for all $s, s' \in D_s, t, t' \in D_t$, and*

$$Cf(s,t) = \iint_{D_s \times D_t} c(s,t,s',t') f(s',t') ds' dt', \qquad f \in L^2(D_s \times D_t, \mathbb{R}),$$

then $\mathrm{Tr}_1(C)$ is the integral operator on $L^2(D_t, \mathbb{R})$ with kernel $k(t,t') = \int_{D_s} c(s,t,s,t') ds$. The analogous result also holds for $\mathrm{Tr}_2(C)$.

Proof. Let $\varepsilon > 0$. By Lemma D.7, we know that there exists an integral operator C' with continuous kernel c' such that $|||C - C'|||_1 \leq \varepsilon/2$ and $\|c - c'\|_\infty \leq \varepsilon/2$, where $C' = \sum_{n=1}^N A_n \tilde{\otimes} B_n$, and each A_n, B_n are finite rank operators, with continuous kernels a_n, respectively b_n, and $\|g\|_\infty = \sup_x |g(x)|$. We have

$$\begin{aligned} \left|\left|\left| \mathrm{Tr}_1(C) - \int_{D_s} c(s,\cdot,s,\cdot) ds \right|\right|\right|_2 \leq {} & |||\mathrm{Tr}_1(C) - \mathrm{Tr}_1(C')|||_2 \\ & + \left|\left|\left| \mathrm{Tr}_1(C') - \int_{D_s} c'(s,\cdot,s,\cdot) ds \right|\right|\right|_2 \\ & + \left|\left|\left| \int_{D_s} c'(s,\cdot,s,\cdot) ds - \int_{D_s} c(s,\cdot,s,\cdot) ds \right|\right|\right|_2 . \end{aligned} \tag{C.3}$$

The first term is bounded $|||\mathrm{Tr}_1(C) - \mathrm{Tr}_1(C')|||_1 \leq |||C - C'|||_1 \leq \varepsilon/2$. The second term is equal to zero since

$$\begin{aligned} \mathrm{Tr}_1\left(\sum_{n=1}^N A_n \tilde{\otimes} B_n\right) &= \sum_{n=1}^N \mathrm{Tr}(A_n) B_n \\ &= \sum_{n=1}^N \int_{D_s} A_n(s,s) ds B_n \\ &= \int_{D_s} \left(\sum_{n=1}^N A_n \tilde{\otimes} B_n \right) (s,\cdot,s,\cdot) ds. \\ &= \int_{D_s} c'(s,\cdot,s,\cdot) ds, \end{aligned}$$

where the second equality comes from the fact that A_n is a finite rank operator (hence trace-class) with continuous kernel. The third term of (C.3) is

$$\left(\iint_{D_t \times D_t} \left(\int_{D_s} \left[c'(s,t,s,t') - c(s,t,s,t') \right] ds \right)^2 dt dt' \right)^{1/2} \leq |D_s| \, |D_t| \, \|c' - c\|_\infty \leq \varepsilon/2,$$

where $|D_s| = \int_{D_s} dx$ and $|D_t| = \int_{D_t} dy$. Therefore,

$$\left|\left|\left| \mathrm{Tr}_1(C) - \int_{D_s} c(s,t,s,t') ds \right|\right|\right|_2 \leq \varepsilon.$$

Since this holds for any $\varepsilon > 0$, $\mathrm{Tr}_1(C)$ is equal to the operator with kernel $k(t,t') = \int_{D_s} c(s,t,s,t')ds$. The proof of the analogous result for $\mathrm{Tr}_2(C)$ is similar. □

The next result states that the partial trace of a Gaussian random trace-class operator is also Gaussian.

proposition C.3. *Let $Z \in \mathcal{S}_1(H_1 \otimes H_2)$ be a Gaussian random element. Then $\mathrm{Tr}_1(Z) \in \mathcal{S}_1(H_2)$ is a Gaussian random element.*

Proof. The proof is finished by noticing that $A \in \mathcal{S}_\infty(H_2)$, we have $\mathrm{Tr}(A\,\mathrm{Tr}_1(Z)) = \mathrm{Tr}\left((\mathrm{Id} \,\widetilde{\otimes}\, A)Z\right)$, where the right-hand side is obviously Gaussian. □

D Background Results

This section presents some background results that are used in the paper. Some references for these results are Zhu (2007), Gohberg & Krejn (1971), Gohberg et al. (1990), Kadison & Ringrose (1997*a*,*b*), Ringrose (1971).

D.1 Tensor Products Hilbert Spaces, and Hilbert–Schmidt Operators

Let H_1, H_2 be two real separable Hilbert spaces, whose inner products are denoted by $\langle\cdot,\cdot\rangle_1$ and $\langle\cdot,\cdot\rangle_2$, respectively. Let $H_1 \otimes H_2$ denote the Hilbert space obtained as the completion of the space of finite linear combinations of simple tensors $u \otimes v, u \in H_1, v \in H_2$ under the inner product

$$\langle u \otimes v, u' \otimes v' \rangle = \langle u, u' \rangle_1 \langle v, v' \rangle_2 .$$

The Hilbert space $H_1 \otimes H_2$ is actually isometrically isomorphic to the space of *Hilbert–Schmidt operators* from H_2 to H_1, denoted by $\mathcal{S}_2(H_2, H_1)$, which consists of all continuous linear operators $T : H_2 \to H_1$ satisfying

$$|||T|||_2^2 = \sum_{n \geq 1} \|Te_n\|^2,$$

where the sum extends over any orthonormal basis $(e_n)_{n\geq 1}$ of H_2. The norm $|||\cdot|||$ is actually induced by the inner-product inner product

$$\langle T, S \rangle_{\mathcal{S}_2} = \sum_{n\geq 1} \langle Te_n, Se_n \rangle_1 , \quad T, S \in \mathcal{S}_2(H_2, H_1),$$

which is independent of the choice of the basis (the space $\mathcal{S}_2(H_2, H_1)$ is therefore itself a Hilbert space). The isomorphism between $H_1 \otimes H_2$ and $\mathcal{S}_2(H_2, H_1)$ is given by the mapping $\Phi : H_1 \otimes H_2 \to \mathcal{S}_2(H_2, H_1)$, defined by $\Phi(u \otimes v) = u \otimes_2 v$ for all $u \in H_1, v \in H_2$, where $u \otimes_2 v(v') = \langle v', v \rangle u$ for $u \in H_1, v, v' \in H_2$. We therefore identify these two spaces, and might write $u \otimes v$ instead of $u \otimes_2 v$ hereafter.

Notice that since $\mathcal{H} = \mathcal{S}_2(H_1) = \mathcal{S}_2(H_1, H_1)$ is itself a Hilbert space, if $A, B \in \mathcal{H}$, the operator $A \bigotimes_2 B \in \mathcal{S}_2(\mathcal{H}, \mathcal{H})$ is defined by $(A \bigotimes_2 B)(C) = \langle C, B \rangle_{\mathcal{S}_2} A$, for $A, B, C \in \mathcal{H}$. Here are some properties of the tensor product $\cdot \otimes_2 \cdot$:

proposition D.1. *Let H be a real separable Hilbert space. For any $u, v, f, g \in H$, $A, B \in \mathcal{S}_2(H)$*

1. *$\cdot \otimes_2 \cdot$ is linear on the left, and conjugate-linear on the right,*

2. $\langle u \otimes_2 v, f \otimes_2 g \rangle_{\mathcal{S}_2} = \langle u, f \rangle \langle g, v \rangle = \langle (u \otimes_2 v) g, f \rangle$,
3. $\langle A, u \otimes_2 v \rangle_{\mathcal{S}_2} = \langle Av, u \rangle = \left\langle v \otimes_2 u, A^\dagger \right\rangle_{\mathcal{S}_2}$,
4. $\operatorname{Tr}(u \otimes_2 v) = \langle u, v \rangle$,
5. $|||u \otimes_2 v|||_1 = |||u \otimes_2 v|||_2 = \|u\| \|v\|$,
6. $(u \otimes_2 v)(f \otimes_2 g) = \langle f, v \rangle \, u \otimes_2 g$,
7. $(u \otimes_2 v)^\dagger = v \otimes_2 u$,
8. $(A \bigotimes_2 B)^\dagger = B \bigotimes_2 A$.

Proof. The proof follows from the definition and the properties of the inner product, and is therefore omitted. □

Recall that for $A \in \mathcal{S}_\infty(H_1), B \in \mathcal{S}_\infty(H_2)$, the operator $(A \ \tilde{\otimes} \ B) \in \mathcal{S}_\infty(H_1 \otimes H_2)$ is defined by the linear extension of

$$(A \ \tilde{\otimes} \ B)(u \otimes v) = Au \otimes Bv, u \in H_1, v \in H_2$$

Furthermore, we have $(A \ \tilde{\otimes} \ B)^\dagger = A^\dagger \ \tilde{\otimes} \ B^\dagger$, $\left|\left|\left| A \ \tilde{\otimes} \ B \right|\right|\right|_\infty = |||A|||_\infty |||B|||_\infty$, and $(A \ \tilde{\otimes} \ B)(C \ \tilde{\otimes} \ D) = (AC \ \tilde{\otimes} \ CD)$ for $A, C \in \mathcal{S}_\infty(H_1), B, D \in \mathcal{S}_\infty(H_2)$. For $u, v \in H_1, f, g \in H_2$, $(u \otimes_2 v) \ \tilde{\otimes} \ (f \otimes_2 g) = (u \otimes f) \bigotimes_2 (v \otimes g)$ If $A \in \mathcal{S}_1(H_1), B \in \mathcal{S}_1(H_2)$, then $A \ \tilde{\otimes} \ B \in \mathcal{S}_1(H_1 \otimes H_2)$,

$$\left|\left|\left| A \ \tilde{\otimes} \ B \right|\right|\right|_1 \leq |||A|||_1 |||B|||_1,$$

and $\operatorname{Tr}(A \ \tilde{\otimes} \ B) = \operatorname{Tr}(A) \operatorname{Tr}(B)$.

In the case $H_1 = L^2\left([-S, S]^d, \mathbb{R}\right)$, $H_2 = L^2\left([0, T], \mathbb{R}\right)$, with $S, T > 0$, if $A \in \mathcal{S}_2(H_1)$, $B \in \mathcal{S}_2(H_2)$ are Hilbert–Schmidt operators (hence also integral operators, with kernels $a(s, s), b(t, t)$, respectively), the operator $A \ \tilde{\otimes} \ B \in \mathcal{S}_2(H_1 \otimes H_2) = \mathcal{S}_2(L^2\left([-S, S]^d \times [0, T], \mathbb{R}\right))$ is also an integral operator with kernel $k(s, t, s', t') = a(s, s')b(t, t')$, that is,

$$(A \ \tilde{\otimes} \ B)u(s, t) = \int_0^T \int_{[-S,S]^d} k(s, t, s', t')u(s', t')dsdt.$$

D.2 Random Elements in Banach Spaces

We understand random elements of a separable Banach space $(B, \|\cdot\|)$ in the Bochner sense (e.g. Ryan 2002). A random element $X \in B$ satisfying $\mathbb{E}\|X\| < \infty$ has a mean $\mathbb{E} X \in B$, which satisfies $S(\mathbb{E} X) = \mathbb{E}(SX)$ for all bounded linear operator $S : B \to B'$, where B' is another Banach space.

D.3 Random Trace-class Operators

If $X \in \mathcal{S}_1(H)$ is a random element satisfying $\mathbb{E}|||X|||_1 < \infty$, i.e. a *random trace-class operator*, then $\mathbb{E} \operatorname{Tr}(AX) = \operatorname{Tr}(A \, \mathbb{E} X)$ for any $A \in \mathcal{S}_\infty(H)$. Furthermore, if $X' \in \mathcal{S}_1(H)$ is another random element such that $\mathbb{E}(|||X|||_1 |||X'|||_1) < \infty$, then

$$\begin{aligned}\mathbb{E}\left(\operatorname{Tr}[AX] \operatorname{Tr}[A'X']\right) &= \mathbb{E}\left(\operatorname{Tr}\left[AX \ \widetilde{\bigotimes} \ A'X'\right]\right) \\ &= \mathbb{E}\left(\operatorname{Tr}\left[(A \ \widetilde{\bigotimes} \ A')(X \ \widetilde{\bigotimes} \ X')\right]\right)\end{aligned}$$

$$= \operatorname{Tr}\left[(A \,\widetilde{\bigotimes}\, A')\ \mathbb{E}\left(X \,\widetilde{\bigotimes}\, X'\right)\right],$$

for any $A, A' \in \mathcal{S}_\infty(H)$.

The second-order structure of a random element $X \in \mathcal{S}_1(H)$ satisfying $\mathbb{E}\,|||X|||_1^2 < \infty$ is encoded by the covariance functional $\Gamma : \mathcal{S}_\infty(H) \times \mathcal{S}_\infty(H) \to \mathbb{R}$, which is defined by

$$\Gamma(A, B) = \operatorname{cov}\left(\operatorname{Tr}[AX], \operatorname{Tr}[BY]\right).$$

Since

$$\begin{aligned}\Gamma(A, B) &= \mathbb{E}\left(\operatorname{Tr}[A(X-\mu)]\operatorname{Tr}[B(X-\mu)]\right)\\ &= \operatorname{Tr}\left[(A \,\widetilde{\bigotimes}\, B)\ \mathbb{E}\left((X-\mu) \,\widetilde{\bigotimes}\, (X-\mu)\right)\right],\end{aligned}$$

the second-order structure is also encoded by the *generalized covariance operator*

$$\Gamma = \mathbb{E}\left((X-\mu) \,\widetilde{\bigotimes}\, (X-\mu)\right) \in \mathcal{S}_1(H \otimes H).$$

proposition D.2. *Let H_1, H_2 be real separable Hilbert spaces. Let $Y \in \mathcal{S}_1(H_1)$ be a Gaussian random element such that $\mathbb{E}\,|||Y|||_1^2 < \infty$. Then, for any $T \in \mathcal{S}_1(H_2)$ fixed, $Y \,\widetilde{\otimes}\, T$ is a Gaussian random element of $\mathcal{S}_1(H_1 \otimes H_2)$.*

Proof. We need to show that for all $S \in \mathcal{S}_\infty(H_1 \otimes H_2)$, $\operatorname{Tr}(S(Y \,\widetilde{\otimes}\, T))$ is Gaussian. This can be reduced to showing that $\operatorname{Tr}(S_n(Y \,\widetilde{\otimes}\, T))$ is Gaussian for all $n \geq 1$, where (S_n) is a sequence of operators in $\mathcal{S}_\infty(H_1 \otimes H_2)$ that converges weakly to S. Indeed, letting $D_n = S_n - S$, we have

$$\begin{aligned}\mathbb{E}\left[\left(\operatorname{Tr}(S_n(Y \,\widetilde{\otimes}\, T)) - \operatorname{Tr}(S(Y \,\widetilde{\otimes}\, T)\right)^2\right] &= \mathbb{E}\left[\operatorname{Tr}(D_n(Y \,\widetilde{\otimes}\, T))^2\right]\\ &= \mathbb{E}\operatorname{Tr}\left((D_n \,\widetilde{\otimes}\, D_n)(Y \,\widetilde{\otimes}\, T \,\widetilde{\otimes}\, Y \,\widetilde{\otimes}\, T)\right)\\ &= \operatorname{Tr}\left((D_n \,\widetilde{\otimes}\, D_n)\ \mathbb{E}(Y \,\widetilde{\otimes}\, T \,\widetilde{\otimes}\, Y \,\widetilde{\otimes}\, T)\right),\end{aligned}$$

where the last equality is valid since

$$\mathbb{E}\left|\left|\left|(D_n \,\widetilde{\otimes}\, D_n)(Y \,\widetilde{\otimes}\, T \,\widetilde{\otimes}\, Y \,\widetilde{\otimes}\, T)\right|\right|\right|_1 \leq |||D_n|||_\infty^2 |||T|||_1^2\ \mathbb{E}\,|||Y|||_1^2 < \infty.$$

Lemma D.4 tells us that $D_n \,\widetilde{\otimes}\, D_n$ converges weakly to zero, and since

$$\left|\left|\left|\ \mathbb{E}(Y \,\widetilde{\otimes}\, T \,\widetilde{\otimes}\, Y \,\widetilde{\otimes}\, T)\right|\right|\right|_1 \leq |||T|||_1^2\ \mathbb{E}\,|||Y|||_1^2 < \infty,$$

Lemma D.5 tells us that $\mathbb{E}\left[\left(\operatorname{Tr}(S_n(Y \,\widetilde{\otimes}\, T)) - \operatorname{Tr}(S(Y \,\widetilde{\otimes}\, T)\right)^2\right] \to 0$ as $n \to \infty$. Therefore, since the space of Gaussian random variables is close under the $L^2(\Omega, \mathbb{P})$ norm, $\operatorname{Tr}(S(Y \,\widetilde{\otimes}\, T))$ is Gaussian if $\operatorname{Tr}(S_n(Y \,\widetilde{\otimes}\, T))$ is Gaussian for all $n \geq 1$. Lemma D.3 tells us that we can choose $S_n = \sum_{i=1}^n A_i \,\widetilde{\otimes}\, B_i$, where $A_i \in \mathcal{S}_\infty(H_1)$ and $B_i \in \mathcal{S}_\infty(H_2)$. In this case,

$$\operatorname{Tr}(S_n(Y \,\widetilde{\otimes}\, T)) = \sum_{i=1}^n \operatorname{Tr}\left((A_i \,\widetilde{\otimes}\, B_i)(Y \,\widetilde{\otimes}\, T)\right) = \sum_{i=1}^n \operatorname{Tr}(A_i Y)\operatorname{Tr}(B_i T),$$

which is Gaussian since $\operatorname{Tr}(A_i Y)$ is Gaussian for each $i \geq 1$. □

D.4 Technical results

Recall that $(T_n)_{n\geq 1} \subset \mathcal{S}_\infty(H)$ is said to converge weakly to $T \in \mathcal{S}_\infty(H)$ if for all $u, v \in H$, $\langle T_n u, v\rangle \to \langle Tu, v\rangle$ as $n \to \infty$.

lemma D.3. *Let H_1, H_2 be real separable Hilbert spaces. For any $S \in \mathcal{S}_\infty(H_1 \otimes H_2)$, there exists a sequences of operators $(S_n)_{n\geq 1} \subset \mathcal{S}_\infty(H_1 \otimes H_2)$ of the form $S_n = \sum_{i=1}^n A_i \widetilde{\otimes} B_i$ with $A_i \in \mathcal{S}_\infty(H_1)$, $B_i \in \mathcal{S}_\infty(H_2)$, such that S_n converges weakly to S.*

Proof. For $S \in \mathcal{S}_\infty(H_1 \otimes H_2)$, define

$$\begin{aligned} S_N &= \sum_{n,n',m,m'=1}^{N} \langle Se_n \otimes f_m, e_{n'} \otimes f_{m'}\rangle \, (e_{n'} \otimes f_{m'}) \otimes_2 (e_n \otimes f_m), \\ &= \sum_{n,n',m,m'=1}^{N} \langle Se_n \otimes f_m, e_{n'} \otimes f_{m'}\rangle \, (e_{n'} \otimes e_n) \widetilde{\otimes} (f_{m'} \otimes f_m), \end{aligned}$$

where $(e_n)_{n\geq 1}$ is an orthonormal basis of H_1, and $(f_m)_{m\geq 1}$ is an orthonormal basis of H_2. First, notice that we have the following equality:

$$\begin{aligned} \langle S_N e_i \otimes f_j, e_k \otimes f_l\rangle &= \sum_{n,m=1}^{N} \sum_{n',m'=1}^{N} \langle Se_n \otimes f_m, e_{n'} \otimes f_{m'}\rangle \langle e_n \otimes f_m, e_i \otimes f_j\rangle \langle e_{n'} \otimes f_{m'}, e_k \otimes f_l\rangle \\ &= \langle Se_i \otimes f_j, e_k \otimes f_l\rangle \mathbf{1}_{\{i\leq N\}} \mathbf{1}_{\{j\leq N\}} \mathbf{1}_{\{k\leq N\}} \mathbf{1}_{\{l\leq N\}}. \end{aligned} \tag{D.1}$$

Therefore, for general $g, h \in H_1 \otimes H_2$, $g = \sum_{i,j\geq 1} \alpha_{ij} e_i \otimes f_j$ and

$$h = \sum_{k,l\geq 1} \beta_{kl} e_k \otimes f_l,$$

we have

$$\begin{aligned} \langle S_N h, g\rangle &= \sum_{i,j\geq 1} \sum_{k,l\geq 1} \alpha_{ij} \beta_{kl} \langle S_N(e_i \otimes f_j), e_k \otimes f_l\rangle \\ &= \sum_{i,j\geq 1} \sum_{k,l\geq 1} \alpha_{ij} \beta_{kl} \langle S(e_i \otimes f_j), e_k \otimes f_l\rangle \mathbf{1}_{\{i\leq N\}} \mathbf{1}_{\{j\leq N\}} \mathbf{1}_{\{k\leq N\}} \mathbf{1}_{\{l\leq N\}} \\ &\qquad \text{(Using (D.1))} \\ &= \left\langle S\left(\sum_{i,j=1}^{N} \alpha_{ij} e_i \otimes f_j\right), \sum_{k,l=1}^{N} \beta_{kl} e_k \otimes f_l \right\rangle. \end{aligned}$$

Therefore, by continuity of the inner product and the continuity of S, we have $\lim_{N\to\infty} \langle S_N h, g\rangle = \langle Sh, g\rangle$. □

lemma D.4. *Let $(S_n)_{n\geq 1} \subset \mathcal{S}_\infty(H)$ be a sequence of operators converging weakly to $S \in \mathcal{S}_\infty(H)$. Then, $S_n \widetilde{\otimes} S_n$ converges weakly to $S \widetilde{\otimes} S \in \mathcal{S}_\infty(H \otimes H)$.*

Proof. For $u, v, z, w \in H$, we have

$$\begin{aligned} \lim_{n\to\infty} \left\langle \left(S_n \widetilde{\otimes} S_n\right)(u \otimes v), z \otimes w\right\rangle &= \lim_{n\to\infty} \langle S_n u, z\rangle \langle S_n v, w\rangle \\ &= \langle Su, z\rangle \langle Sv, w\rangle \end{aligned}$$

$$= \left\langle \left(S \mathbin{\widetilde{\otimes}} S\right)(u \otimes v), z \otimes w \right\rangle.$$

Now for general $g, h \in H \otimes H$, let us write $g = \sum_{i,j\geq 1} \alpha_{ij} e_i \otimes e_j$, $h = \sum_{k,l\geq 1} \beta_{kl} e_k \otimes e_l$, where $(e_i)_{i\geq 1}$ is an orthonormal basis of H, and let $g^N = \sum_{i,j=1}^{N} \alpha_{ij} e_i \otimes e_j$, $h^N = \sum_{k,l=1}^{N} \beta_{kl} e_k \otimes e_l$ for $N \geq 1$. Also, let

$$K = \max\{\sup_{n\geq 1} |||S_n|||_\infty, |||S|||_\infty\},$$

and notice that $K < \infty$ by the uniform boundedness principle (e.g. Rudin 1991). We have

$$\begin{aligned}
\left|\left\langle (S_n \mathbin{\widetilde{\otimes}} S_n)g, h\right\rangle - \left\langle (S \mathbin{\widetilde{\otimes}} S)g, h\right\rangle\right| &\leq \left|\left\langle (S_n \mathbin{\widetilde{\otimes}} S_n)g, h\right\rangle\right| + \left|\left\langle (S_n \mathbin{\widetilde{\otimes}} S_n)g^N, h^N\right\rangle\right| \\
&\quad + \left|\left\langle (S_n \mathbin{\widetilde{\otimes}} S_n)g^N, h^N\right\rangle - \left\langle (S \mathbin{\widetilde{\otimes}} S)g^N, h^N\right\rangle\right| \\
&\quad + \left|\left\langle (S \mathbin{\widetilde{\otimes}} S)g^N, h^N\right\rangle - \left\langle (S \mathbin{\widetilde{\otimes}} S)g, h\right\rangle\right| \\
&\leq K^2\|g\|\left\|h - h^N\right\| + K^2\left\|g - g^N\right\|\left\|h^N\right\| \\
&\quad + \Big| \sum_{i,j,k,l=1}^{N} \alpha_{ij}\beta_{kl} \big[\langle (S_n - S)e_i, e_k\rangle \langle S_n e_j, e_l\rangle \\
&\qquad\qquad + \langle S e_i, e_k\rangle \langle (S_n - S)e_j, e_l\rangle \big] \Big| \\
&\quad + K^2\|g\|\left\|h - h^N\right\| + K^2\left\|g^N - g\right\|\|h\| \\
&\leq 4K^2 \max\{\|g\|, \|h\|\} \max\left\{\left\|h - h^N\right\|, \left\|g - g^N\right\|\right\} \\
&\quad + 2KN^4 (\|g\| + \|h\|) \max_{1\leq i,k\leq N} \{|\langle (S_n - S)e_i, e_k\rangle|\}.
\end{aligned}$$

Now, for any $\varepsilon > 0$, choose $N > 1$ such that $\max\left\{\left\|h - h^N\right\|, \left\|g - g^N\right\|\right\} \leq \varepsilon \left(6K^2 \max\{\|g\|, \|h\|\}\right)^{-1}$. Since N is fixed, we can find an $n' \geq 0$ such that

$$\max_{1\leq i,k\leq N} |\langle (S_n - S)e_i, e_k\rangle| \leq \frac{\varepsilon}{6KN^4(\|g\| + \|h\|)}, \quad \text{for all } n \geq n'.$$

Then, for all $n \geq n'$, we have $|\langle (S_n \mathbin{\widetilde{\otimes}} S_n)g, h\rangle - \langle (S \mathbin{\widetilde{\otimes}} S)g, h\rangle| \leq \varepsilon$, therefore $S_n \mathbin{\widetilde{\otimes}} S_n$ converges weakly to $S \mathbin{\widetilde{\otimes}} S$. □

lemma D.5. *Let $(S_n)_{n\geq 1} \subset \mathcal{S}_\infty(H)$ be a sequence of operators converging weakly to $S \in \mathcal{S}_\infty(H)$. Then, for all $T \in \mathcal{S}_1(H)$, we have*

$$\mathrm{Tr}(S_n T) \to \mathrm{Tr}(ST), \quad n \to \infty.$$

Proof. Let $T = \sum_{l\geq 1} \lambda_l u_l \otimes_2 v_l$ be the singular value decomposition of T. Without loss of generality, $(v_l)_{l\geq 1}$ is an orthonormal basis of H. We have

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{Tr}(S_n T) &= \lim_{n\to\infty} \sum_{l\geq 1} \langle v_l, S_n T v_l\rangle \\
&= \lim_{n\to\infty} \sum_{l\geq 1} \lambda_l \langle v_l, S_n u_l\rangle \\
&= \sum_{l\geq 1} \lambda_l \lim_{n\to\infty} \langle v_l, S_n u_l\rangle
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{l\geq 1} \lambda_l \langle v_l, S u_l\rangle \\
&= \mathrm{Tr}(ST),
\end{aligned}
$$

where the third equality is justified by the dominated convergence theorem since $\sum_{l\geq 1} |\lambda_l| |\langle v_l, S_n u_l\rangle| \leq \sup_{n\geq 1}\{|||S_n|||_\infty\}\, |||T|||_1 < \infty$ by the uniform boundedness theorem □

lemma D.6. *The operators of the form $\sum_{n=1}^N A_n \,\widetilde{\otimes}\, B_n$, where $A_n : H_1 \to H_1$ and $B_n : H_2 \to H_2$ are finite rank operators, and $N < \infty$, are dense in the Banach space $\mathcal{S}_1(H_1 \otimes H_2)$.*

Proof. Let $T \in \mathcal{S}_1(H_1 \otimes H_2)$. Then $T = \sum_{n\geq 1} \lambda_n U_n \otimes_2 V_n$, with convergence in trace norm, where $(\lambda_n)_{n\geq 1}$ is a summable decreasing sequence of positive numbers, and $(U_n)_{n\geq 1} \subset H_1 \otimes H_2$ and $(V_n)_{n\geq 1} \subset H_1 \otimes H_2$ are orthonormal sequences. Each U_n can be written as $U_n = \sum_{l\geq 1} \lambda_{n,l} u_{n,l}^{(1)} \otimes u_{n,l}^{(2)}$, with convergence in the norm of $H_1 \otimes H_2$, that we denote by $\|\cdot\|_2$. Similarly, $V_n = \sum_{l\geq 1} \gamma_{n,l} v_{n,l}^{(1)} \otimes v_{n,l}^{(2)}$. Let

$$
U_n^M = \sum_{l=1}^M \lambda_{n,l} u_{n,l}^{(1)} \otimes u_{n,l}^{(2)} \quad \text{and} \quad V_n^M = \sum_{l=1} \gamma_{n,l} v_{n,l}^{(1)} \otimes v_{n,l}^{(2)}.
$$

Fix $\varepsilon > 0$, and choose N such that $\left|\left|\left| T - \sum_{n=1}^N \lambda_n U_n \otimes_2 V_n \right|\right|\right|_1 \leq \varepsilon/2$. For $M \geq 1$ fixed, we have

$$
\begin{aligned}
\left|\left|\left| T - \sum_{n=1}^N \lambda_n U_n^M \otimes_2 V_n^M \right|\right|\right|_1 &\leq \left|\left|\left| T - \sum_{n=1}^N \lambda_n U_n \otimes_2 V_n \right|\right|\right|_1 \\
&\quad + \left|\left|\left| \sum_{n=1}^N \lambda_n \left[(U_n - U_n^M) \otimes_2 V_n + U_n^M \otimes_2 (V_n - V_n^M) \right] \right|\right|\right|_1 \\
&\leq \varepsilon/2 + \sum_{n=1}^N \lambda_n \left[|||(U_n - U_n^M) \otimes_2 V_n|||_1 + |||U_n^M \otimes_2 (V_n - V_n^M)|||_1 \right] \\
&\leq \varepsilon/2 + \sum_{n=1}^N \lambda_n \left(\|U_n - U_n^M\|_2 \|V_n\|_2 + \|U_n^M\|_2 \|V_n - V_n^M\|_2 \right).
\end{aligned}
$$

Take $M \geq 1$ such that

$$
\max_{n=1,\dots,N} \left\{ \|U_n - U_n^M\|_2, \|V_n - V_n^M\|_2 \right\} \leq \min\left\{ \frac{\varepsilon}{6 \,\mathrm{Tr}\, C}, 1 \right\}.
$$

Then, since U_n, V_n have unit length, and $\|U_n^M\|_2 \leq 1 + \|U_n - U_n^M\| \leq 2$ for $n = 1, \dots, N$, we have

$$
\begin{aligned}
\left|\left|\left| T - \sum_{n=1}^N \lambda_n U_n^M \otimes_2 V_n^M \right|\right|\right|_1 &\leq \varepsilon/2 + 3 \max_{n=1,\dots,N} \left\{ \|U_n - U_n^M\|_2, \|V_n - V_n^M\|_2 \right\} \cdot \left(\sum_{n=1}^N \lambda_n \right) \\
&\leq \varepsilon/2 + 3 \frac{\varepsilon}{6 \,\mathrm{Tr}\, C} \mathrm{Tr}\, C \\
&= \varepsilon.
\end{aligned}
$$

Since $\sum_{n=1}^N \lambda_n U_n^M \otimes_2 V_n^M = \sum_{n=1}^N \sum_{j,l=1}^K \left(\lambda_n \lambda_{n,l} \gamma_{n,j} u_{n,l}^{(1)} \otimes v_{n,j}^{(1)} \right) \widetilde{\otimes} \left(u_{n,l}^{(2)} \otimes v_{n,j}^{(2)} \right)$ the proof is finished. □

If $H = L^2(D_s \times D_t, \mathbb{R})$, we can approximate certain integral operators in a stronger sense:

lemma D.7. *Let $D_s \subset \mathbb{R}^p, D_t \subset \mathbb{R}^q$ be compact subsets, and $C \in \mathcal{S}_1(L^2(D_s \times D_t, \mathbb{R}))$ be a positive definite integral operator with symmetric continuous kernel $c = c(s,t,s',t')$, i.e. $c(s,t,s',t') = c(s',t',s,t)$ for all $s, s' \in D_s, t, t' \in D_t$.*

For any $\varepsilon > 0$, there exists an operator $C' = \sum_{n=1}^N A_n \,\tilde{\otimes}\, B_n$, where $A_n : L^2(D_s, \mathbb{R}) \to L^2(D_s, \mathbb{R}), B_n : L^2(D_t, \mathbb{R}) \to L^2(D_t, \mathbb{R})$ are finite rank operators with continuous kernels a_n, respectively b_n, such that

1. $|||C - C'|||_1 \leq \varepsilon$,
2. $\sup_{s,s' \in D_s, t,t' \in D_t} |c(s,t,s',t') - c'(s,t,s',t')| \leq \varepsilon$, *where c' is the kernel of the operator C',*

Proof. By Mercer's Theorem, there exists continuous orthonormal functions $(U_n)_{n\geq 1} \subset L^2(D_s \times D_t, \mathbb{R})$ and $(\lambda_n)_{n \geq 1} \subset \mathbb{R}$ is a summable decreasing sequence of positive numbers, such that

$$c(s,t,s',t') = \sum_{n\geq 1} \lambda_n U_n(s,t) U_n(s',t'), \tag{D.2}$$

where the convergence is uniform in (s,t,s',t').

Let $C^N = \sum_{n=1}^N \lambda_n U_n \otimes_2 U_n$, and let c^N denote its kernel. Fix $\varepsilon > 0$, and let $\|g\|_\infty = \sup_x |g(x)|$. We have that for N large enough, both $\left|\left|\left|C - C^N\right|\right|\right|_1$ and $\left\|c - c^N\right\|_\infty$ are bounded by $\varepsilon/2$, since C is positive and (D.2) is also its singular value decomposition.

We can now approximate each of the continuous functions $U_n(s,t), n = 1, \ldots, N$, by tensor products of continuous functions (Cheney 1986). Let $U_n^M = \sum_{l=1}^M u_{n,l}^{(1)} \otimes u_{n,l}^{(2)}$, where $u_{n,l}^{(1)} \in L^2(D_s, \mathbb{R}), u_{n,l}^{(2)} \in L^2(D_t, \mathbb{R}), l \geq 1$, are *continuous* functions such that

$$\left\|U_n - U_n^M\right\|_\infty \leq \min\left\{\frac{\varepsilon}{6\kappa \operatorname{Tr} C}, \kappa\right\},$$

where $\kappa = \max_{n=1,\ldots,N} \|U_n\|_\infty$ (notice that $\kappa < \infty$ since each U_n is continuous). Writing $C^{N,M} = \sum_{n=1}^N \lambda_n U_n^M \otimes_2 U_n^M$, and denoting by $c^{N,M}$ its kernel, we have

$$\begin{aligned}
\left\|c^N - c^{N,M}\right\|_\infty &\leq \sum_{n=1}^N \lambda_n \left[\left\|U_n - U_n^M\right\|_\infty \|U_n\|_\infty + \left\|U_n^M\right\|_\infty \left\|U_n - U_n^M\right\|_\infty\right] \\
&\leq 3\kappa \max_{n=1,\ldots,N} \left\|U_n - U_n^M\right\|_\infty \cdot \operatorname{Tr}(C) \\
&\leq \varepsilon/2.
\end{aligned}$$

Furthermore, we also have

$$\begin{aligned}
\left|\left|\left|C^N - C^{N,M}\right|\right|\right|_1 &\leq \sum_{n=1}^N \lambda_n \left[\left|\left|\left|(U_n - U_n^M) \otimes_2 U_n\right|\right|\right|_1 + \left|\left|\left|U_n^M \otimes_2 (U_n - U_n^M)\right|\right|\right|_1\right] \\
&\leq \sum_{n=1}^N \lambda_n \left[\left\|U_n - U_n^M\right\|_2 \|U_n\|_2 + \left\|U_n^M\right\|_2 \left\|U_n - U_n^M\right\|_2\right] \\
&\leq \sum_{n=1}^N \lambda_n \left[\left\|U_n - U_n^M\right\|_\infty \|U_n\|_\infty + \left\|U_n^M\right\|_\infty \left\|U_n - U_n^M\right\|_\infty\right]
\end{aligned}$$

$$\leq \varepsilon/2.$$

Since $C^{N,M} = \sum_{n=1}^{N}\sum_{j,l=1}^{K}\left(\lambda_n u_{n,l}^{(1)} \otimes u_{n,j}^{(1)}\right) \tilde{\otimes} \left(u_{n,l}^{(2)} \otimes u_{n,j}^{(2)}\right)$, the proof is finished. $\square$

E Implementation details

All the implementation details described here are implemented in the R package `covsep` (Tavakoli 2016).

In practice, random elements of $H_1 \otimes H_2$ are first projected onto a truncated basis of $H_1 \otimes H_2$. We shall assume that the truncated basis is of the form $(e_i \otimes f_j)_{i=1,\ldots,d_1; j=1,\ldots,d_2}$, for some $d_1, d_2 < \infty$, where $(e_i)_{i\geq 1} \subset H_1$, respectively $(f_j)_{j\geq 1} \subset H_2$, is an orthonormal basis of H_1, respectively H_2. In this way, one can encode (and approximate) an element $X \in H_1 \otimes H_2$ by a $d_1 \times d_2$ matrix $\mathbf{X} \in \mathbb{R}^{d_1 \times d_2}$, whose (k,l)-th coordinate is given by $\mathbf{X}(k,l) = \langle X, e_k \otimes f_l \rangle, k = 1,\ldots,d_1; l = 1,\ldots,d_2$. We therefore assume from now on that only need to describe the implementation of $T_N(r,s)$ for $H_1 \otimes H_2 = \mathbb{R}^{d_1} \otimes \mathbb{R}^{d_2} = \mathbb{R}^{d_1 \times d_2}$. In this case, $\mathbf{X}$ is a random element of $\mathbb{R}^{d_1 \times d_2}$, i.e. a random $d_1 \times d_2$ matrix, and we observe $\mathbf{X}_1,\ldots,\mathbf{X}_n \overset{\text{i.i.d.}}{\sim} \mathbf{X}$. We have

$$\mathbf{C}_{1,N}(k,k') = \frac{\widetilde{\mathbf{C}}_{1,N}(k,k')}{\sqrt{\mathrm{Tr}(\widetilde{\mathbf{C}}_{1,N})}}, \qquad \mathbf{C}_{2,N}(l,l') = \frac{\widetilde{\mathbf{C}}_{2,N}(l,l')}{\sqrt{\mathrm{Tr}(\widetilde{\mathbf{C}}_{2,N})}},$$

and

$$\begin{aligned}
\widetilde{\mathbf{C}}_{1,N}(k,k') &= \frac{1}{N}\sum_{i=1}^{N}\sum_{l=1}^{d_2}\left(\mathbf{X}_i(k,l) - \overline{\mathbf{X}}(k,l)\right)\left(\mathbf{X}_i(k',l) - \overline{\mathbf{X}}(k',l)\right) \\
&= \left(\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^{\mathsf{T}}\right)_{k,k'} = \sum_{l=1}^{d_2}\mathbf{C}_N(k,l,k',l), \\
\widetilde{\mathbf{C}}_{2,N}(l,l') &= \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{d_1}\left(\mathbf{X}_i(k,l) - \overline{\mathbf{X}}(k,l)\right)\left(\mathbf{X}_i(k,l') - \overline{\mathbf{X}}(k,l')\right) \\
&= \left(\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)^{\mathsf{T}}\left(\mathbf{X}_i - \overline{\mathbf{X}}\right)\right)_{l,l'} = \sum_{k=1}^{d_1}\mathbf{C}_N(k,l,k,l'), \\
\overline{\mathbf{X}}(k,l) &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i(k,l), \\
\mathbf{C}_N(k,l,k',l') &= \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{X}_i(k,l) - \overline{\mathbf{X}}(k,l)\right)\left(\mathbf{X}_i(k',l') - \overline{\mathbf{X}}(k',l')\right),
\end{aligned}$$

for all $k,k' = 1,\ldots,d_1; l,l' = 1,\ldots,d_2$.

The computation of $\widetilde{\mathbf{C}}_{1,N}(k,k')$ using the above formula is not efficient in `R`, when implemented using a double `for` loop. However, if we denote by $\mathbf{A}_k$ the $N \times d_2$ matrix with $(\mathbf{A}_k)_{il} = \mathbf{X}_i(k,l) - \overline{\mathbf{X}}(k,l), n = 1,\ldots,N; k = 1,\ldots,d_1; l = 1,\ldots,d_2$, by $\mathrm{Vec}(\mathbf{A}_k)$ the vector obtained by stacking the columns of $\mathbf{A}_k$ into a vector of length Nd_2, and by

Y_n the n-th row of the $Nd_2 \times d_1$ matrix $A = (\mathrm{Vec}(A_1), \ldots, \mathrm{Vec}(A_{d_1}))$, we get $\overline{Y} = (Nd_2)^{-1} \sum_{n=1}^{Nd_2} = 0$ and

$$(Nd_2)^{-1} \sum_{n=1}^{Nd_2} (Y_n)_k (Y_n)_{k'} = (Nd_2)^{-1} \sum_{i=1}^{n} \sum_{l=1}^{d_2} \left(\mathbf{X}_i(k,l) - \overline{\mathbf{X}}_i(k,l) \right).$$

Therefore $\tilde{\mathbf{C}}_{1,N} = \frac{Nd_2-1}{N} \mathtt{cov}(A)$, where $\mathtt{cov}$ is the standard $\mathtt{R}$ function returning the covariance, and the computation is very fast. The computation of $\tilde{\mathbf{C}}_{2,N}$ can be done similarly.

If we denote by $(\hat{\lambda}_r, \hat{\mathbf{u}}_r)$, respectively $(\hat{\gamma}_s, \hat{\mathbf{v}}_s)$, the r-th eigenvalue/eigenvector pair of $\mathbf{C}_{1,N}$, respectively the s-th eigenvalue/eigenvector pair of $\mathbf{C}_{2,N}$, we have

$$T_N(r,s) = T_N(r,s|X_1,\ldots,X_N) = \sqrt{N} \left[\frac{1}{N} \sum_{i=1}^{N} \left(\hat{\mathbf{u}}_r^{\mathsf{T}} (\mathbf{X}_i - \overline{\mathbf{X}}) \hat{\mathbf{v}}_s \right)^2 - \hat{\lambda}_r \hat{\gamma}_s \right],$$

where $\cdot^{\mathsf{T}}$ denotes matrix transposition. The variance of $T_N(r,s|X_1,\ldots,X_N)$ is estimated by

$$\hat{\sigma}^2(r,s|X_1,\ldots,X_N) = \frac{2\hat{\lambda}_r^2\hat{\gamma}_s^2 \left(\mathrm{Tr}(\mathbf{C}_{1,N})^2 + |||\mathbf{C}_{1,N}|||_2^2 - 2\hat{\lambda}_r \,\mathrm{Tr}(\mathbf{C}_{1,N}) \right) \left(\mathrm{Tr}(\mathbf{C}_{2,N})^2 + |||\mathbf{C}_{2,N}|||_2^2 - 2\hat{\gamma}_s \,\mathrm{Tr}(\mathbf{C}_{2,N}) \right)}{\mathrm{Tr}(\mathbf{C}_{1,N})^2 \,\mathrm{Tr}(\mathbf{C}_{2,N})^2}, \tag{E.1}$$

where $|||\mathbf{A}|||_2^2 = \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} [(\mathbf{A})_{ij}]^2$ for a $d_1 \times d_2$ matrix $\mathbf{A}$.

If $\mathcal{I} = \{1,\ldots,p\} \times \{1,\ldots,q\}$, then $(T_N(r,s))_{(r,s)\in\mathcal{I}}$ is asymptotically a mean zero Gaussian random $p \times q$ matrix, with separable covariance. Its left (row) covariances are consistently estimated by the $p \times p$ matrix $\hat{\Sigma}_{L,\mathcal{I}} = \hat{\Sigma}_{L,\mathcal{I}}(X_1,\ldots,X_N)$ with entries

$$\left(\hat{\Sigma}_{L,\mathcal{I}} \right)_{r,r'} = \frac{\sqrt{2\hat{\lambda}_r\hat{\lambda}_{r'}} \left(\delta_{rr'} \,\mathrm{Tr}(\widehat{C}_{1,N})^2 + \left|\left|\left| \widehat{C}_{1,N} \right|\right|\right|_2^2 - (\hat{\lambda}_r + \hat{\lambda}_{r'}) \,\mathrm{Tr}(\widehat{C}_{1,N}) \right)}{\mathrm{Tr}(\widehat{C}_{1,N}) \,\mathrm{Tr}(\widehat{C}_{2,N})}, \tag{E.2}$$

$r, r' \in \{1,\ldots,p\}$, and it right (column) covariances are consistently estimated by the $q \times q$ matrix $\hat{\Sigma}_{R,\mathcal{I}} = \hat{\Sigma}_{R,\mathcal{I}}(X_1,\ldots,X_N)$ with entries

$$\left(\hat{\Sigma}_{R,\mathcal{I}} \right)_{s,s'} = \frac{\sqrt{2\hat{\gamma}_s\hat{\gamma}_{s'}} \left(\delta_{ss'} \,\mathrm{Tr}(\widehat{C}_{2,N})^2 + \left|\left|\left| \widehat{C}_{2,N} \right|\right|\right|_2^2 - (\hat{\gamma}_s + \hat{\gamma}_{s'}) \,\mathrm{Tr}(\widehat{C}_{2,N}) \right)}{\mathrm{Tr}(\widehat{C}_{1,N}) \,\mathrm{Tr}(\widehat{C}_{2,N})} \tag{E.3}$$

$s, s' \in \{1,\ldots,q\}$,

The computation of $|||D_N|||_2$ can be done without storing the full covariance C_N in memory. The following pseudo-code returns $|||D_N|||_2$:

I. Compute and store $\mathbf{C}_{1,N}$ and $\mathbf{C}_{2,N}$, and set $s = 0$.

II. Replace $\mathbf{X}_n$ by $\mathbf{X}_n - \overline{\mathbf{X}}$ for each $n = 1,\ldots,N$.

III. For $i,k = 1,\ldots,d_1; j,l = 1,\ldots,d_2$,

   (a) Compute $y = N^{-1} \sum_{n=1}^{N} \mathbf{X}_n(i,j)\mathbf{X}_n(k,l)$.

   (b) Set $s = s + (y - \mathbf{C}_{1,N}(i,k)\mathbf{C}_{2,N}(j,l))^2$.

IV. Return s.

The computation of $|||D_N^* - D_N|||_2$ requires a slight modification of the pseudo-code. Given $X_1, \dots, X_N$ and $X_1^*, \dots, X_N^*$,

I. Compute and store $\mathbf{C}_{1,N}$ and $\mathbf{C}_{2,N}$, $\mathbf{C}_{1,N}^*$ and $\mathbf{C}_{2,N}^*$, and set $s = 0$.

II. Replace $\mathbf{X}_n$ by $\mathbf{X}_n - \overline{\mathbf{X}}$, and $\mathbf{X}_n^*$ by $\mathbf{X}_n^* - \overline{\mathbf{X}^*}$ for each $n = 1, \dots, N$.

III. For $i, k = 1, \dots, d_1; j, l = 1, \dots, d_2$,

   (a) Compute $y = N^{-1} \sum_{n=1}^{N} (\mathbf{X}_n(i,j)\mathbf{X}_n(k,l) - \mathbf{X}_n^*(i,j)\mathbf{X}_n^*(k,l))$.

   (b) Set $s = s + \left(y - \mathbf{C}_{1,N}^*(i,k)\mathbf{C}_{2,N}^*(j,l) + \mathbf{C}_{1,N}(i,k)\mathbf{C}_{2,N}(j,l)\right)^2$.

IV. Return s.

Finally, Algorithms 1 and 2 describe the procedure to approximate the p-values for the tests based on parametric and empirical bootstrap, respectively.

Algorithm 1 Parametric Bootstrap p-value approximation for H_N

Given $X_1, \dots, X_N$,

I. compute $\overline{X}$, $C_{1,N} \;\widetilde{\otimes}\; C_{2,N}$, and $H_N = H_N(X_1, \dots, X_N)$.

II. For $b = 1, \dots, B$,

   (a) Create bootstrap samples $\mathbf{X}^b = \{X_1^b, \dots, X_N^b\}$, where $X_i^b \overset{\text{i.i.d.}}{\sim} F\left(\overline{X}, C_{1,N} \;\widetilde{\otimes}\; C_{2,N}\right)$.

   (b) Compute $H_N^b = H_N(\mathbf{X}^b)$,

III. Compute the estimated bootstrap p-value

$$p = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}_{\{H_N^b > H_N\}},$$

where $\mathbf{1}_{\{A\}} = 1$ if A is true, and zero otherwise.

F Additional results from the simulation studies

Figure 6 shows the empirical powers empirical bootstrap version of the tests $\widetilde{G}_N(\mathcal{I})$ for increasing projection subspaces, i.e. for $\mathcal{I} = \mathcal{I}_l, l = 1, 2, 3$, where $\mathcal{I}_1 = \{(1,1)\}, \mathcal{I}_2 = \{(i,j) : i, j = 1, 2\}$ and $\mathcal{I}_3 = \{(i,j) : i = 1, \dots, 4; j = 1, \dots, 10\}$, when data are generated from a multivariate t distribution with 6 degrees of freedom (the Non-Gaussian scenario in the paper). Figure 9 shows the empirical power for the asymptotic test, the parametric and empirical bootstrap tests based on the test statistic $\widetilde{G}_N(\mathcal{I}_2)$, as well as parametric and bootstrap tests based on the test statistics $G_N(\mathcal{I}), \widetilde{G}_N^a(\mathcal{I}_2)$ where $\mathcal{I}_2 = \{(i,j) : i, j = 1, 2\}$. Figure 10 shows the analogous results for the projection set $\mathcal{I}_3$. Tables 2, 3 and 4 give the true levels of the tests for $\mathcal{I} = \mathcal{I}_1, \mathcal{I}_2$, and $\mathcal{I} = \mathcal{I}_3$, respectively.

Figure 7 shows the empirical size and power of the separability test, in the Gaussian scenario, as functions of the projection set

$$\mathcal{I}_{r,s} = \{(i,j) : 1 \le i \le r, 1 \le j \le s\},$$

Algorithm 2 Empirical Bootstrap p-value approximation for H_N

Given $\mathbf{X} = \{X_1, \dots, X_N\}$,

I. Compute $C_{1,N} \,\widetilde{\otimes}\, C_{2,N}$, and $H_N = H_N(\mathbf{X})$.

II. For $b = 1, \dots, B$,

(a) Create the bootstrap sample $\mathbf{X}^b = \{X_1^b, \dots, X_N^b\}$ by drawing with repetition from $X_1, \dots, X_N$.

(b) For each bootstrap sample, compute $\Delta_N^b = \Delta_N(\mathbf{X}^b; \mathbf{X})$.

III. Compute the estimated bootstrap p-value

$$p = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}_{\{\Delta_N^b > H_N\}},$$

where $\mathbf{1}_{\{A\}} = 1$ if A is true, and zero otherwise.

---

for all possible choices of (r, s). The test used is $\widetilde{G}_N(\mathcal{I})$, with distribution approximated by the empirical bootstrap with $B = 1000$. Figure 8 is analogous plot for the Non-Gaussian scenario.

Acknowledgments

We wish to thank the editor, associate editor, and the referees for their comments that have led to an improved version of the paper. We also wish to thank Victor Panaretos for interesting discussions.

Table 2: Empirical size of the testing procedures (with $\alpha = 0.05$), for $\mathcal{I} = \mathcal{I}_1$.

	N=10	N=25	N=50	N=100
Asymptotic Distribution	0.17	0.09	0.08	0.06
Gaussian parametric bootstrap (non-Studentized)	0.22	0.10	0.08	0.08
(diag Studentized)	0.04	0.04	0.06	0.05
(full Studentized)	0.02	0.04	0.05	0.05
Empirical bootstrap (non-Studentized)	0.20	0.11	0.08	0.07
(diag Studentized)	0.10	0.05	0.05	0.06
(full Studentized)	0.10	0.07	0.06	0.06
Gaussian parametric Hilbert–Schmidt	0.07	0.08	0.07	0.07
Empirical Hilbert–Schmidt	0.11	0.05	0.03	0.04

(a) *Gaussian* scenario

	N=10	N=25	N=50	N=100
Asymptotic Distribution	0.29	0.20	0.18	0.15
Gaussian parametric bootstrap (non-Studentized)	0.31	0.21	0.18	0.17
(diag Studentized)	0.08	0.13	0.14	0.15
(full Studentized)	0.08	0.12	0.14	0.14
Empirical bootstrap (non-Studentized)	0.20	0.07	0.06	0.08
(diag Studentized)	0.07	0.06	0.04	0.04
(full Studentized)	0.06	0.04	0.03	0.03
Gaussian parametric Hilbert–Schmidt	0.37	0.51	0.55	0.63
Empirical Hilbert–Schmidt	0.06	0.01	0.01	0.01

(b) *Non-Gaussian* scenario

Table 3: Empirical size of the testing procedures (with $\alpha = 0.05$), for $\mathcal{I} = \mathcal{I}_2$.

	N=10	N=25	N=50	N=100
Asymptotic Distribution	0.43	0.19	0.11	0.09
Gaussian parametric bootstrap (non-Studentized)	0.17	0.09	0.08	0.07
(diag Studentized)	0.04	0.05	0.06	0.05
(full Studentized)	0.02	0.04	0.05	0.04
Empirical bootstrap (non-Studentized)	0.12	0.08	0.07	0.07
(diag Studentized)	0.01	0.04	0.04	0.05
(full Studentized)	0.00	0.01	0.02	0.04
Gaussian parametric Hilbert–Schmidt	0.07	0.08	0.07	0.07
Empirical Hilbert–Schmidt	0.11	0.05	0.03	0.04

(a) *Gaussian* scenario

	N=10	N=25	N=50	N=100
Asymptotic Distribution	0.61	0.37	0.32	0.28
Gaussian parametric bootstrap (non-Studentized)	0.26	0.19	0.17	0.16
(diag Studentized)	0.09	0.12	0.14	0.14
(full Studentized)	0.10	0.16	0.20	0.22
Empirical bootstrap (non-Studentized)	0.10	0.04	0.06	0.07
(diag Studentized)	0.00	0.03	0.03	0.03
(full Studentized)	0.00	0.01	0.01	0.01
Gaussian parametric Hilbert–Schmidt	0.37	0.51	0.55	0.63
Empirical Hilbert–Schmidt	0.06	0.01	0.01	0.01

(b) *Non-Gaussian* scenario

Table 4: Empirical size of the testing procedures (with $\alpha = 0.05$), for $\mathcal{I} = \mathcal{I}_3$.

	N=10	N=25	N=50	N=100
Asymptotic Distribution	1.00	0.98	0.79	0.40
Gaussian parametric bootstrap (non-Studentized)	0.18	0.09	0.08	0.07
(diag Studentized)	0.01	0.01	0.04	0.05
(full Studentized)	0.01	0.02	0.03	0.05
Empirical bootstrap (non-Studentized)	0.10	0.08	0.07	0.07
(diag Studentized)	0.00	0.00	0.01	0.01
(full Studentized)	0.00	0.00	0.00	0.00
Gaussian parametric Hilbert–Schmidt	0.07	0.08	0.07	0.07
Empirical Hilbert–Schmidt	0.11	0.05	0.03	0.04

(a) *Gaussian* scenario

	N=10	N=25	N=50	N=100
Asymptotic Distribution	1.00	1.00	0.98	0.94
Gaussian parametric bootstrap (non-Studentized)	0.28	0.19	0.17	0.17
(diag Studentized)	0.03	0.19	0.23	0.30
(full Studentized)	0.07	0.34	0.53	0.64
Empirical bootstrap (non-Studentized)	0.09	0.04	0.05	0.07
(diag Studentized)	0.00	0.00	0.00	0.00
(full Studentized)	0.00	0.00	0.00	0.00
Gaussian parametric Hilbert–Schmidt	0.37	0.51	0.55	0.63
Empirical Hilbert–Schmidt	0.06	0.01	0.01	0.01

(b) *Non-Gaussian* scenario

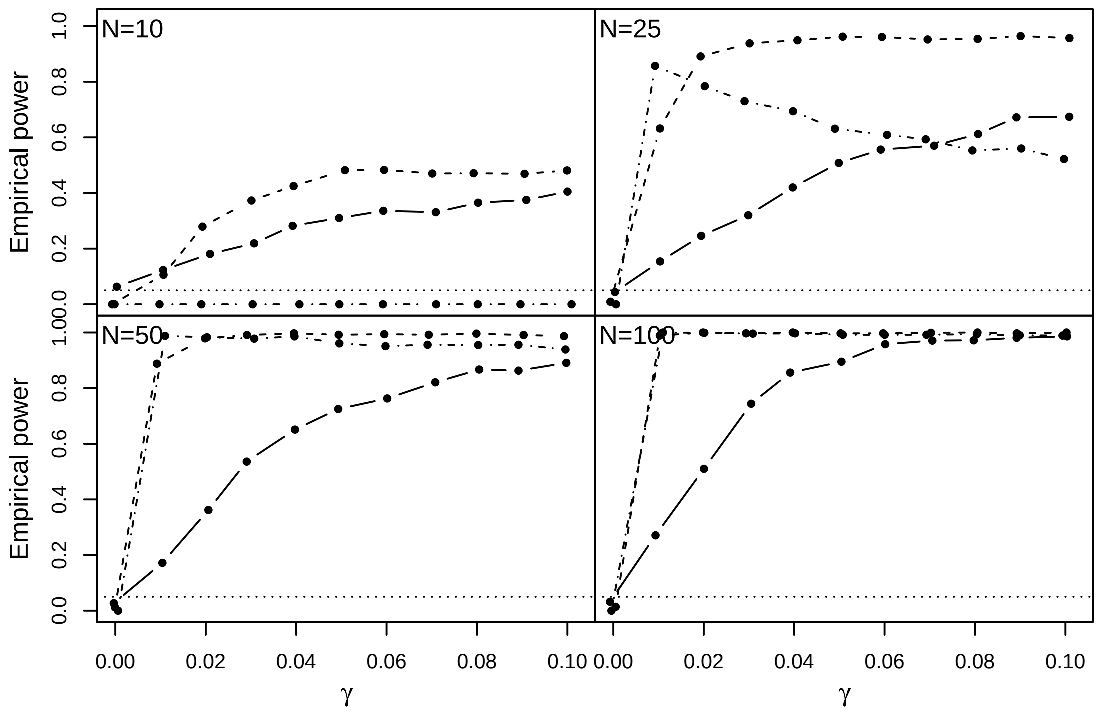


Figure 6: Empirical power of the empirical bootstrap version of $\widetilde{G}_N(\mathcal{I}_l)$, for $l = 1$ (solid line), $l = 2$ (dashed line) and $l = 3$ (dash-dotted line), in the *non-Gaussian* scenario. The horizontal dotted line indicates the nominal level (5%) of the test. Note that the points have been horizontally jittered for better visibility.

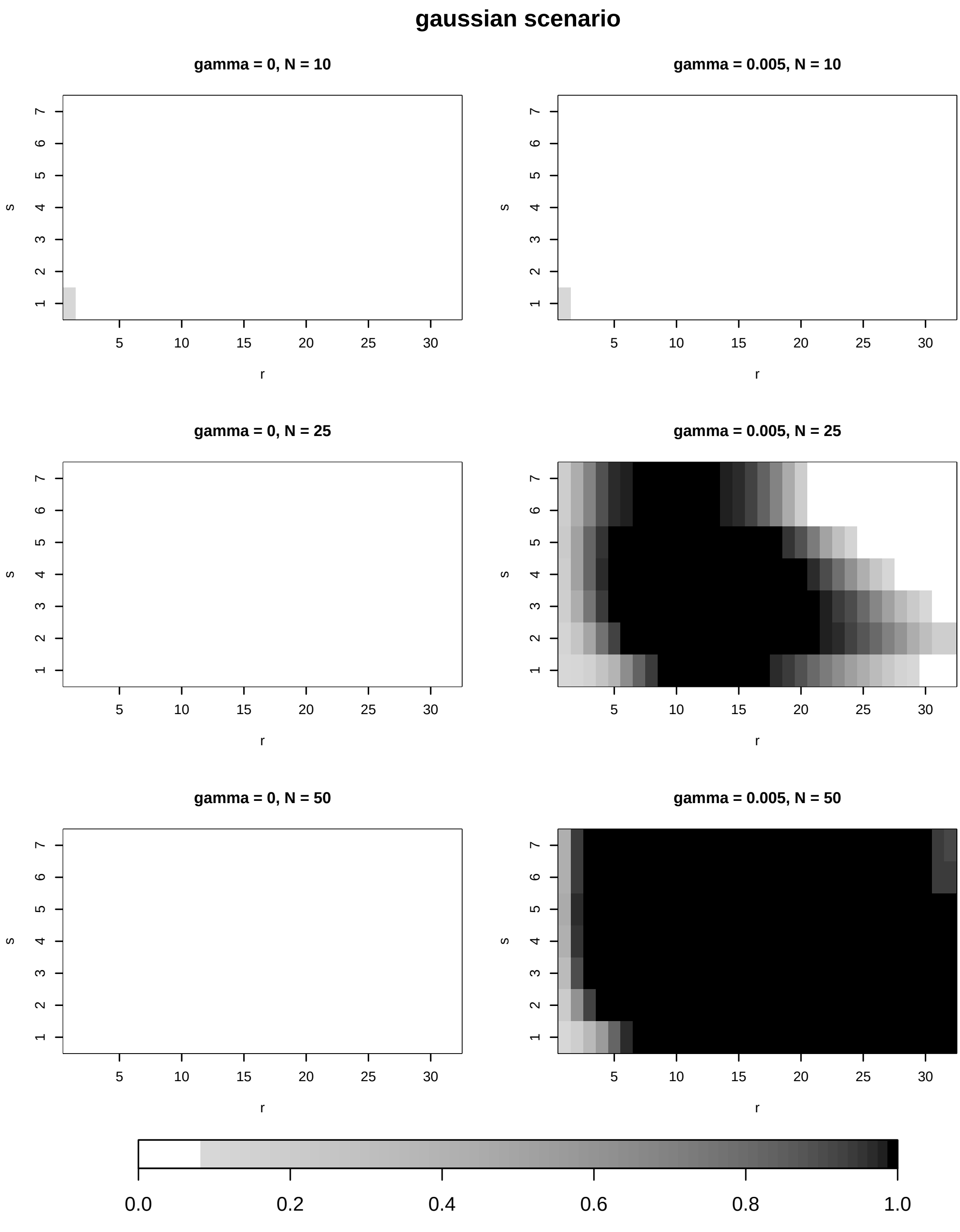


Figure 7: Empirical size (left column) and power (right column) of the separability test as functions of the projection set $\mathcal{I}$. The test used is $\widetilde{G}_N(\mathcal{I})$, with distribution approximated by the empirical bootstrap with $B = 1000$. The left plots, respectively the right plots, were simulated from the Gaussian scenario with $\gamma = 0$, respectively $\gamma = 0.005$. Each row corresponds to a different sample size: $N = 10$ (top), $N = 25$ (middle), $N = 50$ (bottom). Each (r, s) rectangle represents the level/power of the test based on the projection set $\mathcal{I} = \{(i, j) : 1 \leq i \leq r, 1 \leq j \leq s\}$.

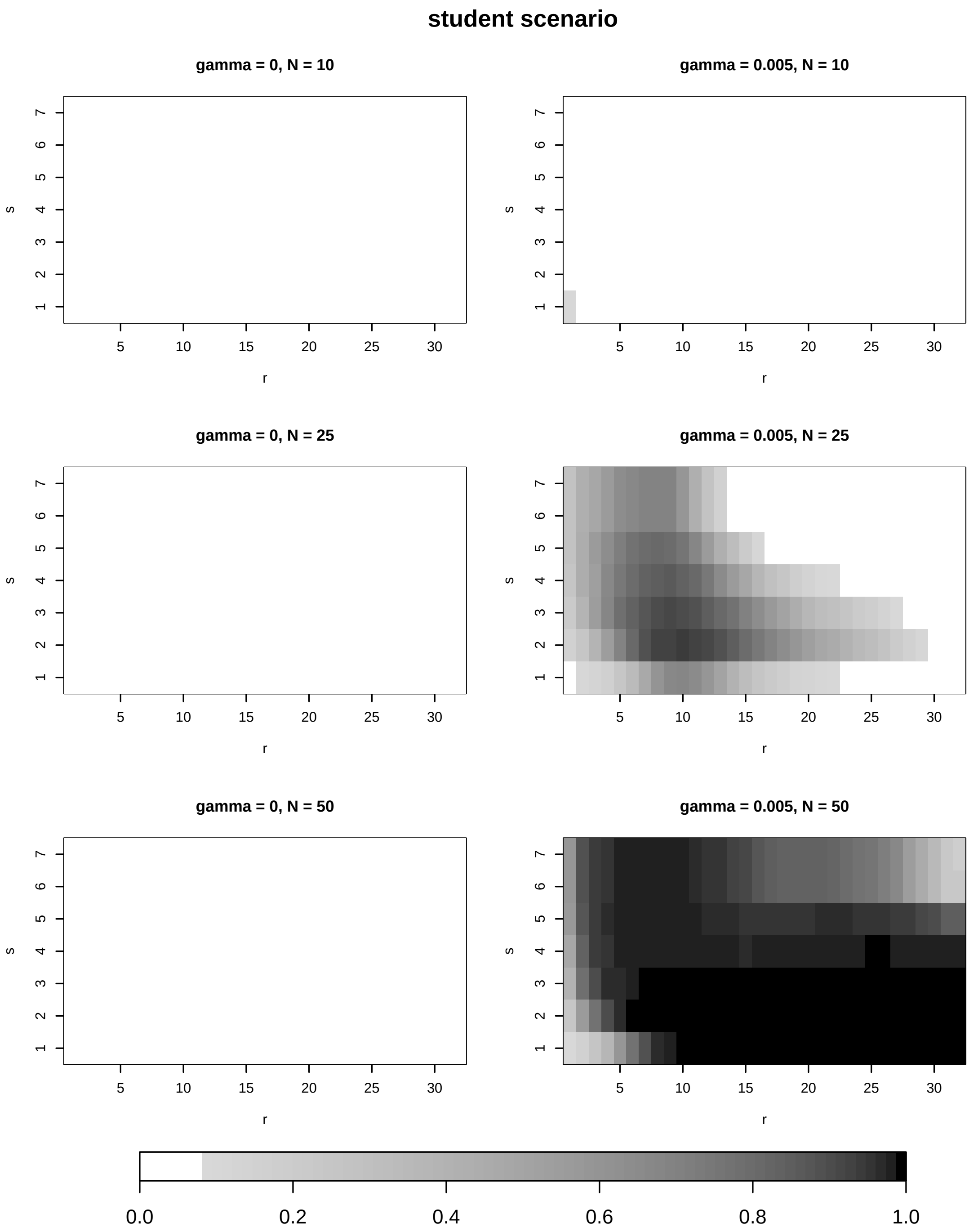
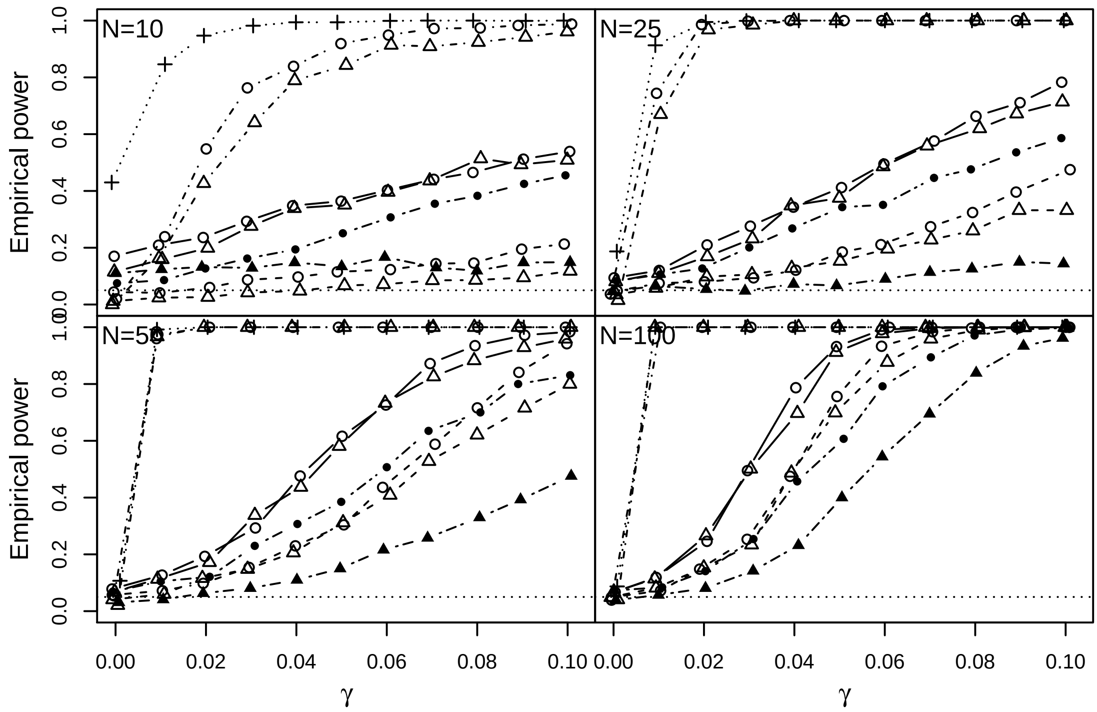
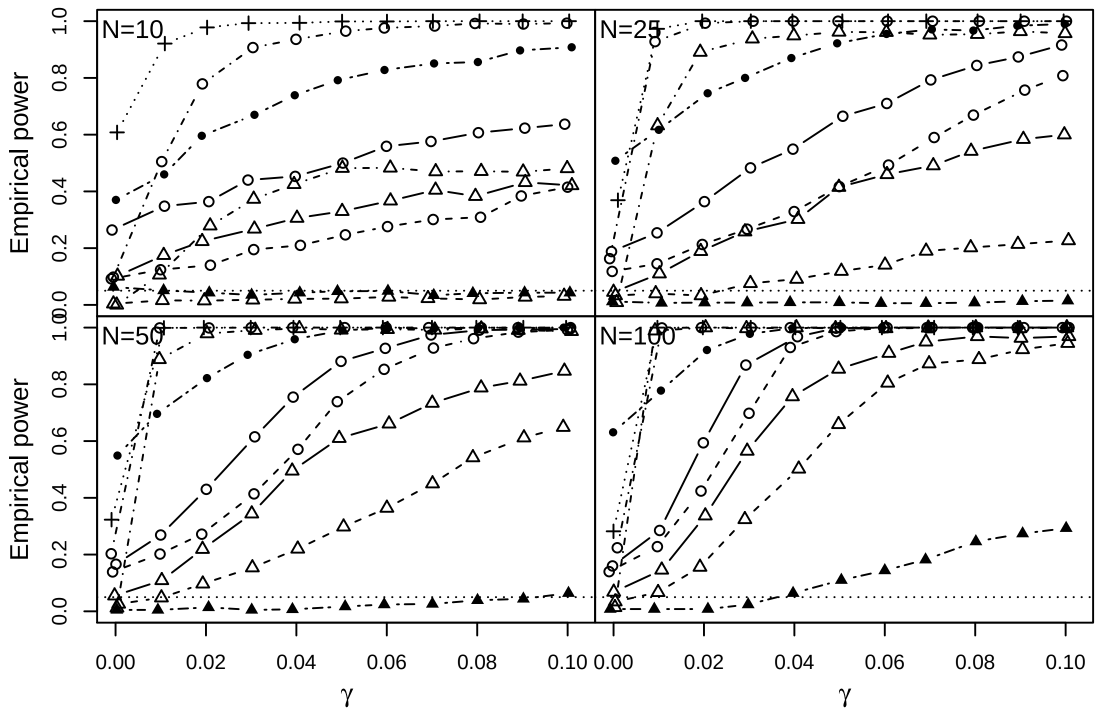


Figure 8: Empirical size (left column) and power (right column) of the separability test as functions of the projection set $\mathcal{I}$. The test used is $\widetilde{G}_N(\mathcal{I})$, with distribution approximated by the empirical bootstrap with $B = 1000$. The left plots, respectively the right plots, were simulated from the Non-Gaussian scenario with $\gamma = 0$, respectively $\gamma = 0.005$. Each row corresponds to a different sample size: $N = 10$ (top), $N = 25$ (middle), $N = 50$ (bottom). Each (r, s) rectangle represents the level/power of the test based on the projection set $\mathcal{I} = \{(i, j) : 1 \leq i \leq r, 1 \leq j \leq s\}$.

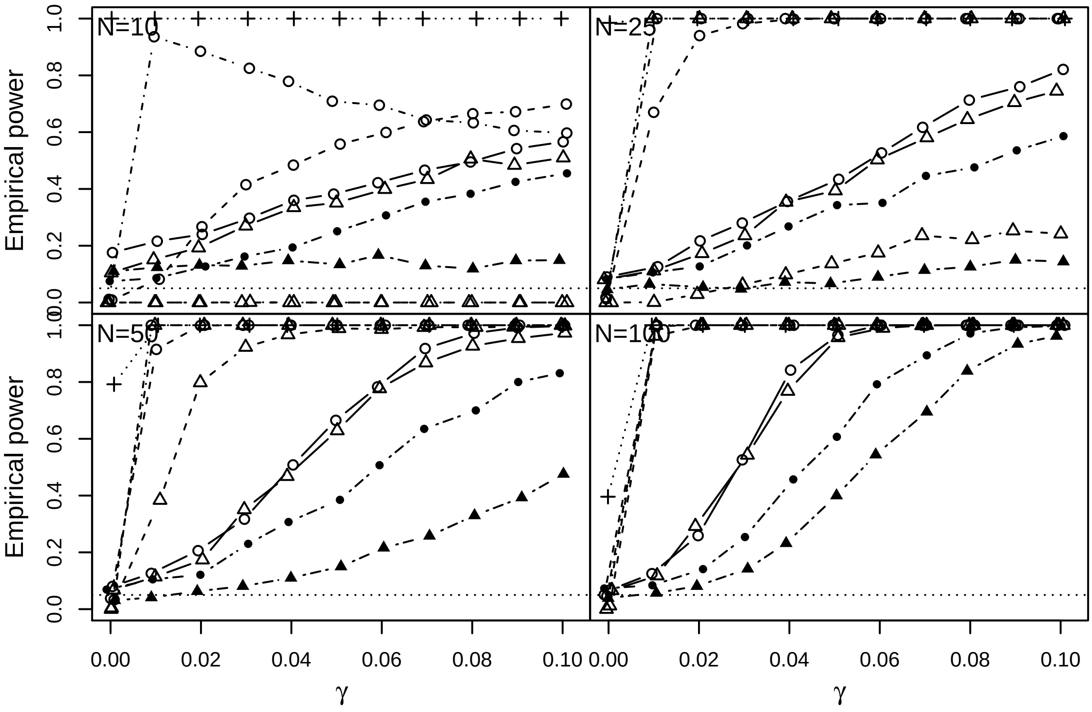


(a) Gaussian scenario

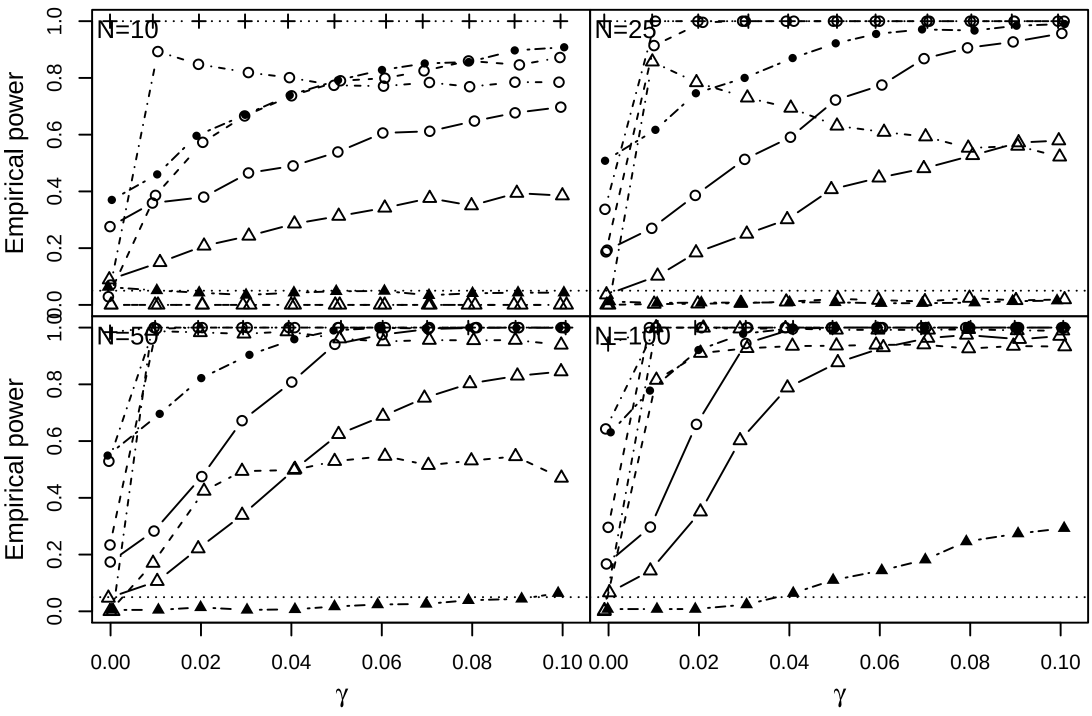


(b) Non-Gaussian scenario

Figure 9: Empirical power of the testing procedures in the *Gaussian* scenario (panel (a)) and *non-Gaussian* scenario (panel (b)), for $N = 10, 25, 50, 100$ and $\mathcal{I} = \mathcal{I}_3$. The results shown correspond to the test (3.1) based on its asymptotic distribution (····+····), the Gaussian parametric bootstrap test $\widetilde{G}_N(\mathcal{I}_2)$ (dash-dotted line with empty circles), $\widetilde{G}^a_N(\mathcal{I}_2)$ (dashed line with empty circles), and $G_N(\mathcal{I}_2)$ (solid line with empty circles), the empirical bootstrap projection tests $\widetilde{G}_N(\mathcal{I}_2)$ (– · –△– · –), $\widetilde{G}^a_N(\mathcal{I}_2)$ (– –△– –), and $G_N(\mathcal{I}_2)$ (—△—), the Gaussian parametric Hilbert–Schmidt test (dash-dotted line with filled circles) and the empirical Hilbert–Schmidt test (dash-dotted line with filled triangles). The horizontal dotted line indicates the nominal level (5%) of the test. Note that the points have been horizontally jittered for better visibility.



(a) Gaussian scenario



(b) Non-Gaussian scenario

Figure 10: Empirical power of the testing procedures in the *Gaussian* scenario (panel (a)) and *non-Gaussian* scenario (panel (b)), for $N = 10, 25, 50, 100$ and $\mathcal{I} = \mathcal{I}_3$. The results shown correspond to the test (3.1) based on its asymptotic distribution (····+····), the Gaussian parametric bootstrap test $\widetilde{G}_N(\mathcal{I}_3)$ (dash-dotted line with empty circles), $\widetilde{G}^a_N(\mathcal{I}_3)$ (dashed line with empty circles), and $G_N(\mathcal{I}_3)$ (solid line with empty circles), the empirical bootstrap projection tests $\widetilde{G}_N(\mathcal{I}_3)$ (– · –△– · –), $\widetilde{G}^a_N(\mathcal{I}_3)$ (– –△– –), and $G_N(\mathcal{I}_3)$ (—△—), the Gaussian parametric Hilbert–Schmidt test (dash-dotted line with filled circles) and the empirical Hilbert–Schmidt test (dash-dotted line with filled triangles). The horizontal dotted line indicates the nominal level (5%) of the test. Note that the points have been horizontally jittered for better visibility.

References

Aston, J. A. D. & Kirch, C. (2012), 'Evaluating stationarity via change-point alternatives with applications to fMRI data', *The Annals of Applied Statistics* **6**(4), 1906–1948.

Billingsley, P. (1999), *Convergence of probability measures*, Wiley Series in Probability and Statistics: Probability and Statistics, second edn, John Wiley & Sons, Inc., New York. A Wiley-Interscience Publication.

Chen, K., Delicado, P. & Müller, H.-G. (2015), 'Modeling function-valued stochastic processes, with applications to fertility dynamics', *Technical report* .

Cheney, E. W. (1986), *Multivariate approximation theory: Selected topics*, SIAM.

Constantinou, P., Kokoszka, P. & Reimherr, M. (2015), 'Testing separability of space–time functional processes', *ArXiv e-prints* .

Cressie, N. & Huang, H.-C. (1999), 'Classes of nonseparable, spatio-temporal stationary covariance functions', *Journal of the American Statistical Association* **94**(448), 1330–1339.

Ferraty, F. & Vieu, P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, Springer.

Fuentes, M. (2006), 'Testing for separability of spatial–temporal covariance functions', *Journal of statistical planning and inference* **136**(2), 447–466.

Garcia, D. (2010), 'Robust smoothing of gridded data in one and higher dimensions with missing values', *Computational Statistics & Data Analysis* **54**(4), 1167–1178.

Genton, M. G. (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., Genton, M. G. & Guttorp, P. (2007), 'Geostatistical space-time models, stationarity, separability, and full symmetry', *Monographs On Statistics and Applied Probability* **107**, 151.

Gohberg, I. C. & Krejn, M. G. (1971), *Introduction à la théorie des opérateurs linéaires non auto-adjoints dans un espace hilbertien*, Dunod, Paris. Traduit du russe par Guy Roos, Monographies Universitaires de Mathématiques, No. 39.

Gohberg, I., Goldberg, S. & Kaashoek, M. A. (1990), *Classes of linear operators. Vol. I*, Operator Theory: Advances and Applications, Birkhäuser Verlag, Basel.

Hall, P. & Hosseini-Nasab, M. (2006), 'On properties of functional principal components analysis', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(1), 109–126.

Horváth, L. & Kokoszka, P. (2012*a*), *Inference for functional data with applications*, Springer series in statistics, Springer, New York, NY.

Horváth, L. & Kokoszka, P. (2012*b*), *Inference for functional data with applications*, Vol. 200, Springer Science & Business Media.

Kadison, R. & Ringrose, J. (1997*a*), *Fundamentals of the Theory of Operator Algebras: Elementary theory*, number v. 1 *in* 'Fundamentals of the Theory of Operator Algebras', American Mathematical Society.

Kadison, R. V. & Ringrose, J. R. (1997*b*), *Fundamentals of the theory of operator algebras. Vol. II*, Vol. 16 of *Graduate Studies in Mathematics*, American Mathematical Society, Providence, RI. Advanced theory, Corrected reprint of the 1986 original.

Lindquist, M. A. (2008), 'The statistical analysis of fMRI data', *Statistical Science* **23**(4), 439–464.

Liu, C., Ray, S. & Hooker, G. (2014), 'Functional Principal Components Analysis of Spatially Correlated Data', *ArXiv e-prints* **1411.4681**.

Lu, N. & Zimmerman, D. L. (2005), 'The likelihood ratio test for a separable covariance matrix', *Statistics & probability letters* **73**(4), 449–457.

Mas, A. (2006), 'A sufficient condition for the CLT in the space of nuclear operators—Application to covariance of random functions', *Statistics & probability letters* **76**(14), 1503–1509.

Mitchell, M. W., Genton, M. G. & Gumpertz, M. L. (2005), 'Testing for separability of space–time covariances', *Environmetrics* **16**(8), 819–831.

Pigoli, D., Aston, J. A. D., Dryden, I. L. & Secchi, P. (2014), 'Distances and inference for covariance operators', *Biometrika* **101**(2), 409–422.

Rabiner, L. R. & Schafer, R. W. (1978), *Digital processing of speech signals*, Vol. 100, Prentice-hall Englewood Cliffs.

Ramsay, J., Graves, S. & Hooker, G. (2009), *Functional Data Analysis with R and MATLAB*, Springer, New York.

Ramsay, J. O. & Silverman, B. W. (2002), *Applied functional data analysis*, Springer Series in Statistics, Springer-Verlag, New York. Methods and case studies.

Ramsay, J. O. & Silverman, B. W. (2005), *Functional data analysis*, Springer Series in Statistics, second edn, Springer, New York.

Ringrose, J. R. (1971), *Compact non-self-adjoint operators*, Van Nostrand Reinhold Co., London.

Rudin, W. (1991), *Functional analysis. 2nd ed.*, 2nd ed. edn, New York, NY: McGraw-Hill.

Ryan, R. A. (2002), *Introduction to tensor products of Banach spaces*, Springer Monographs in Mathematics, Springer-Verlag London, Ltd., London.

Secchi, P., Vantini, S. & Vitelli, V. (2015), 'Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan', *Statistical Methods & Applications* . in press.

Simpson, S. L. (2010), ‘An adjusted likelihood ratio test for separability in unbalanced multivariate repeated measures data’, *Statistical Methodology* **7**(5), 511–519.

Simpson, S. L., Edwards, L. J., Styner, M. A. & Muller, K. E. (2014), ‘Separability tests for high-dimensional, low-sample size multivariate repeated measures data’, *Journal of applied statistics* **41**(11), 2450–2461.

Tang, R. & Müller, H.-G. (2008), ‘Pairwise curve synchronization for functional data’, *Biometrika* **95**(4), 875–889.

Tavakoli, S. (2016), *covsep: Tests for Determining if the Covariance Structure of 2-Dimensional Data is Separable*. R package version 1.0.0.
URL: *https://CRAN.R-project.org/package=covsep*

Tibshirani, R. J. (2014), In praise of sparsity and convexity, *in* X. Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott & J.-L. Wang, eds, ‘Past, Present, and Future of Statistical Science’, Chapman and Hall, pp. 497–506.

Wang, J.-L., Chiou, J.-M. & Mueller, H.-G. (2015), ‘Review of Functional Data Analysis’.
URL: *http://arxiv.org/abs/1507.05135v1*

Worsley, K. J., Marrett, S., Neelin, P., Vandal, A. C., Friston, K. J. & Evans, A. C. (1996), ‘A unified statistical approach for determining significant signals in images of cerebral activation’, *Human brain mapping* **4**(1), 58–73.

Yao, F., Müller, H.-G. & Wang, J.-L. (n.d.), ‘Functional Data Analysis for Sparse Longitudinal Data’, *Journal of the American Statistical Association* **100**(470), 577–590.

Zhu, K. (2007), *Operator theory in function spaces*, Mathematical Surveys and Monographs, second edn, American Mathematical Society, Providence, RI.