Functional data analysis by matrix completion

THÈSE N^{O} 7616 (2017)

PRÉSENTÉE LE 13 JUIN 2017

À LA FACULTÉ DES SCIENCES DE BASE
CHAIRE DE STATISTIQUE MATHÉMATIQUE
PROGRAMME DOCTORAL EN MATHÉMATIQUES

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Marie-Hélène DESCARY

acceptée sur proposition du jury:

Prof. S. Morgenthaler, président du jury
Prof. V. Panaretos, directeur de thèse
Prof. J. Aston, rapporteur
Prof. G. Hooker, rapporteur
Prof. A. Davison, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2017

Remerciements

J'aimerais tout d'abord remercier chaleureusement mon directeur de thèse Victor Panaretos qui m'a tant aidé et encouragé pendant ces quatre belles années à travailler sous sa supervision.

I also want to thank the members of my thesis committee, Professor John Aston, Professor Giles Hooker and Professor Anthony Davison for their thoughtful and constructive comments, and Professor Stephan Morgenthaler who kindly accepted to be the president of my jury.

Pendant mon passage à l'EPFL, j'ai eu la chance de côtoyer des gens formidables, en particulier tous mes (ex-)collègues du groupe SMAT : Andrea, Anirvan, Kate, Mikael, Pavol, Shahin, Susan, Tomas, Tung, Valentina et Yoav. Je vous remercie pour votre soutien et pour les bons moments passés ensemble. Je pense aussi au groupe d'extrémistes : Alix, Claudio, Hélène, Léo, Linda, Raphaël, Sebastian, Sophie et Thomas, qui ont contribué à rendre nos diners en SV aussi agréables. J'aimerais aussi exprimer ma reconnaissance à Yoav et Shahin pour leur amitié et pour l'aide précieuse qu'ils m'ont apportée tout au long de mon doctorat ; à Raphaël et Rémy pour nos belles soirées en trio à Sat ; à Alix, Hélène, Linda, Sophie et Yousra pour nos soupers de filles délirants ; à Susan pour tous nos voyages rocambolesques ; et finalement, à Valentina pour son écoute et ses bons conseils.

Ma vie en Suisse n'aurait jamais été aussi épanouie sans mes amis hors de la communauté statisticienne. Alex, Daniele et Martina, merci d'avoir rendu les deux dernières années de mon doctorat tout simplement inoubliables. Un immense merci à Géraud et Loïc, qui, grâce à un weekend de ski mémorable organisé par l'ACIDE, sont devenus mes «caveurs» préférés. Je tiens aussi à remercier Samuel pour être fou et Valaisan, et à dire un gros merci à Diane pour être la colocataire idéale.

Je remercie également tous mes amis du Québec qui malgré la distance sont toujours bien présents dans ma vie. Merci pour les nombreuses conversations skype, votre amitié est tellement importante pour moi.

Finalement, je souhaite remercier ma famille, qui au cours de ces quatre années passées en Suisse, a su faire mentir le dicton «loin des yeux, loin du coeur», et plus particulièrement mes parents pour leur soutien inconditionnel.

Lausanne, 28 Février 2017 Marie-Hélène Descary

Abstract

Traditional approaches to analysing functional data typically follow a two-step procedure, consisting in first smoothing and then carrying out a functional principal component analysis. The idea underlying this procedure is that functional data are well approximated by smooth functions, and that rough variations are due to noise. However, it may very well happen that localised features are rough at a global scale but still smooth at some finer scale. In this thesis we put forward a new statistical approach for functional data arising as the sum of two uncorrelated components: one smooth plus one rough. We give non-parametric conditions under which the covariance operators of the smooth and of the rough components are jointly identifiable on the basis of discretely observed data: the covariance operator corresponding to the smooth component must be of finite rank and have real analytic eigenfunctions, while the one corresponding to the rough component must have a banded covariance function. We construct consistent estimators of both covariance operators without assuming knowledge of the true rank or bandwidth. We then use them to estimate the best linear predictors of the the smooth and the rough components of each functional datum. In both the identifiability and the inference part, we do not follow the usual strategy used in functional data analysis which is to first employ smoothing and work with continuous estimate of the covariance operator. Instead, we work directly with the covariance matrix of the discretely observed data, which allows us to use results and tools from linear algebra. In fact, we show that the whole problem of uniquely recovering the covariance operator of the smooth component given the one of the raw data can be seen as a low-rank matrix completion problem, and we make great use of a classical relation between the rank and the minors of a matrix to solve this matrix completion problem. The finite-sample performance of our approach is studied by means of simulation study.

Key words: analyticity, banded covariance function, functional principal component analysis, low-rank, matrix completion, scale, smoothing.

Contents

List of Figures

List of Tables

Introduction

Functional data analysis (FDA) is a branch of statistics concerned with complex data objects, such as curves and surfaces, that can be seen as realizations of a random function. Functional data arise naturally in many applications, and are becoming more and more accessible and consequently prevalent, due to the advances of technology in the last decades. Growth curves and temperature curves are two classical examples of functional data defined on an interval, while satellite images are examples of functional data defined on a surface, which in this case would be a subset of the sphere. The two main features of functional data are that they belong to an infinite-dimensional space, and that they are assumed to have some smoothness properties. The second feature compensates for the first in the sense that smoothness transforms the curse of dimensionality induced by the intrinsic infinite nature of the data to a blessing of dimensionality. The term *functional data analysis* was introduced by Ramsay [Ram82] and Ramsay and Dalzell [RD91], but its origin can be traced back to Grenander [Gre50] and Rao [Rao58].

The classical setup in FDA is to suppose that we have a collection of independent realizations of an $L^2([0,1])$-valued random function X, and that each of these realizations is observed at discrete points with measurement errors. The standard model for the measurements errors is to consider them as i.i.d. random variables, or equivalently as the discrete observations of a white noise process. There exist two main traditional approaches to estimating the covariance operator (or equivalently the covariance function) of the random function X from such a sample. Both consist in a two-step procedure, where the first step is a smoothing step to transform the discretely observed data into functions, and the second step carries out a functional principal component analysis in order to approximate each curve of the sample by its r-term Karhunen-Loève expansion. The idea underlying this two-step procedure is that the observed random function X arises as the sum of two processes: a true signal Y which is a process of smoothness class C^k, $k \geq 2$, having essentially a finite rank r, plus a noise component W which is essentially a white noise process. This means that any variations that are not at least of smoothness class C^2 or that correspond to fluctuations around the eigenfunctions of order at least $r+1$ are attributed to the noise component and thus discarded. However, it might very well happen that a rough process is not pure noise but has some structure at a finer scale which represents local variations in the data.

In this thesis, we want to take into account such rough variations, and not simply consider

them as noise, as is usually done in FDA. We thus introduce a new model where the random function X is equal to the sum of two uncorrelated processes: a process Y which is smooth and represents the global variations of the data, and a process W which is rough and represents the local variations, or equivalently the variations at scale at most $\delta \in (0,1)$, of the data. It turns out that to handle this new model, a new statistical approach is needed (Section 1.3). The main contribution of this thesis is to develop such a new approach in terms of both theoretical (Chapter 2) and practical (Chapter 3) aspects. These results, as presented in Chapter 2 and 3, are largely based on Descary and Panaretos [DP16].

In Chapter 1, we give an overview of functional data analysis. We first formally define random variables taking values in a function space, and we then draw the connection between random functions and continuous stochastic processes (Section 1.1.2). In Section 1.1.3, we define the basic objects of FDA such as the mean function and the covariance operator of a random function, and its most important tool, the functional principal component analysis which is based on the celebrated Karhunen–Loève expansion. The problem of statistical inference based on a sample of fully observed curves is presented in Section 1.1.4. Section 1.2 is devoted to the presentation of the standard model for discretely observed data and of two main approaches to carry out statistical inference based on a sample of densely or sparsely observed curves. We conclude the chapter with a discussion on the limits of the standard model, which highlights the need for a more flexible model and for a new approach to analyzing functional data (Section 1.3).

In the second chapter of this thesis, we treat theoretical aspects, which are two-fold. We first determine nonparametric conditions under which the smooth and the rough variations of the data are jointly identifiable, given a K-resolution version of X, i.e., given K measurements on the function X (Section 2.2). Once we have a well-posed problem, we consider the statistical inference part, which is done given a sample of n independent realizations of X, each observed on the same grid of K points. In Section 2.3, we define consistent estimators of the covariance operators of Y and W. Using their spectral decomposition then allows us to separate the smooth part Y from the rough part W of the data (Section 2.4). In both the identifiability and the inference part, we do not follow the usual strategy used in FDA, which is to first employ smoothing and work with a continuous estimate of the covariance operator. Instead, we work directly with the covariance matrix of the discretely observed data, which allows us to use results and tools from linear algebra. In fact, we show that the whole problem of uniquely recovering the covariance operator of Y given the one of X can be seen as a low-rank matrix completion problem, and we make great use of a classical relation between the rank and the minors of a matrix to solve this problem.

Chapter 3 is devoted to the practical aspects and to the study of the performance of our new approach. We first describe how our new method can be implemented in practice to analyse functional data (Section 3.1). The estimator of the covariance of Y is obtained with a two-step procedure: we first select its rank with a scree plot approach, and we then compute the estimator using a minimisation problem easily solved in MATLAB. On the other hand,

the estimator for the covariance operator of W is obtained using an alternating projection algorithm. In Section 3.2, we study the finite-sample performance of our method. On the one hand, we carry out a sensitivity analysis to probe the effect of the starting value on the minimisation algorithm used to compute the estimator of the covariance of Y, study the effect of the sample size n and of the grid size K on the performance of our method, and study the performance of the scree plot approach for the selection of the rank. On the other hand, we study the performance of our method by comparing our results with those obtained by three other methods used in FDA. We conclude the chapter with the analysis of a real dataset.

1 Functional Data Analysis

In this chapter we give a succinct overview of some key aspects of Functional Data Analysis. We first introduce the basic concepts and tools in Section 1.1, we then present two general approaches to analyzing discretely observed functional data (Section 1.2). In Section 1.3, we conclude the chapter with a discussion on the restrictions of the standard model in FDA, which leads to the formulation of a new model for functional data, for which a novel approach will be developed in this thesis. The literature on functional data analysis has now become very rich, and for a detailed presentation of the subject one can consult, for example, Ramsay and Silverman [RS05], Ferraty and Vieu [FV06], Horvath and Kokoszka [HK12] and Hsing and Eubank [HE15]; for a brief review one can consult Wang et al. [WCM16].

1.1 The basis of FDA

Functional data analysis is concerned with observations that can be seen as arising from a random variable which takes its values in a function space which is typically the separable Hilbert space $L^2([0,1])$. In Section 1.1.1 we give a brief overview of operator theory in a Hilbert space, we then give the formal definition of random functions in $L^2([0,1])$ and draw the connection between them and continuous time stochastic processes (Section 1.1.2). In Sections 1.1.3 and 1.1.4, we define the functional version of Principal Component Analysis and its empirical counterpart, which play a far more important role in FDA than in multivariate analysis.

1.1.1 Brief overview of operator theory in Hilbert spaces

Let $\mathscr{H}$ be a separable Hilbert space with the inner product $\langle \cdot, \cdot \rangle_{\mathscr{H}}$ which induces the norm $\| \cdot \|_{\mathscr{H}}$. A linear operator $\mathscr{T} : \mathscr{H} \to \mathscr{H}$ is said to be bounded, or equivalently continuous, if

$$\| \mathscr{T} \|_B := \sup\{ \| \mathscr{T} x \|_{\mathscr{H}} : x \in \mathscr{H} \text{ and } \| x \|_{\mathscr{H}} \leq 1 \} < \infty.$$

The space of bounded operators on $\mathscr{H}$ equipped with the norm $\|\cdot\|_B$ forms a Banach space, which we denote by $B(\mathscr{H})$. An operator $\mathscr{T} \in B(\mathscr{H})$ is called

1. *self-adjoint* if it is equal to its adjoint $\mathscr{T}^* \in B(\mathscr{H})$ defined as the unique operator such that
$$\langle \mathscr{T} f, g\rangle_{\mathscr{H}} = \langle f, \mathscr{T}^* g\rangle_{\mathscr{H}}, \text{ for any } f, g \in \mathscr{H},$$
2. *positive semidefinite* if
$$\langle \mathscr{T} f, f\rangle_{\mathscr{H}} \geq 0, \text{ for any } f \in \mathscr{H},$$
3. *compact* if for any bounded sequence $\{f_n\}$ in $\mathscr{H}$, there exists a convergent subsequence of $\{\mathscr{T} f_n\}$.

The eigen-decomposition of a self-adjoint, positive semidefinite and compact operator $\mathscr{T} \in B(\mathscr{H})$ is defined in the following proposition.

Proposition 1.1.1. *Let $\mathscr{T} \in B(\mathscr{H})$ be self-adjoint, positive semidefinite and compact. Then its spectral decomposition is given by*

$$\mathscr{T} = \sum_{i=1}^{\infty} \lambda_i (v_i \otimes v_i) = \sum_{i=1}^{\infty} \lambda_i \langle v_i, \cdot\rangle v_i, \tag{1.1.1}$$

where $\{\lambda_i\}_{i=1}^{\infty}$ is the set of eigenvalues of $\mathscr{T}$ such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, and $\{v_i\}_{i=1}^{\infty}$ is the set containing the corresponding orthonormal eigenfunctions, i.e., $\mathscr{T} v_i = \lambda_i v_i$ for any i. Moreover, the set of non-zero eigenvalues is either finite or forms a sequence that decreases toward zero and the associated set of eigenvalues form a basis for $\overline{Im(\mathscr{T})}$.

An important subclass of the collection of bounded and compact operators is the *Hilbert–Schmidt operators*, where an operator $\mathscr{T} \in B(\mathscr{H})$ is Hilbert-Schmidt (HS) if it satisfies the inequality

$$\|\mathscr{T}\|_{HS}^2 := \sum_{i=1}^{\infty} \|\mathscr{T} e_i\|_{\mathscr{H}}^2 < \infty,$$

for an arbitrary orthonormal basis $\{e_i\}$ of $\mathscr{H}$. The space of Hilbert–Schmidt operators, denoted by $B_{HS}(\mathscr{H})$, is itself a separable Hilbert space when equipped with the inner product defined as

$$\langle \mathscr{T}_1, \mathscr{T}_2\rangle_{HS} = \sum_{i=1}^{\infty} \langle \mathscr{T}_1 e_i, \mathscr{T}_2 e_i\rangle_{\mathscr{H}},$$

where again $\{e_i\}$ can be taken as any orthonormal basis of $\mathscr{H}$.

Moreover, a HS operator $\mathscr{T}$ is said to be a *trace class operator*, or a *nuclear operator*, if it satisfies the inequality

$$\|\mathscr{T}\|_{TR} := \sum_{i=1}^{\infty} \langle (\mathscr{T}^* \mathscr{T})^{1/2} e_i, e_i\rangle_{\mathscr{H}} < \infty,$$

for an arbitrary orthonormal basis $\{e_i\}$ of $\mathscr{H}$, where the square root of the positive semidefinite operator $\mathscr{T}^*\mathscr{T} \in B(\mathscr{H})$ is defined as the unique positive semidefinite operator $\mathscr{A} \in B(\mathscr{H})$, such that $\mathscr{A}^2 = \mathscr{T}^*\mathscr{T}$. The space of trace class operators, denoted by $B_{TR}(\mathscr{H})$, equipped with the norm $\|\mathscr{T}\|_{TR}$ forms a Banach space.

The relations between the three spaces of operators described in this section and their associated norm are given respectively by $B_{TR}(\mathscr{H}) \subset B_{HS}(\mathscr{H}) \subset B(\mathscr{H})$ and $\|\mathscr{T}\|_{TR} \geq \|\mathscr{T}\|_{HS} \geq \|\mathscr{T}\|_B$. Note that if $\mathscr{T}$ can be written as in (1.1.1), then taking the set of eigenfunctions $\{v_i\}$ as the orthonormal basis $\{e_i\}$ yields $\|\mathscr{T}\|_{HS} = \sqrt{\sum_{i=1}^{\infty} \lambda_i^2}$ and $\|\mathscr{T}\|_{TR} = \sum_{i=1}^{\infty} \lambda_i$.

1.1.2 The space L^2 and its random elements

As already mentioned, the typical approach in FDA is to assume that the random variables of interest are random elements of the separable Hilbert space $L^2([0,1])$, which we recall is the space of real-valued measurable and square integrable functions on $[0,1]$, with the inner product defined as $\langle f, g\rangle_{L^2} = \int_0^1 f(t)g(t)dt$. An $L^2([0,1])$-valued random element X is defined formally as a measurable map from a probability space $(\Omega, \mathscr{A}, \mathbb{P})$ to $L^2([0,1], B([0,1]), \mu)$, with $B([0,1])$ the Borel σ-field and μ the Lebesgue measure. Recall that the space L^2 is in fact a quotient space, which means that $f \in L^2([0,1])$ represents an equivalence class of functions and thus that $f = g$ in $L^2([0,1])$ means that

$$\|f - g\|_{L^2}^2 = \langle f - g, f - g\rangle_{L^2} = \int_0^1 [f(t) - g(t)]^2 dt = 0.$$

We thus have that an $L^2([0,1])$-valued random function X is an abstract object for which it does not make sense to consider its evaluation $X(t)$ for a specific point $t \in [0,1]$. For the expression $X(t)$ to make sense, i.e., to have well-defined point evaluations of X, one needs to either assume that X belongs to a Reproducing Kernel Hilbert Space (RKHS) (Berlinet and Thomas-Agnan [BTA11]), or to consider the function X as a stochastic process, which is the approach that we adopt and that we now describe.

Let $X : \Omega \to L^2([0,1])$ be a continuous stochastic process on a probability space $(\Omega, \mathscr{A}, \mathbb{P})$, such that $X(\omega) = \{X_\omega(t) : t \in [0,1]\}$. We say that X is a *second-order process* if its mean and covariance functions, defined respectively by

$$m(t) = \mathbb{E}[X(t)], \qquad \rho(s,t) = \mathbb{E}[(X(t) - m(t))(X(s) - m(s))], \quad s,t \in [0,1], \tag{1.1.2}$$

are well-defined. A second-order process X is said to be *mean-square continuous* if

$$\lim_{n\to\infty} \mathbb{E}[X(t_n) - X(t)]^2 = 0,$$

for any $t \in [0,1]$ and any sequence $\{t_n\}$ in $[0,1]$ converging to t, or equivalently, if its mean and covariance functions are continuous. Note that X being a stochastic process implies that $X(t)$ is a random variable for any $t \in [0,1]$, but does not imply that X is itself a random

element of the function space $L^2([0,1])$ equipped with the Borel σ-algebra. However, if X is a mean-square continuous stochastic process with continuous sample paths $t \to X_\omega(t)$ for any $\omega \in \Omega$, then X is jointly measurable, and by Hsing and Eubank ([HE15, Theorem 7.4.1]) we know that X is a random element of $L^2([0,1])$.

In the remainder of this thesis, we always suppose that an $L^2([0,1])$-valued random function can also be seen as a second-order mean-square continuous process with continuous sample paths. This means that point evaluation of a $L^2([0,1])$-valued random function X is well defined, and thus that its mean and covariance functions can be expressed as in (1.1.2), where $m \in L^2([0,1])$ and $\rho \in L^2([0,1] \times [0,1])$ are continuous functions. Note that X being a second-order process implies that $\mathbb{E}\|X\|_{L^2}^2 < \infty$.

1.1.3 Functional principal component analysis

Principal Component Analysis (Jolliffe [Jol02]) is a common and useful tool used in multivariate analysis to explain the principal modes of variation of a random vector. The idea of a functional analogue, called functional PCA (fPCA), was introduced by Grenander [Gre50], Karhunen [Kar46], Loève [Loe46] and Rao [Rao58] and developed in a statistical framework in Dauxois [DPR82]. Besides explaining the main sources of variation of a random function, fPCA is mainly used as a dimension reduction tool, and as such plays a central role in FDA. In an analoguous way to multivariate PCA, which is based on the eigenstructure of a covariance matrix, fPCA is based on the eigenstructure of a covariance operator. This object thus plays a very important role in FDA and is defined as follow.

Definition 1.1.1. *Let X be an $L^2([0,1])$-valued random variable ($\mathbb{E}\|X\|_{L^2}^2 < \infty$) with mean function m and covariance function ρ. Then the covariance operator $\mathscr{R}$ of X is the integral operator*

$$\begin{aligned} \mathscr{R} : L^2([0,1]) &\to L^2([0,1]) \\ f &\to \int_0^1 \rho(\cdot, s) f(s) ds, \end{aligned}$$

where the covariance function ρ plays the role of a kernel function. We can also express it as $\mathscr{R} = \mathbb{E}[(X-m) \otimes (X-m)]$, where $(u \otimes v) f = \langle u, f \rangle_{L^2} v$, for $u, v, f \in L^2([0,1])$.

The operator $\mathscr{R}$ is compact (e.g., Hsing and Eubank [HE15, Theorem 4.6.2]). Moreover, since the covariance function ρ is positive semidefinite ($\int_0^1 \int_0^1 \rho(s,t) f(s) f(t) dt ds \geq 0$ for any $f \in L^2([0,1])$) and symmetric ($\rho(s,t) = \rho(t,s)$), $\mathscr{R}$ is positive semidefinite and self-adjoint. Using Proposition 1.1.1, we can write

$$\mathscr{R} = \sum_{i=1}^{\infty} \theta_i (\phi_i \otimes \phi_i) = \sum_{i=1}^{\infty} \theta_i \langle \phi_i, \cdot \rangle \phi_i, \tag{1.1.3}$$

where $\{\theta_i, \phi_i\}$ is the set of eigenvalues and eigenvectors of $\mathscr{R}$. The following result is known as Mercer's Theorem, and tells us that we can write the covariance function ρ with an expansion similar as the one for $\mathscr{R}$, in a strong uniform convergence sense.

Proposition 1.1.2 (Mercer's Theorem, e.g., Grenander [Gre81])**.** *Let $\mathscr{R}$ be the covariance operator associated with the continuous covariance function ρ, and let $\{\theta_i, \phi_i\}_{i=1}^{\infty}$ be its spectrum. Then*

$$\rho(s,t) = \sum_{i=1}^{\infty} \theta_i \phi_i(s)\phi_i(t),\ s,t \in [0,1],$$

where the convergence of the sum holds uniformly and absolutely on $[0,1]^2$.

For simplicity of notation and since we will be primarily interested in second-order structure, from now on we will assume without lost of generality that the mean function m is equal to zero. Using the representation given in Proposition 1.1.2 for ρ, one can show easily that $\|\mathscr{R}\|_{HS}^2 = \sum_{i=1}^{\infty} \theta_i^2 = \int_0^1 \int_0^1 \rho^2(s,t)\,ds\,dt$, and from $\mathbb{E}\|X\|_{L^2}^2 < \infty$, we have that the right hand side of this expression is finite, and then that $\mathscr{R}$ is a Hilbert–Schmidt operator. Moreover, since $\mathscr{R}\phi_i = \theta_i\phi_i$, the eigenvalues can be written as $\theta_i = \langle \mathscr{R}\phi_i, \phi_i\rangle_{L^2} = \mathbb{E}[\langle X, \phi_i\rangle_{L^2}^2]$ and then

$$\sum_{i=1}^{\infty} \theta_i = \sum_{i=1}^{\infty} \mathbb{E}[\langle X, \phi_i\rangle_{L^2}^2] = \mathbb{E}\|X\|_{L^2}^2 < \infty,$$

where the last equality comes from Parseval's identity. The covariance operator $\mathscr{R}$ is thus a trace-class operator.

We are now ready to construct a PCA version for $L^2([0,1])$-valued random variables, which is given by the celebrated Karhunen–Loève expansion (Karhunen [Kar46] and Loève [Loe46]).

Proposition 1.1.3 (Karhunen–Loève expansion)**.** *Let X be a mean zero mean-square continuous stochastic process with continuous sample paths, or equivalently an $L^2([0,1])$-valued random function of mean zero, with a covariance operator $\mathscr{R}$ that can be written as in (1.1.3). Then*

$$X(t) = \sum_{i=1}^{\infty} \langle X, \phi_i\rangle_{L^2}\phi_i(t) = \sum_{i=1}^{\infty} \xi_i \phi_i(t), \tag{1.1.4}$$

where the ξ_i are mean zero random variables such that $\mathbb{E}[\xi_i\xi_j] = \lambda_i\delta_{ij}$, with $\delta_{ij} = 1$ if $i = j$ and zero otherwise. The above expansion converges uniformly in mean-square, i.e.,

$$\lim_{n\to\infty} \sup_{t\in[0,1]} \mathbb{E}\left[X(t) - \sum_{i=1}^{n} \xi_i\phi_i(t)\right]^2 = 0. \tag{1.1.5}$$

The expansion in (1.1.4) is interesting for many reasons. It gives an elegant representation of the data, where the random part, encapsulated in the principal component scores ξ_i, is separated from the deterministic part, encapsulated in the principal components ϕ_i. Moreover, the shape and the smoothness of the first principal components (the ones with the largest

impact) give information on the shape that we are expecting from a realization of X. Perhaps, though, the more useful fact about it is the following. The r-term Karhunen–Loève expansion defined as the truncated sum $\sum_{i=1}^{r} \xi_i \phi_i$, is the best r-dimensional linear approximation of X in the following mean-square sense

$$\mathbb{E}\left[\left\| X - \sum_{i=1}^{r} \langle X, e_i \rangle e_i \right\|_{L^2}^2\right] \geq \mathbb{E}\left[\left\| X - \sum_{i=1}^{r} \langle X, \phi_i \rangle \phi_i \right\|_{L^2}^2\right],$$

for any orthonormal basis $\{e_i\}_{i=1}^{\infty}$ of $L^2([0,1])$. This last result is very important, since in practice we have to work with finite representations of random functions, it is thus common practice to approximate a function X by its r-term Karhunen–Loève expansion. It is the case for example in functional linear model (see e.g. Müller [Mul05] and Morris [Mor15]), in classification and clustering problems (see e.g. Leng and Muller [LM06], Peng and Müller [PM08] and Chiou and Li [CL08]), and in hypothesis testing for two sample tests for covariance function (see e.g. Panaretos et al. 2010 [PKM10]). In the next subsection we explain how to estimate this decomposition from a sample of random curves and we also mention different techniques to choose the value of r.

1.1.4 Statistical inference

Let $X_1, \ldots, X_n$ be a sample of independent realizations of an $L^2([0,1])$-valued random function X of mean and covariance functions m and ρ, with $\mathbb{E}\|X\|_{L^2}^2 < \infty$. Using such a sample, the mean and covariance functions of X are estimated respectively by

$$m_n(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t), \qquad \rho_n(s,t) = \frac{1}{n}\sum_{i=1}^{n} (X_i(t) - m_n(t))(X_i(s) - m_n(s)),$$

and the estimator $\mathscr{R}_n$ of the covariance operator $\mathscr{R}$ of X is the integral operator defined through the kernel function ρ_n.

The consistency and the asymptotic normality of the estimators m_n and $\mathscr{R}_n$ can be derived directly from the functional version of the strong law of large numbers and of the central limit theorem. These two important results are presented in Proposition 1.1.4.

Proposition 1.1.4 (e.g. Bosq [Bos00])**.** *Let $Y_1, \ldots, Y_n$ be independent copies of Y, a random element of a separable Hilbert space $\mathscr{H}$ with mean μ and covariance operator $\mathscr{F}$. If $\mathbb{E}\|Y\|_{\mathscr{H}} < \infty$, then by the strong law of large numbers*

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{a.s.} \mu, \text{ as } n \to \infty.$$

If moreover $\mathbb{E}\|Y\|_{\mathscr{H}}^2 < \infty$, then by the central limit theorem

$$n^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \mu\right) \xrightarrow{d} Z, \text{ as } n \to \infty,$$

where Z is a Gaussian random element of $\mathscr{H}$ with mean μ and covariance operator $\mathscr{F}$.

First, recall that Z is said to be a Gaussian random element of $\mathscr{H}$ if the random variable $\langle Z, f\rangle_{\mathscr{H}} \in \mathbb{R}$ has a Gaussian distribution for every $f \in \mathscr{H}$ (as a convention, a degenerate distribution at zero is considered as a centred Gaussian variable of variance zero). By letting $\mathscr{H}$ be the space $L^2([0,1])$, and by setting $\mu = m$ and $\mathscr{F} = \mathscr{R}$ in Proposition 1.1.4, we obtain that m_n converges almost surely to m, and is asymptotically distributed as a Gaussian random variable when suitably normalised. In a similar way, if we let $\mathscr{H}$ be the space $B_{HS}(L^2([0,1]))$ and set $Y = (X-m)\otimes(X-m)$, we obtain that $\mathscr{R}_n$ converges almost surely to $\mathscr{R}$ if $\mathbb{E}\|X\|^2_{L^2} < \infty$, and if moreover $\mathbb{E}\|X\|^4_{L^2} < \infty$, that $n^{1/2}(\mathscr{R}_n - \mathscr{R})$ is asymptotically distributed as a mean zero Gaussian random element Z in $B_{HS}(L^2([0,1]))$, with covariance operator

$$\mathbb{E}[Z\otimes Z] = \mathbb{E}[\{((X-m)\otimes(X-m)) - \mathscr{R}\}\tilde{\otimes}\{((X-m)\otimes(X-m)) - \mathscr{R}\}],$$

where $\tilde{\otimes}$ is the tensor product on $B_{HS}(L^2([0,1]))$, which for $\mathscr{A}, \mathscr{B}$ and $\mathscr{C} \in B_{HS}(L^2([0,1]))$, is defined as $(\mathscr{A}\tilde{\otimes}\mathscr{B})\mathscr{C} = \langle \mathscr{A}, \mathscr{C}\rangle_{HS}\mathscr{B}$.

Now that we have defined consistent estimators for the mean and the covariance operator of a random function, we are interested in finding estimators for the eigenvalues and the eigenfunctions $\{\theta_i, \phi_i\}$ of $\mathscr{R}$. A natural choice is to use the spectrum $\{\hat{\theta}_i, \hat{\phi}_i\}_{i=1}^n$ of $\mathscr{R}_n$, which is defined as

$$\int_0^1 \rho_n(s,t)\hat{\phi}_i(s)ds = \hat{\theta}_i\hat{\phi}_i(t), \quad i = 1,\dots,n.$$

Note however that the estimation of the j-th eigenfunction of $\mathscr{R}$ only makes sense if the corresponding eigenvalue θ_j has multiplicity one, since otherwise the j-th eigenfunction would not be identifiable (since there is no ordering of the eigenfunctions within an eigenspace). This is why instead of trying to estimate the eigenfunctions themselves, we estimate the eigenprojections $\mathscr{P}_j$ by $\hat{\mathscr{P}}_j$, where $\mathscr{P}_j$ (resp. $\hat{\mathscr{P}}_j$) is the projection operator onto the eigenspace spanned by the eigenfunctions of $\mathscr{R}$ (resp. $\mathscr{R}_n$) associated with the eigenvalue θ_j(resp. $\hat{\theta}_j$). By combining results from perturbation theory (Kato [Kat80]) and results from the convergence of $\mathscr{R}_n$, it can be shown (Dauxois et al. [DPR82], Bosq [Bos00] and Hall and Hosseini-Nasab [HHN06]) that $n^{1/2}(\hat{\theta}_j - \theta_j)$ and $n^{1/2}(\hat{\mathscr{P}}_j - \mathscr{P}_j)$ are asymptotically distributed as mean-zero Gaussian random elements. Note that the variance of both limiting Gaussian random elements depends on the spacing of the distinct eigenvalues of $\mathscr{R}$. The impact of these spacings have been studied in Hall and Hosseini-Nasab [HHN06] for the case where all the eigenvalues of $\mathscr{R}$ are distinct (i.e. $\mathscr{P}_j = \phi_j \otimes \phi_j$), and they show that eigenvalues spacings have a first order effect on the estimators of the eigenfunctions but only a second order effect on eigenvalue estimators.

Note that if we assume that the eigenvalues of a covariance operator up to a certain order p are all distinct, then we obtain that the corresponding eigenfunctions are identifiable up to a sign change. The following proposition is very useful to prove consistency of eigenvalues and eigenfunctions estimators in this context, and will be of great use in the next chapter.

Proposition 1.1.5 (e.g. Hall and Hosseini-Nasab [HHN06], Horváth and Kokoszka [HK12])**.** *Let $\mathscr{R}$ be the covariance operator of an $L^2([0,1])$-valued random variable X such that $\mathbb{E}\|X\|_{L^2}^2 < \infty$, and $\mathscr{R}_n$ be an estimator of it based on a sample of n independent copies of X. Let $\{\theta_i, \phi_i\}$ be the eigenvalues and eigenfunctions of $\mathscr{R}$ where the first p eigenvalues are distinct and bigger than zero, and let $\{\hat{\theta}_i, \hat{\phi}_i\}$ be the eigenvalues and eigenfunctions of $\mathscr{R}_n$. Then*

$$|\theta_i - \hat{\theta}_i| \leq \|\mathscr{R} - \mathscr{R}_n\|_{HS}$$

and

$$\|sign(\langle \phi_i, \hat{\phi}_i \rangle_{L^2})\phi_i - \hat{\phi}_i\|_{L^2} \leq \frac{2\sqrt{2}}{\alpha_i}\|\mathscr{R} - \mathscr{R}_n\|_{HS},$$

for $i = 1, \ldots, p$ and where $\alpha_1 = \theta_1 - \theta_2$ and $\alpha_j = \min\{\theta_{j-1} - \theta_j, \theta_j - \theta_{j+1}\}$ for $2 \leq j \leq p$.

To simplify the presentation, we will now assume that the eigenfunctions of $\mathscr{R}$ are identifiable up to their sign, i.e. that $\theta_1 < \theta_2 < \ldots$, and to avoid dealing with this sign issue, we will suppose that the eigenfunctions of $\mathscr{R}_n$ are correctly estimating this sign, i.e., that $\langle \phi_i, \hat{\phi}_i \rangle_{L^2} \geq 0$. We will moreover assume that X has mean zero, or equivalently that the mean m_n has been subtracted from the curves $X_1, \ldots, X_n$. The empirical version of the Karhunen–Loève expansion for a function X_i is given by

$$X_i(t) = \sum_{j=1}^{n} \langle X_i, \hat{\phi}_j \rangle_{L^2} \hat{\phi}_j(t) = \sum_{j=1}^{n} \hat{\xi}_{ij} \hat{\phi}_j(t),$$

where the eigenfunctions $\{\hat{\phi}_i\}_{i=1}^n$ of $\mathscr{R}_n$ are called the empirical principal components and the inner product $\hat{\xi}_{ij} = \langle X_i, \hat{\phi}_j \rangle_{L^2}$ the empirical j-th score of X_i, where as before $\hat{\xi}_{ij}$ represents the amplitude of the contribution of $\hat{\phi}_j$ to the curve X_i. As discussed in Section 1.1.3, in practice we use a truncation of the Karhunen-Loève expansion at a level r (usually much smaller than n) to obtain a concise, but accurate approximation of the data $X_1, \ldots, X_n$. There are many possible procedures to choose the level of truncation r. A popular one is the graphical method based on the so-called scree plot, which consists in plotting the eigenvalues $\hat{\theta}_i$ against i, and selecting r as the value i beyond which the decrease of the eigenvalues levels off. Another method is based on the cumulative percentage of total variance (CPV) explained by the first p empirical principal components which is given by $CPV(p) = \sum_{i=1}^{p} \hat{\theta}_i / \sum_{i=1}^{n} \hat{\theta}_i$. The idea is to choose r such that $CPV(r)$ is bigger than a certain threshold, and the recommended value for this threshold is usually 85%. It is also possible to use approaches based on pseudo-versions of AIC and BIC (Yao et al. [YMW05]) and on cross-validation (Rice and Silverman [RS91]).

We conclude this section by pointing out that the empirical principal components $\hat{\phi}_i$ have behaviour similar to the principal components ϕ_i, in the sense that they form the optimal empirical orthonormal basis to represent the functions $X_1, \ldots, X_n$. In fact, if we minimize the function

$$f(v_1, v_2, \ldots, v_r) = \sum_{i=1}^{n} \left\| X_i - \sum_{j=1}^{r} \langle X_i, v_j \rangle_{L^2} v_j \right\|_{L^2}^2,$$

over the set of orthonormal basis $\{v_j\}_{j=1}^r$, then the minimum is attained by letting $\{v_j\}_{j=1}^r = \{\hat{\phi}_j\}_{j=1}^r$.

1.2 How to deal with discrete measurements?

In the previous section, we introduced the basic objects and concepts of functional data analysis, and also their empirical counterparts based on a sample formed by independent realizations of a random function. These realizations were assumed to be fully observed, i.e. we were observing the whole functions on the interval $[0,1]$. However, it is clearly impossible to observe a whole function in practice, and thus one has to deal with discrete measurements, i.e, to deal with the fact that every function in the sample is observed on a discrete grid of points. Moreover, it is common to assume that what we observed has been corrupted by random noise, resulting, for example, from measurement errors. The following model is commonly adopted to deal with functional data (e.g., Shi et al. [SWT96], Staniswalis and Lee [SL98], James and Hastie [JH01], Diggle et al. [DHLZ02], Müller [Mul05]). We observe a sample of vectors $X_1, \dots, X_n$ where the entries of the vector X_i of dimension n_i are such that

$$X_i[j] = Y_i(t_{ij}) + \varepsilon_{ij}, \quad \text{for} \quad t_{i1}, \dots, t_{in_i} \in [0,1], \tag{1.2.1}$$

where the random variables ε_{ij} have mean zero, variance σ^2, are independent across i and j and are uncorrelated with the Y_i's. The functions $Y_1, \dots, Y_n$ are independent copies of Y, an $L^2([0,1])$-valued random variable, and the grid of points $\{t_{ij}\}_{j=1}^{n_i}$ on which the function Y_i is observed can be random or deterministic. Even if the model (1.2.1) gives the appearance of a model for multivariate data, it is far from being one. In fact, the vectors X_i's can be vectors of different length, and even if they were of same length, it does not mean that their respective component would represent a realization of the same part of the curve Y, since the grid of points on which the functions Y_i's are observed can be different for every curve. The main difference comes from the fact that each observed vector X_i is assumed to be the noisy discrete realization of an underlying continuous process Y_i, where this latter is assumed to have a certain degree of smoothness. Indeed, this smoothness is a key feature of FDA.

As we already emphasized in Section 1.1, covariance operators, or equivalently covariance functions, are key objects in FDA, since they allow us to carry out functional PCA. In this subsection, we will mainly focus on how to estimate them from a sample of vectors defined as in (1.2.1). There are two main general approaches to do that , namely the "smooth-then-estimate" and the "estimate-then-smooth" approaches. The two approaches are respectively introduced in Section 1.2.1 and in Section 1.2.2.

1.2.1 "Smooth-then-estimate" approach

The "smooth-then-estimate" approach has been popularised by Ramsay and Silverman [RS05], and consists in first carrying out a preprocessing step where one smooths the data, i.e.,

transforms the raw data into smooth functions, and then uses the resulting functions to estimate the mean and covariance functions as explained in Section 1.1.4. This approach is most suitable for densely-observed data, i.e. when the numbers $n_1, \dots, n_n$ of observations per curve are relatively large and well-dispersed in $[0,1]$.

The objective of preprocessing the data is to recover the whole function Y_i from the observed vector X_i. Since this preprocessing step is done independently for each curve, we will now drop the index i, and consider only the problem of recovering a function Y, from the observed vector X of length K, with entries $X[j] = Y(t_j) + \varepsilon_j$, for a grid $t_1, \dots, t_K \in [0,1]$. A common method to estimate Y from the vector X is to approximate the function Y by a linear combination of a finite number p of basis functions $v_1, \dots, v_p$:

$$\tilde{Y}(t) = \sum_{i=1}^{p} c_i v_i(t). \tag{1.2.2}$$

The choice of the basis is quite important here since the estimate $\tilde{Y}$ will inherit of the characteristics, such as continuity and differentiability, of the chosen basis. A common approach is to use a Fourier basis if the data are periodic, and to use B-spline bases otherwise. Wavelet bases have also gained in popularity in the last years, since they are fast to compute and allow one to capture local features in the data with only a few basis elements (e.g. Mallat [Mal09]).

The classical method to determinate the coefficients c_i in (1.2.2) is to choose them such that they minimize the least squares criterion

$$\sum_{j=1}^{K} \left(X[j] - \sum_{i=1}^{p} c_i v_i(t_j) \right)^2 = (X - Vc)^\top (X - Vc),$$

where V is the $K \times p$ matrix such that $V(j,i) = v_i(t_j)$ and c is the vector containing the p coefficients to estimate. Such coefficients are given by $\tilde{c} = (V^\top V)^{-1} V^\top X$, which leads to $\tilde{Y}(t) = \sum_{i=1}^{p} \tilde{c}_i v_i(t)$. Note that choosing p is not easy and it leads to the classical bias/variance trade-off. If p is too large then $\tilde{Y}(t)$ will fit the data well but will certainly have a high variance and a wiggly appearance since it will likely be affected by the noise observed in the data. On the other hand, if p is too small, then there is a risk that $\tilde{Y}(t)$ will be too smooth (small variance) to capture local features of the data that could be of interest. In summary, the value of p controls the degree of smoothness of our estimator $\tilde{Y}$.

The least squares method that we just described provides the basis of many other smoothing methods such as weighted least squares and localized least squares methods. We will now briefly present the penalized least squares method, since according to Ramsay and Silverman [RS05], adding a roughness penalty to the least squares criterion usually gives better results, and has the advantage that the degree of smoothing is controlled in a continuous manner instead of with a discrete parameter (p). This method consists of estimating Y by the function

$\tilde{Y}$ that minimizes the penalized least squares criterion

$$\tilde{Y} = \underset{f \in C^2([0,1])}{\operatorname{argmin}} \sum_{j=1}^{K} \big(X[j] - f(t_j)\big)^2 + \lambda \int_0^1 \left(\frac{d^2 f(t)}{dt^2}\right)^2 dt. \tag{1.2.3}$$

The integral of the second derivative of f is used here to characterize its roughness, since it is a measure of the total curvature of f (for a discussion on the use of roughness penalty in statistics, see, e.g., Green and Silverman [GS94]). The parameter $\lambda \geq 0$ is called a smoothing parameter and as its name suggests, it controls the degree of smoothing or equivalently the tradeoff between the bias and the variance of the estimator. Indeed, if λ is big then $\tilde{Y}$ will be smooth but will not necessary fit the data well, whereas if λ is very small, $\tilde{Y}$ will be an approximate twice differentiable interpolant. Remarkably it turns out that the function $\tilde{Y}$ defined in (1.2.3) is a cubic spline (de Boor [dB01]).

As we already mentioned, once we have the smoothed data $\tilde{Y}_1, \ldots, \tilde{Y}_n$, we can obtain an estimator $\tilde{\mathscr{R}}$ of the covariance operator of Y. Using this estimator we can carry out a functional principal components analysis as explained in Section 1.1.4, and then approximate the functions $\{Y_i\}$ by their r-term Karhunen-Loève expansion. This approach of smoothing first the data, and then doing an fPCA has been studied in Besse and Ramsay [BR86], Ramsay and Dalzell [RD91] and Besse, Cardot and Ferraty [BCF97]. In Hall et al. 2006 [HMW06], they show that if the numbers $n_1, \ldots, n_n$ of observations per curve are such that $n_i \gg n^{1/4}$ as $n \to \infty$, and if each curve has been smoothed with a local linear smoother (Fan and Gijbels [FG96]), then the estimators of the mean function, of the covariance function, of the eigenfunctions and of the eigenvalues are $\sqrt{n}$-consistent. On the other hand, Paul and Peng [PP11] show that if the number of observations per curve is not sufficiently large, then the diagonal of the covariance function obtained with this approach will suffer of an intrinsic bias. In Rice and Silverman [RS91], Pezzuli and Silverman [PS93] and Silverman [Sil96], it is suggested to incorporate the smoothing part into the fPCA procedure by imposing a roughness penalty on the eigenfunctions, though this is done when considering a fully-observed curves setup.

1.2.2 "Estimate-then-smooth" approach

The "estimate-then-smooth" approach consists of using the raw data to obtain an estimator of the discrete (multivariate) version of the mean and covariance functions, and then to smooth these discrete objects using a nonparametric method in order to transform them into functions. As for the previous approach, once we have an estimator of the covariance operator, we can do fPCA in order to estimate the eigenfunctions and then predict the functions $\{Y_i\}$.

If the data are densely observed, the procedure is quite straightforward, as we present below. However, if the data fall into the category of longitudinal data, i.e. when each underlying curve Y_i is only observed on a sparse and irregular grid, the procedure has to be slightly modified. In this case, one needs to pool all the curves together in order to construct scatter plots that are used as surrogates of the empirical multivariate mean and covariance function (Staniswalis

and Lee [SL98], Yao et al. (2003) [YM$^+$03] and Yao et al. (2005) [YMW05]).

We first consider the case where all the curves of the sample are observed on the same dense grid of points $t_1, \ldots, t_K$. Let $X_1, \ldots, X_n$ be a sample of vectors where each vector is defined as $X_i = (X_i[1], \ldots, X_i[K])$, with $X_i[j] = Y_i(t_j) + \varepsilon_{ij}$ and let m, ℓ and $\mathscr{L}$ be respectively the mean function, the covariance function and the covariance operator of the random function Y. Since the sample we are working with is actually constructed as a multivariate sample, we can calculate its empirical mean vector and covariance matrix, respectively given by

$$m_n^K[a] = \frac{1}{n}\sum_{i=1}^{n} X_i[a], \quad R_n^K[a,b] = \frac{1}{n}\sum_{i=1}^{n}\big(X_i[a] - m_n^K[a]\big)\big(X_i[b] - m_n^K[b]\big), \quad \text{for } 1 \le a, b \le K.$$

By smoothing these estimators, with a kernel smoothing technique for example, we can obtain an estimator of m on the interval $[0,1]$ and of ℓ on the interval $[0,1]^2$. This procedure works well to obtain an estimator of the mean function, however, we have to be careful when we consider transforming R_n^K into a surface, since the effect of the measurement errors is different on the diagonal than it is elsewhere on $[0,1]^2$. In fact, we have that

$$\begin{aligned} \operatorname{Cov}(X_i[j], X_i[l]) &= \operatorname{Cov}(Y_i(t_j), Y_i(t_l)) + \operatorname{Cov}(\varepsilon_{ij}, \varepsilon_{il}) \\ &= \begin{cases} \operatorname{Cov}(Y_i(t_j), Y_i(t_l)), & j \neq l, \\ \operatorname{Cov}(Y_i(t_j), Y_i(t_l)) + \sigma^2, & j = l, \end{cases} \end{aligned}$$

and thus $R_n^K[a,a]$, $1 \le a \le K$, is a biased estimator of $\operatorname{Var}[Y(a)]$, since $\mathbb{E}\big[\big(X_i[a] - m_n^K[a]\big)^2\big] \approx \operatorname{Var}[Y(a)] + \sigma^2$. The strategy to obtain a good estimator of the covariance function of Y is then to remove the diagonal of R_n^K before smoothing it. This diagonal removal procedure has been first introduced in Staniswalis and Lee [SL98]. The estimators of the mean and covariance functions obtained with this approach can achieve consistency only if $K = K_n$, the number of observations per curve, grows with the sample size n.

Let's now consider the more challenging setup where each underlying curve Y_i is observed on a sparse and irregular grid. The objective is again to estimate the mean function m and the covariance function ℓ of the $L^2([0,1])$-valued random function Y from the sample $X_1, \ldots, X_n$ defined as in (1.2.1). Since the functions $Y_1, \ldots, Y_n$ are not necessarily observed on the same points, we cannot calculate an empirical multivariate version of the mean (respectively covariance) function as before, and even if the curves were all observed on the same grid of points, since the grid is sparse we could not obtain good estimates on the whole interval $[0,1]$ (resp. $[0,1]^2$) by applying a smoothing technique. The idea is then to pool all the curves together in order to borrow information from neighbors of a given point from all the curves in the sample. More precisely, the estimator $\hat{m}$ of the mean function is obtained by smoothing the scatter plot

$$\{(X_i[j], t_{ij}), i = 1, \ldots, n, \text{and } j = 1, \ldots, n_i\},$$

across time. This smoothing can be done with a nonparametric smoother such as a local

polynomial smoother. Similarly, the estimator $\hat{\ell}$ of the covariance function is obtained by smoothing the 2D scatter plot

$$\{((X_i[j]-\hat{m}(t_{ij})(X_i[k]-\hat{m}(t_{ik})), t_{ij}, t_{ik}),\ i=1,\dots,n, \text{ and } j,k=1,\dots,n_i, j\neq k\},$$

from which we have removed the diagonal since, as in the dense case, it is contaminated by the variance of the noise component. The asymptotic setup in sparse sampling is that $\max_{1\leq i\leq n} n_i \leq C$, where C is a constant, as $n\to\infty$. Both estimators $\hat{m}$ and $\hat{\ell}$ can achieve consistency only if the grid of pooled points $\{t_{ij}, i=1,\dots,n \text{ and } j=1,\dots,n_j\}$ is getting dense in $[0,1]$ as $n\to\infty$, and if so, their rate of convergence will be nonparametric.

In the scatter plot procedure that we just described, each curve does not have the same weight in the smoothing step, since this weight is proportional to the number of points on which a curve is observed. Li and Hsing [LH10] studied an alternative method that unifies the dense and the sparse framework by assigning to each curve the same weight. Zhang and Wang [ZW16] push this work forward by studying in greater depth the effect of either putting the same weight on each curve or on each observation. Their work yields a classification of sampling schemes into three categories, namely ultra dense sampling, dense sampling and sparse sampling, accordingly to the bias and the rate of convergence of the corresponding mean and covariance function estimators. More work has been done to study the effect of the sampling scheme in FDA. For example, Cai and Yuan [CY11] study the effect of the design of the grid (deterministic or random) on the convergence rate of the estimator of the mean function obtained with a nonparametric approach. On the other hand, Amini and Wainwright [AW12] study the effect of the sample size n and functional size (number of grid points) on the estimation of the subspace span by the eigenfunctions of the covariance operator of Y, where it is assumed that the curves $Y_1,\dots,Y_n$ lie in a RKHS.

We conclude this subsection by describing the Principal components Analysis through Conditional Expectation (PACE) method introduced in Yao et al. (2005) [YMW05]. This method has been developed to extend fPCA to longitudinal data but can also be used with densely observed data. Its main novelty is to use conditional expectation to provide estimators of the principal component scores of the random functions $Y_1,\dots,Y_n$. Using these scores, it is then possible to predict the curves themselves by estimating their r-term Karhunen–Loève expansion.

The setting is as before, i.e., we observe a sample of vectors $X_1,\dots,X_n$, where $X_i[j]=Y_i(t_j)+\varepsilon_{ij}$, where $\mathrm{Var}(\varepsilon_{ij})=\sigma^2$. The estimators of the mean and covariance functions m and ℓ of Y are obtained using local linear smoothers as follows. For each $t\in[0,1]$, $\hat{m}_{PACE}(t)=\hat{\beta}_0$, where $\hat{\beta}_0$ is such that

$$(\hat{\beta}_0,\hat{\beta}_1)=\underset{\beta_0,\beta_1}{\arg\min}\sum_{i=1}^{n}\sum_{j=1}^{n_i}k_1\left(\frac{t-t_{ij}}{h_1}\right)(X_i[j]-\beta_0-\beta_1(t-t_{ij}))^2,$$

where k_1 is a kernel and h_1 is a bandwidth. For each $(s,t)\in[0,1]^2$, $\hat{\ell}_{PACE}(s,t)=\hat{\gamma}_0$, where $\hat{\gamma}_0$ is

such that

$$(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_2) = \underset{\gamma_0,\gamma_1,\gamma_2}{\arg\min} \sum_{i=1}^{n} \sum_{1 \le j \ne l \le n_i} k_2\left(\frac{s-t_{ij}}{h_2}, \frac{t-t_{il}}{h_2}\right) \{G_i(t_{ij}, t_{il}) - \gamma_0 - \gamma_1(s-t_{ij}) - \gamma_2(t-t_{il})\}^2,$$

where $G_i(t_{ij}, t_{il}) = \{X_i[j] - \hat{m}_{PACE}(t_{ij})\}\{X_i[l] - \hat{m}_{PACE}(t_{il})\}$, k_2 is a bivariate kernel and h_2 is a bandwidth. The estimator of σ^2 is defined as

$$\hat{\sigma}^2 = 2\int_{1/4}^{3/4} (\hat{V}(t) - \hat{\ell}_{PACE}(t,t))\,dt,$$

where $\hat{V}(t)$ is an estimator of the diagonal $\ell(t,t)+\sigma^2$ obtained by using a local linear smoother on $\{G_i(t_{ij}, t_{ij}) | 1 \le i \le n, 1 \le j \le n_i\}$, and the integral is not defined on the entire interval $[0,1]$ to avoid boundary issues.

Let the Karhunen–Loève expansion of Y_i be $Y_i(t) = m(t) + \sum_{a=1}^{\infty} \langle Y_i, \eta_a \rangle_{L^2} \eta_a(t)$, and denote the a-th principal component score of Y_i by ξ_{ia}. The best linear prediction of ξ_{ia} given the vector X_i is

$$\mathbb{E}[\xi_{ia}|X_i] = \lambda_a \eta_{ia}^\top \Sigma_{X_i}^{-1}(X_i - m_i),$$

where λ_a is the eigenvalue associated with the eigenfunction η_a of $\mathscr{L}$, $\eta_{ia}^\top = (\eta_a(t_{i1}), \ldots, \eta_a(t_{in_i}))$, $\Sigma_{X_i}[j,l] = \mathrm{Cov}(X_i[j], X_i[l])$ and $m_i = (m(t_{i1}), \ldots, m(t_{in_i}))^\top$. From the covariance operator defined through $\hat{\ell}_{PACE}$, one can obtain the estimators $\{\hat{\lambda}_a\}$ and $\{\hat{\eta}_a\}$ of the eigenvalues and eigenfunctions of $\mathscr{L}$, and then the estimator of the best linear prediction of ξ_{ia} is given by

$$\hat{\xi}_{ia} = \hat{\lambda}_a \hat{\eta}_{ia}^\top \hat{\Sigma}_{X_i}^{-1}(X_i - \hat{m}_i)$$

where $\hat{\Sigma}_{X_i}[j,l] = \hat{\ell}_{PACE}(t_{ij}, t_{il}) + \hat{\sigma}^2$. Finally, the estimator of the curve Y_i is given by

$$\hat{Y}_i(t) = \hat{m}(t) + \sum_{a=1}^{r} \hat{\xi}_{ia} \hat{\eta}_a(t),$$

where the number r is chosen by cross-validation.

1.3 What if roughness is not just pure noise?

In this section, we first discuss the assumptions and the limits of the standard model for discretely observed data (Section 1.3.1). We then study in Section 1.3.2 the effects of including the variations of a random function at finer scales into the standard model on the two main approaches that have been introduced in the previous section. We conclude the section with a description of the objective of this thesis (Section 1.3.3).

1.3.1 Limits of the standard model

Let's first recall the classical setup in FDA. We are interested in carrying out statistical inferences on an $L^2([0,1])$-valued random function Y using a collection of independent realizations $Y_1,\dots,Y_n$. Since in practice we cannot observe the complete functions Y_i, we have to work with a sample of vectors $X_1,\dots,X_n$, for which we assume the following model:

$$X_i[j] = Y_i(t_j) + \varepsilon_{ij}, \quad j = 1,\dots,K. \tag{1.3.1}$$

For simplicity we have assumed that each function Y_i is observed on the same deterministic grid of K points $(t_1,\dots,t_K)$. The ε_{ij} are identically and independently distributed (i.i.d) random variables of mean zero and variance σ^2, and they are uncorrelated with the functions $Y_1,\dots,Y_n$. Note that as K increases, the vector $\varepsilon_i = (\varepsilon_{i1},\dots,\varepsilon_{iK})^\top$ is becoming a white noise process. In fact, the model (1.3.1) is equivalent to the following model where each random variable ε_{ij} is viewed as the realization of a white noise process W_i at time t_j as follows

$$X_i[j] = Y_i(t_j) + W_i(t_j), \quad j = 1,\dots,K. \tag{1.3.2}$$

We are committing a slight abuse by calling the process W white noise. In fact, we know that such a process is not realizable as a proper stochastic process, as it would not be trace class (Ramsay and Silverman [RS05, Section 3.2.4]). This means that W must have some dependence structure at some very fine scale, so its covariance function is supported on a band $\{|s-t|<\delta\}$ for some infinitesimally small δ.

As explained in the previous section, there are two general approaches to deal with such discretely observed data. While different, these approaches share the same two-step structure: a smoothing step followed by a step consisting in doing fPCA, or equivalently a step consisting in approximating the functions $\{Y_i\}$ by a r-term Karhunen–Loève expansion. Moreover, they both make the same implicit assumption in the smoothing step, which is that the random function Y in which we are interested (the "true" signal) is of smoothness class C^k with $k \geq 2$, while in the fPCA step, they both assume that Y can be well-approximated by a process of rank r. It follows that any variations that are of smoothness class less than C^2, or that correspond to fluctuations around eigenfunctions of order at least $r+1$, are consigned to the noise process W. However it seems restrictive to consider any rough variation in the data as noise, since a rough process could still have some structure, but at a finer scale. This observation is confirmed by the discussion of Ramsay and Silverman [RS05, Section 3.2.4] on the standard model for the errors ε_{ij}. They claim that the assumption that W has a covariance function with only a σ^2-ridge on the diagonal is often too restrictive to hold with functional data. In fact, it is really common to see autocorrelation in the functional residuals $\{W_i\}$, which indicates that the process W is probably smooth at some fine resolution. This means that instead of modeling W as a white noise process, we should probably consider it as a mean-square continuous process having a banded covariance function $b(s,t) = b(s,t)\mathbf{1}\{|s-t|<\delta\}$, for b a non-constant continuous function and $\delta > 0$ non-negligible. In this case, W would be a rough process that

captures the local variations, or equivalently the variations at scale at most δ, of the data. As stated by Ramsay and Silverman [RS05, Section 3.2.4], this modification to the standard model is worthy of consideration:

"*Nevertheless, a model specifically for variance heterogeneity and/or autocorrelation can pay off in terms of better estimation, and this type of structure may be in itself interesting. A thoughtful application of functional data analysis will always be open to these possibilities*".

1.3.2 Effects of a less restrictive model on the two main approaches in FDA

We now study the consequences of allowing the covariance function b of W to be banded (instead of having a trivial structure as in the standard model) on the two approaches described in the previous section.

We first investigate the effect on the smoothing step. To accommodate for this new definition of the process W, we would need to modify the "smooth-then-estimate" approach (Section 1.2.1) by replacing the objective function (1.2.3) by a weighted penalized least square criterion:

$$\tilde{Y}_i = \underset{f \in C^2([0,1])}{\operatorname{argmin}} \, (X_i - f_v)^T \Omega (X_i - f_v)^T + \lambda \int_0^1 (f''(t))^2 dt,$$

where $f_v = (f(t_1), \ldots, f(t_K))^T$ and $\Omega = \Sigma_\varepsilon^{-1}$, with Σ_ε the covariance matrix of the vector $\varepsilon = (W(t_1), \ldots, W(t_K))^\top$. However, there are two problems that arise with this new formulation of the objective function. First, the matrix Σ_ε is unknown, and second, we would need to add some parametric assumption for the processes Y_i and W_i to be jointly identifiable (Opsomer et al. [OWY01]). On the other hand, for the "estimate-then-smooth" approach (Section 1.2.2), instead of removing the diagonal of the empirical covariance matrix R_n^K before smoothing it, we would now need to remove a band $\{(s,t) \in [0,1]^2 : |s-t| < \delta\}$. As with the other approach, this would again lead to identifiability issues.

To study the effect of the new model for W on the fPCA step, we first study its consequences at the population level. Let $X = Y + W$, and $\{\theta_i, \phi_i\}_{i=1}^\infty$, $\{\lambda_i, \eta_i\}_{i=1}^r$ and $\{\beta_i, \psi_i\}_{i=1}^\infty$ be the respective spectrum of the covariance operator $\mathscr{R}$ of X, $\mathscr{L}$ of Y and $\mathscr{B}$ of W, where we assumed $\mathscr{L}$ to be of rank $r < \infty$. We are interested in the relationship between the set of eigenfunctions $\{\phi_i\}_{i=1}^\infty$, $\{\eta_i\}_{i=1}^r$ and $\{\psi_i\}_{i=1}^\infty$, since we want to study when it is possible to separate the processes Y and W from X using its Karhunen–Loève expansion, or to simply recover the eigenfunctions of $\mathscr{L}$. This is a relevant question, because now that we do not consider W as a white noise process, it is not necessarily true that the first r eigenfunctions of $\mathscr{R}$ and $\mathscr{L}$ will be equal. With the standard model, this was not a problem because we know that if $\mathscr{R} = \mathscr{L} + \mathscr{D}$, with $\mathscr{D}$ a covariance operator defined through the covariance function $d(s,t) = \sigma^2 \mathbf{1}\{|s-t| \leq \delta\}$ for $\delta = 0$, then the first eigenfunctions of $\mathscr{R}$ and $\mathscr{L}$ will be exactly the same.

To help clarify things, let's write the Karhunen–Loève expansion of $X = Y + W$:

$$X = \sum_{i=1}^{\infty} \langle X, \phi_i \rangle \phi_i = \sum_{i=1}^{r} \langle Y, \eta_i \rangle \eta_i + \sum_{i=1}^{\infty} \langle W, \psi_i \rangle \psi_i,$$

and the Mercer decomposition of its covariance function r:

$$r(s,t) = \sum_{i=1}^{\infty} \theta_i \phi_i(s) \phi_i(t) = \sum_{i=1}^{r} \lambda_i \eta_i(s) \eta_i(t) + \sum_{i=1}^{\infty} \beta_i \psi_i(s) \psi_i(t).$$

If the systems $\{\eta_i\}_{i=1}^r$ and $\{\psi_i\}_{i=1}^\infty$ are orthogonal, then we can recover the expected principal components of variation since the set $\{\phi_i\}_{i=1}^\infty$ will be equal to $\{\eta_i\}_{i=1}^r \cup \{\psi_i\}_{i=1}^\infty$. If moreover the eigenvalues are "well-ordered", i.e. $\lambda_r > \beta_1$, then $\{\phi_i\}_{i=1}^r = \{\eta_i\}_{i=1}^r$, and since W is orthogonal to the functions $\{\eta_i\}_{i=1}^r$, we obtain that Y is exactly equal to the r-term Karhunen–Loève expansion of X, which means that for this particular case, the process W does not affect the fPCA step. However, there is no reason why the systems $\{\eta_i\}_{i=1}^r$ and $\{\psi_i\}_{i=1}^\infty$ should be orthogonal or "well-ordered", and if they are not, then it might well happen that neither $\{\eta_i\}$ nor $\{\psi_i\}$ are eigenfunctions of $\mathscr{R}$. Assuming for example that no pair $\{\eta_i, \psi_j\}$ is orthogonal, then even in the "well-ordered" situation where $\lambda_r > \beta_1$, it is not necessarily true that $\phi_i = \eta_i$ for $i = 1, \ldots, r$. In fact depending on the spacings of $\{\lambda_i\}_{i=1}^r \cup \{\beta_i\}_{i=1}^\infty$, it could happen that $\phi_i, i \leq r$, is a linear combination of some η_j and ψ_j. A similar situation is illustrated in Figure 1.1 with a simulated example. We have simulated data such that $X = Y + W$, where the covariance operator $\mathscr{L}$ of Y is of rank 5 and has trigonometric eigenfunctions, and where the eigenfunctions of $\mathscr{B}$ are locally supported on non-overlapping intervals of length $\delta = 0.05$ (which implies that the covariance function b has a banded structure). Moreover, the eigenvalues have been chosen such that $\lambda_4 > \beta_1 > \lambda_5$. The figure shows 10 realizations of Y, W and X ((a) to (c)), the covariance functions ℓ, b and ρ ((d) to (f)) and the eigenfunctions of each operator ((g) to (i)). In graph (i) we can see that ϕ_5 (green) has been affected by ψ_1, since it is clearly a distorted version of η_5. But it is not the only eigenfunction of $\mathscr{R}$ that has been affected, since in fact, even if $\beta_1 < \lambda_4 < \lambda_3$, we can see that eigenfunctions ϕ_3 (blue) and ϕ_4 (yellow) are slightly modified versions of η_3 and η_4 .

The smoothing step which is taken before the fPCA step can also have some confounding effects in practice. In fact, since spline smoothing is approximately equivalent to kernel smoothing (Silverman [Sil84]), smoothing either using spline bases or a local linear smoother implies approximately convolving the discrete data with a kernel of bandwidth h. If this bandwidth is comparable with δ, then the variation of W would propagate to larger scales, and then it would be impossible to separate the covariance operators $\mathscr{L}$ and $\mathscr{B}$ from $\mathscr{R}$.

1.3.3 Objective of the thesis

It seems clear from the discussion of Section 1.3.2 that the two common approaches in FDA cannot handle, without major modifications, a model where the rough part (variations at fine scales) of a random function is not considered as white noise, and thus there is a need for

a new approach. Developing the theory and the practical implementation of such a novel approach is the main objective of this thesis.

In Chapter 2, we formulate the new model rigorously, and we then find nonparametric conditions under which there exists a unique decomposition of the covariance operator $\mathscr{R}$ into $\mathscr{L} + \mathscr{B}$, or equivalently under which $\mathscr{L}$ and $\mathscr{B}$ are jointly identifiable. We then consider the more challenging problem of finding conditions to ensure the uniqueness of the discrete analogue of this decomposition, i.e., of the decomposition $\mathscr{R}^K = \mathscr{L}^K + \mathscr{B}^K$, where $\mathscr{R}^K$ is the K-resolution version of $\mathscr{R}$ obtained from the evaluation of the curve X on a grid of K points. The idea is to translate the discrete version of the decomposition into a matrix completion problem, and then translate the identifiability conditions at the continuum level into conditions ensuring the uniqueness of the matrix completion problem. Once we have conditions ensuring that our problem is well-posed, we define consistent estimators of the covariance operators $\mathscr{L}$ and $\mathscr{B}$, without assuming knowledge of the rank r of $\mathscr{L}$ or of the bandwidth δ of $\mathscr{B}$. Using these estimators, we then define consistent estimators of the best linear predictors of Y and W respectively. In Chapter 3, we describe how to obtain in practice the estimators defined in Chapter 2, and we study the finite-sample performance of our new approach by means of a simulation study and on a real dataset.

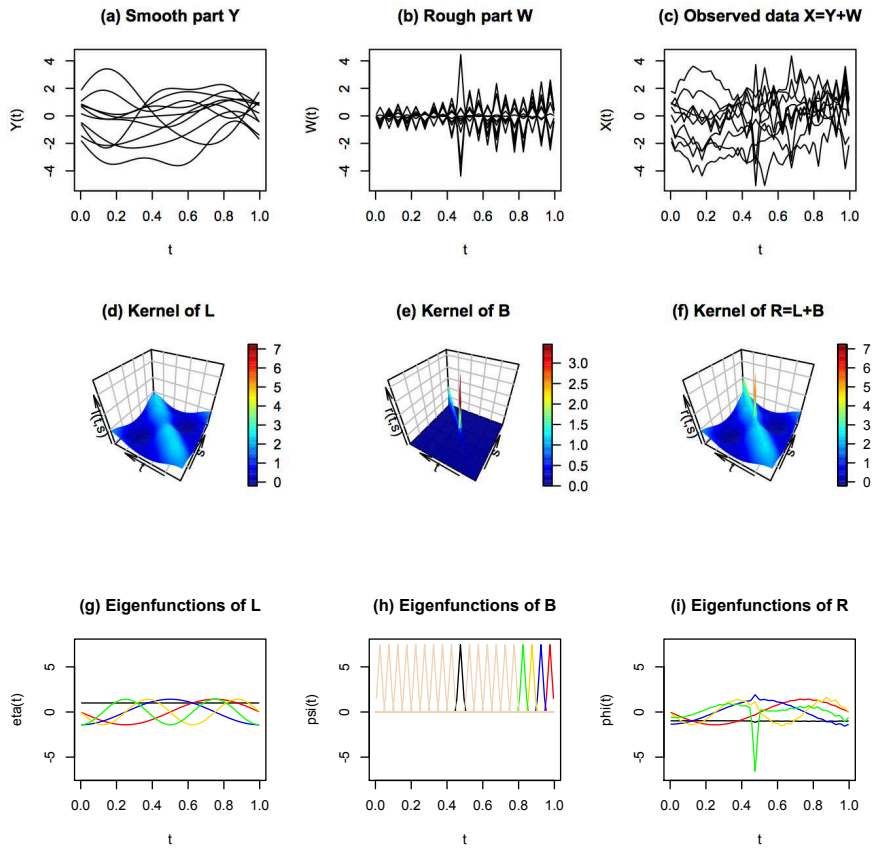

Figure 1.1 – Ten smooth curves $Y_1,\ldots,Y_{10}$ (plot (a)), ten rough curves $W_1,\ldots,W_{10}$ (plot (b)) and their sum resulting in ten observations $X_1,\ldots,X_{10}$ (plot (c)). The kernel of the smooth operator $\mathscr{L}$ of rank 5 is illustrated in plot (d) with eigenfunctions η_j (plot (g)), that of $\mathscr{B}$ in plot (e) with eigenfunctions ψ_j (plot (h)) and finally that of $\mathscr{R}=\mathscr{L}+\mathscr{B}$ in plot (f) with its first 5 eigenfunctions ϕ_j represented in plot (i). The first 5 eigenfunctions of each operator are represented respectively in black, red, blue, yellow and green.

2 A new approach to analysing functional data

In this chapter we develop a new approach to analyze functional data that allows the possibility that the rough part observed in a random curve is not simply due to random noise, but is due to a rough process that has a smooth structure at some fine scale and thus that is itself a signal. We first introduce our new model and the practical setup under consideration (Section 2.1), then in Section 2.2 we give conditions under which our model is unique and identifiable for both continuous and discrete versions of the covariance operators of the smooth and the rough processes. We then define estimators of these covariance operators and prove their consistency (Section 2.3), and we conclude the chapter by separating the smooth part from the rough part of the data in Section 2.4 by estimating the best linear unbiased predictors of these two processes.

2.1 The model : smooth plus rough components

As we already mentioned, we want to develop a model that can deal with data that arise as the sum of a smooth process and a rough process where this latter is not necessary a noise component. Our new model is defined as follows. We assume that $X \in L^2([0,1])$ is a mean-zero mean square continuous random function that can be decomposed as

$$X(t) = Y(t) + W(t), \quad t \in [0,1], \tag{2.1.1}$$

where Y and W are two uncorrelated processes. From Equation (2.1.1) and the fact that Y and W are uncorrelated random variables we obtain the following additive structure of the covariance operator $\mathscr{R}$ of X and of its covariance function $\rho(s,t) = \mathbb{E}[X(s)X(t)]$, $s,t \in [0,1]$

$$\begin{aligned} \mathscr{R} &= \mathscr{L} + \mathscr{B} \\ \rho(s,t) &= \ell(s,t) + b(s,t), \quad t,s \in [0,1], \end{aligned}$$

where $\mathscr{L}$ and $\mathscr{B}$ are the covariance operators of Y and W, respectively defined by the covariance functions

$$\ell(s,t) = \mathbb{E}[Y(s)Y(t)] - \mathbb{E}[Y(s)]\mathbb{E}[Y(t)], \quad b(s,t) = \mathbb{E}[W(s)W(t)] - \mathbb{E}[W(s)]\mathbb{E}[W(t)].$$

In this model, the process Y represents the smooth part of X and the process W represents its rough part. Smooth versus rough processes can be interpreted as having variation of coarse scale or fine scale respectively. Let $\{\lambda_j, \eta_j\}_{j=1}^r$ be the spectrum of the operator $\mathscr{L}$, which leads to the following Mercer decomposition of ℓ

$$\ell(s,t) = \sum_{j=1}^{r} \lambda_j \eta_j(s)\eta_j(t).$$

To ensure that Y represents variations that occur at a global scale, we will assume that the eigenfunctions $\{\eta_j\}_{j=1}^r$ are smooth and finite in number, i.e., that $\mathscr{L}$ is of finite rank ($r < \infty$). We assume that r is finite because it is usual in FDA to suppose that there are finitely many non-negligable modes of variation, and also to ensure that the process Y is smooth. We will specify later (Section 2.2.1) what precise smoothness conditions we need to impose on the eigenfunctions of $\mathscr{L}$. From now on we will refer to $\mathscr{L}$ as the smooth operator.

Another way to think of a process representing variations that occur at local scale is to see it as a short-range dependence process. Letting the range of dependence of W be of order $\delta \in (0,1)$, we expect its covariance function b to have a δ-banded structure, i.e.

$$b(s,t) = 0 \quad \text{for} \quad s,t \in [0,1] \quad \text{such that} \quad |s-t| > \delta, \tag{2.1.2}$$

with Mercer decomposition given by

$$b(s,t) = \sum_{j=1}^{\infty} \beta_j \psi_j(s)\psi_j(t) = \mathbf{1}\{|t-s| < \delta\} \sum_{j=1}^{\infty} \beta_j \psi_j(s)\psi_j(t),$$

where $\{\beta_j\}_{j=1}^{\infty}$ and $\{\psi_j\}_{j=1}^{\infty}$ are respectively the eigenvalues and the eigenfunctions of $\mathscr{B}$. Note that in the sequel we will refer to $\mathscr{B}$ as the banded operator and to b as the δ-banded function.

The main objective of this chapter is to estimate $\mathscr{L}$ and $\mathscr{B}$ based on a sample of curves observed discretely. Throughout the chapter, we will work with a sample $X_1, \ldots, X_n$ of n iid random functions defined as in (2.1.1), where each function X_i is observed on the grid of K points

$$(t_1, \ldots, t_K) \in \mathscr{T}_K = \left\{(x_1, \ldots, x_K) \in \mathbb{R}^K : x_1 \in I_{1,K}, \ldots, x_K \in I_{K,K}\right\},$$

where $\{I_{j,K}\}_{j=1}^K$ is the partition of $[0,1]$ into intervals of length $1/K$. We consider the set of grids $\mathscr{T}_K$ because it imposes a regularity on the grid that will be very helpful when we will work with the discrete version of the covariance function b, and this without being too restrictive (the points of the grid don't have to be equally-spaced for example). Using this kind of sample, we will derive estimators of the covariance operators $\mathscr{L}$ and $\mathscr{B}$ and prove their consistency in

Section 2.3. We will then work on the second objective of the chapter which is to recover the curves Y and W from X, and to do this we will define consistent estimators of the best linear unbiased predictors for both curves Y and W (Section 2.4). But before doing all this, we study in the next section the well-posedness of the decomposition of $\mathscr{R}$ into $\mathscr{L}+\mathscr{B}$, i.e. we study if the decomposition is unique and if it is identifiable given a sample of discretely observed curves constructed as we just described.

2.2 Well-Posedness : uniqueness and identifiability

In this section, we first study under which conditions the decomposition $\mathscr{R}=\mathscr{L}+\mathscr{B}$ is unique (Section 2.2.1). We then study in Section 2.2.2 the uniqueness of the discrete version of this decomposition, which is defined using the K-resolution version of the covariance operators.

2.2.1 Uniqueness of the decomposition $\mathscr{R}=\mathscr{L}+\mathscr{B}$

Recall that we have a covariance operator $\mathscr{R}$ that can be decomposed as the sum of a smooth operator $\mathscr{L}$ and a banded operator $\mathscr{B}$, and we would like to estimate these two latter operators. It is clear that we need to make more assumptions on at least one of them if we want the decomposition to be unique, since a priori there is no reason to suspect that only two operators sum up to $\mathscr{R}$.

It turns out that the key assumption we need to make for this decomposition to be unique is that the covariance function ℓ of $\mathscr{L}$ is a bivariate real analytic function. Recall that a function $f:U\to\mathbb{R}$, where U is an open subset of $\mathbb{R}^k$, is analytic if for each $\alpha\in U$ the function may be represented by a convergent power series in some neighborhood V of α, i.e. $f(x)=\sum_{n=0}^{\infty}P_n(x-\alpha),\forall x\in V$, where P_n is a homogeneous polynomial of degree n in k variables. Using analytic continuation, we know that if the function ℓ is analytic on an open subset $\mathscr{U}$ containing the band $\{(s,t)\in[0,1]^2:|s-t|\leq\delta\}$, then knowing ℓ on the domain $\{(s,t)\in[0,1]^2:|s-t|>\delta\}$ is equivalent to knowing it on the whole domain $[0,1]^2$. This last fact is sufficient to obtain uniqueness of the decomposition, as stated in the following theorem.

Theorem 2.2.1 (Uniqueness)**.** *Let $\mathscr{L}_1,\mathscr{L}_2:L^2[0,1]\to L^2[0,1]$ be trace-class covariance operators of rank $r_1<\infty$ and $r_2<\infty$, respectively. Let $\mathscr{B}_1,\mathscr{B}_2:L^2[0,1]\to L^2[0,1]$ be banded trace-class covariance operators of bandwidth $\delta_1<1$ and $\delta_2<1$ respectively. If the eigenfunctions of $\mathscr{L}_1$ and $\mathscr{L}_2$ are real analytic, then we have the equivalence*

$$\mathscr{L}_1+\mathscr{B}_1=\mathscr{L}_2+\mathscr{B}_2\iff\mathscr{L}_1=\mathscr{L}_2\quad\&\quad\mathscr{B}_1=\mathscr{B}_2.$$

Proof. Since the eigenfunctions of $\mathscr{L}_1$ and $\mathscr{L}_2$ are analytic and $\max\{r_1,r_2\}<\infty$, it follows that the corresponding covariance functions ℓ_1 and ℓ_2 are bivariate analytic functions on $[0,1]^2$ (Krantz and Parks [KP02]).

The zero set of a real bivariate analytic function is at most 1-dimensional, unless the function is

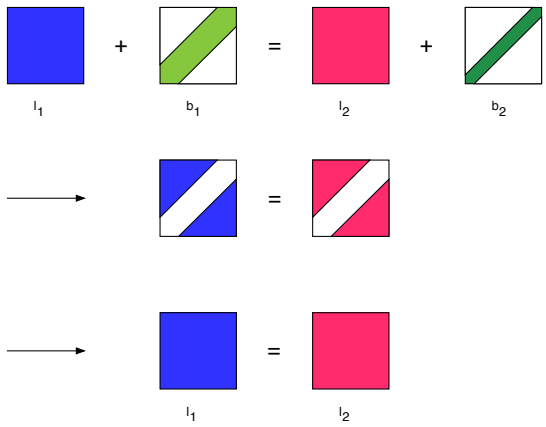

Figure 2.1 – Schematic illustration of the idea underlying the proof of Theorem 2.2.1.

uniformly zero (Krantz and Parks [KP02, Thm 6.33]). Since analytic functions are closed under subtraction, this implies that if ℓ_1 and ℓ_2 coincide on an open subset U of $[0,1]^2$, then the function $\ell_1 - \ell_2$ has to be uniformly zero, which implies that ℓ_1 and ℓ_2 coincide everywhere on $(0,1)^2$, and thus on $[0,1]^2$ by continuity. This will in turn imply that $\mathscr{B}_1$ and $\mathscr{B}_2$ also coincide. Finding such a subset U will then conclude the proof.

Without lost of generality, assume that $\delta_1 \geq \delta_2$. Define

$$U = \Big(\delta_1, 1\Big) \times \Big(0, 1-\delta_1\Big).$$

Since $\mathscr{L}_1 + \mathscr{B}_1 = \mathscr{L}_2 + \mathscr{B}_2$, but $\mathscr{B}_1 = \mathscr{B}_2 = 0$ on U, it must be that the covariance functions of $\mathscr{L}_1$ and $\mathscr{L}_2$ coincide on the open set U, and the proof is complete. □

An schematic illustration of the idea of the proof is presented on Figure 2.1; the squares represent covariance functions and the white color is used to represent zero values.

In the previous theorem, it would have sufficed to assume that the covariance function ℓ is a real analytic function on some open subset $\mathscr{U} \supset \{(s,t) \in [0,1]^2 : |s-t| \leq \delta\}$. We present the theorem with the stronger assumption that the smooth operator $\mathscr{L}$ is of finite rank and has real analytic eigenfunctions since this stronger assumption will be required when we will consider the discrete version of this decomposition problem in the next subsection.

It is interesting to note that the condition of the analyticity of ℓ on an open subset $\mathscr{U}$ containing the δ-band is essentially a necessary condition. Indeed, it is impossible to relax it without stronger assumptions on the covariance operator $\mathscr{B}$, since even assuming that ℓ is C^∞ on the whole domain $[0,1]^2$ is not sufficient to guarantee the uniqueness of the decomposition. This follows from the fact that it is possible to construct δ-banded functions that are C^∞. Let s be C^∞, h be a δ-banded function and m be C^∞ and δ-banded. Then if we consider the function

$r(s,t) = c(s,t) + h(s,t) + m(s,t)$, we have more than one way to write it as a smooth component plus a banded one. We can either set $\ell = c + m$ and $b = h$ or $\ell = c$ and $b = h + m$. An example of a C^∞ and δ-banded function is given by the self-convolution of the bump function defined as

$$b(s,t) = \begin{cases} \exp\left(-\frac{1}{1-\left(\frac{s-t}{2\delta}\right)^2}\right) & \text{if } |s-t| < 2\delta, \\ 0 & \text{otherwise.} \end{cases}$$

On the other hand, the assumptions of finite rank and real analytic eigenfunctions that we make on $\mathscr{L}$ are not as restrictive as they may seem. Indeed, the class of real analytic functions includes polynomials, trigonometric functions, exponential and logarithmic functions, and it is closed under the most common operations such as linear combinations, products, compositions and differentiation. Moreover, the following proposition tells us that the set of finite rank r covariance operators with analytic eigenfunctions is dense in the set of all finite rank r covariance operators. This means that the rank r Mercer approximation of the covariance function, which is optimal in a mean-square sense, can be approximated at any level of precision by a rank r covariance operator with analytic eigenfunctions. Consequently, if we do believe that a covariance operator is approximately of finite rank, then it can be very well approximated by an operator of the same rank having analytic eigenfunctions:

Proposition 2.2.1. *Let Z be an $L^2[0,1]$-valued random function with a trace class covariance $\mathscr{G}$ of rank $r < \infty$. Then, for any $\epsilon > 0$ there exists a random function Y whose covariance $\mathscr{L}$ has analytic eigenfunctions and rank $q \leq r$, such that*

$$\mathbb{E}\|Z - Y\|_{L^2}^2 < \epsilon \quad \textit{and} \quad \|\mathscr{G} - \mathscr{L}\|_{TR} < \epsilon.$$

If additionally $\mathscr{G}$ has C^1 eigenfunctions on $[0,1]$, then we have the stronger result that for any $\epsilon > 0$, there exists a random function Y whose covariance $\mathscr{L}$ has analytic eigenfunctions and rank $q \leq r$, such that

$$\sup_{t\in[0,1]} \mathbb{E}|Z(t) - Y(t)|^2 < \epsilon \quad \textit{and} \quad \sup_{s,t\in[0,1]} \left|g(s,t) - \ell(s,t)\right| < \epsilon,$$

where g and ℓ are the kernels of $\mathscr{G}$ and $\mathscr{L}$, respectively.

Proof. We will first prove the results referring to the processes Z and Y, and then those referring to their covariances, $\mathscr{G}$ and $\mathscr{L}$. Let μ be the mean function of Z and $\mathscr{G} = \sum_{n=1}^r \theta_n \phi_n \otimes \phi_n$, be the spectral representation of its covariance operator, with $\{\theta_n, \phi_n\}$ the corresponding eigenvalues/eigenfunctions. The Karhunen-Loève expansion of Z is given by

$$Z = \mu + \sum_{n=1}^{r} \langle Z - \mu, \phi_n \rangle_{L^2} \phi_n.$$

Now let $\epsilon > 0$ be arbitrary, and define $\gamma = \epsilon / \|\{\mathscr{G}\|_{TR}$. Define the function $f_{n,J}$ to be the order

J Fourier series approximation of ϕ_n, and note that this is an analytic function for all $J < \infty$ (and of course all n). Since Fourier series are dense in L^2, we know that there exists $J_1, ..., J_r$ such that

$$\|\phi_n - f_{n,J_n}\|_{L^2} < \gamma.$$

In particular, if we we pick $J_* = \max\{J_1, ..., J_r\}$ and define $f_n = f_{n,J^*}$, we have that

$$\sup_{1 \le n \le r} \|\phi_n - f_n\|_{L^2} < \gamma.$$

The functions f_n are, of course, analytic. Finally, define a new random function Y via the random series

$$Y = \mu + \sum_{n=1}^{r} \underbrace{\langle Z - \mu, \phi_n \rangle_{L^2}}_{=\xi_n} f_n = Z + \sum_{n=1}^{r} \xi_n e_n,$$

where $e_n = f_n - \phi_n$ satisfies $\|e_n\|_{L^2} < \gamma$. Note that since the $\{f_n\}_{n=1}^r$ are analytic and finitely many, their span consists of analytic functions. Thus the eigenfunctions of the covariance of Y (which are not necessarily exactly equal to the f_n) are analytic too. Furthermore, the rank of Y can clearly not exceed r, whatever the value of ϵ. Now, since the $\{\xi_n\}$ are mean-zero uncorrelated random variables of variance $\{\theta_n\}$, we may write

$$\begin{aligned} \mathbb{E}\|Z - Y\|_{L^2}^2 &= \mathbb{E}\int_0^1 \left(\sum_{n=1}^{r} \xi_n e(t) \right)^2 dt = \int_0^1 \mathbb{E}\left(\sum_{n=1}^{r} \xi_n e(t) \right)^2 dt = \int_0^1 \sum_{n=1}^{r} \theta_n e_n^2(t) dt \\ &= \sum_{n=1}^{r} \theta_n \|e_n\|_{L^2}^2 < \gamma \sum_{n=1}^{r} \theta_n = \gamma \|\mathscr{G}\|_{TR} = \epsilon. \end{aligned}$$

If we happen to know that $\{\phi_n\}$ are C^1, we may define again $\gamma = \epsilon / \|\mathscr{G}\|_{TR}$, but now re-define f_n to be trigonometric functions such that

$$\sup_{1 \le n \le r} \|\phi_n - f_n\|_{\infty} < \gamma^{1/2},$$

where $\|\phi_n\|_{\infty} = \sup\{|\phi_n(t)| : t \in [0,1]\}$. This is possible, since the eigenfunctions $\{\phi_n\}$ are C^1, and thus can be uniformly approximated by Fourier series. Define Y and e_n as before, but with the new definition of f_n in place. Once again, since the $\{\xi_n\}$ are mean-zero uncorrelated random variables of variance $\{\theta_n\}$, we have that for any $t \in [0,1]$,

$$\mathbb{E}(Z(t) - Y(t))^2 = \mathbb{E}\left[\sum_{n=1}^{r} \xi_n e_n(t) \right]^2 = \sum_{i=1}^{r} \theta_n e_n^2(t) < \gamma \|\mathscr{G}\|_{TR} = \epsilon.$$

Now let us focus on the approximation of $\mathscr{G}$ itself. Let $\epsilon > 0$, and define $\gamma = \epsilon / (2\|\mathscr{G}\|_{TR})$. Define the function $f_{n,J}$ to be the order J Fourier series approximation of the eigenfunction ϕ_n, as before. Again, there exist $J_1, ..., J_r$ such that

$$\|\phi_n - f_{n,J_n}\|_{L^2} < \gamma.$$

Set $J_* = \max\{J_1, ..., J_r\}$ and define $f_n = f_{n,J^*}$, so that

$$\sup_{1\le n\le r} \|\phi_n - f_n\|_{L^2} < \gamma.$$

The functions f_n are, of course, analytic. Now define the operator $\mathscr{L}$ to be

$$\mathscr{L} = \sum_{n=1}^{r} \theta_n f_n \otimes f_n.$$

This operator is analytic and has rank at most r. Furthermore, its eigenfunctions are analytic since they lie in the range of $\mathscr{L}$, which is spanned by the analytic functions f_n. We now have:

$$\begin{aligned}
\|\mathscr{G} - \mathscr{L}\|_{TR} &\le \sum_{n=1}^{r} \theta_n \|\phi_n \otimes \phi_n - f_n \otimes f_n\|_{TR} \\
&= \sum_{n=1}^{r} \theta_n \|\phi_n \otimes \phi_n - \phi_n \otimes f_n + \phi_n \otimes f_n - f_n \otimes f_n\|_{TR} \\
&\le \sum_{n=1}^{r} \theta_n \left\{\|\phi_n \otimes (\phi_n - f_n)\|_{TR} + \|(\phi_n - f_n) \otimes f_n\|_{TR}\right\} \\
&= \sum_{n=1}^{r} \theta_n \left\{\|\phi_n\|_{L^2}\|\phi_n - f_n\|_{L^2} + \|\phi_n - f_n\|_{L^2}\|f_n\|_{L^2}\right\} \\
&= \sum_{n=1}^{r} \theta_n (1 + \|f_n\|_{L^2})\|\phi_n - f_n\|_{L^2} \\
&< 2\gamma\|\mathscr{G}\|_{TR} = \epsilon,
\end{aligned}$$

where we used the fact that $\|f_n\|_{L^2} < 1$ and where the fourth line is obtained from the equality

$$\begin{aligned}
\|u \otimes v\|_{TR} &= \sum_{i=1}^{\infty} \langle (u \otimes v)^* (u \otimes v)^{1/2} e_i, e_i\rangle_{L^2} \\
&= \sum_{i=1}^{\infty} \langle \|v\|_{L^2}\|u\|_{L^2}^{-1} (u \otimes u) e_i, e_i\rangle_{L^2} \\
&= \sum_{i=1}^{\infty} \|v\|_{L^2}\|u\|_{L^2}^{-1} \langle u, e_i\rangle_{L^2}^2 = \|v\|_{L^2}\|u\|_{L^2},
\end{aligned}$$

where $\{e_i\}$ is an arbitrary orthonormal basis of $L^2([0,1])$. If we know that the eigenfunctions $\{\phi_n\}$ of $\mathscr{G}$ are C^1, the Fourier series expansion of each $\phi_n(t)$ converges *uniformly* and *absolutely*. Let $c_1 < \infty$ be the maximum of the ℓ_1 norms of the Fourier coefficients of $\phi_1, ..., \phi_r$ ($c_1 < \infty$ by absolute convergence of the respective Fourier series). Re-define

$$\gamma = \epsilon \times \left[\left(c_1 + \sup_{1\le n\le r} \|\phi_n\|_\infty\right)\|\mathscr{G}\|_{TR}\right]^{-1}.$$

Following the same steps as before, we can choose a J^* sufficiently large, such that setting

$f_n = f_{n,J^*}$ we have

$$\sup_{1\le n\le r} \|\phi_n - f_n\|_\infty < \gamma.$$

It now follows that

$$\begin{aligned}
\|g-\ell\|_\infty &\le \sum_{n=1}^{r} \theta_n \sup_{s,t} |\phi_n(s)\phi_n(t) - f_n(s)f_n(t)| \\
&= \sum_{n=1}^{r} \theta_n \sup_{s,t} |\phi_n(s)\phi_n(t) - \phi_n(s)f_n(t) + \phi_n(s)f_n(t) - f_n(s)f_n(t)| \\
&\le \sum_{n=1}^{r} \theta_n \left\{ \sup_t \sup_s |\phi_n(s)| \left|\phi_n(t) - f_n(t)\right| + \sup_s \sup_t |f_n(t)| \left|\phi_n(s) - f_n(s)\right| \right\} \\
&\le \sum_{n=1}^{r} \theta_n (c_1 + \sup_t |\phi_n(t)|) \|\phi_n - f_n\|_\infty \\
&< \left(c_1 + \sup_{1\le n\le r} \|\phi_n\|_\infty \right) \gamma \|\mathscr{G}\|_{TR} = \epsilon.
\end{aligned}$$

Finally, for any $\epsilon > 0$, we can replace the specific truncation $J^*(\epsilon)$ used in each of the four parts of the proof, by the largest of all these $J^*(\epsilon)$, and so ϵ can be chosen to be the same in all the approximation results. This concludes the proof.

□

2.2.2 Uniqueness of the decomposition $\mathscr{R}^K = \mathscr{L}^K + \mathscr{B}^K$

As we already mentioned, in practice we can't observe the function X on the whole interval $[0,1]$ but only on a grid $(t_1,\dots,t_K) \in \mathscr{T}_K$ of K points. This means that we need to work with a discrete version of the objects $\mathscr{R}, \mathscr{L}$ and $\mathscr{B}$, i.e., with their K-resolution representations that is denoted respectively by $\mathscr{R}^K, \mathscr{L}^K$ and $\mathscr{B}^K$, and defined respectively through the covariance functions

$$\begin{aligned}
\rho^K(x,y) &= \sum_{i,j=1}^{K} \rho(t_i,t_j)\mathbf{1}\{(x,y) \in I_{i,K} \times I_{j,K}\}, \\
\ell^K(x,y) &= \sum_{i,j=1}^{K} \ell(t_i,t_j)\mathbf{1}\{(x,y) \in I_{i,K} \times I_{j,K}\}, \\
b^K(x,y) &= \sum_{i,j=1}^{K} b(t_i,t_j)\mathbf{1}\{(x,y) \in I_{i,K} \times I_{j,K}\}.
\end{aligned}$$

Since these three K-resolution covariance functions are uniquely defined by $K \times K$ coefficients, we use the following matrices to represent them

$$R^K(i,j) = \rho(t_i,t_j), \quad L^K(i,j) = \ell(t_i,t_j), \quad B^K(i,j) = b(t_i,t_j).$$

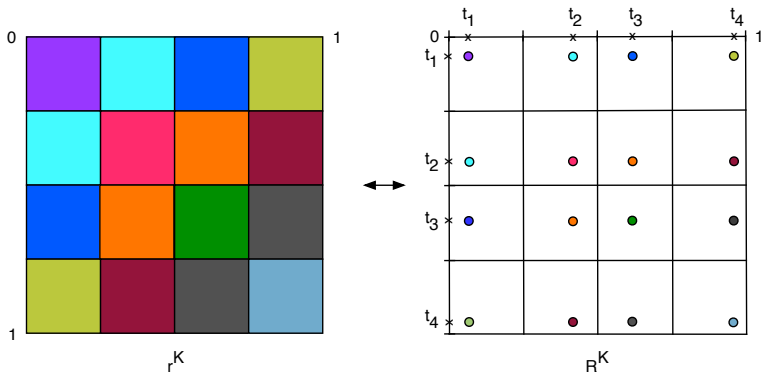

Figure 2.2 – Illustration of a K-resolution covariance function and of its matrix representation for the case $K = 4$

Figure 2.2 illustrates the concept of a K-resolution covariance function and of its matrix representation. Note that the Hilbert–Schmidt norm of a K-resolution covariance operator can be written in terms of the Frobenius norm of its matrix representation as

$$\begin{aligned} \|\mathscr{R}^K\|_{HS}^2 = \int_0^1 \int_0^1 (\rho^K(s,t))^2 dtds &= \sum_{i=1}^K \sum_{j=1}^K \int_{I_{i,K}} \int_{I_{j,K}} (\rho(t_i,t_j))^2 dtds \\ &= \sum_{i=1}^K \sum_{j=1}^K \frac{1}{K^2} (\rho(t_i,t_j))^2 = \frac{1}{K^2} \|R^K\|_F^2. \end{aligned}$$

We are now interested in finding conditions under which the decomposition $\mathscr{R}^K = \mathscr{L}^K + \mathscr{B}^K$ is identifiable, and for this, we will work with the equivalent matrix decomposition given by $R^K = L^K + B^K$. From the assumption $\text{rank}(\mathscr{L}) = r$, we have that the matrix L^K is at most of rank r. On the other hand, since the covariance function b of $\mathscr{B}$ is δ-banded and due to the regularity of the grids contain in $\mathscr{T}_K$, we have that the matrix B^K is a banded matrix such that $B^K(i,j) = 0$ if $|i-j| > \lceil \delta \cdot K \rceil$, and from now on we will refer to it as the δ-banded matrix. To sum up, we are looking for conditions under which it is possible to uniquely recover the matrices L^K and B^K from R^K, when L^K is at most of rank r and B^K is δ-banded, which is a particular case of a sparse structure. We are essentially dealing with a decomposition of the type "low-rank + sparse". This kind of problem has received a lot of attention in recent years, see Remark 2.2.1 below, and it is also related to low-rank matrix completion. In fact, an appealing aspect of our problem is that we know the sparsity pattern of the matrix B^K exactly, so we can translate our "low-rank + sparse" decomposition into a low-rank matrix completion problem in the following way. Define the matrix $R^K_{s_\delta}$ as the matrix R^K from which we have removed the values on the band $s_\delta = \{(i,j) \in \{1,\dots,K\}^2 | |i-j| \leq \lceil \delta \cdot K \rceil\}$ (by removing the values we mean that we now consider them as unknown). We are then looking for conditions under which a completion of $R^K_{s_\delta}$ of at most rank r will be unique and equal to L^K. It turns out

that the conditions to ensure such a unique completion are those that ensure the unique decomposition of $\mathscr{R} = \mathscr{L} + \mathscr{B}$, plus a condition on the relation between the resolution K, the rank r of the smooth operator and the bandwidth δ of the banded operator. The result is stated in Theorem 2.2.2 below .

Remark 2.2.1. *In recent years, considerable attention has been dedicated to two statistical problems involving the estimation of a low-rank matrix from partial or contaminated information: namely the low-rank matrix completion problem and the low-rank plus sparse decomposition problem. In the first problem, some entries of the data matrix are unobserved and in the second one, some of its entries have been corrupted by noise of arbitrary magnitude. The classical example of an incompletely observed low-rank matrix is the matrix obtained from users ranking of movies on Netflix ([Net]) and examples of data that can be effectively modelled as the sum of a low-rank matrix plus a sparse one are abundant in imaging data (Basri and Jacobs [BJ03], Candès et al. [CLMW11]).*

The problem of low-rank matrix completion can be translated into mathematical terms as follows. Let M_0 be a $n \times n$ low-rank matrix and let M be equal to M_0 for the entries $(i, j) \in \Omega$, where Ω is a subset of $[n] \times [n]$, with $[n] = \{1, 2, \ldots, n\}$, and be unobserved for the other entries. The goal is to recover M_0 from M, and for this one would like to solve the optimisation problem

$$\min_{\theta \in \mathbb{R}^{n \times n}} \operatorname{rank}(\theta) \qquad \textit{subject to} \qquad M_{ij} = \theta_{ij} \;\; \forall (i, j) \in \Omega, \tag{2.2.1}$$

and hope that there is a unique solution which is equal to M_0. Unfortunately, this problem is computationally intractable. But we know from the work of Fazel [Faz02] that the nuclear norm of a matrix gives a good approximation of its rank (more precisely that the nuclear norm is the convex envelope of the rank function), where we recall that the nuclear norm of a matrix A is the sum of its singular values and is denoted by $\|A\|_*$. Using the nuclear norm as a surrogate of the rank in (2.2.1) gives a convex relaxed version of the low-rank matrix completion problem and this version has received considerable attention. A key contribution on this subject is that of Candès and Recht [CR09], who give conditions to ensure that the solution of the relaxed problem is unique and equal to M_0 with high probability. These conditions can be summarised as follows. The low rank matrix M_0 has to satisfy the standard and strong incoherence properties which among other things, imply that M_0 is not sparse. Moreover the number m of missing values has to be smaller than a given bound and the subset Ω has to be uniformly distributed over all subset of m values of $[n] \times [n]$. Using these assumptions, considerable work has been done to improve the bound on m (Candès and Tao [CT10], Recht [Rec11], Gross [Gro11]) and to extend the model to the case where the observed entries of M are corrupted versions of those of M_0 (Candès and Plan [CP10], Chen at al. [CJSC13]). On the other hand, Chen [Che15] shows that strong incoherence is not a necessary condition. More importantly still, Chen et al. [CBSW14] show that neither incoherence property (standard or strong) is necessary if, instead of being chosen uniformly at random, the revealed entries are chosen proportionally to the local row and column coherence of the matrix, which roughly speaking means that the more a row or a column is sparse, the more of its entries should be revealed.

Let's now focus on the decomposition of a matrix into a low-rank matrix plus a sparse matrix. Let M be a $n \times n$ matrix equal to $L_0 + S_0$ where L_0 is a low-rank matrix and S_0 is a sparse one. Using the observed matrix M, the goal is to recover both L_0 and S_0, and to do this one would like to solve the optimisation problem

$$\min_{L,S \in \mathbb{R}^{n\times n}} \operatorname{rank}(L) + \lambda \|S\|_0 \qquad \textit{subject to} \qquad M = L + S, \tag{2.2.2}$$

where λ is a tuning parameter and the pseudo-norm $\|S\|_0$ yields the number of non-zero entries of S. Again this problem is computationally intractable, so a common approach is to work with the relaxed version given as follows. The rank is replaced by the nuclear norm as before, and the ℓ_0 pseudo-norm is replaced by the ℓ_1-norm. Note that the ℓ_1-norm is known to be a good surrogate when dealing with cardinality minimisation since it is the convex envelope of the ℓ_0 pseudo norm. Recall that the ℓ_1-norm of a $n \times n$ matrix A is given by $\|A\|_1 = \sum_{i,j=1}^{n} |a_{ij}|$. In Candès et al. [CLMW11], conditions on L_0 and B_0 that guarantee that the solution $(\hat{L}, \hat{S})$ of the convex relaxed version of (2.2.2) is unique and such that $(\hat{L}, \hat{S}) = (L_0, S_0)$ with high probability are provided. These conditions are quite similar to those for the low-rank matrix completion problem, namely the matrix L_0 has to satisfy the standard and the strong incoherence properties and the sparsity pattern of S_0 has to be selected uniformly at random. An extension of this paper is Zhou et al. [ZLW+ 10], which considers the more general model $M = L_0 + S_0 + Z$, where Z is a noise matrix. Instead of assuming that the sparsity pattern of S_0 is random, Chandrasekaran et al. [CaSPW11] develop deterministic conditions to ensure that the non-zero entries are "well spread". The work in Chen et al. [CJSC13] unifies the low-rank matrix completion and the low-rank plus sparse decomposition problems where both the missing values (the erasures) pattern and the sparsity pattern of the errors can be either random or deterministic. Particular cases of their main unified result recover the results of Candès and Recht [CR09], Candès and Tao [CT10], Chandrasekaran et al. [CaSPW11] and Candès et al. [CLMW11].

Theorem 2.2.2 (Discrete Identifiability)**.** *Let $\mathscr{L}_1$ and $\mathscr{L}_2$ be covariance operators of finite ranks $r_1 < \infty$ and $r_2 < \infty$, respectively, and assume without loss of generality that $r_1 \geq r_2$. Let $\mathscr{B}_1$ and $\mathscr{B}_1$ be two banded continuous covariance operators of bandwidth $\delta_1 < 1/2$ and $\delta_2 < 1/2$ respectively. Given $(t_1, \ldots, t_K) \in \mathscr{T}_K$, define their K-resolution matrix coefficients to be $(L_1^K, B_1^K, L_2^K, B_2^K) \in \mathbb{R}^{K\times K}$,*

$$L_m^K(i,j) = \ell_m(t_i, t_j) \quad \textit{and} \quad B_m^K(i,j) = b_m(t_i, t_j), \quad i,j \in \{1,\ldots,K\}, \quad \textit{for } m = 1,2.$$

If the eigenfunctions of $\mathscr{L}_1$ and $\mathscr{L}_2$ are all real analytic, and

$$K \geq K^* = \max\left(\frac{2r_1+2}{1-2\delta_1}, \frac{2r_1+2}{1-2\delta_2}\right),$$

then we have the equivalence

$$L_1^K + B_2^K = L_2^K + B_2^K \iff L_1^K = L_2^K \quad \& \quad B_1^K = B_2^K,$$

almost everywhere on $\mathscr{T}_K$ with respect to Lebesgue measure.

We now have a result that holds almost everywhere on $\mathscr{T}_K$ but not pointwise on it. This means that it is valid for almost all grids in $\mathscr{T}_K$, or equivalently, that if we choose a grid uniformly at random in $\mathscr{T}_K$, then the result hold with probability one. This restriction comes from the key result stated in the following theorem, which is fundamental in proving Theorem 2.2.2.

Theorem 2.2.3. *Let $\mathscr{L}$ be a covariance operator associated with the covariance function $\ell(s,t) = \sum_{i=1}^{r} \lambda_i \eta_i(s)\eta_i(t)$ with $r < \infty$ and real analytic orthonormal eigenfunctions $\{\eta_1, \ldots, \eta_r\}$. If $K > r$, then the minors of order r of the matrix $L^K = \{\ell(t_i, t_j)\}_{i,j=1}^{K}$ are all non-zero, almost everywhere on $\mathscr{T}_K$.*

Proof. First notice that from $\ell(s,t) = \sum_{i=1}^{r} \lambda_i \eta_i(s)\eta_i(t)$, we have

$$L^K(j,l) = \sum_{i=1}^{r} \lambda_i \eta_i(t_j)\eta_i(t_l).$$

Thus, L^K can be written as $U^K \Sigma (U^K)^\top$, where

$$U^K = \begin{pmatrix} \eta_1(t_1) & \eta_2(t_1) & \cdots & \eta_r(t_1) \\ \eta_1(t_2) & \eta_2(t_2) & \cdots & \eta_r(t_2) \\ \vdots & \vdots & & \vdots \\ \eta_1(t_K) & \eta_2(t_K) & \cdots & \eta_r(t_K) \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_r \end{pmatrix}.$$

Any $r \times r$ submatrix of L^K obtained by deleting rows and columns, can then be written as

$$U_F^K \Sigma (U_{F'}^K)^\top,$$

where U_F^K (resp. $U_{F'}^K$) is an $r \times r$ matrix obtained by deleting rows of U^K whose indices are not included in $F \subseteq \{1, \ldots, K\}$ (resp. F'). The condition that any minor of order r of L^K be non-zero is then equivalent to the condition that

$$\det\left[U_F^K \Sigma (U_{F'}^K)^\top\right] = \det[U_F^K]\det[\Sigma]\det[U_{F'}^K] \neq 0,$$

for any subset $F, F' \subseteq \{1, \ldots, K\}$ of cardinality r. By construction $\det(\Sigma) \neq 0$, so the minor condition is then equivalent to requiring that $\det(U_F^K) \neq 0$ for any subset $F \subseteq \{1, \ldots, K\}$ of cardinality r.

We will show that this is indeed the case almost everywhere on $\mathscr{T}_K$. Note first that $\mathscr{T}_K$ can be seen as a hypercube of dimension K with side length $1/K$, where each point in the hypercube represents a different grid, and so we can define the Lebesgue measure μ on $\mathscr{T}_K$. Let $F = \{1, \ldots, r\}$, without loss of generality (so that U_F^K is formed by keeping the first r rows of U^K). Using the Leibniz formula, we have that $\det(U_F^K)$ can be written as the function

$$D(t_1, \ldots, t_r) = \sum_{\sigma \in S_r} \varepsilon(\sigma) \prod_{i=1}^{r} \eta_i(t_{\sigma(i)}),$$

where S_r is the symmetric group on r elements and $\varepsilon(\sigma)$ is the signature of the permutation σ. The function D is real analytic on $(0,1)^r$, by virtue of each η_i being real analytic on $(0,1)$.

We will now proceed by contradiction. Assume that

$$\mu\{(x_1,\dots,x_K)\in\mathscr{T}_K : D(x_1,\dots,x_r)=0\}>0.$$

Since μ is Lebesgue measure, it follows that the Hausdorff dimension of the set $A=\{(x_1,\dots,x_r)\in(0,1)^r : D(x_1,\dots,x_r)=0\}$ is equal to r. However, since D is analytic, Theorem 6.33 of Krantz and Parks [KP02] implies the dichotomy: either D is constant everywhere on $(0,1)^r$, or the set A is at most of dimension $r-1$. Thus it must be that D is everywhere constant on $(0,1)^r$, the constant being of course zero:

$$D(x_1,\dots,x_r)=\sum_{\sigma\in S_r}\varepsilon(\sigma)\prod_{i=1}^{r}\eta_i(x_{\sigma(i)})=0,\ \forall\ (x_1,\dots,x_r)\in(0,1)^r.$$

Now fix $(x_1,\dots,x_{r-1})$ and apply to D (viewed as a function of x_r only) the continuous linear functional $T_{\eta_r}(f)=\langle f,\eta_r\rangle$. We obtain that for all $(x_1,\dots,x_{r-1})\in(0,1)^r$:

$$\begin{aligned}0=\langle D,\eta_r\rangle &= \sum_{\sigma\in S_r}\varepsilon(\sigma)\left[\prod_{i:\sigma(i)\neq r}\eta_i(x_{\sigma(i)})\right]\langle\eta_{\sigma^{-1}(r)},\eta_r\rangle\\ &= \sum_{\sigma\in S_{r-1}}\varepsilon(\sigma)\prod_{i=1}^{r-1}\eta_i(x_{\sigma(i)}).\end{aligned}$$

Applying the continuous linear functionals $T_{\eta_j}(f)=\langle f,\eta_j\rangle$ iteratively to D while keeping $(x_1,\dots,x_{j-1})$ fixed then leads to

$$\eta_1(y)=0,\ \forall\ y\in(0,1).$$

This last equality contradicts the fact that η_1 is of norm one, and allows us to conclude that $\mu\{(x_1,\dots,x_K)\in\mathscr{T}_K : D(x_1,\dots,x_r)=0\}=0$.

□

The proof of Theorem 2.2.2 below is inspired by Proposition 2.12 of Király and Tomioka [KT12] and makes great use of Theorem 2.2.3.

Proof of Theorem 2.2.2. Let $\delta=\max\{\delta_1,\delta_2\}$ and assume without loss of generality that $r_1\geq r_2$. Let Ω be the set of indices on which both B_1 and B_2 vanish, which is $\Omega=\{(i,j)\in\{1,\dots,K\}^2 : |i-j|>\lceil\delta\cdot K\rceil\}$ due to the bandwidth assumption imposed on $\mathscr{B}_1$ and $\mathscr{B}_2$ and the fact that $(t_1,\dots,t_K)\in\mathscr{T}_K$. From $L_1+B_1=L_2+B_2$, we obtain that $L_1(i,j)=L_2(i,j),\forall(i,j)\in\Omega$. Let Ω_A be the set of indices of a submatrix formed by the first r_1 rows and the last r_1 columns of a $K\times K$ matrix, the condition $K\geq K^*=\frac{2r_1+2}{1-2\delta}$ implies that $\Omega_A\subset\Omega$, which in turn implies that the matrices L_1 and L_2 contain a common submatrix A of dimension $r_1\times r_1$.

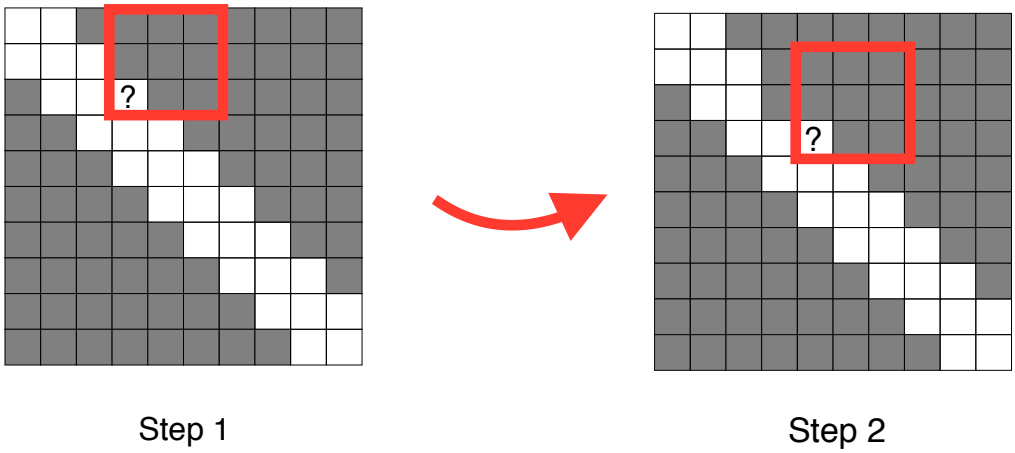

Figure 2.3 – Illustration of two possible first steps of the iterative procedure explained in the proof of Theorem 2.2.2. The gray squares represent known entries and the white ones unknown entries. In this example, $K = 10, \delta = 0.2$ and $r_1 = 2$ and the red squares represent a $r_1 \times r_1$ matrix with only one unknown entry.

Assume that all minors of order r_1 of L_1 are non-zero. Then, the determinant of A is non-zero, which implies that the rank of L_2 is also r_1. We thus establish that L_1 and L_2 are two rank r_1 matrices equal on Ω. Let L^* be a matrix equal to L_1 on Ω, but unknown at those indices that do not belong to Ω. We will now show that there exists a unique rank r_1 completion of L^*. Due to the banded pattern of the unobserved entries of L^* and the inequality $K \geq K^* = \frac{2r_1+2}{1-2\delta}$, it is possible to find a submatrix of L^* of dimension $(r_1+1) \times (r_1+1)$ with only one unobserved entry, denoted x^*. Using the fact that the determinant of any square submatrix of dimension bigger than $r_1 + 1$ is zero, we obtain a linear equation of the form $ax^* + b = 0$, where a is equal to the determinant of a submatrix of dimension $r_1 \times r_1$. Since we assume that any minor of order r_1 is non-zero, we have that $a \neq 0$ and the previous equation has a unique solution. It is then possible to impute the value of x^*. Applying this procedure iteratively until all missing entries are determined allows us to uniquely complete the matrix L^* into a rank r_1 matrix. Note that this iterative procedure is illustrated in Figure 2.3 .

In summary, we have demonstrated that when all minors of order r_1 of L_1 are non-zero, $L^* = L_1 = L_2$ and hence $B_1 = B_2$. Theorem 2.2.3 assures us that L_1 indeed has non-vanishing minors of order r_1 almost everywhere on $\mathcal{T}_K$, and so we conclude that it must be that $L_1 = L_2$ and $B_1 = B_2$ almost everywhere on $\mathcal{T}_K$. □

We now comment in more detail on the result stated in Theorem 2.2.2. As already mentioned, when we pass from the continuous to the discrete case, we still have a unique decomposition, but this time for almost every grid in $\mathcal{T}_K$ and only if this following relation between the number

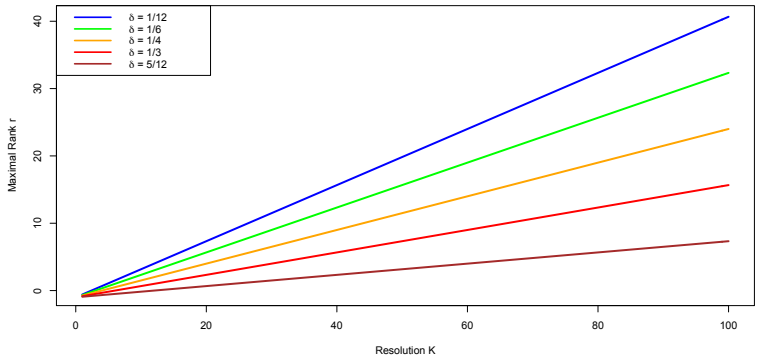

Figure 2.4 – Illustration of the relation between the resolution K, the rank r of the smooth operator and the bandwidth δ of the rough operator. For different values of δ, the maximal possible rank r is plotted as a function of the resolution K.

of grid points K, the rank r of $\mathscr{L}$ and the bandwidth δ of $\mathscr{B}$ holds:

$$r \leq K\left(\frac{1}{2} - \delta\right) - 1.$$

As depicted in Figure 2.4, we can see that this last condition is not very restrictive, since in the practice of FDA, one almost always assumes that the rank of the smooth operator is quite small (≤ 5), so it allows us to have identifiability for quite large values of δ and modest values of K.

Another interesting aspect of this theorem is that we can use it to derive directly an optimisation problem in order to uniquely recover L^K and B^K from R^K. This result is presented in Corollary 2.2.1 and will be very helpful in the next section in order to define estimators of $\mathscr{L}$ and $\mathscr{B}$.

Corollary 2.2.1. *Let $\mathscr{L}$ be a rank $r < \infty$ covariance operator with analytic eigenfunctions and kernel ℓ, and $\mathscr{B}$ a banded continuous covariance operator of bandwidth δ with kernel b. For $(t_1, \ldots, t_K) \in \mathscr{T}_K$, let*

$$L^K = \{\ell(t_i, t_j)\}_{ij}, \quad B^K = \{b(t_i, t_j)\}_{ij},$$

and $R^K = L^K + B^K$. Assume that

$$\delta < \frac{1}{4} \qquad \text{and} \qquad K \geq 4r + 4.$$

Define the matrix $P^K \in \mathbb{R}^{K\times K}$ by $P^K(i,j) = \mathbf{1}\left\{|i-j| > \lceil K/4 \rceil\right\}$. Then, for almost all grids in $\mathscr{T}_K$:

1. *The matrix L^K is the unique solution to the optimisation problem*

$$\min_{\theta \in \mathbb{R}^{K\times K}} \operatorname{rank}\{\theta\} \qquad \text{subject to } \left\| P^K \circ (R^K - \theta) \right\|_F^2 = 0. \tag{2.2.3}$$

2. *Equivalently, in penalised form,*

$$L^K = \underset{\theta \in \mathbb{R}^{K\times K}}{\arg\min} \left\{ \left\| P^K \circ (R^K - \theta) \right\|_F^2 + \tau \operatorname{rank}(\theta) \right\}, \tag{2.2.4}$$

for all $\tau > 0$ *sufficiently small.*

Here "$\circ$" *denotes the Hadamard product (entrywise product).*

Instead of letting $\delta < 1/2$ and $K \geq (2r+2)/(1-2\delta)$ as in Theorem 2.2.2, we have made the more restrictive assumption that $\delta < 1/4$ and $K \geq 4r+4$. Passing from $\delta < 1/2$ to $\delta < 1/4$ allows us to uniquely solve the optimisation problems (2.2.3) and (2.2.4) without knowing neither r nor δ, which is very appealing. Note also that the minimisation problem (2.2.3) is exactly a low-rank matrix completion. Indeed, it asks to find the matrix with the smallest rank that is equal to R^K everywhere except on the band $s_{1/4} = \{(i,j) \in \{1,\dots,K\}^2 | |i-j| \leq \lceil K/4 \rceil\}$, and this is exactly equivalent to seeking to complete the matrix $R^K_{s_{1/4}}$ (R^K without the band $s_{1/4}$) such that it has the smallest possible rank.

Proof. Since $\delta < 1/4$ and $K \geq 4r+4$ implies $K \geq \frac{2r+2}{1-2\delta}$, we can directly use Theorem 2.2.2, which tells us that L^K is the only matrix of rank at most r such that $P^K \circ R^K = P^K \circ L^K$, and then such that $\|P_K \circ (R^K - L^K)\|_F^2 = 0$. The objective function (2.2.3) thus achieves its minimal value of r at L^K. We now prove that

$$L^K = \underset{\theta \in \mathbb{R}^{K\times K}}{\arg\min} \left\{ \left\| P^K \circ (R^K - \theta) \right\|_F^2 + \tau \operatorname{rank}(\theta) \right\},$$

for all $\tau > 0$ sufficiently small. Since we have established that L^K uniquely solves

$$\min_{\theta \in \mathbb{R}^{K\times K}} \operatorname{rank}\{\theta\} \qquad \text{subject to } \left\| P^K \circ (R^K - \theta) \right\|_F^2 = 0,$$

it follows that for all $\tau > 0$ and any $\theta \in \mathbb{R}^{K\times K}$ of rank greater or equal to r, we have that

$$\left\| P^K \circ (R^K - L^K) \right\|_F^2 + \tau \operatorname{rank}(L^K) < \left\| P^K \circ (R^K - \theta) \right\|_F^2 + \tau \operatorname{rank}(\theta).$$

We thus concentrate on matrices $\theta \in \mathbb{R}^{K\times K}$ of rank at most $r-1$, for $r > 1$. Let

$$\mu = \min_{\theta \in \mathbb{R}^{K\times K},\ \operatorname{rank}(\theta) \leq r-1} \left\{ \left\| P^K \circ (R^K - \theta) \right\|_F^2 \right\} > 0.$$

Now let $\tau_* = \frac{\mu}{r-1}$. Then, for any $\tau < \tau_*$, and any θ of rank less than r,

$$\left\| P^K \circ (R^K - L^K) \right\|_F^2 + \tau \operatorname{rank}(L^K) = \tau r < \mu + \tau \leq \left\| P^K \circ (R^K - \theta) \right\|_F^2 + \tau \operatorname{rank}(\theta).$$

In summary, putting our results together, we have shown that for all $\tau \in (0, \tau_*)$,

$$L^K = \underset{\theta \in \mathbb{R}^{K\times K}}{\arg\min} \left\{ \left\| P^K \circ (R^K - \theta) \right\|_F^2 + \tau \operatorname{rank}(\theta) \right\}.$$

Finally, it is worth pointing out that although τ_* depends on r, this does not mean that the objective function depends on unknowns: r can be shown (using Theorem 2.2.3) to be equal to the rank of the submatrix formed by the first $\lceil K/4 \rceil$ rows and the last $\lceil K/4 \rceil$ columns of R^K, and thus we can determine τ_* directly from the matrix R^K. This completes the proof. □

2.3 Estimation of $\mathscr{L}$ and $\mathscr{B}$

In this section we will define estimators of $\mathscr{L}$ and $\mathscr{B}$, and we will prove that they are consistent. In order to do this, we will first define estimators of the matrices L^K and B^K. In the previous section, we gave conditions under which there exists a unique decomposition of the matrix R^K into a low-rank matrix L^K plus a banded matrix B^K, and Corollary 2.2.1 gave us two equivalent optimisation problems that one can solve to recover L^K from R^K. The idea to define our estimator of L^K is simply to use the optimisation problem (2.2.4) but where we have replaced the unknown matrix R^K by its empirical version as it is presented in the following definition.

Definition 2.3.1. *Let $(X_1, \dots, X_n)$ be i.i.d. copies of $X = Y + W$. Let $(t_1, \dots, t_K) \in \mathscr{T}_K$ and assume we observe*

$$X_i(t_j) = Y_i(t_j) + W_i(t_j), \quad \text{for } i = 1, \dots, n \text{ and } j = 1, \dots, K.$$

Let $R_n^K \in \mathbb{R}^{K\times K}$ be the empirical covariance matrix of our sample, i.e.

$$R_n^K(i,j) = \frac{1}{n} \sum_{a=1}^{n} (X_a(t_i) - \hat{\mu}(t_i))(X_a(t_j) - \hat{\mu}(t_j)), \text{ where } \hat{\mu}(t_i) = \frac{1}{n} \sum_{a=1}^{n} X_a(t_i).$$

We define the estimator $\hat{L}_n^K$ of L^K to be an approximate minimum of

$$\min_{0 \preceq \theta \in \mathbb{R}^{K\times K}} \quad \frac{1}{K^2} \left\{ \left\| P^K \circ (R_n^K - \theta) \right\|_F^2 + \tau_n \operatorname{rank}(\theta) \right\}, \tag{2.3.1}$$

where $P^K \in \mathbb{R}^{K\times K}$ is defined as $P^K(i,j) = \mathbf{1}\{|i-j| > \lceil K/4 \rceil\}$ and $\tau_n > 0$ is a tuning parameter such that $\tau_n \to 0$ as $n \to \infty$. By approximate minimum it is meant that the value of the functional at $\hat{L}_n^K$ is within $O_{\mathbb{P}}(n^{-1})$ of the value of the overall minimum.

Since the function to be minimized in (2.3.1) is not convex, it is not immediately clear how to solve the optimisation problem in practice, and it is also not clear how to choose the tuning parameter τ_n. We will discuss in detail how to get around these issues in Section 3.1. Now that we have an estimator of L^K, we could simply define the estimator of B^K as being $\Delta_n^K = R_n^K - \hat{L}_n^K$. However, we have no guarantee that this matrix will be positive definite or banded. We thus define our estimator of B^K as follows.

Definition 2.3.2. *Let R_n^K and $\hat{L}_n^K$ be as in Definition 2.3.1. We define the estimator $\hat{B}_n^K$ of B_n^K to be the projection of $\Delta_n^K = R_n^K - \hat{L}_n^K$ onto the convex set of non-negative banded $K \times K$ matrices of bandwidth at most $\lceil K/4 \rceil$.*

Once again, we will discuss how to obtain this estimator in practice in Section 3.1. Now that we have the estimators $\hat{L}_n^K$ and $\hat{B}_n^K$, we define their sum to be an estimator of R^K:

Definition 2.3.3. *Let $\hat{L}_n^K$ and $\hat{B}_n^K$ be as in Definitions 2.3.1 and 2.3.2. We define the plug-in estimator $\hat{R}_n^K$ of R^K to be $\hat{R}_n^K = \hat{L}_n^K + \hat{B}_n^K$.*

The estimator $\hat{R}_n^K$ should be preferred to the raw empirical estimator R_n^K, since due to the way we constructed it, it should be less noisy. Now that we have the estimators $(\hat{L}_n^K, \hat{B}_n^K, \hat{R}_n^K)$, we can define the K-resolution estimators $(\hat{\mathscr{L}}_n^K, \hat{\mathscr{B}}_n^K, \hat{\mathscr{R}}_n^K)$ of $(\mathscr{L}, \mathscr{B}, \mathscr{R})$ to have respectively the covariance functions

$$\begin{aligned}
\hat{\ell}_n^K(x,y) &= \sum_{j=1}^K \hat{L}_n^K(i,j)\mathbf{1}\{(x,y) \in I_{i,K} \times I_{j,K}\}, \\
\hat{b}_n^K(x,y) &= \sum_{j=1}^K \hat{B}_n^K(i,j)\mathbf{1}\{(x,y) \in I_{i,K} \times I_{j,K}\}, \\
\hat{\rho}_n^K(x,y) &= \sum_{j=1}^K \hat{R}_n^K(i,j)\mathbf{1}\{(x,y) \in I_{i,K} \times I_{j,K}\},
\end{aligned}$$

and the spectral decompositions

$$\hat{\mathscr{L}}_n^K = \sum_{j=1}^{\hat{r}} \hat{\lambda}_j \hat{\eta}_j^K \otimes \hat{\eta}_j^K, \quad \hat{\mathscr{B}}_n^K = \sum_{j=1}^K \hat{\beta}_j \hat{\psi}_j^K \otimes \hat{\psi}_j^K, \quad \hat{\mathscr{R}}_n^K = \sum_{j=1}^K \hat{\theta}_j \hat{\varphi}_j^K \otimes \hat{\varphi}_j^K,$$

where $\hat{r}$ is the rank of $\hat{\mathscr{L}}_n^K$. We place the subscript K on the eigenfunctions to emphasize that these functions are K-resolution estimators, i.e., piecewise constant functions, of the original eigenfunctions. In particular, the functions $\hat{\eta}_j^K, j = 1, \ldots, r$ are estimators of analytic functions, so if one wishes to have smoother estimators than piecewise constant functions, two strategies are possible. One could either smooth directly the eigenfunctions $\hat{\eta}_j^K, j = 1, \ldots, r$, but there would be no guarantee that the resulting functions would be orthogonal to each other. Another possibility is to smooth the covariance function ℓ_n^K, and then to work with the eigenfunctions of the resulting smooth operator (these will inherit the corresponding smoothness properties).

In the rest of this section we will establish the consistency of the estimators $\hat{\mathscr{L}}_n^K$, $\hat{\mathscr{B}}_n^K$ and $\hat{\mathscr{R}}_n^K$ and of their corresponding spectra. Of course our asymptotic framework will be that the number of observed curves is growing to infinity ($n \to \infty$), and that the grid of points on which we observe each curve is getting dense ($K \to \infty$). However to simplify the proof

of the consistency of $\hat{\mathscr{L}}_n^K$ (which is quite subtle), we will consider nested grids of points obtained from refinement of a grid of critical resolution $K^* = 4r+4$. Moreover, we will use the convention that the signs of the eigenfunctions are correctly identified. Our main consistency result, which is the consistency of our estimator of $\mathscr{L}$ and of its spectrum is presented in Theorem 2.3.1.

Theorem 2.3.1. *Let $X = Y + W$, $\mathscr{R} = \mathscr{L} + \mathscr{B}$ be defined as in Section 2.1, and let the $r < \infty$ eigenvalues of $\mathscr{L}$ be of multiplicity one, and the corresponding eigenfunctions be real analytic. Let $\mathbb{E}\|X\|_{L^2}^4 < \infty$, $\delta < \frac{1}{4}$, and define $K^* = 4r+4$ to be the critical resolution. Provided that $\tau_n \to 0$, it holds almost everywhere on $\mathscr{T}_K$ that*

$$\left\| \hat{\mathscr{L}}_n^K - \mathscr{L} \right\|_{\mathrm{HS}}^2 \leq O_{\mathbb{P}}(n^{-1}) + 4K^{-2} \sup_{x,y\in[0,1]} \|\nabla \ell(x,y)\|_2^2, \tag{2.3.2}$$

$$\left\| \hat{\eta}_j^K - \eta_j \right\|_{L^2}^2 \leq O_{\mathbb{P}}(n^{-1}) + 2K^{-2}\|\eta_j'\|_\infty^2, \quad j \in \{1,\dots,r\}, \tag{2.3.3}$$

$$\sup_{j\geq 1} |\hat{\lambda}_j^K - \lambda_j|^2 = O_{\mathbb{P}}(n^{-1}) + 4K^{-2} \sup_{x,y\in[0,1]} \|\nabla \ell(x,y)\|_2^2, \tag{2.3.4}$$

for any refinement $K = m \times K^*$, $m \geq 1$. All three $O_{\mathbb{P}}(n^{-1})$ terms are uniform in K. Furthermore, the rank of $\hat{\mathscr{L}}_n^K$ satisfies,

$$n\tau_n K^{-2} |\mathrm{rank}(\hat{\mathscr{L}}_n^K) - r| = O_{\mathbb{P}}(1). \tag{2.3.5}$$

This theorem tells us that we obtain a parametric rate of convergence for $\hat{\mathscr{L}}_n^K$, irrespective of the convergence rate of τ_n to zero, if the number of grid points K is at least of the order $\sqrt{n}$. On the other hand, if we want the rank estimator $\mathrm{rank}(\hat{\mathscr{L}}_n^K)$ to be consistent, K has to be $o(\sqrt{n\tau_n})$ with $n\tau_n \to \infty$. The take home message then seems to be that dense grids will lead to accurate estimation of the covariance operator but that sparse grid are more appropriate to accurately estimate the rank.

Proof. We begin by the usual bias/variance decomposition

$$\begin{aligned} \left\| \hat{\mathscr{L}}_n^K - \mathscr{L} \right\|_{\mathrm{HS}}^2 &\leq 2 \left\| \hat{\mathscr{L}}_n^K - \mathscr{L}^K \right\|_{\mathrm{HS}}^2 + 2 \left\| \mathscr{L}^K - \mathscr{L} \right\|_{\mathrm{HS}}^2 \\ &= 2K^{-2} \left\| \hat{L}_n^K - L^K \right\|_{\mathrm{F}}^2 + 2 \left\| \mathscr{L}^K - \mathscr{L} \right\|_{\mathrm{HS}}^2. \end{aligned}$$

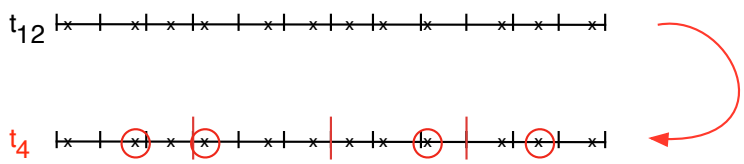

Figure 2.5 – Illustration of how to obtain a grid $\boldsymbol{t}_{K^*}$ from $\boldsymbol{t}_K$, for $K^* = 4$, and $K = 3 \cdot K^* = 12$.

For the second term (bias), we note that by a Taylor expansion

$$\begin{aligned}
\int_0^1 \int_0^1 (\ell(x,y) - \ell^K(x,y))^2 dxdy &= \sum_{i,j=1}^K \int_{I_{i,K}} \int_{I_{j,K}} (\ell(x,y) - \ell(t_i,t_j))^2 dxdy \\
&\leq \sum_{i,j=1}^K \int_{I_{i,K}} \int_{I_{j,K}} 2K^{-2} \sup_{(x,y)\in I_{i,K}\times I_{j,K}} \|\nabla \ell(x,y)\|_2^2 \\
&= \sum_{i,j=1}^K 2K^{-4} \sup_{(x,y)\in I_{i,K}\times I_{j,K}} \|\nabla \ell(x,y)\|_2^2 \\
&\leq 2K^{-2} \sup_{(x,y)\in[0,1]^2} \|\nabla \ell(x,y)\|_2^2.
\end{aligned}$$

It now remains to be shown that $K^{-2}\left\|\hat{L}_n^K - L^K\right\|_{\mathrm{F}}^2 = O_{\mathbb{P}}(n^{-1})$, and that the $O_{\mathbb{P}}(n^{-1})$ is uniform in K. To show that this is valid almost everywhere on $\mathscr{T}_K$ with respect to the Lebesgue measure, we will show equivalently that it is valid almost surely, when choosing a grid uniformly at random from $\mathscr{T}_K$. Using the fact that $K = m \times K^*$, $m \geq 1$,we consider a simple coupling construction, as this will be convenient for our proof:

1. Choose $t_{j,K}$ uniformly at random from $I_{j,K}$, independently for all $j = 1,\ldots,K$, writing $\boldsymbol{t}_K = \{t_{j,K}\}_{j=1}^K$.

2. Choose t_{j,K^*} uniformly at random among the m elements of $\boldsymbol{t}_K \cap I_{j,K^*}$, and write $\boldsymbol{t}_{K^*} = \{t_{j,K^*}\}_{j=1}^{K^*}$.

3. Define the set $q = \{j \in \{1,\ldots,K\} : t_{j,K} \in \boldsymbol{t}_{K^*}\}$ of cardinality K^*.

Figure 2.5 illustrates how a grid $\boldsymbol{t}_{K^*}$ with $K^* = 4$ is constructed from a grid $\boldsymbol{t}_K$ with $K = 12$. In this particular example $q = \{2,4,9,11\}$. Note that our construction guarantees that $\boldsymbol{t}_{K^*} \subseteq \boldsymbol{t}_K$, with $\boldsymbol{t}_{K^*}$ being marginally uniformly distributed on $\mathscr{T}_{K^*}$. Define Θ_K to be the space of $K \times K$ covariance operators of rank smaller than $K/4$. We consider the functionals

$$\mathbb{S}_{n,K} : \Theta_K \to [0,\infty), \qquad \mathbb{S}_{n,K}(\theta) = \underbrace{K^{-2}\|P^K \circ (\theta - R_n^K)\|_{\mathrm{F}}^2}_{\mathbb{M}_{n,K}(\theta)} + \tau_n K^{-2}\mathrm{rank}(\theta),$$

$$S_{n,K}:\Theta_K\to[0,\infty),\qquad S_{n,K}(\theta)=\underbrace{K^{-2}\|P^K\circ(\theta-R^K)\|_{\mathrm{F}}^2}_{M_K(\theta)}+\tau_n K^{-2}\mathrm{rank}(\theta),$$

where $P_K(i,j)=\mathbf{1}\{|i-j|>\lceil K/4\rceil\}$. We will show that for n sufficiently large, L^K is the unique minimizer of $S_{n,K}(\cdot)$ on Θ_K. Note that, since $K\geq 4r+4$, Theorem 2.2.2 implies that L^K is (almost surely) the unique matrix for which the functional $M_K(\theta)$ attains the minimal value of zero on Θ_K^r, where Θ_K^r is defined as the space of $K\times K$ covariance operators of rank at most r. It follows that for any $H\in\Theta_K$ different than L^K and of rank $p\in\{r,\dots,K/4-1\}$,

$$S_{n,K}(L^K)=M_K(L^K)+\frac{r\tau_n}{K^2}=\frac{r\tau_n}{K^2}\leq\frac{p\tau_n}{K^2}<M_K(H)+\frac{p\tau_n}{K^2}=S_{n,K}(H),$$

for all $n\geq 1$, almost surely. Thus the addition of the penalty term distinguishes L^K as the unique minimizer among all matrices of rank in $\{r,\dots,K/4\}$, almost surely. We will now show that for n sufficiently large, L^K almost surely uniquely minimises $S_{n,K}$, even when considering matrices of rank in $\{1,\dots,r-1\}$.

Let $\theta\in\mathbb{R}^{K\times K}$, and $\theta_q\in\mathbb{R}^{K^*\times K^*}$ be a matrix obtained by keeping only rows and columns of θ with indices in the index set q and define the function $\tau:[K^*]\to q$ such that $\theta_q(i,j)=\theta(\tau(i),\tau(j))$. Then, by the coupling construction of $\boldsymbol{t}_K$ and $\boldsymbol{t}_{K^*}$, it almost surely holds that

$$\begin{aligned}
M_K(\theta) &= \sum_{\substack{i,j\in[K]\\ |i-j|>\lceil K/4\rceil}} K^{-2}\Big|\theta(i,j)-r(t_{i,K},t_{j,K})]\Big|^2\\
&= \sum_{\substack{i,j\in q\\ |\tau^{-1}(i)-\tau^{-1}(j)|>\lceil K^*/4\rceil}} K^{-2}\Big|\theta(i,j)-r(t_{i,K},t_{j,K})]\Big|^2+\\
&\qquad\sum_{\substack{i,j\in q\\ |\tau^{-1}(i)-\tau^{-1}(j)|\leq\lceil K^*/4\rceil\\ |i-j|>\lceil K/4\rceil}} K^{-2}\Big|\theta(i,j)-r(t_{i,K},t_{j,K})]\Big|^2+\\
&\qquad\qquad+\sum_{\substack{i,j\notin q\\ |i-j|>\lceil K/4\rceil}} K^{-2}\Big|\theta(i,j)-r(t_{i,K},t_{j,K})]\Big|^2\\
&\geq \sum_{a,b\in[K^*]:|a-b|>\lceil K^*/4\rceil} K^{-2}\Big|\theta_q(a,b)-r(t_{a,K^*},t_{b,K^*})]\Big|^2\\
&= \left(\frac{K^*}{K}\right)^2 M_{K^*}(\theta_q).
\end{aligned}$$

As θ ranges over $\mathbb{R}^{K\times K}$ matrices of rank at most $r-1$, the sub-matrix θ_q ranges over $\mathbb{R}^{K^*\times K^*}$ matrices of rank at most $r-1$. It follows that

$$\inf_{\substack{\theta\in\Theta_K\\ \mathrm{rank}(\theta)<r}} M_K(\theta)\geq\left(\frac{K^*}{K}\right)^2\inf_{\substack{\theta\in\Theta_{K^*}\\ \mathrm{rank}(\theta)<r}} M_{K^*}(\theta)=\left(\frac{K^*}{K}\right)^2\epsilon(\boldsymbol{t}_{K^*}),\quad\text{almost surely.}$$

Since $K^* = 4r+4$, $\epsilon(\boldsymbol{t}_{K^*}) > 0$ almost surely, as $\boldsymbol{t}_{K^*}$ is uniformly distributed on $\mathcal{T}_{K^*}$. Thus, if n is sufficiently large so that $\tau_n < (K^*)^2\epsilon(\boldsymbol{t}_{K^*})/r$ a.s., it holds that

$$\left(\frac{K^*}{K}\right)^2 \epsilon(\boldsymbol{t}_{K^*}) > r\tau_n K^{-2}, \qquad \text{almost surely.}$$

Note that the critical $N(\boldsymbol{t}_{K^*})$ that n needs to surpass for this to hold true does not depend on K, but only on the distribution of $\boldsymbol{t}_{K^*}$, which is the uniform distribution on $\mathcal{T}_{K^*}$, and on the precise rate of convergence of τ_n to zero (which is for us to choose). Consequently, $N(\boldsymbol{t}_{K^*})$ *does not grow* with K. Summarising, provided that n is sufficiently large,

$$\inf_{\substack{\theta\in\Theta_K \\ \text{rank}(\theta)<r}} M_K(\theta) > r\tau_n K^{-2} = S_{n,K}(L^K), \quad \text{almost surely.}$$

But

$$\inf_{\substack{\theta\in\Theta_K \\ \text{rank}(\theta)<r}} S_{n,K}(\theta) = \inf_{\substack{\theta\in\Theta_K \\ \text{rank}(\theta)<r}} \left\{M_K(\theta) + \tau_n K^{-2}\text{rank}(\theta)\right\} > \inf_{\substack{\theta\in\Theta_K \\ \text{rank}(\theta)<r}} M_K(\theta), \text{ almost surely,}$$

which shows L^K to be the *unique* minimizer of $S_{n,K}(\cdot)$ on Θ_K almost surely, for n sufficiently large. Rephrasing, we have shown that for almost any grid $\boldsymbol{t}_K \in \mathcal{T}_K$, there exists an N (independent of K), such for all $n \geq N$, the matrix L^K is the unique minimum of $S_{n,K}$.

From now on, take $n \geq N(\boldsymbol{t}_{K^*})$, and note that all statements are valid almost surely with respect to the random grid $\boldsymbol{t}_K$. Our strategy will be to make use of Theorem 3.4.1 of van der Vaart and Wellner [vdVW96]. To this aim, define $d_n : \Theta_K \times \Theta_K \to [0,\infty)$ as

$$d_n(\theta_1,\theta_2) = \sqrt{K^{-2}\|\theta_1-\theta_2\|_F^2 + \tau_n K^{-2}|\text{rank}(\theta_1) - \text{rank}(\theta_2)|},$$

noting that d_n *does not* need to be a metric. To obtain the result, we need to show that for $\gamma > 0$,

(i) $\sup_{\theta\in\Theta_K : d_n(L^K,\theta)<\gamma} |\Delta(\theta)| \leq \gamma^2$, with $\Delta(\theta) = S_{n,K}(\theta) - S_{n,K}(L^K)$,

(ii) $\mathbb{E}\left\{\sqrt{n}\sup_{\theta\in\Theta_K : d_n(\theta,L^K)<\gamma} |D(\theta)|\right\} \lesssim \gamma$, with $D(\theta) = \mathbb{S}_{n,K}(\theta) - S_{n,K}(\theta) - \mathbb{S}_{n,K}(L^K) + S_{n,K}(L^K)$.

To simplify the proof, we will work with vector representations of the matrices involved in the functions Δ and D, and we will indicate this modification at the level of the matrices and of the functions themselves by the addition of a tilde in the notation. Recall that the vector representation of a $K \times K$ matrix is obtained by serially stacking its columns over each other. Let's first show (i). Consider the difference

$$\tilde{\Delta}(\tilde{\theta}) = \tilde{S}_{n,K}(\tilde{\theta}) - \tilde{S}_{n,K}(\tilde{L}^K) = \tilde{M}_K(\tilde{\theta}) - \tilde{M}_K(\tilde{L}^K) + \tau_n \text{rank}(\theta)K^{-2} - \tau_n r K^{-2}.$$

By doing a second order Taylor expansion with Lagrange remainder of the function $\tilde{M}_K$ around

$\tilde{L}^K$, we obtain that for some $p_* \in [0,1]$ and $\tilde{\theta}_* = p_* \tilde{L}^K + (1-p_*)\tilde{\theta}$ we have

$$\tilde{\Delta}(\tilde{\theta}) = \langle \tilde{M}'_K(\tilde{L}^K), (\tilde{\theta} - \tilde{L}^K) \rangle + \frac{1}{2}(\tilde{\theta} - \tilde{L}^K)^\top \tilde{M}''_K(\tilde{\theta}_*)(\tilde{\theta} - \tilde{L}^K) + \tau_n K^{-2}(\text{rank}(\theta) - r),$$

where $\tilde{M}'_K(\tilde{\theta}) = 2K^{-2}\tilde{P}^K \circ (\tilde{\theta} - \tilde{R}^K)$ and $\tilde{M}''_K(\tilde{\theta}_*)(\tilde{\theta} - \tilde{L}^K) = 2K^{-2}\tilde{P}^K \circ (\tilde{\theta} - \tilde{L}^K)$. Noticing that $\tilde{M}'_K(\tilde{L}^K) = 0$ (by Theorem 2.2.2), we have

$$\begin{aligned}
\tilde{\Delta}(\tilde{\theta}) &= K^{-2}\langle \tilde{P}^K \circ (\tilde{\theta} - \tilde{L}^K), \tilde{\theta} - \tilde{L}^K \rangle + \tau_n K^{-2}(\text{rank}(\theta) - r) \\
&= K^{-2}\langle \tilde{P}^K \circ (\tilde{\theta} - \tilde{L}^K), \tilde{P}^K \circ (\tilde{\theta} - \tilde{L}^K) \rangle + \tau_n K^{-2}(\text{rank}(\theta) - r) \\
&= K^{-2}\|P^K \circ (\theta - L^K)\|_{\mathrm{F}}^2 + \tau_n K^{-2}(\text{rank}(\theta) - r) \\
&\leq K^{-2}\|\theta - L^K\|_{\mathrm{F}}^2 + \tau_n K^{-2}|\text{rank}(\theta) - r|.
\end{aligned}$$

It follows that $\sup_{\theta \in \Theta_K : d_n(L^K, \theta) < \gamma} |\Delta(\theta)| \leq \gamma^2$. To prove (ii), consider:

$$\begin{aligned}
D(\theta) &= \mathbb{S}_{n,K}(\theta) - S_{n,K}(\theta) - \mathbb{S}_{n,K}(L^K) + S_{n,K}(L^K) \\
&= \mathbb{M}_{n,K}(\theta) - M_K(\theta) - \mathbb{M}_{n,K}(L^K) + M_K(L^K).
\end{aligned}$$

We expand $(\tilde{\mathbb{M}}_{n,K} - \tilde{M}_K)$ in a first order Taylor expansion with Lagrange remainder, around $\tilde{L}^K$, which gives for a certain $p^* \in [0,1]$ and $\tilde{\theta}^* = p^* \tilde{L}^K + (1-p^*)\tilde{\theta}$:

$$\begin{aligned}
\tilde{D}(\tilde{\theta}) &= \langle \tilde{\mathbb{M}}'_{n,K}(\tilde{\theta}^*), \tilde{\theta} - \tilde{L}^K \rangle - \langle \tilde{M}'_K(\tilde{\theta}^*), \tilde{\theta} - \tilde{L}^K \rangle \\
&= K^{-2}\langle 2\tilde{P}^K \circ (\tilde{\theta}^* - \tilde{R}_n^K), \tilde{\theta} - \tilde{L}^K \rangle - K^{-2}\langle 2\tilde{P}^K \circ (\tilde{\theta}^* - \tilde{R}^K), (\tilde{\theta} - \tilde{L}^K) \rangle \\
&= K^{-2}\langle 2\tilde{P}^K \circ \tilde{\theta}^* - 2\tilde{P}^K \circ \tilde{\theta}^* - 2\tilde{P}^K \circ \tilde{R}_n^K + 2\tilde{P}^K \circ \tilde{R}^K, \tilde{\theta} - \tilde{L}^K \rangle \\
&\leq K^{-2}\|2P^K \circ (R_n^K - R^K)\|_{\mathrm{F}} \|\theta - L^K\|_{\mathrm{F}} \leq 2K^{-1}\|R_n^K - R^K\|_{\mathrm{F}} K^{-1}\|\theta - L^K\|_{\mathrm{F}}.
\end{aligned}$$

Since X is a process with continuous sample paths, $\mathbb{E}\|X\|_{L^2}^4 < \infty$ implies that the variance of $X(s)X(t)$ is finite for all $(s,t) \in [0,1]^2$. Assume without loss of generality that $\mathbb{E}X = 0$. Since the observations $X_i(t_j)$ are independent for distinct i, and since $X_m(t_j)X_m(t_j)$ is an unbiased estimator of $\mathbb{E}[X(t_j)X(t_j)]$, we have

$$\begin{aligned}
K^{-2}\mathbb{E}\|R_n^K - R^K\|_F^2 &= \sum_{i=1}^K \sum_{j=1}^K K^{-2}\mathbb{E}\left[\frac{1}{n}\sum_{m=1}^n X_m(t_{i,K})X_m(t_{j,K}) - \mathbb{E}\left[X(t_{i,K})X(t_{j,K})\right]\right]^2 \\
&= \frac{K^{-2}}{n}\sum_{i=1}^K \sum_{j=1}^K \text{Var}[X(t_{i,K})X(t_{j,K})] \\
&\leq \frac{1}{n}\sup_{(s,t)\in[0,1]^2} \text{Var}[X(s)X(t)] = \frac{C}{n},
\end{aligned}$$

and $C = \sup_{[0,1]^2} \text{Var}[X(s)X(t)] < \infty$. Note that C is uniform in K. In summary we may con-

clude that

$$\mathbb{E}\left\{\sup_{\theta\in\Theta_K: d_n(\theta,L^K)<\gamma}|D(\theta)|\right\} \le 2\gamma K^{-1}\mathbb{E}\|R_n^K - R^K\|_{\mathrm{F}} \le 2\gamma\sqrt{K^{-2}\mathbb{E}\|R_n^K - R^K\|_F^2} = 2\gamma\sqrt{\frac{C}{n}}.$$

Now that we have proved (i) and (ii), we use Theorem 3.4.1 of van der Vaart and Wellner [vdVW96] with $r_n = \sqrt{n}$ and $\phi_n(\beta) = \beta$ for $\beta \in (0,\infty) = (\delta_n, \eta)$. It follows that if $\hat{L}_n^K$ is such that $\mathbb{S}_{n,K}(\hat{L}_n^K) \le \mathbb{S}_{n,K}(L^K) + O_{\mathbb{P}}(n^{-1})$, which is guaranteed by our definition of $\hat{L}_n^K$ as an approximate minimiser of $\mathbb{S}_{n,K}$, then it holds that

$$n\,d_n^2(\hat{L}_n^K, L^K) = nK^{-2}\|\hat{L}_n^K - L^K\|_F^2 + n\left(\tau_n K^{-2}|\mathrm{rank}(\hat{L}_n^K) - r|\right) = O_{\mathbb{P}}(1),$$

where the $O_{\mathbb{P}}(1)$ term does not depend on K. We may conclude that

$$\begin{aligned} n\|\hat{\mathscr{L}}_n^K - \mathscr{L}^K\|_{\mathrm{HS}}^2 &= nK^{-2}\|\hat{L}_n^K - L^K\|_F^2 = O_{\mathbb{P}}(1), \\ n\tau_n K^{-2}|\mathrm{rank}(\hat{\mathscr{L}}_n^K) - r| &= O_{\mathbb{P}}(1), \end{aligned}$$

where $O_{\mathbb{P}}(1)$ is uniform in K.

We now turn our attention to the estimated eigenfunctions. Since these are finitely many, we will omit the index indicating the order of an eigenfunction for tidiness, and consider an eigenfunction η. Let η^K be the K-resolution step function approximation of η,

$$\eta^K(x) = \sum_{j=1}^{K}\eta(t_{j,K})\mathbf{1}\{x \in I_{j,K}\}.$$

Then, by Taylor expansion,

$$\int_0^1\left(\eta(x) - \eta^K(x)\right)^2 dx = \sum_{j=1}^{K}\int_{I_{j,K}}\left(\eta(x) - \eta(t_{j,K})\right)^2 dx \le \sum_{j=1}^{K}K^{-3}\|\eta'\|_\infty^2 = \frac{\|\eta'\|_\infty^2}{K^2}.$$

It follows that

$$\begin{aligned} \left\|\hat{\eta}^K - \eta\right\|_{L^2}^2 &\le 2\left\|\hat{\eta}^K - \eta^K\right\|_{L^2}^2 + 2\left\|\eta^K - \eta\right\|_{L^2}^2 \\ &\le c\|\hat{\mathscr{L}}_n^K - \mathscr{L}^K\|_{\mathrm{HS}}^2 + \frac{2\|\eta'\|_\infty^2}{K^2} = O_{\mathbb{P}}(n^{-1}) + \frac{2\|\eta'\|_\infty^2}{K^2}, \end{aligned}$$

and the $O_{\mathbb{P}}(n^{-1})$ has been shown to be uniform in K. The constant c can be chosen uniformly over the order of eigenfunction, since there are only $r < \infty$ eigenfunctions to consider. The convergence rate for $\sup_j|\hat{\lambda}_j^K - \lambda_j|$ follows from the inequality $\sup_j|\hat{\lambda}_j^K - \lambda_j| \le \|\hat{\mathscr{L}}_n^K - \mathscr{L}\|_{\mathrm{HS}}$ (Proposition 1.1.5).

□

The consistency of $\hat{\mathscr{B}}_n^K$ is implied directly by that of $\hat{\mathscr{L}}_n^K$. However, if we want to have convergence rates, we need to assume more than only continuity of the kernel b. For example, if we assume that b is C^1 on the band $\{|t-s|<\delta\}$ or on the whole domain $[0,1]^2$, we obtain the convergence rates that are presented in Corollary 2.3.1.

Corollary 2.3.1. *Let the first $p+1$ eigenvalues of $\mathscr{B}$ be such that $\beta_1 > \beta_2 > \ldots, \beta_p > \beta_{p+1}$. Under the same conditions as in Theorem 2.3.1, and for any $\epsilon, \gamma > 0$ there exist $M_1(\epsilon,\gamma) > 0$ such that*

$$\mathbb{P}\left\{\left\|\hat{\mathscr{B}}_n^K - \mathscr{B}\right\|_{\mathrm{HS}} > \epsilon\right\} < \gamma, \quad \textit{and} \quad \mathbb{P}\left\{\sup_{j\geq 1}|\hat{\beta}_j^K - \beta_j| > \epsilon\right\} < \gamma \qquad \forall\, K, n > M_1(\epsilon,\gamma),$$

and $M_2(\epsilon,\gamma,j) > 0$ such that

$$\mathbb{P}\left\{\left\|\hat{\psi}_j^K - \psi_j\right\|_{L^2} > \epsilon\right\} < \gamma, \quad j = 1,\ldots,p, \qquad \forall\, K, n > M_2(\epsilon,\gamma,j),$$

for almost every grid $(t_1,\ldots,t_K) \in \mathscr{T}_K$. If the covariance function $b(s,t) : [0,1]^2 \to \mathbb{R}$ associated with $\mathscr{B}$ is additionally assumed to be continuously differentiable on $\{|t-s|<\delta\}$, we obtain the rates of convergence

$$\left\|\hat{\mathscr{B}}_n^K - \mathscr{B}\right\|_{\mathrm{HS}}^2 = O_{\mathbb{P}}(n^{-1}) + O(K^{-1}), \tag{2.3.6}$$

$$\left\|\hat{\psi}_j^K - \psi_j\right\|_{L^2}^2 = \frac{O_{\mathbb{P}}(n^{-1}) + O(K^{-1})}{\sigma_j^2}, \quad j = 1,\ldots,p, \tag{2.3.7}$$

$$\sup_{j\geq 1}|\hat{\beta}_j^K - \beta_j|^2 = O_{\mathbb{P}}(n^{-1}) + O(K^{-1}), \tag{2.3.8}$$

for any refinement $K = m \times K^*$, $m \geq 1$, and almost every grid $(t_1,\ldots,t_K) \in \mathscr{T}_K$, where

$$\sigma_1 = \beta_1 - \beta_2, \quad \& \quad \sigma_j = \min\{\beta_{j-1} - \beta_j, \beta_j - \beta_{j+1}\}, \qquad 2 \leq j \leq p,$$

and all three $O_{\mathbb{P}}(n^{-1})$ terms are uniform in K. If b is continuously differentiable on the whole unit square $[0,1]^2$, then the $O(K^{-1})$ terms in Equations (2.3.6), (2.3.7), and (2.3.8) can be improved to $O(K^{-2})$.

Proof. We have

$$\begin{aligned}\left\|\hat{\mathscr{B}}_n^K - \mathscr{B}\right\|_{\mathrm{HS}} &\leq \left\|\hat{\mathscr{B}}_n^K - \mathscr{B}^K\right\|_{\mathrm{HS}} + \left\|\mathscr{B}^K - \mathscr{B}\right\|_{\mathrm{HS}} \\ &= K^{-1}\left\|\hat{B}_n^K - B^K\right\|_{\mathrm{F}} + \left\|\mathscr{B}^K - \mathscr{B}\right\|_{\mathrm{HS}},\end{aligned} \tag{2.3.9}$$

where the second term on the right hand side converges in probability to zero as K goes to infinity. Let P be the projection onto the set of non-negative banded $K \times K$ matrices of bandwidth at most $\lceil K/4 \rceil$. The first term on the right hand side of equality (2.3.9) can be

written as

$$\begin{aligned} K^{-1}\left\|\hat{B}_n^K - B^K\right\|_{\mathrm{F}} &= K^{-1}\left\|P(R_n^K - \hat{L}_n^K) - P(R^K - L^K)\right\|_{\mathrm{F}} \\ &\leq K^{-1}\left\|(R_n^K - \hat{L}_n^K) - (R^K - L^K)\right\|_{\mathrm{F}}, \end{aligned}$$

where from the consistency of $\hat{L}_n^K$ and the definition of R_n^K we obtain that the right hand side term of the last inequality converge to zero as n goes to infinity. The consistency of the eigenfunctions $\hat{\psi}_m$ follows directly from the consistency of $\hat{\mathscr{B}}_n^K$. This completes the first part of the theorem. Now assume that b is continuously differentiable on the δ-band. In what follows, we will write $I_{ij} = I_{i,K} \times I_{j,K}$ for tidiness. We start with the decomposition:

$$\left\|\hat{\mathscr{B}}_n^K - \mathscr{B}\right\|_{\mathrm{HS}}^2 \leq 2\left\|\hat{\mathscr{B}}_n^K - \mathscr{B}^K\right\|_{\mathrm{HS}}^2 + 2\left\|\mathscr{B}^K - \mathscr{B}\right\|_{\mathrm{HS}}^2.$$

Since the kernel b is C^1 inside the δ-band and zero outside it, we write,

$$\begin{aligned} \int_0^1\int_0^1 (b(x,y) - b^K(x,y))^2 dxdy &= \iint_A (b(x,y) - b^K(x,y))^2 dxdy \\ &\quad + \iint_{A^c} (b(x,y) - b^K(x,y))^2 dxdy, \end{aligned}$$

where $A = \bigcup_{i,j}\left\{I_{ij} : \{|x-y| = \delta\} \cap I_{ij} \neq \emptyset\right\}$ is the union of rectangles I_{ij} that intersect with the boundary of the δ-band. On this set, $(b - b^K)^2$ is bounded by $\sup_{x,y,z,w}|b(x,y) - b^K(z,w)|^2 = C < \infty$, and so $\iint_A (b(x,y) - b^K(x,y))^2 dxdy$ is bounded by $|A|K^{-2}C$, where $|A| = \#\{(i,j) : I_{ij} \subset A\}$. Noting that $|A|$ is of the order of K, we have

$$\iint_A (b(x,y) - b^K(x,y))^2 dxdy = O(K^{-1}).$$

For the other term, we decompose the integral and use Taylor's theorem to see that

$$\begin{aligned} \iint_{A^c} (b(x,y) - b^K(x,y))^2 dxdy &= \iint_{\{|x-y|\leq\delta\}\cap A^c} (b(x,y) - b^K(x,y))^2 dxdy \\ &\quad + \iint_{\{|x-y|\geq\delta\}\cap A^c} (b(x,y) - b^K(x,y))^2 dxdy \end{aligned}$$

$$
\begin{aligned}
&= \iint_{\{|x-y|\le\delta\}\cap A^c} (b(x,y)-b^K(x,y))^2\,dx\,dy \\
&\le \sum_{i,j=1}^{K} \mathbf{1}\Big\{I_{ij}\subset\big\{\{|x-y|\le\delta\}\cap A^c\big\}\Big\}\iint_{I_{ij}} 2K^{-2}\sup_{(x,y)\in I_{ij}}\|\nabla b(x,y)\|_2^2\,dx\,dy \\
&\le \nu(K)2K^{-4}\sup_{|x-y|<\delta}\|\nabla b(x,y)\|_2^2 = O(K^{-2}),
\end{aligned}
$$

where $\nu(K)$ is the number of rectangles within the δ-band, and thus satisfies $\nu(K)=O(K^2)$. In conclusion,

$$\left\|\mathscr{B}^K-\mathscr{B}\right\|_{\mathrm{HS}}^2 = O(K^{-1}).$$

For the other term, we note that, almost everywhere on $\mathscr{T}_K$,

$$\|\hat{\mathscr{B}}_n^K-\mathscr{B}^K\|_{\mathrm{HS}}^2 \le \|\mathscr{D}_n^K-\mathscr{B}^K\|_{\mathrm{HS}}^2 \le 2\|\hat{\mathscr{L}}_n^K-\mathscr{L}^K\|_{\mathrm{HS}}^2+2\|\mathscr{R}_n^K-\mathscr{R}^K\|_{\mathrm{HS}}^2 = O_{\mathbb{P}}(n^{-1}),$$

where $\mathscr{D}_n^K$ is the operator corresponding to the matrix $\Delta_n^K=R_n^K-\hat{L}_n^K$ and, with the $O_{\mathbb{P}}(n^{-1})$ being uniform in K (as has been shown in the proof of Theorem 2.3.1). The convergence rate for the estimated eigenfunctions is now obtained by the fact that $\|\hat{\psi}_j^K-\psi_j\|_{L^2}\le 2\sqrt{2}\sigma_j^{-1}\|\hat{\mathscr{B}}_n^K-\mathscr{B}\|_{\mathrm{HS}}$ (Proposition 1.1.5).

To complete the proof, we need to show that if $b\in C^1([0,1]^2)$, then $\left\|\mathscr{B}^K-\mathscr{B}\right\|_{\mathrm{HS}}^2 = O(K^{-2})$. In this case we may Taylor expand to write

$$
\begin{aligned}
\int_0^1\int_0^1 (b(x,y)-b^K(x,y))^2\,dx\,dy &\le \sum_{i,j=1}^{K}\iint_{I_{ij}} 2K^{-2}\sup_{(x,y)\in I_{ij}}\|\nabla b(x,y)\|_2^2\,dx\,dy \\
&= \sum_{i,j=1}^{K} 2K^{-4}\sup_{(x,y)\in I_{ij}}\|\nabla b(x,y)\|_2^2 = O(K^{-2}),
\end{aligned}
$$

since there are at most order K^2 terms of the form $\sup_{(x,y)\in I_{ij}}\|\nabla b(x,y)\|_2^2$ that are non-zero, by the bandedness of the kernel b.

The convergence rates for $\sup_j|\hat{\beta}_j^K-\beta_j|$ follow from the inequality $\sup_j|\hat{\beta}_j^K-\beta_j|\le\|\hat{\mathscr{B}}_n^K-\mathscr{B}\|_{\mathrm{HS}}$ (Proposition 1.1.5). □

We state the following result without proof since it is derived directly from Theorem 2.3.1 and Corollary 2.3.1.

Corollary 2.3.2. *Under the same conditions as in Theorem 2.3.1, and if the kernel $b(s,t):[0,1]^2\to\mathbb{R}$ of $\mathscr{B}$ is assumed to be continuously differentiable on $\{|t-s|<\delta\}$, we obtain the rate of convergence*

$$\left\|\hat{\mathscr{R}}_n^K-\mathscr{R}\right\|_{\mathrm{HS}}^2 = O_{\mathbb{P}}(n^{-1})+O(K^{-1}), \tag{2.3.10}$$

for any refinement $K = m \times K^*$, $m \geq 1$, and almost all grids in $\mathscr{T}_K$. The $O_{\mathbb{P}}(n^{-1})$ term is uniform in K. If b is continuously differentiable on the whole unit square $[0,1]^2$, then the $O(K^{-1})$ term in Equation (2.3.10) can be improved to $O(K^{-2})$.

2.4 Prediction of the smooth curves

In this section we want to make use of the estimators defined in the previous section in order to recover the smooth part Y_i and the rough part W_i from the function X_i, for $i = 1, \dots, n$. In the rest of the section we will drop the index i since we are working with only one curve at a time. In order for this curve decomposition to be identifiable, we need to assume that the mean function of at least one of the two random variables Y and W is equal to zero. We will thus suppose that the rough process W has mean zero, and that the mean of Y, which is in this case equal to the mean of X, has been subtracted from the data, so at the end we have $\mathbb{E}[Y] = \mathbb{E}[W] = 0$.

Since in practice we only observe a discrete version of X, i.e. $X(t_1), \dots, X(t_K)$ for $\{t_j\}_{j=1}^K \in \mathscr{T}_K$, we will work with the piecewise constant function X^K defined from these observations as follows

$$X^K(t) = \sum_{j=1}^{K} \mathbf{1}(t \in I_{j,K}) X(t_j).$$

Using the function X^K, our goal is to find predictors of $Y^K = \sum_{j=1}^K \mathbf{1}(t \in I_{j,K}) Y(t_j)$ and $W^K = \sum_{j=1}^K \mathbf{1}(t \in I_{j,K}) W(t_j)$.

Let $\mathscr{R}^K = \sum_{i=1}^q \theta_i^K \varphi_i^K \otimes \varphi_i^K$ be the covariance operator of X^K and $\mathscr{L}^K = \sum_{j=1}^r \lambda_j^K \eta_j^K \otimes \eta_j^K$ that of Y^K. From [Bos14, Prop. 3.1], we know that the best linear predictor of Y^K given X^K, i.e. the linear function $\Pi(X^K)$ that minimises $\mathbb{E}\|Y^K - \Pi_0(X^K)\|_{L^2}^2$, is given by

$$\Pi(X^K) = \mathbb{E}[Y^K | X^K] = \sum_{j=1}^{r} \sum_{i=1}^{q} \gamma_{ij} (\varphi_i^K \otimes \eta_j^K) X^K, \tag{2.4.1}$$

where

$$\gamma_{ij} = \frac{\mathbb{E}\langle X^K, \varphi_i^K \rangle \langle Y^K, \eta_j^K \rangle}{\mathbb{E}\langle X^K, \varphi_i^K \rangle^2} = \frac{\left\langle \mathbb{E}\langle Y^K + W^K, \varphi_i^K \rangle Y^K, \eta_j^K \right\rangle}{\theta_i^K} = \frac{\langle \mathscr{L}^K \varphi_i^K, \eta_j^K \rangle}{\theta_i^K} = \frac{\langle \varphi_i^K, \mathscr{L}^K \eta_j^K \rangle}{\theta_i^K} = \frac{\lambda_j^K \langle \varphi_i^K, \eta_j^K \rangle}{\theta_i^K}.$$

The third equality in the previous equation comes from the fact that Y^K and W^K are uncorrelated, thus $\mathbb{E}\langle W^K, \varphi_i^K \rangle Y^K = 0$. By plugging the expression of γ_{ij} in (2.4.1), we obtain the

following expression for Π

$$\Pi(X^K) = \sum_{j=1}^{r} \sum_{i=1}^{q} \frac{\lambda_j^K}{\theta_i^K} \langle \varphi_i^K, \eta_j^K \rangle \langle \varphi_i^K, X^K \rangle \eta_j^K = \sum_{j=1}^{r} \xi_j \eta_j^K.$$

Since the Karhunen–Loève representation of Y^K is given by $\sum_{j=1}^{r} \langle Y^K, \eta_j^K \rangle \eta_j^K$, we can see that ξ_j is in fact a predictor of the j^{th} functional principal component score of Y^K. We will now show how close this predictor is to the one defined in Yao et al. [YMW05]. Manipulating the term ξ_j, we obtain

$$\xi_j = \left\langle \lambda_j^K \sum_{i=1}^{q} \frac{1}{\theta_i^K} (\varphi_i^K \otimes \varphi_i^K) \eta_j^K, X^K \right\rangle = \left\langle \lambda_j^K \left(\mathscr{R}^K \right)^\dagger \eta_j^K, X^K \right\rangle,$$

where $\left(\mathscr{R}^K \right)^\dagger$ is the generalised inverse of $\mathscr{R}^K$. Translating this latter expression of the predictor into matrix notation, we obtain

$$\xi_j = \frac{1}{K} \lambda_j^K (\tilde{X}^K)^\top \left(R^K \right)^\dagger \tilde{\eta}_j^K = \frac{1}{K} \lambda_j^K (\tilde{X}^K)^\top \left(L^K + B^K \right)^\dagger \tilde{\eta}_j^K,$$

where $\tilde{X}^K = [X^K(t_1), \ldots, X^K(t_K)]^\top$ and $\tilde{\eta}_j^K = [\eta_j^K(t_1), \ldots, \eta_j^K(t_K)]^\top$. We thus obtain almost the same expression as in Yao et al. [YMW05], except that here instead of having a diagonal matrix of the form $\sigma^2 \mathbb{I}_K$ as the covariance matrix of the rough process, we have the banded matrix B^K.

Now that we have defined the best linear predictor of Y^K given X^K, we obtain directly the best linear predictor of W^K given X^K which is defined by

$$\Psi(X^K) = \mathbb{E}[W^K | X^K] = \mathbb{E}[X^K - Y^K | X^K] = X^K - \mathbb{E}[Y^K | X^K] = X^K - \Pi(X^K). \tag{2.4.2}$$

We finally estimate the predictors Π and Ψ by plugging our estimators $\hat{\mathscr{L}}_n^K = \sum_{j=1}^{r} \hat{\lambda}_j^K \hat{\eta}_j^K \otimes \hat{\eta}_j^K$ and $\hat{\mathscr{R}}_n^K = \sum_{i=1}^{\hat{q}} \hat{\theta}_i^K \hat{\varphi}_i^K \otimes \hat{\varphi}_i^K$ of $\mathscr{L}^K$ and $\mathscr{R}^K$ in equations (2.4.1) and (2.4.2), which gives

$$\hat{Y}_n^K = \sum_{j=1}^{\hat{r}} \sum_{i=1}^{\hat{q}} \frac{\hat{\lambda}_j^K}{\hat{\theta}_i^K} \langle \hat{\varphi}_i^K, \hat{\eta}_j^K \rangle \langle \hat{\varphi}_i^K, X^K \rangle \hat{\eta}_j^K \quad \text{and} \quad \hat{W}_n^K = X^K - \hat{Y}_n^K.$$

The consistency of these two estimators is stated in the following corollary.

Corollary 2.4.1. *Let $X = Y + W$ and $\mathscr{R} = \mathscr{L} + \mathscr{B}$ be defined as in Section 2.1 and let the $r < \infty$ eigenfunctions of $\mathscr{L}$ be real analytic. Let $K = m \times K^*$ $(m \geq 1)$ be a refinement of the critical grid size $K^* = 4r + 4$. If $\mathscr{R}^K$ is of rank K, and if the kernel $b(s,t) : [0,1]^2 \to \mathbb{R}$ of $\mathscr{B}$ is continuously differentiable on $\{|t-s| < \delta\}$, then*

$$\| \hat{Y}_n^K - \Pi(X^K) \|_{L^2} = O_{\mathbb{P}}(n^{-1/2}), \quad \textit{and} \quad \| \hat{W}_n^K - \Psi(X^K) \|_{L^2} = O_{\mathbb{P}}(n^{-1/2}),$$

almost everywhere on $\mathscr{T}_K$.

Proof. Since $K \geq K^*$, it must be that

$$\|\hat{\mathscr{L}}_n^K - \mathscr{L}^K\|_{\mathrm{HS}} = O_{\mathbb{P}}(n^{-1/2}), \quad \& \quad \|\hat{\mathscr{B}}_n^K - \mathscr{B}^K\|_{\mathrm{HS}} = O_{\mathbb{P}}(n^{-1/2}),$$

almost everywhere on $\mathscr{T}_K$ (as has been shown in the proof of Theorem 2.3.1 and of Corollary 2.3.1). Consequently, for almost all grids in $\mathscr{T}_K$,

$$\|\hat{\mathscr{R}}_n^K - \mathscr{R}^K\|_{\mathrm{HS}} \leq \|\hat{\mathscr{L}}_n^K - \mathscr{L}^K\|_{\mathrm{HS}} + \|\hat{\mathscr{B}}_n^K - \mathscr{B}^K\|_{\mathrm{HS}} = O_{\mathbb{P}}(n^{-1/2}).$$

It thus holds true that, for almost all grids in $\mathscr{T}_K$,

$$|\hat{\theta}_i^K - \theta_i^K| = O_{\mathbb{P}}(n^{-1/2}), \quad i = 1, \ldots, K,$$

where $\hat{\theta}_i^K$ (resp. θ_i^K) is the ith eigenvalue of $\hat{\mathscr{R}}_n^K$ (resp. $\mathscr{R}^K$). Since $\mathrm{rank}(\mathscr{R}^K) = K$, it must be that $\theta_1^K, \ldots, \theta_K^K > 0$. Letting $g(x) = x^{-1}\mathbf{1}\{x > 0\}$, and noting that g is differentiable at $\{\theta_i^K\}_{i=1}^K$, the delta method thus implies

$$\left| \frac{\mathbf{1}\{\hat{\theta}_i^K > 0\}}{\hat{\theta}_i^K} - \frac{\mathbf{1}\{\theta_i^K > 0\}}{\theta_i^K} \right| = \left| \frac{\mathbf{1}\{\hat{\theta}_i^K > 0\}}{\hat{\theta}_i^K} - \frac{1}{\theta_i^K} \right| = O_{\mathbb{P}}(n^{-1/2}), \quad i = 1, \ldots, K,$$

for almost all grids in $\mathscr{T}_K$. Now observe that

$$\hat{Y}_n^K := \hat{\Pi}_n(X^K) = \sum_{j=1}^{\hat{r}} \sum_{i=1}^{K} \mathbf{1}\{\hat{\theta}_i^K > 0\} \frac{\hat{\lambda}_j^K}{\hat{\theta}_i^K} \langle \hat{\varphi}_i^K, \hat{\eta}_j^K \rangle \langle \hat{\varphi}_i^K, X^K \rangle \hat{\eta}_j^K = \sum_{i=1}^{K} \mathbf{1}\{\hat{\theta}_i^K > 0\} \left(\frac{\langle \hat{\varphi}_i^K, X^K \rangle}{\hat{\theta}_i^K} \right) \hat{\mathscr{L}}_n^K \hat{\varphi}_i^K. \tag{2.4.3}$$

By the continuous mapping theorem, we know that the right hand side converges in probability to

$$\sum_{i=1}^{K} \left(\frac{\langle \varphi_i^K, X^K \rangle}{\theta_i^K} \right) \mathscr{L}^K \varphi_i^K = \sum_{j=1}^{r} \sum_{i=1}^{K} \frac{\lambda_j^K}{\theta_i^K} \langle \varphi_i^K, \eta_j^K \rangle \langle \varphi_i^K, X^K \rangle \eta_j^K = \Pi(X^K),$$

for almost all grids in $\mathscr{T}_K$, as $n \to \infty$. The fact that the rate of convergence is $O_{\mathbb{P}}(n^{-1/2})$ follows directly from the fact that each term of the summands in the right hand side of Equation (2.4.3) has been shown to converge at the rate $O_{\mathbb{P}}(n^{-1/2})$. The corresponding result follows for $\|\hat{W}_n^K - \Psi(X^K)\|_{L^2}$ by writing

$$\|\hat{W}_n^K - \Psi(X^K)\|_{L^2} = \|X^K - \hat{Y}_n^K - (X^K - \Pi(X^K))\|_{L^2} = \|\hat{Y}_n^K - \Pi(X^K)\|_{L^2}.$$

□

3 Implementation and simulation study

The purpose of this chapter is to present how our new approach to analyse functional data can be implemented in practice, and to study its finite-sample performance. We first present in Section 3.1 how to obtain in practice the estimators $\hat{\mathscr{L}}_n^K$ and $\hat{\mathscr{B}}_n^K$ defined in Section 2.3. Using simulated data, we then study the performance of our method in Section 3.2; we investigate our rank selection procedure (Section 3.2.4), compare our estimator of $\mathscr{L}$ to the one obtained with three other classical methods (Section 3.2.5), and probe the predictive accuracy of our estimator of the smooth part of the data (Section 3.2.6). In section 3.3, we conclude the chapter with an illustration of our method on a real dataset.

3.1 Practical implementation

In this section, we will present how to obtain the estimators $\hat{L}_n^K$ and $\hat{B}_n^K$ that have been introduced in Section 2.3, in practice. We first consider the estimator $\hat{L}_n^K$ that was defined in Definition 2.3.1 as an approximate minimiser of

$$\min_{0\preceq\theta\in\mathbb{R}^{K\times K}} \frac{1}{K^2}\|P^K\circ(R_n^K-\theta)\|_F^2+\tau_n\text{rank}(\theta). \tag{3.1.1}$$

Solving this minimisation problem is equivalent to the following two-step procedure:

(i) Solve the minimisation problem

$$\min_{0\preceq\theta\in\mathbb{R}^{K\times K}} \|P^K\circ(R_n^K-\theta)\|_F^2 \quad \text{such that} \quad \text{rank}(\theta)\leq\omega, \tag{3.1.2}$$

for $\omega=1,\dots,K$, obtaining minimisers $\hat{\theta}_1,\dots,\hat{\theta}_K$.

(ii) Let $\tilde{f}:\{1,\dots,K\}\to\mathbb{R}$ be the function that gives the minimum attained by every minimiser, i.e. $\tilde{f}(\omega)=\|P^K\circ(R_n^K-\hat{\theta}_\omega)\|_F^2$. Define functions $f,g:\{1,\dots,K\}\to\mathbb{R}$ such that $f(\omega)=\tilde{f}(\omega)-\tilde{f}(K)$ and $g(\omega)=f(\omega)+\tau\omega$. Set the matrix $\hat{L}_n^K$ equal to $\hat{\theta}_{\omega^*}$, where ω^* is the

minimiser of the function g over $\omega = 1, \ldots, K$. In other words, ω^* is such that

$$\omega^* = \underset{\omega \in \{1,\ldots,K\}}{\arg\min} \{f(\omega) + \tau\omega\}. \tag{3.1.3}$$

There are some comments to be made about this new formulation of (3.1.1). First we have removed the coefficient K^{-2} since its purpose is to obtain a consistent estimator, but for a fixed K, the absence or presence of this factor is inconsequential. Secondly, step (ii) can be seen as the step where we pick the rank $\hat{r} = \omega^*$ of our estimator for a given τ, since by construction the matrix $\hat{\theta}_\omega$ will have the biggest possible rank which is here equal to ω. We will show below that this step is in fact equivalent to choosing the rank in the following way, which as we will see later can be seen as a scree plot approach:

(ii') Let c be a given constant. Set the matrix $\hat{L}_n^K$ equal to $\hat{\theta}_{\omega^{*'}}$, where $\omega^{*'}$ is the smallest value for which $f(\omega^{*'}) < c$. In other words, $\omega^{*'}$ is such that

$$\omega^{*'} = \left\{ \begin{array}{l} \underset{\omega \in \{1,\ldots,K\}}{\arg\min}\ \omega \\ \text{s.t. } f(\omega) < c \end{array} \right\} = \left\{ \begin{array}{l} \underset{\omega \in \{1,\ldots,K\}}{\arg\min} f(\omega) \\ \text{s.t. } \omega \leq f^{-1}(c) \end{array} \right\}, \tag{3.1.4}$$

where $f^{-1}(c) = \min\{x \in \{1, \ldots, K\} | f(x) \leq c\}$.

Steps (ii) and (ii') are equivalent, i.e. they lead to the same rank $\omega^* = \omega^{*'}$ and then to the same estimator, for certain values of τ and c, as it is presented in the following lemma.

Lemma 3.1.1. *Extend the function f defined in step (ii) to the interval $[1, K]$ by linear interpolation and call this new function $\bar{f}$. If $x \mapsto \bar{f}(x)$ is strictly convex, then, for any constant $c > 0$, the problem (3.1.4) with constraint parameter c is equivalent to (3.1.3) with a tuning parameter in the range*

$$\max\{\Delta_j : j \geq f^{-1}(c)\} < \tau < \min\{\Delta_j : j \leq f^{-1}(c) - 1\},$$

where $\Delta_i := f(i) - f(i+1) \geq 0$, for $i = 1, \ldots, K-1$. Furthermore, τ can be made arbitrarily small by choosing c to be arbitrarily small.

Proof. Choose $c > 0$ and let $q = f^{-1}(c)$. If we can choose a value of τ that simultaneously satisfies

$$\begin{aligned} \tau(q-j) + f(q-j) &> \tau q + f(q), \quad \forall\, j \in \{1, \ldots, q-1\} \text{ and} \\ \tau(q+j) + f(q+j) &> \tau q + f(q), \quad \forall\, j \in \{1, \ldots, K-q\}, \end{aligned}$$

then the function $g(\omega)$ reaches its minimal value at $\omega = q$ and so $\omega^* = q$. Moreover, since the function f is decreasing, it will imply that it reaches its minimal value at $\omega = q$ when ω is constrained to be smaller than q, and then that $\omega^{*'} = q$. If such a τ exists, it then means that the problem (3.1.4) with constraint parameter c is equivalent to (3.1.3).

Let's now examine when choosing such a τ is feasible. Notice that the two conditions that τ must satisfy are equivalent to:

$$\tau < \frac{f(q-j)-f(q)}{j}, \quad \forall\, 1 \le j < q \qquad \& \qquad \tau > \frac{f(q)-f(q+j)}{j}, \quad \forall\, 1 \le j \le K-q.$$

And so, by telescoping,

$$\frac{f(q-j)-f(q)}{j} = \underbrace{\frac{f(q-j)-f(q-j+1)}{j} + \cdots + \frac{f(q-1)-f(q)}{j}}_{j \text{ terms}},$$

and

$$\frac{f(q)-f(q+j)}{j} = \underbrace{\frac{f(q)-f(q+1)}{j} + \cdots + \frac{f(q+j-1)-f(q+j)}{j}}_{j \text{ terms}}.$$

We may thus re-write the conditions on τ as

$$\begin{aligned} \tau &< \frac{f(q-j)-f(q-j+1)}{j} + \cdots + \frac{f(q-1)-f(q)}{j}, \quad \forall\, 1 \le j < q, \\ \tau &> \frac{f(q)-f(q+1)}{j} + \cdots + \frac{f(q+j-1)-f(q+j)}{j}, \quad \forall\, 1 \le j \le K-q. \end{aligned}$$

By convexity of arithmetic averaging, a sufficient condition for the above to be true is to require

$$\begin{aligned} \tau &< f(i)-f(i+1) := \Delta_i, \quad \forall\, 1 \le i < q, \\ \tau &> f(i)-f(i+1) = \Delta_i, \quad \forall\, q \le i \le K-1. \end{aligned}$$

Since $x \mapsto \bar{f}(x)$ is strictly convex, the sequence Δ_i is strictly decreasing in i. It follows that the last two conditions are compatible, and we may choose any τ in the range

$$\max\{\Delta_j : j \ge f^{-1}(c)\} < \tau < \min\{\Delta_j : j \le f^{-1}(c)-1\},$$

while retaining the same optima for the two problems. Furthermore, since Δ_j can be made arbitrarily small for $j \ge f^{-1}(c)$ by choosing c to be sufficiently small, we see that τ can be taken to be arbitrarily small by appropriate choice of c.

□

Note that if $x \mapsto \bar{f}(x)$ is convex, then it will almost surely be strictly convex since $\{f(i)\}_{i \ge 1}$ are continuous random variables.

As we already mentioned, step (ii') can be seen as a scree plot approach. Indeed it consists in plotting the function f and picking the rank ω^* as the first value for which the function f

is below a given threshold c. We will now discuss how we can choose c (or equivalently $\hat{r}$) in practice. The idea is based on the following fact: if instead of observing f, we were able to observe $f_0(\omega) = \|P^K \circ (R^K - \hat{\Theta}_\omega)\|_F^2$, where $\hat{\Theta}_\omega$ is a minimiser of

$$\min_{0 \leq \Theta \in \mathbb{R}^{K \times K}} \|P^K \circ (R^K - \Theta)\|_F^2, \quad \text{such that} \quad \operatorname{rank}(\Theta) \leq \omega,$$

for $\omega = 1, \ldots, K$, then we would have by Corollary 2.2.1 that

$$f_0(\omega) > 0 \text{ for } \omega = 1, \ldots, r-1 \quad \text{and} \quad f_0(\omega) = 0 \text{ for } \omega \geq r, \tag{3.1.5}$$

where r is the rank of L^K. Since the function f can be seen as an approximation of f_0, we expect it to have a similar behaviour to the one described in (3.1.5), and then to level out beyond r. In practice, we will then plot the function f, and we will set $\hat{r}$ to be the point where f is levelling out, or equivalently to be the point representing the "elbow" of the function. The next lemma tells us that this procedure is consistent, which means that $\hat{r}$ converges to r as $n \to \infty$.

Lemma 3.1.2. *Assume the same conditions and context as in Lemma 3.1.1 and Corollary 2.2.1. Then for almost all grids in $\mathscr{T}_K$ it holds that*

$$\limsup_{n \to \infty} f(i) = 0 \qquad \textit{almost surely,}$$

for all $i \geq r$ whereas

$$\liminf_{n \to \infty} f(i) > \frac{1}{2} \sum_{j=i+1}^{r} \zeta_j^2 > 0 \qquad \textit{almost surely,}$$

for all $i < r$, whenever $r > 1$. Here $r = \operatorname{rank}(L^K)$ is the true rank of $\mathscr{L}$, and $\{\zeta_i\}_{i=1}^r$ are non-zero eigenvalues of the symmetric $K \times K$ matrix U^K, obtained by retaining the top-right and bottom-left $r \times r$ submatrices of L^K, and setting all other entries equal to zero.

Proof. We will write $f_n(i)$ instead of $f(i)$ in order to highlight the dependence on n. Let $\mathscr{A}_K \subseteq \mathscr{T}_K$ be the set of grids for which Corollary 2.2.1 is valid, and fix a grid $\mathbf{t}_K \in \mathscr{A}_K$. Note that this suffices for the purposes of the proof, since $\mathscr{A}_K$ is of full Lebesgue measure. Now, note that

$$f_n(r) \leq \left\| P^K \circ (R_n^K - L^K) \right\|_F^2 \xrightarrow{\text{a.s.}} \left\| P^K \circ (R^K - L^K) \right\|_F^2 = 0,$$

where $r = \operatorname{rank}(L^K)$. Consequently, $f_n(j) \leq f_n(r) \xrightarrow{\text{a.s.}} 0$ for all $j \geq r$, and obviously

$$\limsup_{n \to \infty} f_n(i) = 0 \qquad \text{almost surely,}$$

for all $i \geq r$. We now turn to the second assertion. We will consider the case $i = r - 1$ (the remaining cases follow similarly). Write $\zeta = \zeta_r > 0$ for the smallest eigenvalue of U^K. First, note that this must be non-zero, since Theorem 2.2.2 implies that all $r \times r$ minors of L^K are of full rank r.

We will argue by contradiction: suppose that the event $\{f_n(r-1) < \zeta^2/2 \text{ infinitely often}\}$ has positive probability. It follows that there exists a sequence θ_k of rank $r-1$ random matrices and a subsequence $\{R^K_{n_k}\}$ of $\{R^K_n\}$ such that

$$\left\| P^K \circ (R^K_{n_k} - \theta_k) \right\|_F^2 = \left\| P^K \circ R^K_{n_k} - P^K \circ \theta_k \right\|_F^2 < \zeta^2/2, \qquad \forall\, k \geq 1,$$

with positive probability. On the other hand, we know that

$$\left\| P^K \circ R^K_{n_k} - P^K \circ R^K \right\|_F^2 \xrightarrow{\text{a.s.}} 0.$$

Consequently, since $P^K \circ (L^K - R^K) = 0$, it follows that for all k sufficiently large,

$$\left\| P^K \circ \theta_k - P^K \circ L^K \right\|_F^2 < \zeta^2/2 + \zeta^2/2 = \zeta^2,$$

with positive probability. Now let ϑ_k denote the symmetric matrix formed by retaining the bottom-left and top-right $r \times r$ minors of θ_k, and setting the remaining elements equal to zero. Since our assumptions entail that $K \geq K^* = 4r+4$, we now have:

1. By Theorem 2.2.3, U^K is of rank r, and of course ϑ_k is of rank at most $r-1$, for all k, with probability 1.

2. The event $\left\| P^K \circ \theta_k - P^K \circ L^K \right\|_F^2 < \zeta^2$ has positive probability, and thus the event $\|\vartheta_k - U^K\|_F^2 < \zeta^2$ also has positive probability.

These two conclusions constitute a contradiction: the closest element to U^K from within the set $\{\theta : \text{rank}(\theta) = r-1\}$ is the $(r-1)$-spectral truncation of U^K, and this has squared Frobenius distance from U^K equal to ζ^2. This concludes the proof.

□

Now that we have explained how to deal with the second step of our procedure, we need to discuss how to solve the minimisation problem in step (i). If we had to exactly solve the same problem but without the presence of the matrix P^K in the norm, then we would have a classical PCA problem, and we would know that the minimiser is given by a truncated version of the singular value representation of R^K_n. However, due to the presence of the matrix P^K, we are looking for a rank ω matrix that fits best the matrix R^K_n outside a given band, and this complicates the problem — in particular the problem no longer has a closed form solution. To circumvent this problem, we use the fact that any $K \times K$ positive semidefinite matrix A of rank smaller or equal to ω can be written as $A = CC^\top$, where C is a $K \times \omega$ matrix. The optimisation problem (3.1.2) in step (i) can then be written as

$$\hat{\theta}_\omega = \hat{C}_\omega \hat{C}_\omega^\top, \quad \text{where} \quad \hat{C}_\omega \in \underset{C \in \mathbb{R}^{K \times \omega}}{\arg\min} \|P^K \circ (R^K_n - CC^\top)\|_F^2, \tag{3.1.6}$$

where we solve the latter problem using the function `fminunc` of the optimisation toolbox in MATLAB [MAT12]. This uses a subspace trust-region method based on the interior-reflective Newton method described in Coleman and Li ([CL94] and [CL96]) to perform the optimisation. The drawback with this procedure is that the minimisation problem is not convex in C, and so we don't have guarantees that the algorithm will converge to a global optimum. However, note that our problem can be seen as factorized matrix completion, and Chen and Wainwright [CW15] recently showed that projected descent algorithms have a high probability of yielding a "good" local optimum in this kind of problem, if the starting value is reasonable. Although we are not using the same exact optimisation method, we are in a similar setup and thus expect to obtain good results in practice. Our stating value C_0 is constructed such that $C_0 C_0^\top$ is the projection of R_n^K into the space of matrices of rank ω (i.e. $C_0 = U_i \Sigma_i^{1/2}$, where: $U\Sigma U^\top$ is the singular value decomposition of R_n^K; U_i is the $n \times i$ matrix obtained by keeping the first i columns of U; and Σ_i is the $i \times i$ matrix obtained by keeping the first i lines and columns of Σ), which seems to be the natural choice of starting value for this problem. In fact we perform a sensitivity analysis in Section 3.2.3, and it seems that our minimisation method is not particularly sensitive to the choice of starting value.

Now that we have described how to obtain $\hat{L}_n^K$ in practice, we will discuss the case of $\hat{B}_n^K$ defined in Definition 2.3.2. We don't have a closed form for the projection P of $\Delta_n^K = R_n^K - \hat{L}_n^K$ onto $C \cap D$, where C is the set of banded $K \times K$ matrices of bandwidth at most $\lceil K/4 \rceil$ and D is the set of non-negative definite $K \times K$ matrices, but we can approximate it using either an alternating projection algorithm (also called projections onto convex sets) (Bauschke and Borwein [BB96]) or the Dykstra's projection algorithm (Boyle and Dykstra [BD86]), since we have a closed form for the projections P_C and P_D and the sets C and D are closed and convex. In the results presented in the following sections, we used an alternating projection algorithm to obtain $\hat{B}_n^K$.

3.2 Simulation study

In this section we present a simulation study designed to study the performance of our methodology for a broad class of setups, and also to compare our results to those obtained with three classical methods. We first present the different types of data we are considering. We then carry out a sensitivity analysis to probe the effect of the starting value on the algorithm used to solve (3.1.6), study the effect of the sample size n and of the grid size K on the performance of our method and study the performance of our scree plot approach to estimate the rank of the matrix $\mathscr{L}$. We conclude the section by presenting the results of our comparison study with existing methods and some results related to the prediction part.

Note that in the rest of this section, we consider the maximal value of the bandwidth δ to be $1/10$ and not $1/4$ as in the theory, since we rarely expect a rough process to have such a long memory. The matrix P^K is then now constructed as $P^K(i,j) = \mathbf{1}\{|i-j| > \lceil K/10 \rceil\}$.

3.2.1 Simulation of the data

To test our method on a variety of data, we consider 3 different ways to simulate the smooth curve Y and 3 different ways to simulate the rough curve W, which leads to 9 different scenarios. For each of these scenarios, we simulate samples of n iid mean-zero functions Y_i and n iid mean-zero functions W_i on the grid $T = \{1/2K, 3/2K, \dots, (2K-1)/2K\}$ of K equally spaced points on the interval $[0,1]$. For each sample, we compute the empirical covariance matrix for the smooth and the rough curves, which are respectively given by :

$$L_n^K(a,b)\frac{1}{n}\sum_{i=1}^{n} Y_i(t_a)Y_i(t_b) \text{ and } B_n^K(a,b)\frac{1}{n}\sum_{i=1}^{n} W_i(t_a)W_i(t_b),$$

for $a, b \in \{1,\dots,K\}$, and then set $R_n^K = L_n^K + B_n^K$.

The smooth curves Y_i are constructed as independent copies of the random function Y which is defined such that its covariance function given by $\ell(s,t) = \sum_{j=1}^{r} \lambda_j \eta_j(s)\eta_j(t)$ has real analytic eigenfunctions $\eta_1,\dots,\eta_r$. For this, we set $Y(t) = \sum_{j=1}^{r} c_j \lambda_j^{1/2}\eta_j(t)$, where $\lambda_1,\dots,\lambda_r$ are positive scalars and $c_j \sim \mathrm{N}(0,1)$. The eigenfunctions are constructed in one of the three following ways :

(FB) as the first r Fourier basis elements, except for the special case $r = 1$, where instead of using $\eta_1(t) = 1$, we use $\eta_1(t) = \sin(2\pi t)$;

(AC) as being the Gram-Schmidt orthogonalisation of the first r analytic functions from the following list:

$$\begin{aligned} &\eta_1(t) = 5t\sin(2\pi t), && \eta_3(t) = 5t + \sin(2\pi t) - 2, && \eta_5(t) = \tfrac{\Gamma(4)}{\Gamma(2)\Gamma(2)}\, t(1-t).\\ &\eta_2(t) = t\cos(2\pi t) - 3, && \eta_4(t) = \cos(4\pi t) + (t/2)^2, \end{aligned}$$

(LP) as the first r shifted Legendre polynomials $\tilde{P}_i(x)$ defined as :

$$\begin{aligned} &\tilde{P}_0(t) = 1, && \tilde{P}_2(t) = 6t^2 - 6t + 1, && \tilde{P}_4(t) = 70t^4 - 140t^3 + 90t^2 - 20t + 1.\\ &\tilde{P}_1(t) = 2t - 1 && \tilde{P}_3(t) = 20t^3 - 30t^2 + 12t - 1, \end{aligned}$$

except for the special case $r = 1$, where instead of using $\eta_1(t) = \tilde{P}_0(t) = 1$, we use $\eta_1(t) = \tilde{P}_3(t)$.

The rough curves W_i are constructed as independent copies of the random function W which is defined such that its covariance function $b(s,t) = \sum_{j=1}^{d} \beta_j \psi_j(s)\psi_j(t)$ is δ-banded. For this, we set $W(t) = \sum_{j=1}^{d} b_j \beta_j^{1/2}\psi_j(t)$ where the eigenfunctions $\psi_1,\dots,\psi_d$ are supported on non overlapping intervals of length at most δ (and then d is at most δ^{-1}), $\beta_1,\dots,\beta_d$ are positive scalars and $b_j \sim \mathrm{N}(0,1)$. The eigenfunctions are constructed in one of the two following ways :

(TRI) each function ψ_j is defined as a triangular function of norm 1 with support $[(j-1)\delta, j\delta]$ which attains its maximal value at $(j-1/2)\delta$.

(RBB) each function ψ_j is a realisation of a reflected Brownian bridge defined on $[(j-1)\delta, j\delta]$.

For simplicity of presentation, we define the third way we used to construct a rough process W directly at a discretised level :

(MA) For each $t_j \in T$, we set $W(t_j) = \sum_{a=0}^{q} \theta_a \epsilon_{j-a}$, where $q = \lceil K\delta \rceil$, $\theta_0 = 1$, $\theta_1 \ldots, \theta_q \in (-1,1)$ are scalars and $\varepsilon_j \overset{iid}{\sim} N(0,1)$.

Note that a vector $(W(t_1), \ldots, W(t_K))$ generated as described in (MA) has a banded covariance matrix since it is in fact a moving average of order q.

The 9 scenarios resulting from our different ways to construct the smooth and the rough functions are described in Table 3.1.

Scenario	A	B	C	D	E	F	G	H	I
Y_i	FB	AC	LP	FB	AC	LP	FB	AC	LP
W_i	MA	MA	MA	TRI	TRI	TRI	RBB	RBB	RBB

Table 3.1 – Scenarios for the simulation study.

For each of these 9 scenarios, we consider 6 different combinations for the value of the rank parameter r and of the bandwidth parameter δ that are summarised in Table 3.2.

Combination	1	2	3	4	5	6
r	1	1	3	3	5	5
δ	0.05	0.1	0.05	0.1	0.05	0.1

Table 3.2 – Different values of the rank and bandwidth parameter.

Finally, we also consider two different regimes for the choice of the eigenvalues $\lambda_1 < \cdots < \lambda_r$ of $\mathscr{L}$ and $\beta_1 < \cdots < \beta_d$ of $\mathscr{B}$; the first one can be seen as the easier case where there is a clear ordering distinction between the two sets, i.e. $\lambda_r > \beta_1$ (Regime 1); the second one is the interlaced case, when $\lambda_r < \beta_1 < \lambda_{r-1}$ (Regime 2). In regime 1, the r eigenvalues λ are equally spaced between $\lambda_1 = 1.45$ and $\lambda_r = 0.25$, and we use $\lambda_1 = 0.25$ for $r = 1$. In regime 2, the eigenvalues $\{\lambda_1, \ldots, \lambda_r\}$ are equally spaced between $\lambda_1 = 1$ and $\lambda_r = 0.04$. In both regimes, the rough processes are simulated with $\beta_1 = 0.09$. The remaining eigenvalues for the scenarios (TRI) or (RBB) are smaller than 0.04 and decreasing toward zero, while those for the scenario (MA) are slowly decreasing toward zero, which yields to a challenging situation in the regime 2 since in this case there is more than one eigenvalue of the rough process that exceeds the smallest eigenvalue of the smooth process. For each combination (r, δ) with $r > 1$ of Table 3.2, we consider each of the two regimes and for the particular case $r = 1$, we consider only regime 1. In total, we consider 10 different cases in each of the nine simulation scenarios.

3.2.2 Impact of the sample size and of the resolution

As we just mentioned in Subsection 3.2.1, in our simulation study we consider 10 different cases (rank/bandwidth/regime) for every of our nine scenarios (A-I) which leads to 90 simulation setups. Still, we need to consider the effect of the sample size n and grid resolution K. Since we already have a fair number of simulation setups, we first studied the impact of different values of n and K on the performance of our method for the 10 different cases of Scenario A. To study the impact of n and K we consider the 7 different combinations given in Table 3.3. For every possible combination of parameters n and K, of parameters r and δ and of regime, we simulated 100 samples from Scenario A. For each of these 100 samples we calculated the relative error

$$\mathrm{RE}(u) = \frac{\|u - L^K\|_F}{\|L^K\|_F},$$

of our estimator $\hat{L}_n^K$, and we reporte the median, the first and the third quartiles of these errors in Table 3.4. The rank of L^K is assumed to be known in these simulations. By looking at Table 3.4, we see that our method works quite well (less than 15% error) for the majority of the setups. In fact we only have unsatisfactory results in the cases with $K = 25$ and $r = 5$, and this makes sense since in these cases we barely satisfy the condition $K \geq 4r + 4 = 24$, which is necessary to recover L^K uniquely (Corollary 2.2.1). Moreover this condition applies to the population version, so in a finite sample, we can imagine that K would have to be much bigger than $4r + 4$ to obtain good results. For the rest of this section, we have chosen to work with $n = 300$ and $K = 100$.

n	100	100	100	300	300	300	300
K	25	50	100	25	50	100	150

Table 3.3 – Different values of the sample size and the grid size.

3.2.3 Sensitivity analysis

In this section we carry out a sensitivity analysis in order to probe the effect of the starting value on the algorithm we use to solve the minimisation problem (3.1.6). The idea is to solve the same optimisation problem several times but with different starting values and to look at the variance of the resulting optima. To do so, we consider the 10 possible combinations of rank (1,3 and 5), bandwidth (0.05 and 0.1) and regime (1 and 2) of Scenario A. For each combination $i \in \{1,\dots,10\}$, we have simulated 100 samples which gave us 100 empirical covariance matrices $R_{i1},\dots,R_{i100}$. Let r_i be the true rank associated with the combination i, and let S_{ij} for $j = 1,\dots,100$ be a $K \times r_i$ matrix such that $S_{ij}S_{ij}^\top$ is the projection of the matrix R_{ij} into the space of matrices of rank r_i. For each combination i and sample j, we executed the following two steps.

Scenario A			
(n,K)	(r,δ)	Regime 1	Regime 2
$n=100, K=25$	$(1,0.05)$	0.14 (0.10, 0.20)	
	$(1,0.10)$	0.15 (0.12, 0.19)	
	$(3,0.05)$	0.15 (0.13, 0.19)	0.18 (0.13, 0.21)
	$(3,0.10)$	0.15 (0.12, 0.19)	0.18 (0.14, 0.23)
	$(5,0.05)$	0.30 (0.24, 0.37)	0.30 (0.24, 0.37)
	$(5,0.10)$	0.27 (0.24, 0.31)	0.27 (0.23, 0.32)
$n=100, K=50$	$(1,0.05)$	0.10 (0.07, 0.16)	
	$(1,0.10)$	0.15 (0.10, 0.22)	
	$(3,0.05)$	0.17 (0.13, 0.22)	0.17 (0.14, 0.22)
	$(3,0.10)$	0.17 (0.12, 0.21)	0.21 (0.19, 0.26)
	$(5,0.05)$	0.21 (0.18, 0.24)	0.21 (0.18, 0.25)
	$(5,0.10)$	0.31 (0.23, 0.34)	0.32 (0.29, 0.35)
$n=100, K=100$	$(1,0.05)$	0.12 (0.06, 0.19)	
	$(1,0.10)$	0.14 (0.09, 0.20)	
	$(3,0.05)$	0.15 (0.12, 0.19)	0.19 (0.16, 0.23)
	$(3,0.10)$	0.16 (0.13, 0.22)	0.18 (0.15, 0.22)
	$(5,0.05)$	0.21 (0.18, 0.24)	0.25 (0.21, 0.29)
	$(5,0.10)$	0.22 (0.18, 0.24)	0.22 (0.19, 0.25)
$n=300, K=25$	$(1,0.05)$	0.08 (0.06, 0.11)	
	$(1,0.10)$	0.09 (0.07, 0.12)	
	$(3,0.05)$	0.08 (0.07, 0.11)	0.11 (0.09, 0.15)
	$(3,0.10)$	0.10 (0.08, 0.13)	0.11 (0.09, 0.14)
	$(5,0.05)$	0.15 (0.14, 0.18)	0.19 (0.16, 1.41)
	$(5,0.10)$	0.15 (0.13, 0.18)	0.19 (0.15, 0.82)
$n=300, K=50$	$(1,0.05)$	0.07 (0.04, 0.11)	
	$(1,0.10)$	0.07 (0.05, 0.10)	
	$(3,0.05)$	0.11 (0.08, 0.13)	0.10 (0.09, 0.13)
	$(3,0.10)$	0.09 (0.07, 0.12)	0.14 (0.12, 0.16)
	$(5,0.05)$	0.12 (0.10, 0.14)	0.14 (0.12, 0.15)
	$(5,0.10)$	0.13 (0.11, 0.15)	0.15 (0.13, 0.17)
$n=300, K=100$	$(1,0.05)$	0.05 (0.03, 0.09)	
	$(1,0.10)$	0.06 (0.05, 0.10)	
	$(3,0.05)$	0.09 (0.07, 0.12)	0.11 (0.10, 0.13)
	$(3,0.10)$	0.08 (0.06, 0.11)	0.11 (0.09, 0.13)
	$(5,0.05)$	0.12 (0.10, 0.14)	0.14 (0.12, 0.15)
	$(5,0.10)$	0.12 (0.10, 0.14)	0.13 (0.12, 0.15)
$n=300, K=150$	$(1,0.05)$	0.13 (0.12, 0.16)	
	$(1,0.10)$	0.08 (0.06, 0.11)	
	$(3,0.05)$	0.10 (0.08, 0.12)	0.12 (0.10, 0.14)
	$(3,0.10)$	0.10 (0.07, 0.12)	0.10 (0.09, 0.13)
	$(5,0.05)$	0.13 (0.11, 0.15)	0.15 (0.13, 0.17)
	$(5,0.10)$	0.13 (0.10, 0.14)	0.13 (0.12, 0.15)

Table 3.4 – Table containing the median (the first and third quartiles are in parentheses) of the relative errors of our method for different combinations of n and K.

1. Find the optima $O_{ij1},\dots,O_{ij100}$, where O_{ijk} is obtained by solving the problem

$$\min_{C\in\mathbb{R}^{K\times r_j}} \|P^K \circ (R_{ij} - CC^\top)\|_F^2,$$

using the starting value S_{ik}.

2. Calculate the relative standard deviation (Rstd), also called coefficient of variation, of the optima $O_{ij1},\dots,O_{ij100}$

$$\text{Rstd}_{ij} = \frac{1}{\bar{O}_{ij}}\sqrt{\frac{1}{100}\sum_{k=1}^{100}(O_{ijk}-\bar{O}_{ij})^2},$$

where $\bar{O}_{ij} = 1/100\sum_{k=1}^{100} O_{ijk}$.

This gave us 100 relative standard errors for each combination, and the distribution of these errors are illustrated using boxplots in Figure 3.1. It seems that the only parameter that has an effect on the convergence behaviour of our algorithm is the regime. In fact, the choice of the starting value seems to have no effect in Regime 1 while it has a small effect in Regime 2.

3.2.4 Rank selection with the scree plot approach

As discussed in Section 3.1, in order to select the rank $\hat{r}$ of our estimator $\hat{L}_n^K$ we use a scree plot approach, i.e., we plot the function $f(\omega) = \|P^K \circ (R_n^K - \hat{C}_\omega\hat{C}_\omega^\top)\|_F^2$, where $\hat{C}_\omega$ is defined as in (3.1.6), and we pick $\hat{r}$ such that the function levels out beyond it. In this subsection we study the performance of this procedure on one sample from each of the 90 possible combinations of scenario/combination/regime. For each sample, we have to evaluate the function $f(\cdot)$ at $\omega = 1,\dots,K$, which is quite computationally intensive, especially for high value of ω. Since in practice we expect the rank of $\mathscr{L}$ to be low, we only calculate $f(\omega)$ for $\omega = 1,\dots,10$. We can see on the resulting plots of f that doing this is not restrictive since the functions become flat much before reaching $\omega = 10$. Figure 3.2 presents the results for scenarios A to E and Figure 3.3 those for F to I, where each graph represents a different combination of scenario and regime. Since we want the functions $f(\cdot)$ for different values of r and δ to appear on a same graph, we normalised each function $f(\cdot)$ by $\|P^K \circ R_n^K\|$ in order to put them on the same scale. The functions associated with a true rank of 1 (respectively of 3 and 5) are represented in blue (respectively in red and in black). To help analysing whether the results obtain with our method are accurate, a red vertical dotted line has been drawn at $\omega = 3$ and a black one at $\omega = 5$. We would then pick the right rank if the elbow of a curve (red or black) is located at the intersection with the vertical line matching its colour. In the case of a blue curve (rank equals to 1), we would select the right rank if the curve is flat.

By looking at the plots on the left of Figures 3.2 and 3.3, we see that in the Regime 1 we would select the right rank for all scenarios, except maybe for Scenario C, when the true rank r is either 1 or 3. For $r = 5$, we do well for scenarios A, D and G, which correspond to scenarios

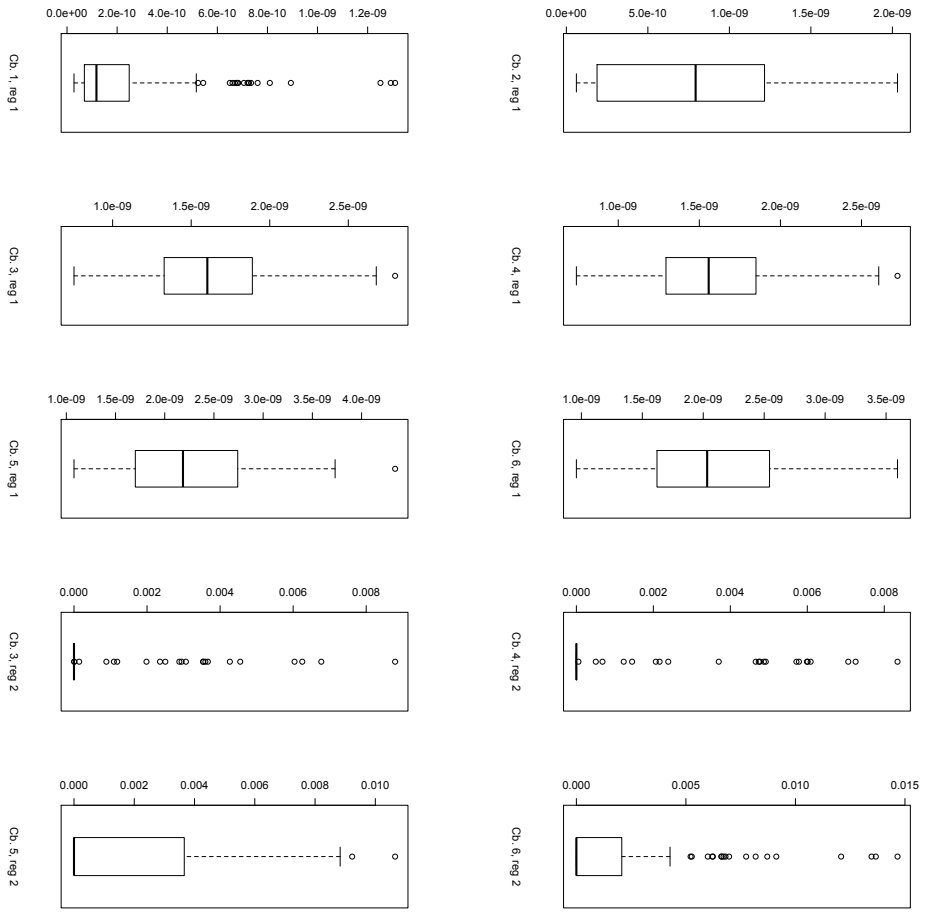

Figure 3.1 – Each boxplot is constructed from the 100 relative standard deviations obtained for a given combination of rank/bandwidh (1-6) and of regime (1 : top six, 2: bottom four) of Scenario A.

where the eigenfunctions of the operator $\mathscr{L}$ are Fourier bases, but we slightly underestimate the rank (by choosing 3 or 4 instead of 5) for the other scenarios. When we look at the plots on the right of the two figures, we notice that for the Regime 2 we underestimate the real rank for almost all cases.

In order to study the impact of rank misspecification, we carried out the following simulation study. We have simulated 100 samples of every possible combination of scenario A, of combinations 3 to 6 of Table 3.2 and of the two regimes. For every sample coming from a combination with $r = 3$, we have run our method 4 times, i.e., we have solved the problem (3.1.6) for $\omega = 2,3,4$ and 5, which gave us the corresponding estimators $\hat{L}_3^{\omega}$. We calculated the relative error $\mathrm{RE}(u) = \|u - L^K\|_F / \|L^K\|_F$ of each of the 4 resulting estimators, and then

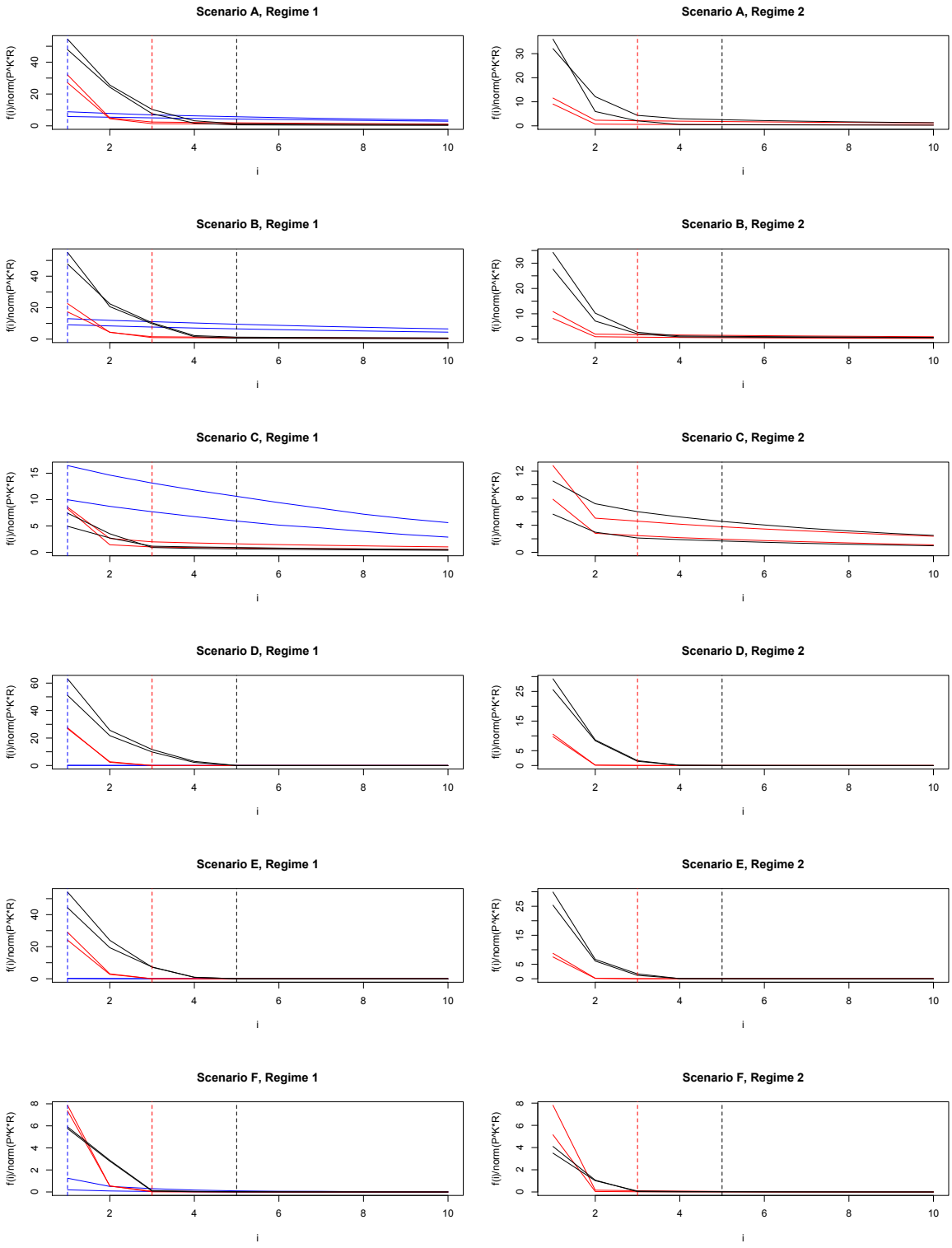

Figure 3.2 – Plots of the function $f(\cdot)$ normalised by $\|P^K \circ R_n^K\|$ for a given scenario (A-F), a given combination of parameters and a given regime. The curves in black correspond to a setting with $r = 5$, those in red to a setting with $r = 3$ and those in blue to a setting with $r = 1$.

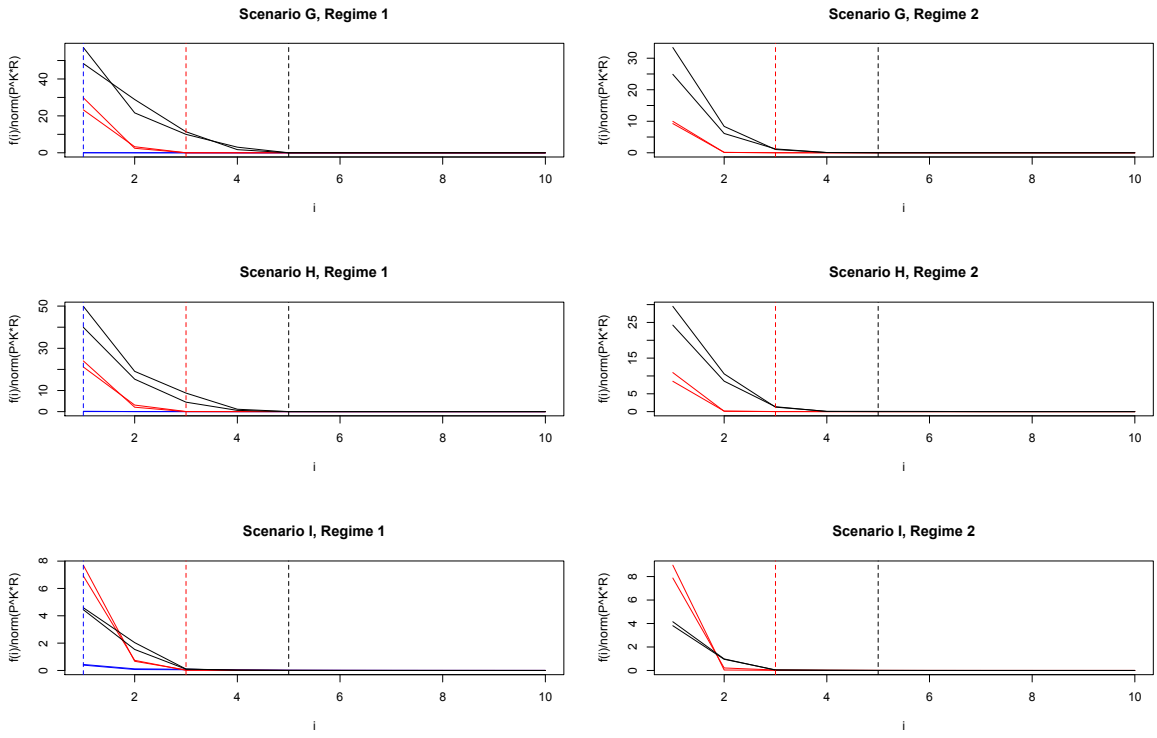

Figure 3.3 – Plots of the function $f(\cdot)$ normalised by $\|P^K \circ R_n^K\|$ for a given scenario (G-I), a given combination of parameters and a given regime. The curves in black correspond to a setting with $r = 5$, those in red to a setting with $r = 3$ and those in blue to a setting with $r = 1$.

construct the relative error ratio

$$\mathrm{RRE}(u) = \frac{\mathrm{RE}(u)}{\mathrm{RE}(\hat{L}_3^3)} = \frac{\|u - L^K\|_F}{\|\hat{L}_3^3 - L^K\|_F},$$

for $u = \hat{L}_3^2, \hat{L}_3^4$ and $\hat{L}_3^5$, which indicates the performance of the estimators obtained with a misspecified rank compared to the one obtains with the rank specified correctly. If this ratio is smaller than or equal to 1, the misspecification does not have a negative impact on our estimation, and otherwise it means that indeed the estimators obtained with a misspecified rank are not as good as those obtained with the rank specified correctly. We applied exactly the same procedure for the samples coming from a combination with $r = 5$, except that this time we solved the problem (3.1.6) for $\omega = 3, 4, 5, 6$ and 7. For each simulation setup with $r = 3$ we then have 3×100 relative error ratios, and for the one with $r = 5$ we have 4×100 relative error ratios. We illustrate the distribution of these ratios on Figure 3.4 using boxplots. For every setup with the regime 1, we see that underestimation has much more negative impacts than overestimation, while for the setups with the regime 2, an error of ± 1 on the rank choice is not very impactful, we thus recommend to chose a larger rank when in doubt.

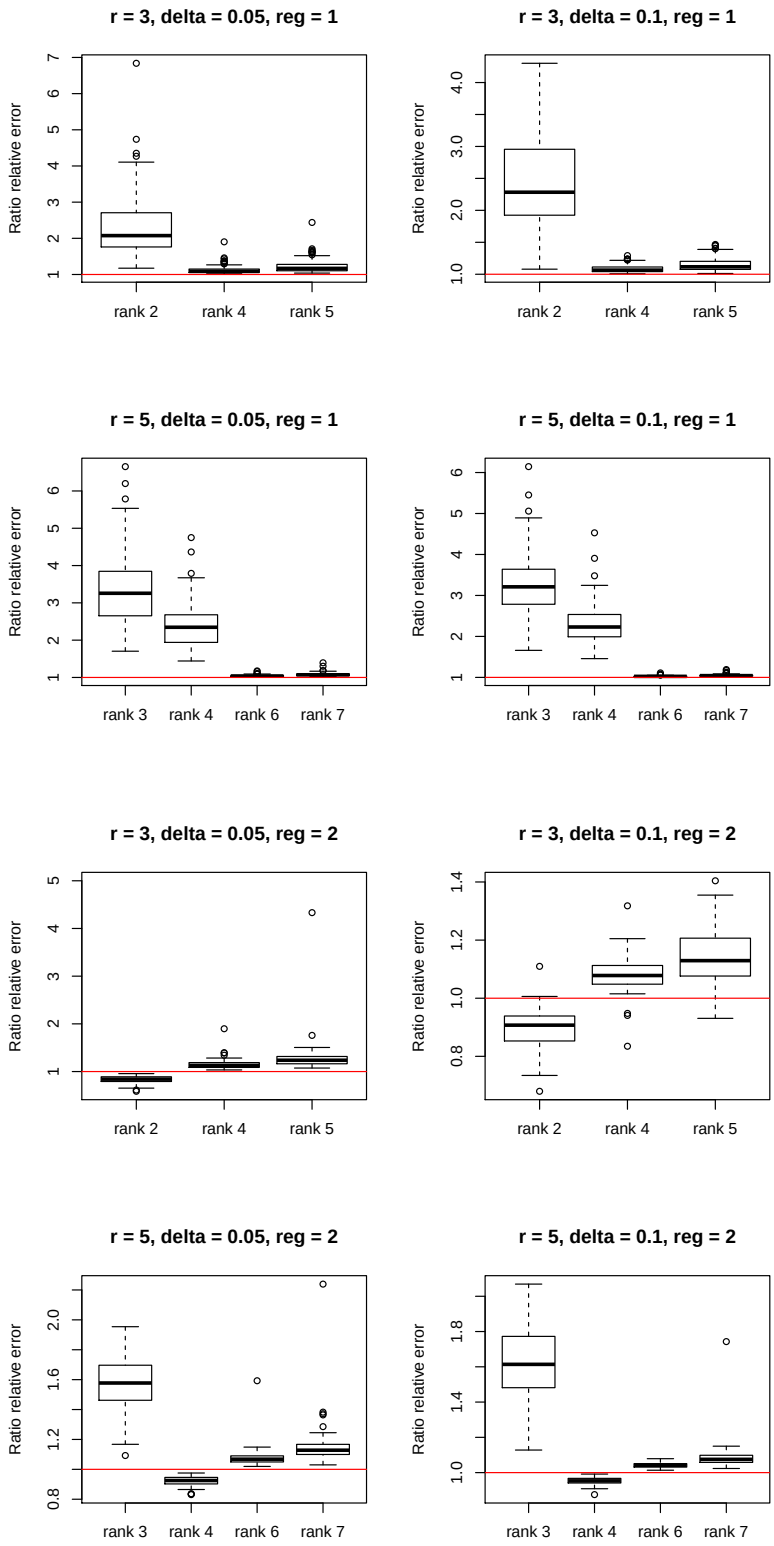

Figure 3.4 – Illustration of the impact of rank misspecification using Scenario A, combinations 3-6 of regime 1 (top four) and of regime 2 (bottom four). The impact is negative for values smaller than 1 (this threshold is represented by the red horizontal line).

3.2.5 Comparison with existing methods

In this subsection we study the performance of our methodology by comparing our results to those obtained with the three following methods:

1. The method popularised by Ramsay and Silverman [RS05] of smoothing first the data which leads to a sample of curves $\tilde{X}_1, \ldots, \tilde{X}_n$, and then to set the estimator of L^K to be $\hat{L}^K_{RS}(a,b) = \frac{1}{n}\sum_{i=1}^n \tilde{X}_i(a)\tilde{X}_i(b)$. The smoothing is done with Fourier basis for scenarios A, D and G since for these scenarios the data are periodic, and with B-splines for all the other scenarios (for more details look at Section 1.2.1).

2. The PACE method (Yao et al. [YMW05]) where the estimator $\hat{L}^K_{PACE}$ of L^K is obtained by local polynomial smoothing of the covariance matrix R^K_n from which we have removed the diagonal (for more details look at Section 1.2.2).

3. The naive method consisting in truncating the empirical Karhunen-Loève (KL) expansion, i.e. we calculate the spectral decomposition of R^K_n and we set the estimator $\hat{L}^K_{KL}$ of L^K to be the truncation of the above decomposition at a level r_K, where r_K is chosen such that the variance explained is at least of 95%.

Before performing our comparison study, let's first mention that throughout this subsection we use the real rank r of $\mathscr{L}^K$ when we apply our method. We justify this choice by the fact that as we have shown in the previous subsection, our scree plot approach gives reasonable results and moreover, since the simulations are computationally intensive, it would be impossible to use an automatic selection method. Furthermore, since we have a huge number of replications, it would also be impossible to make a choice based on the inspection of the scree plot.

We are now ready to present our comparison study. For every possible combination of scenario (A-I), of combination of r and δ (1-6) and of regime (1 or 2), we have simulated 100 samples of $n = 300$ curves evaluated on a grid T of $K = 100$ points. For each of these 100 samples, we have obtained our estimator $\hat{L}^K_n$ and the estimators $\hat{L}^K_{RS}, \hat{L}^K_{PACE}$ and $\hat{L}^K_{KL}$. We then do a two fold comparison, i.e. we compare the resulting estimators through their capacity to recover

1. the population covariance matrix L^K;

2. the empirical covariance matrix L^K_n.

We are interested in the recovery of the matrix L^K_n since besides being the best unbiased estimator of L^K that we could hope for if we could observed the functions $Y_1, \ldots, Y_n$ on the grid T, it is in fact the target of our method since we are not including any smoothing step in our procedure. For this reason, when we consider the recovery of L^K, instead of using our estimator $\hat{L}^K_n$, we use a very slightly smoothed version thereof, denoted by $\tilde{L}^K_n$. We compare the estimators on their performance to recover L^K through the relative error RE, and to recover

L_n^K through the empirical relative error

$$\mathrm{RE}_n(u) = \frac{\|u - L_n^K\|_F}{\|L_n^K\|_F}.$$

Once we have calculated RE and RE_n for the four estimators, we construct the ratios

$$\mathrm{RRE}(u) = \frac{\mathrm{RE}(u)}{\mathrm{RE}(\tilde{L}_n^K)} = \frac{\|u - L^K\|_F}{\|\hat{L}_n^K - L^K\|_F}, \quad \mathrm{RRE}_n(u) = \frac{\mathrm{RE}_n(u)}{\mathrm{RE}_n(\hat{L}_n^K)} = \frac{\|u - L_n^K\|_F}{\|\hat{L}_n^K - L_n^K\|_F},$$

for $u = \hat{L}_{RS}^K, \hat{L}_{PACE}^K$ and $\hat{L}_{KL}^K$. For each of the simulation setups, we then have 100 ratios RRE and 100 ratios RRE_n for each of the three methods of comparison, and we have reported the median, the first and the third quartiles of the ratios RRE in Tables 3.5, 3.6 and 3.7 and of the ratios RRE_n in Tables 3.8, 3.9 and 3.10. All the median values exceeding 1 (in bold in the tables) indicate that our method works at least as well as the method of comparison at least 50% of the time. The distribution of the relative errors RE and RE_n for every simulation setup are illustrated using boxplots in Appendix A.

By looking at Tables 3.5, 3.6 and 3.7, we notice that our method works better than the other methods when the real rank of the smooth covariance operator is 1 and better or comparably well in the other cases, except for the case where the smooth components were simulated with the first 5 Legendre polynomials (scenarios C, F and I with $r = 5$). In this case, due to the particular shape of the matrix L^K, which has values of large magnitude on the band compared to the ones outside it, our optimisation procedure was unstable and then sometimes led to poor estimates, as it is indicated in our tables by the very small values of the first quartiles.

Tables 3.8, 3.9 and 3.10 indicate that our method performs typically much better or comparably well than the other methods, except in two situations. The first is the same as we explained before i.e., for scenarios C, F and I with $r = 5$, and the second is for scenarios A and B with regime 2 where the estimator L_{RS} outperforms ours. However, by looking at Figures A.3 and A.6 in Appendix A, we can see that in fact our method works quite well in these setups, having a median relative error smaller than 0.05 for almost all of them, and that it is really the other method that is giving surprisingly outstanding results.

All these "good" results must be taken with a grain of salt, since we are comparing our method that has been especially designed to deal with the kind of data we have simulated, to methods that assume that the rough part of the data can be seen as white noise. Moreover, let's not forget that we have assumed that we knew the real rank of the matrix L^K, and that the other methods also have tuning parameters that affect their performance. These comparisons should thus be viewed more as a benchmark, than as claim to superiority.

As we already mentioned, our method has been designed to deal with the case where the rough part of the data is not simply due to pure noise. However, we would like it to still work well when the rough process is indeed pure noise, since in practice it might be difficult to know in which situation we are. We thus finish this subsection by comparing the performance

of our method using samples containing 300 curves $X_i = Y_i + W_i$ evaluated on a grid T of 100 points, where the smooth functions Y are simulated with r Fourier basis (FB) and the rough processes W such that $W(t) \overset{\text{iid}}{\sim} \mathrm{N}(0,\sigma^2)$, for $t \in T$. The results are presented in Table 3.11 for two different values of σ^2. We can see that even now our method performs at least as well as the other methods. We were expecting our method to give reasonable results also in this situation since the model considered here is that the matrix B^K is diagonal which is a particular case of being banded, but it seems at first a bit surprising that it works so well compared to the other methods, since one would expect that we are losing information by doing the completion of the matrix R_n^K to which we have removed a band instead of only the diagonal like in the PACE method for example. However, in practice we deal with B_n^K the empirical version of B^K, which won't be exactly diagonal, and it might explain why our method still gives interesting results for this scenario.

3.2.6 Prediction of the smooth curves

In this subsection we want to probe the performance of our estimator $\hat{Y}_n^K$ of the best linear unbiased predictor $\Pi(X^K)$. To do this, for every possible combination of r and δ and of regime (1 or 2) of Scenario E, we have simulated 100 samples of $n = 300$ curves evaluated on a grid T of $K = 100$ points. For each of these samples, we calculated the average of the approximation of the normalised mean integrated squared error of $\hat{Y}_n^K$:

$$\text{relMISE} = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{j=1}^{K}[\hat{Y}_{n,i}^K(t_j) - \Pi(X_i^K)(t_j)]^2}{\sum_{j=1}^{K}[\Pi(X_i^K)(t_j)]^2}.$$

The distribution of the resulting 100 values are illustrated using boxplots in Figure 3.5. We can see that the results seem satisfactory and that not surprisingly, the method performs a little bit better for regime 1 than regime 2.

3.3 Application to real data

We illustrate our method using data related to the air quality in the city of Geneva, Switzerland (the data can be found at the address "http://ge.ch/air/qualite-de-lair/requete-de-donnees"). Measurements of the quantity of nitrogen dioxide (NO2) in the air (in micrograms per cubic meter) have been recorded every hour at the station L'Ile from the second Monday of September to the second Sunday of November from 2005 to 2011. Our dataset is then composed of $n = 62$ curves corresponding to the different weeks, and each curve is evaluated at $K = 168$ points, corresponding to 7 days (from Monday to Sunday) times 24 hours. The raw dataset and its empirical covariance function are presented in Figure 3.6.

For these particular data, we are expecting the covariance function ℓ of our model to capture the variation at the time-scale of a week, and the banded covariance function b to capture the variation at the time-scale of a day. We thus decided to choose an upper bound of 0.14

Regime 1				
Scenario	(r,δ)	PACE	KL	RS
A	(1, 0.05)	**6.55** (3.67, 11.2)	**3.30** (1.94, 5.71)	**1.01** (0.92, 1.07)
	(1, 0.10)	**3.98** (2.64, 5.70)	**3.12** (2.08, 4.32)	**1.38** (1.15, 1.88)
	(3, 0.05)	**1.22** (1.05, 1.41)	**1.14** (1.01, 1.25)	**1.00** (0.98, 1.01)
	(3, 0.10)	**1.13** (1.06, 1.30)	**1.13** (1.02, 1.23)	**1.00** (0.95, 1.04)
	(5, 0.05)	**1.16** (1.07, 1.23)	**1.11** (1.05, 1.22)	**1.00** (0.98, 1.02)
	(5, 0.10)	**1.09** (1.02, 1.14)	**1.07** (1.03, 1.18)	**1.00** (0.98, 1.04)
B	(1, 0.05)	**6.03** (3.33, 8.66)	**2.95** (1.99, 4.67)	**1.87** (1.34, 2.60)
	(1, 0.10)	**3.36** (2.35, 4.13)	**2.67** (1.76, 3.28)	**1.68** (1.39, 2.26)
	(3, 0.05)	**1.36** (1.08, 1.49)	**1.19** (0.89, 1.31)	**1.02** (1.00, 1.05)
	(3, 0.10)	**1.17** (1.01, 1.31)	**1.11** (0.91, 1.25)	**1.04** (0.99, 1.08)
	(5, 0.05)	**1.15** (1.05, 1.24)	**1.10** (1.00, 1.17)	**1.00** (0.98, 1.02)
	(5, 0.10)	**1.08** (0.99, 1.16)	**1.04** (0.95, 1.10)	**1.00** (0.98, 1.03)
C	(1, 0.05)	**1.59** (1.57, 1.62)	0.58 (0.52, 0.63)	0.47 (0.45, 0.48)
	(1, 0.10)	**1.65** (0.75, 6.26)	0.68 (0.41, 2.79)	**1.08** (0.53, 4.94)
	(3, 0.05)	**1.45** (1.07, 1.76)	**1.19** (1.01, 1.43)	**1.03** (0.77, 1.12)
	(3, 0.10)	**1.25** (1.11, 1.48)	**1.12** (1.03, 1.27)	**1.12** (1.02, 1.21)
	(5, 0.05)	0.78 (0.71, 0.86)	0.79 (0.69, 0.85)	0.64 (0.53, 0.76)
	(5, 0.10)	0.67 (0.58, 0.76)	0.73 (0.67, 0.83)	0.63 (0.55, 0.75)
D	(1, 0.05)	**5.21** (3.04, 7.67)	**5.20** (3.03, 7.66)	**3.75** (2.23, 5.65)
	(1, 0.10)	**6.12** (4.26, 10.6)	**6.09** (4.24, 10.6)	**3.74** (2.66, 6.37)
	(3, 0.05)	**1.18** (1.09, 1.32)	**1.21** (1.07, 1.35)	**1.11** (1.02, 1.23)
	(3, 0.10)	**1.24** (1.11, 1.42)	**1.23** (1.10, 1.40)	**1.08** (1.00, 1.22)
	(5, 0.05)	**1.08** (1.04, 1.12)	**1.08** (1.01, 1.13)	**1.05** (0.99, 1.09)
	(5, 0.10)	**1.12** (1.07, 1.18)	**1.11** (1.06, 1.18)	**1.06** (1.01, 1.11)
E	(1, 0.05)	**4.32** (3.04, 6.50)	**4.31** (3.04, 6.51)	**3.88** (2.73, 5.67)
	(1, 0.10)	**6.51** (4.51, 11.2)	**6.48** (4.49, 11.2)	**4.74** (3.43, 8.12)
	(3, 0.05)	**1.18** (1.06, 1.30)	**1.18** (1.05, 1.32)	**1.15** (1.01, 1.26)
	(3, 0.10)	**1.24** (1.15, 1.40)	**1.23** (1.13, 1.44)	**1.11** (1.05, 1.27)
	(5, 0.05)	**1.06** (1.00, 1.13)	**1.07** (0.98, 1.16)	**1.04** (0.98, 1.12)
	(5, 0.10)	**1.11** (1.05, 1.16)	**1.11** (1.04, 1.20)	**1.06** (1.01, 1.13)
F	(1, 0.05)	**8.05** (6.29, 9.41)	**8.21** (6.40, 9.56)	**7.11** (5.60, 8.24)
	(1, 0.10)	**14.4** (12.3, 19.1)	**14.4** (12.3, 19.2)	**10.6** (8.90, 13.9)
	(3, 0.05)	**1.58** (1.30, 1.88)	**1.63** (1.34, 1.96)	**1.45** (1.23, 1.73)
	(3, 0.10)	**1.85** (1.51, 2.21)	**1.86** (1.54, 2.26)	**1.45** (1.27, 1.81)
	(5, 0.05)	0.67 (0.25, 0.92)	0.68 (0.26, 0.97)	0.60 (0.23, 0.89)
	(5, 0.10)	0.97 (0.59, 1.25)	0.96 (0.58, 1.27)	0.89 (0.53, 1.14)
G	(1, 0.05)	**2.74** (1.69, 5.85)	**2.70** (1.71, 5.82)	**1.90** (1.33, 4.05)
	(1, 0.10)	**8.16** (4.56, 13.0)	**8.12** (4.55, 13.1)	**6.01** (3.35, 9.43)
	(3, 0.05)	**1.06** (1.03, 1.12)	**1.06** (0.99, 1.13)	**1.02** (0.98, 1.07)
	(3, 0.10)	**1.37** (1.17, 1.68)	**1.40** (1.15, 1.66)	**1.25** (1.07, 1.43)
	(5, 0.05)	**1.02** (1.00, 1.05)	**1.03** (0.99, 1.06)	**1.01** (0.99, 1.04)
	(5, 0.10)	**1.14** (1.08, 1.29)	**1.13** (1.07, 1.27)	**1.08** (1.02, 1.19)

Table 3.5 – Table containing the median (the first and third quartiles are in parentheses) of the ratios for the three methods we compared our method with and for scenarios A-G with the regime 1. We highlight in bold the medians that exceed 1.

Regime 1				
Scenario	(r,δ)	PACE	KL	RS
H	$(1,0.05)$	**2.94** $(1.90,6.05)$	**2.92** $(1.89,6.08)$	**2.43** $(1.79,5.41)$
	$(1,0.10)$	**7.85** $(4.72,12.3)$	**7.97** $(4.72,12.3)$	**7.02** $(3.97,10.6)$
	$(3,0.05)$	**1.05** $(1.01,1.12)$	**1.05** $(1.00,1.12)$	**1.04** $(0.99,1.10)$
	$(3,0.10)$	**1.32** $(1.14,1.90)$	**1.29** $(1.14,1.97)$	**1.26** $(1.12,1.85)$
	$(5,0.05)$	**1.02** $(1.00,1.06)$	**1.04** $(0.99,1.08)$	**1.01** $(0.99,1.05)$
	$(5,0.10)$	**1.13** $(1.05,1.29)$	**1.13** $(1.04,1.28)$	**1.12** $(1.02,1.23)$
I	$(1,0.05)$	**8.18** $(5.87,10.6)$	**8.42** $(6.09,10.9)$	**7.51** $(5.07,9.31)$
	$(1,0.10)$	**13.6** $(9.71,17.7)$	**13.7** $(9.82,17.8)$	**11.8** $(8.31,15.0)$
	$(3,0.05)$	**1.14** $(1.05,1.32)$	**1.16** $(1.03,1.31)$	**1.09** $(1.02,1.22)$
	$(3,0.10)$	**1.93** $(1.36,2.90)$	**1.98** $(1.40,3.07)$	**1.93** $(1.28,2.53)$
	$(5,0.05)$	0.98 $(0.85,1.08)$	**1.03** $(0.86,1.12)$	0.96 $(0.80,1.05)$
	$(5,0.10)$	0.93 $(0.25,1.20)$	0.96 $(0.26,1.26)$	0.84 $(0.24,1.14)$

Table 3.6 – Table containing the median (the first and third quartiles are in parentheses) of the ratios for the three methods we compared our method with and for scenarios H-I with the regime 1. We highlight in bold the medians that exceed 1.

for δ, since it corresponds to removing a band of width $\lceil \delta K \rceil = 24$ hours in the discrete setup. In order to pick the rank r, we solved the optimisation problem (3.1.6) for $\omega = 1,\dots,7$, and we plotted the functions $f(\omega) = \|P^K \circ (R_n^K - \hat{C}_\omega \hat{C}_\omega^\top)\|_F^2$ and the ratio $r(j) = f(j)/f(j+1)$, for $j = 1,\dots,6$ on Figure 3.7. Our estimated rank should be the point i where the function f levels out, or equivalently, the point j for which the ratio r becomes a constant close to 1. After inspection of the Figure 3.7, we picked $\hat{r} = 3$ and our estimator of L^K is then given by $\hat{L}_n^K = \hat{C}_3 \hat{C}_3^\top$.

A slightly smoothed version $\tilde{L}$ of $\hat{L}_n^K$ is illustrated on Figure 3.8, as well as its three corresponding eigenfunctions that represent variation that propagates globally throughout the week. Since the first principal eigenfunction is negative throughout the week, it seems to indicate fluctuation of the overall level of NO2 on a weekly basis. Moreover, since the negative weight put on the afternoon of every day is around three times bigger than that put on every morning, it seems that the quantity of NO2 in the air is more variable during the afternoon. The second eigenfunction puts a negative weight on the first days of the week and a positive one on the last days, it then appears to capture early/late week effects. Similarly, the third eigenfunction is positive on periods corresponding to the day, and negative on periods corresponding to the night, so it can be seen as capturing the periodic day/night effects, interpreted as a measure of uniformity of the quantity of NO2 between the day and the night. The estimation of the covariance function b and of its three first eigenfunctions are presented in Figure 3.9. We can notice that as expected they are locally supported even if we didn't enforce this behaviour in our method. They can be interpreted as follows, the first represents the mode of variation during the weekend, the second that during the beginning of the week, and finally the last represents the variability of the middle of the week. Note that their corresponding eigenvalues

Regime 2				
Scenario	(r,δ)	PACE	KL	RS
A	$(3,0.05)$	**1.27** $(1.14,1.41)$	**1.12** $(0.99,1.23)$	0.88 $(0.82,0.92)$
	$(3,0.10)$	**1.09** $(1.01,1.20)$	**1.09** $(0.99,1.21)$	0.92 $(0.86,1.02)$
	$(5,0.05)$	**1.22** $(1.12,1.30)$	**1.18** $(1.07,1.25)$	0.92 $(0.87,0.96)$
	$(5,0.10)$	**1.09** $(1.01,1.17)$	**1.09** $(1.01,1.16)$	0.94 $(0.89,1.00)$
B	$(3,0.05)$	**1.45** $(1.21,1.74)$	**1.32** $(0.90,1.45)$	0.91 $(0.82,0.97)$
	$(3,0.10)$	**1.29** $(0.84,1.49)$	**1.26** $(0.78,1.47)$	0.94 $(0.85,1.01)$
	$(5,0.05)$	**1.14** $(0.98,1.33)$	**1.07** $(0.88,1.23)$	0.89 $(0.82,0.96)$
	$(5,0.10)$	**1.05** $(0.89,1.18)$	**1.02** $(0.89,1.13)$	0.88 $(0.79,0.94)$
C	$(3,0.05)$	**2.36** $(1.93,2.98)$	**1.40** $(1.04,1.76)$	**1.17** $(1.08,1.32)$
	$(3,0.10)$	**1.52** $(1.30,1.76)$	**1.20** $(1.02,1.42)$	**1.26** $(1.08,1.49)$
	$(5,0.05)$	**1.32** $(1.19,1.44)$	**1.05** $(0.93,1.15)$	0.85 $(0.76,0.94)$
	$(5,0.10)$	0.97 $(0.87,1.08)$	0.92 $(0.83,1.00)$	0.87 $(0.77,1.00)$
D	$(3,0.05)$	**1.40** $(1.24,1.65)$	**1.42** $(1.22,1.67)$	**1.23** $(1.06,1.39)$
	$(3,0.10)$	**1.59** $(1.33,2.05)$	**1.57** $(1.33,2.03)$	**1.27** $(1.12,1.54)$
	$(5,0.05)$	**1.21** $(1.12,1.29)$	**1.18** $(1.11,1.29)$	**1.09** $(1.02,1.18)$
	$(5,0.10)$	**1.27** $(1.16,1.40)$	**1.26** $(1.13,1.39)$	**1.08** $(1.02,1.20)$
E	$(3,0.05)$	**1.39** $(1.21,1.61)$	**1.36** $(1.20,1.62)$	**1.29** $(1.10,1.53)$
	$(3,0.10)$	**1.58** $(1.31,1.88)$	**1.58** $(1.33,1.87)$	**1.34** $(1.15,1.59)$
	$(5,0.05)$	**1.07** $(0.95,1.18)$	**1.05** $(0.90,1.17)$	**1.02** $(0.88,1.12)$
	$(5,0.10)$	**1.26** $(1.16,1.40)$	**1.29** $(1.16,1.48)$	**1.16** $(1.04,1.28)$
F	$(3,0.05)$	**1.46** $(0.18,2.13)$	**1.49** $(0.19,2.25)$	**1.35** $(0.16,1.93)$
	$(3,0.10)$	**1.96** $(0.33,2.88)$	**1.99** $(0.34,2.95)$	**1.52** $(0.27,2.22)$
	$(5,0.05)$	**1.07** $(0.14,1.41)$	**1.08** $(0.15,1.45)$	0.98 $(0.13,1.31)$
	$(5,0.10)$	**1.10** $(0.18,1.72)$	**1.11** $(0.18,1.75)$	0.89 $(0.13,1.43)$
G	$(3,0.05)$	**1.12** $(1.03,1.24)$	**1.12** $(1.02,1.27)$	**1.06** $(0.98,1.13)$
	$(3,0.10)$	**2.14** $(1.50,3.04)$	**2.18** $(1.50,3.01)$	**1.67** $(1.30,2.21)$
	$(5,0.05)$	**1.06** $(1.01,1.11)$	**1.10** $(1.05,1.16)$	**1.03** $(0.99,1.07)$
	$(5,0.10)$	**1.49** $(1.26,1.76)$	**1.49** $(1.25,1.76)$	**1.25** $(1.10,1.45)$
H	$(3,0.05)$	**1.12** $(1.04,1.22)$	**1.12** $(1.01,1.23)$	**1.07** $(0.99,1.20)$
	$(3,0.10)$	**1.78** $(1.38,2.51)$	**1.78** $(1.39,2.56)$	**1.60** $(1.30,2.25)$
	$(5,0.05)$	**1.04** $(1.00,1.13)$	**1.08** $(1.01,1.18)$	**1.03** $(0.99,1.09)$
	$(5,0.10)$	**1.29** $(1.05,1.62)$	**1.31** $(1.05,1.65)$	**1.23** $(1.01,1.49)$
I	$(3,0.05)$	**1.35** $(1.16,1.78)$	**1.37** $(1.14,1.92)$	**1.27** $(1.07,1.61)$
	$(3,0.10)$	**2.25** $(0.38,3.54)$	**2.31** $(0.39,3.64)$	**2.01** $(0.34,3.10)$
	$(5,0.05)$	0.93 $(0.25,1.15)$	0.93 $(0.27,1.19)$	0.90 $(0.23,1.12)$
	$(5,0.10)$	**1.29** $(0.34,1.74)$	**1.33** $(0.35,1.79)$	**1.16** $(0.27,1.58)$

Table 3.7 – Table containing the median (the first and third quartiles are in parentheses) of the ratios for the three methods we compared our method with and for the 9 scenarios we considered with the regime 2. We highlight in bold the medians that exceed 1.

Regime 1				
Scenario	(r,δ)	PACE	KL	RS
A	(1, 0.05)	**16.2** (15.2, 17.3)	**8.02** (7.53, 8.72)	**1.03** (0.93, 1.14)
	(1, 0.10)	**5.53** (5.13, 6.11)	**4.23** (3.91, 4.71)	**1.70** (1.56, 1.90)
	(3, 0.05)	**8.09** (7.63, 8.50)	**6.04** (5.62, 6.34)	0.94 (0.80, 1.07)
	(3, 0.10)	**4.12** (3.88, 4.47)	**3.46** (3.22, 3.74)	**1.59** (1.41, 1.78)
	(5, 0.05)	**4.55** (4.17, 5.04)	**3.72** (3.32, 4.31)	0.53 (0.45, 0.60)
	(5, 0.10)	**3.28** (3.08, 3.51)	**2.89** (2.63, 3.14)	**1.09** (0.98, 1.25)
B	(1, 0.05)	**13.4** (12.5, 14.1)	**6.74** (6.24, 7.06)	**3.85** (3.52, 4.18)
	(1, 0.10)	**5.13** (4.79, 5.65)	**3.97** (3.56, 4.39)	**2.64** (2.37, 2.83)
	(3, 0.05)	**7.70** (7.23, 8.16)	**4.85** (4.47, 5.28)	**2.25** (1.96, 2.55)
	(3, 0.10)	**4.40** (4.09, 4.65)	**3.26** (2.93, 3.56)	**2.22** (2.02, 2.38)
	(5, 0.05)	**3.26** (2.86, 3.76)	**2.33** (2.03, 2.77)	**1.03** (0.84, 1.13)
	(5, 0.10)	**2.61** (2.23, 2.88)	**2.02** (1.72, 2.29)	**1.28** (1.13, 1.45)
C	(1, 0.05)	**1.60** (1.54, 1.68)	0.58 (0.54, 0.61)	0.47 (0.43, 0.50)
	(1, 0.10)	**1.85** (0.77, 8.53)	0.84 (0.36, 4.15)	**1.37** (0.57, 6.35)
	(3, 0.05)	**6.25** (1.12, 8.22)	**3.64** (1.05, 5.03)	**2.20** (0.41, 3.01)
	(3, 0.10)	**4.66** (3.93, 5.58)	**3.02** (2.62, 3.79)	**3.59** (2.97, 4.05)
	(5, 0.05)	0.63 (0.60, 0.66)	0.61 (0.58, 0.63)	0.23 (0.21, 0.26)
	(5, 0.10)	0.41 (0.38, 0.44)	0.56 (0.54, 0.59)	0.30 (0.27, 0.32)
D	(1, 0.05)	**11.7** (9.89, 12.8)	**11.7** (9.89, 12.8)	**6.61** (5.62, 7.25)
	(1, 0.10)	**21.0** (18.3, 26.5)	**21.9** (18.2, 26.4)	**10.5** (8.78, 12.5)
	(3, 0.05)	**6.83** (5.98, 7.41)	**6.66** (5.85, 7.33)	**5.29** (4.40, 5.72)
	(3, 0.10)	**11.2** (9.62, 12.9)	**10.8** (9.10, 12.4)	**7.85** (6.56, 9.20)
	(5, 0.05)	**4.51** (3.91, 5.18)	**4.27** (3.68, 4.95)	**3.86** (3.20, 4.59)
	(5, 0.10)	**7.50** (6.20, 8.65)	**7.11** (5.65, 8.24)	**6.16** (4.80, 7.18)
E	(1, 0.05)	**7.77** (6.97, 9.13)	**7.76** (6.97, 9.12)	**7.03** (6.17, 8.01)
	(1, 0.10)	**15.1** (12.6, 18.0)	**15.0** (12.6, 18.0)	**11.0** (9.41, 13.4)
	(3, 0.05)	**5.55** (5.05, 6.31)	**5.73** (5.15, 6.61)	**4.88** (4.45, 5.60)
	(3, 0.10)	**9.15** (7.81, 10.7)	**9.36** (8.00, 11.0)	**7.08** (5.98, 8.25)
	(5, 0.05)	**2.83** (2.26, 3.62)	**3.03** (2.39, 3.95)	**2.54** (1.95, 3.12)
	(5, 0.10)	**5.40** (4.31, 6.71)	**5.55** (4.56, 7.09)	**4.30** (3.34, 5.30)
F	(1, 0.05)	**8.91** (7.56, 10.2)	**9.05** (7.69, 10.3)	**7.78** (6.77, 9.08)
	(1, 0.10)	**18.2** (14.6, 24.5)	**18.3** (14.7, 24.6)	**13.3** (10.9, 17.9)
	(3, 0.05)	**5.43** (4.58, 6.31)	**5.67** (4.82, 6.67)	**4.69** (3.89, 5.51)
	(3, 0.10)	**9.84** (8.83, 11.2)	**10.2** (9.12, 11.5)	**7.47** (6.51, 8.43)
	(5, 0.05)	0.51 (0.18, 0.86)	0.52 (0.19, 0.91)	0.44 (0.15, 0.72)
	(5, 0.10)	**1.03** (0.47, 2.11)	**1.07** (0.49, 2.20)	0.73 (0.36, 1.57)
G	(1, 0.05)	**13.5** (10.2, 17.0)	**13.4** (10.2, 16.8)	**7.45** (5.69, 9.11)
	(1, 0.10)	**17.2** (13.0, 24.6)	**17.2** (13.0, 25.0)	**9.95** (7.57, 13.5)
	(3, 0.05)	**9.78** (8.17, 11.8)	**9.21** (7.38, 11.2)	**6.99** (5.93, 8.65)
	(3, 0.10)	**9.76** (7.94, 12.2)	**9.34** (7.58, 12.2)	**7.80** (6.34, 9.47)
	(5, 0.05)	**7.05** (6.07, 8.36)	**7.15** (5.93, 8.67)	**5.64** (4.52, 7.21)
	(5, 0.10)	**6.93** (5.68, 8.23)	**6.44** (5.37, 8.03)	**6.26** (5.06, 7.55)

Table 3.8 – Table containing the median (the first and third quartiles are in parentheses) of the empirical ratios for the three methods we compared our method with and for scenarios A-G with the regime 1. We highlight in bold the medians that exceed 1.

Regime 1				
Scenario	(r,δ)	PACE	KL	RS
H	$(1,0.05)$	**11.0** $(8.29,13.8)$	**10.9** $(8.49,13.8)$	**9.29** $(7.66,11.8)$
	$(1,0.10)$	**14.2** $(10.4,18.2)$	**14.2** $(10.5,18.2)$	**11.7** $(8.96,16.0)$
	$(3,0.05)$	**7.76** $(6.74,9.89)$	**8.72** $(7.00,10.2)$	**6.89** $(5.65,7.85)$
	$(3,0.10)$	**8.67** $(6.83,11.2)$	**8.63** $(6.88,11.3)$	**7.95** $(6.19,10.2)$
	$(5,0.05)$	**4.80** $(3.41,6.20)$	**6.01** $(4.49,8.14)$	**4.03** $(2.94,5.44)$
	$(5,0.10)$	**5.36** $(3.82,6.89)$	**5.60** $(3.89,7.17)$	**4.67** $(3.38,5.95)$
I	$(1,0.05)$	**11.1** $(9.31,13.7)$	**11.7** $(9.68,14.2)$	**9.87** $(8.21,12.4)$
	$(1,0.10)$	**16.0** $(11.4,20.6)$	**16.2** $(11.5,20.7)$	**13.8** $(9.87,17.4)$
	$(3,0.05)$	**7.13** $(6.00,9.25)$	**7.61** $(6.49,10.0)$	**6.03** $(5.21,7.29)$
	$(3,0.10)$	**7.72** $(6.29,9.58)$	**8.17** $(6.49,9.99)$	**6.76** $(5.46,8.43)$
	$(5,0.05)$	**1.06** $(0.65,1.53)$	**1.33** $(0.72,1.92)$	0.88 $(0.53,1.27)$
	$(5,0.10)$	0.94 $(0.18,1.77)$	0.99 $(0.19,1.82)$	0.78 $(0.15,1.54)$

Table 3.9 – Table containing the median (the first and third quartiles are in parentheses) of the empirical ratios for the three methods we compared our method with and for scenarios H-I with the regime 1. We highlight in bold the medians that exceed 1.

are of the same order of magnitude, so their effects should also be of the same order.

We also analyse the dataset using the classical approach of first smoothing the curves, and then calculating the empirical covariance function and its spectrum from these smoothed curves. Since the data are periodic we use a Fourier basis smoothing with a roughness penalty approach to smooth the original data. The six leading eigenfunctions of the resulting estimated covariance function are plotted in Figure 3.10. We can see that the first three eigenfunctions are very similar to that obtained by our method. However, since the three subsequent eigenfunctions need to be orthogonal to the three first ones, they are not locally supported and are very difficult to interpret. Moreover, the first three eigenfunctions only account for 52% of the total variance explained, and the three subsequent ones for 16%, meaning that one needs to consider at least all of them to perform the analysis.

Regime 2				
Scenario	(r,δ)	PACE	KL	RS
A	(3, 0.05)	**2.05** (1.93, 2.20)	**1.53** (1.47, 1.61)	0.23 (0.19, 0.26)
	(3, 0.10)	**1.44** (1.35, 1.52)	**1.41** (1.32, 1.49)	0.56 (0.50, 0.63)
	(5, 0.05)	**2.06** (1.94, 2.18)	**1.83** (1.72, 1.93)	0.22 (0.19, 0.25)
	(5, 0.10)	**1.62** (1.52, 1.73)	**1.58** (1.48, 1.67)	0.54 (0.48, 0.59)
B	(3, 0.05)	**2.12** (2.08, 2.25)	**1.71** (1.48, 1.84)	0.57 (0.53, 0.61)
	(3, 0.10)	**1.67** (1.54, 1.79)	**1.59** (1.50, 1.70)	0.72 (0.66, 0.79)
	(5, 0.05)	**1.38** (1.29, 1.45)	**1.07** (0.95, 1.21)	0.39 (0.35, 0.42)
	(5, 0.10)	**1.15** (1.06, 1.22)	**1.06** (0.98, 1.10)	0.48 (0.45, 0.53)
C	(3, 0.05)	**5.88** (5.61, 6.23)	**2.89** (2.69, 3.14)	**1.90** (1.74, 2.12)
	(3, 0.10)	**3.06** (2.87, 3.35)	**2.00** (1.79, 2.18)	**2.17** (2.02, 2.35)
	(5, 0.05)	**1.68** (1.59, 1.76)	**1.05** (0.99, 1.11)	0.57 (0.53, 0.61)
	(5, 0.10)	0.94 (0.87, 1.02)	0.76 (0.71, 0.82)	0.69 (0.65, 0.74)
D	(3, 0.05)	**5.70** (5.06, 6.62)	**5.59** (5.03, 6.65)	**4.12** (3.54, 4.94)
	(3, 0.10)	**10.7** (8.66, 12.2)	**10.5** (8.48, 12.2)	**6.89** (5.45, 8.07)
	(5, 0.05)	**3.58** (3.10, 4.18)	**3.48** (3.05, 4.03)	**2.79** (2.33, 3.36)
	(5, 0.10)	**6.81** (5.64, 8.09)	**6.63** (5.54, 7.72)	**4.90** (4.04, 5.83)
E	(3, 0.05)	**4.60** (3.89, 5.43)	**4.66** (3.96, 5.45)	**4.16** (3.60, 4.81)
	(3, 0.10)	**8.59** (6.96, 10.2)	**8.65** (7.00, 10.2)	**6.51** (5.22, 7.80)
	(5, 0.05)	**2.09** (1.11, 2.76)	**2.14** (1.13, 2.82)	**1.84** (0.94, 2.45)
	(5, 0.10)	**3.96** (3.15, 5.46)	**4.24** (3.33, 5.72)	**3.12** (2.42, 4.27)
F	(3, 0.05)	**1.13** (0.06, 2.74)	**1.17** (0.07, 2.83)	0.99 (0.06, 2.47)
	(3, 0.10)	**3.45** (0.16, 7.03)	**3.55** (0.16, 7.20)	**2.61** (0.11, 5.21)
	(5, 0.05)	0.78 (0.07, 1.43)	0.81 (0.07, 1.50)	0.66 (0.06, 1.27)
	(5, 0.10)	0.70 (0.09, 2.85)	0.71 (0.09, 2.95)	0.52 (0.07, 2.13)
G	(3, 0.05)	**7.87** (6.60, 9.69)	**7.31** (6.22, 9.55)	**5.30** (4.43, 6.77)
	(3, 0.10)	**8.05** (6.46, 9.91)	**8.02** (6.41, 9.92)	**5.95** (4.92, 7.42)
	(5, 0.05)	**5.73** (4.73, 6.52)	**7.03** (5.95, 8.53)	**4.45** (3.53, 5.20)
	(5, 0.10)	**5.87** (4.77, 7.88)	**5.75** (4.69, 7.92)	**5.01** (3.86, 6.49)
H	(3, 0.05)	**7.10** (6.07, 8.22)	**6.99** (5.73, 8.16)	**6.06** (5.13, 7.17)
	(3, 0.10)	**7.51** (6.03, 9.43)	**7.61** (6.09, 9.53)	**6.74** (5.63, 8.19)
	(5, 0.05)	**3.84** (3.16, 4.91)	**5.26** (4.11, 6.90)	**3.40** (2.64, 4.14)
	(5, 0.10)	**3.89** (1.76, 5.46)	**4.30** (1.82, 5.84)	**3.53** (1.47, 5.02)
I	(3, 0.05)	**4.94** (3.27, 6.13)	**5.32** (3.48, 6.54)	**4.41** (3.12, 5.30)
	(3, 0.10)	**3.11** (0.20, 6.11)	**3.16** (0.20, 6.24)	**2.87** (0.17, 5.12)
	(5, 0.05)	0.59 (0.06, 1.47)	0.67 (0.07, 1.58)	0.49 (0.05, 1.24)
	(5, 0.10)	**1.16** (0.14, 2.54)	**1.20** (0.15, 2.60)	**1.02** (0.11, 2.38)

Table 3.10 – Table containing the median (the first and third quartiles are in parentheses) of the empirical ratios for the three methods we compared our method with and for the 9 scenarios we considered with the regime 2. We highlight in bold the medians that exceed 1.

regime 1 and $\sigma^2 = 1$			
r	PACE	KL	RS
1	**1.41** (1.08, 1.78)	**1.36** (1.05, 1.68)	**2.05** (1.36, 2.68)
3	**1.04** (1.00, 1.11)	**1.06** (0.97, 1.16)	**1.04** (0.97, 1.09)
5	**1.03** (0.99, 1.07)	**1.03** (0.99, 1.07)	**1.01** (0.97, 1.06)
Regime 2 and $\sigma^2 = 1.7$			
r	PACE	KL	RS
3	**1.06** (0.98, 1.18)	**1.13** (1.04, 1.28)	**1.11** (0.97, 1.27)
5	**1.01** (0.97, 1.08)	**1.06** (1.00, 1.13)	**1.03** (0.94, 1.10)

Table 3.11 – Table containing the median (the first and third quartiles are in parentheses) of the ratios for the three methods we compared our method with for the classical scenario where the rough component is a white noise. We highlight in bold the results that exceed 1.

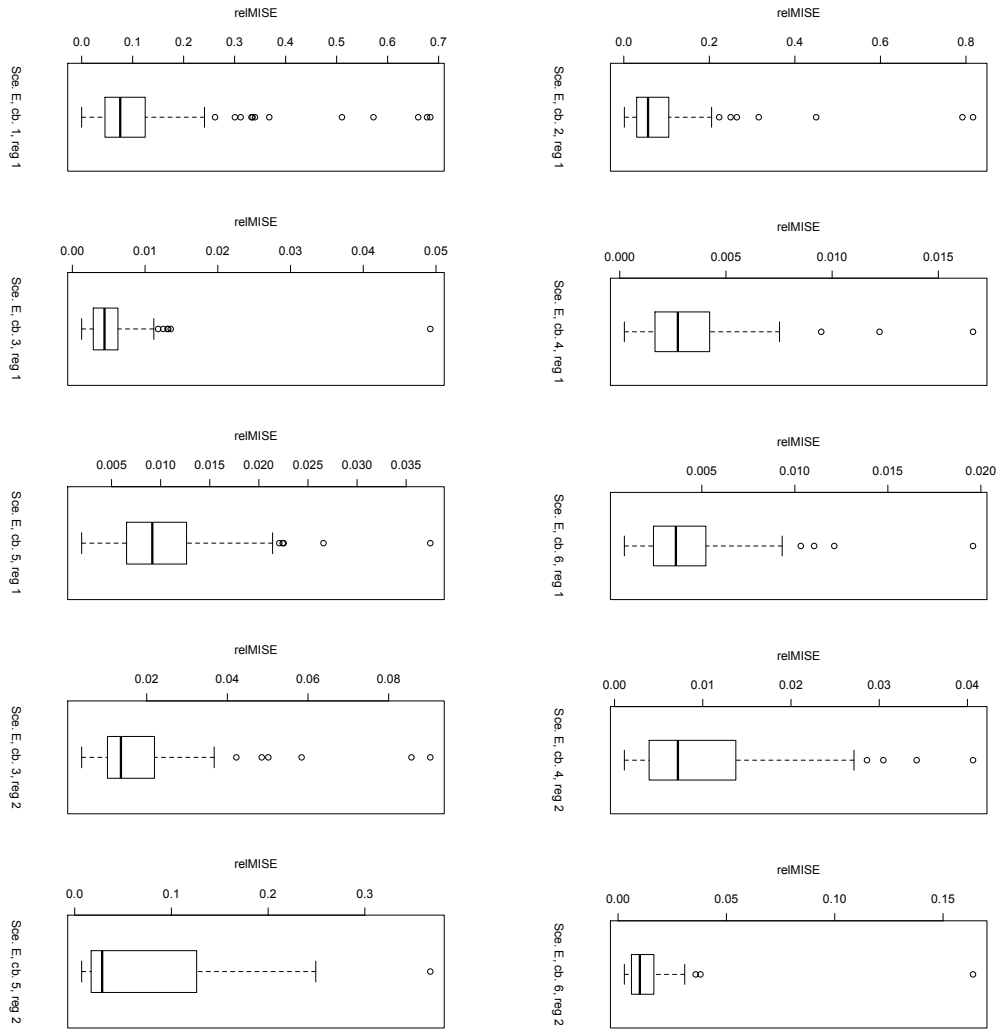

Figure 3.5 – Distributions of relMISE for Scenario E. The first 3 rows represent results for combination with $r = 1, 3$ and 5 (from top to bottom) of regime 1, and the 2 last rows represent combination with $r = 3$ and 5 (from top to bottom) of regime 2. For each row, the boxplot on the left correponds to $\delta = 0.05$ and on the right to $\delta = 0.1$.

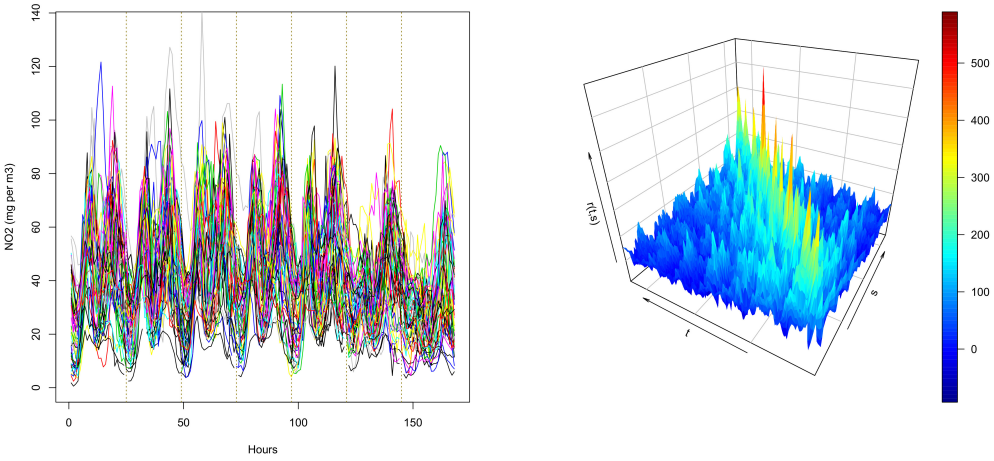

Figure 3.6 – Illustration of the dataset on the left, and of its empirical covariance function on the right.

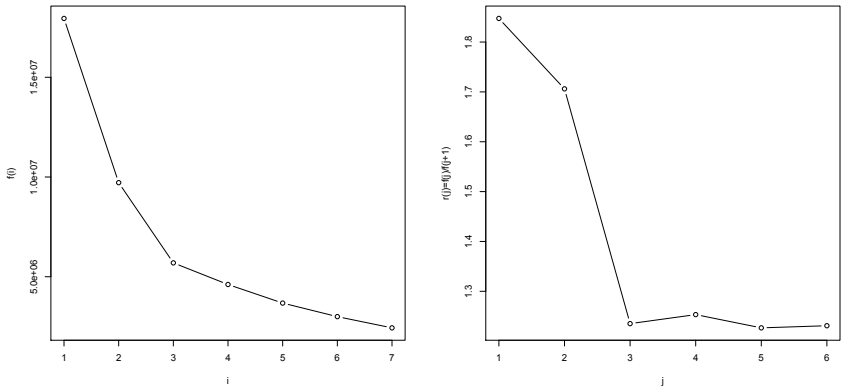

Figure 3.7 – Illustration of the scree plot approach to pick the rank. On the left we plotted the function $f(\omega) = \|P^K \circ (R_n^K - \hat{C}_\omega \hat{C}_\omega^\top)\|_F^2$ for $\omega = 1,\dots,7$, and on the right the ratio $r(j) = f(j)/f(j+1)$ for $j = 1,\dots,6$.

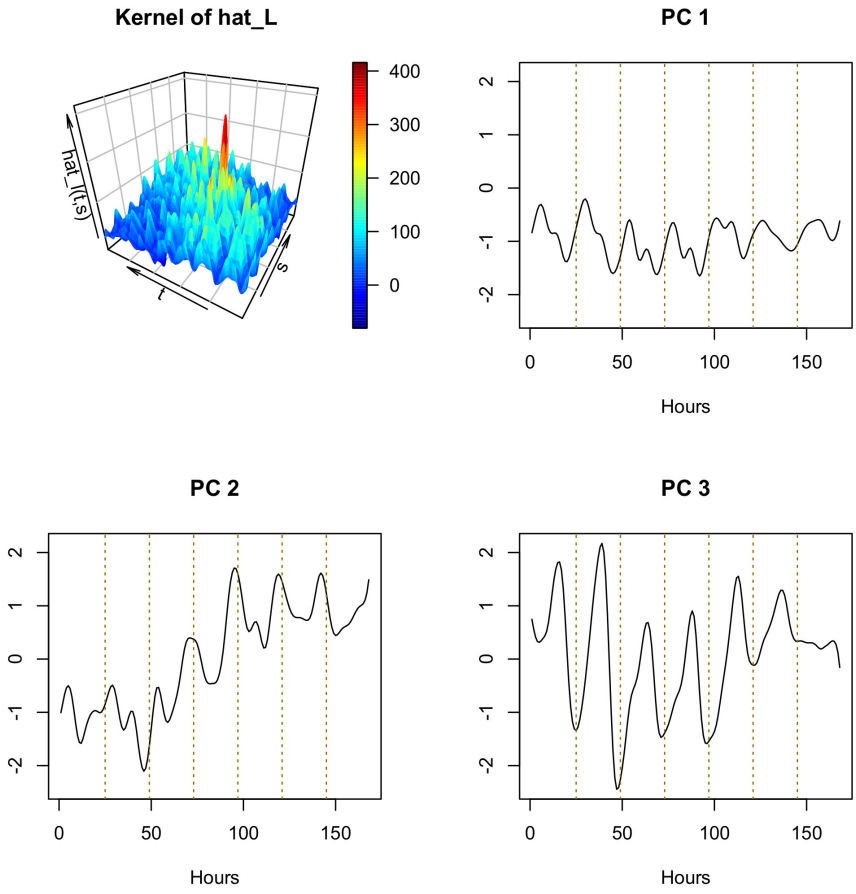

Figure 3.8 – Smoothed estimation of the covariance function ℓ and of its three eigenfunctions. On each plot of the eigenfunction, the dotted lines indicate the different days of the week.

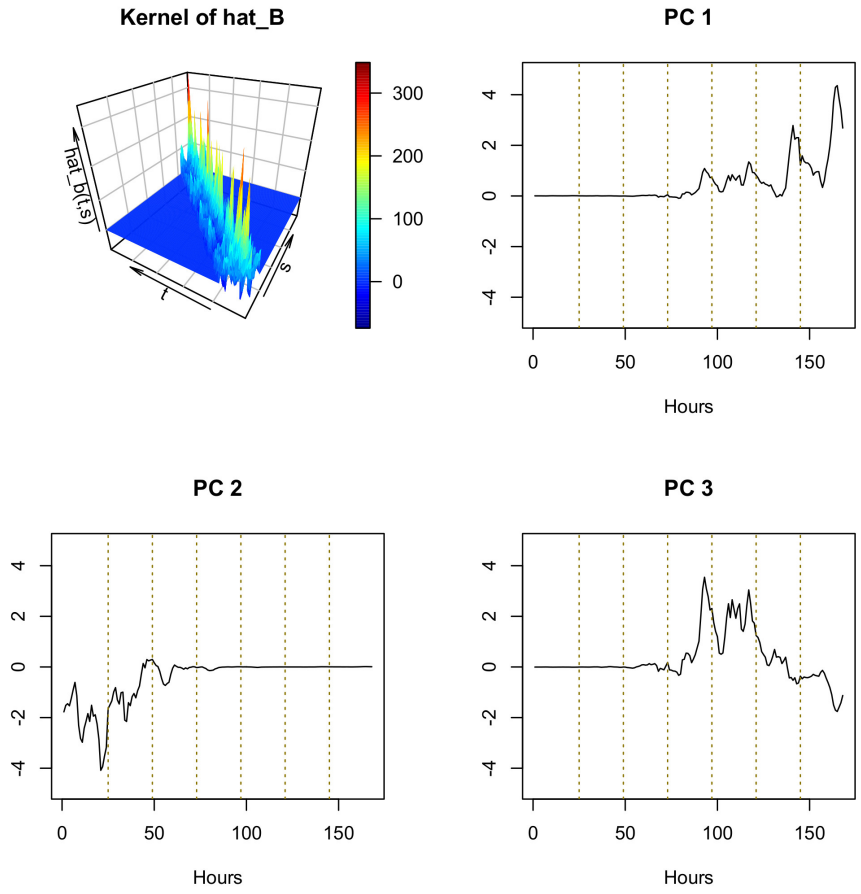

Figure 3.9 – Estimation of the covariance function b and of its first three eigenfunctions. On each plot of the eigenfunction, the dotted lines indicate the different days of the week.

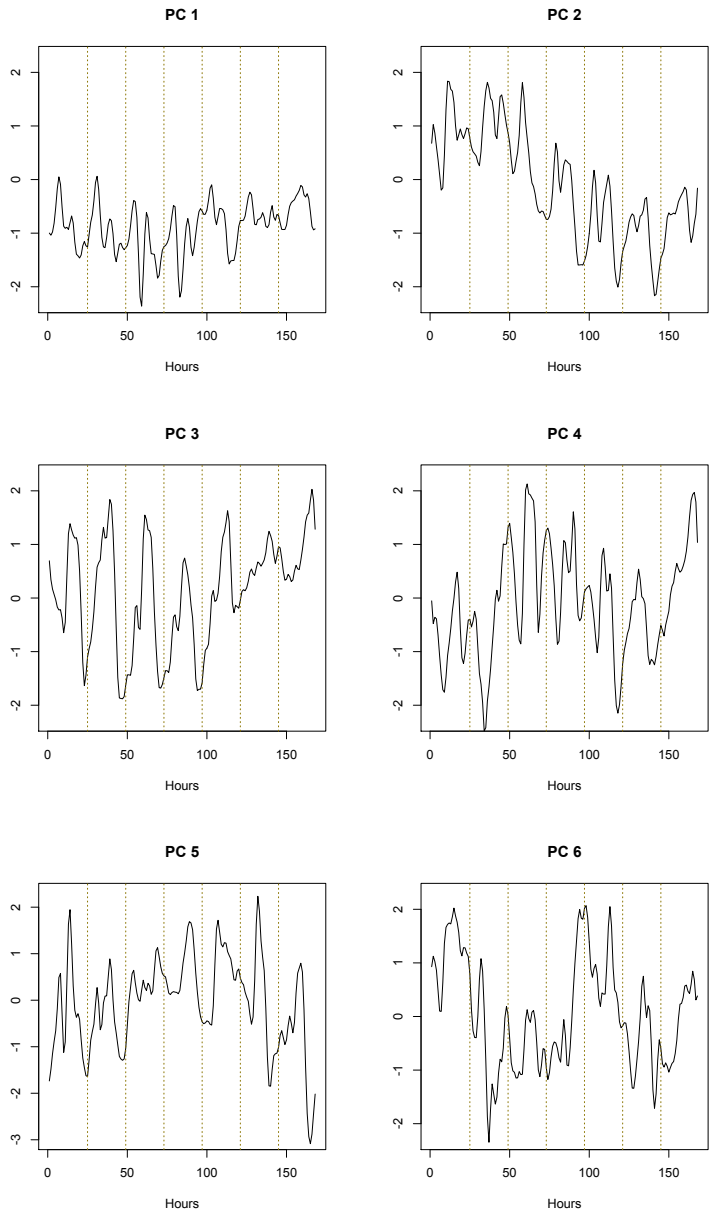

Figure 3.10 – The first six eigenfunctions of the estimate of l obtained by smoothing with a roughness penalty the empirical covariance matrix. The dotted lines indicate the different days of the week.

A Detailed results of the comparison study

This section contains boxplots of the distribution of the relative errors $\mathrm{RE}(u) = \|u - L^K\|_F / \|L^K\|_F$ and of the empirical relative errors $\mathrm{RE}_n(u) = \|u - L_n^K\|_F / \|L_n^K\|_F$ for our method, the PACE method, the truncation of the Karhunen-Loève (KL) expansion method and the spline/Fourier smoothing method. These results are complementary to those in Tables 3.5, 3.6, 3.7, 3.8, 3.9 and 3.10 of Section 3.2.5. In each of figures A.1 to A.27, the distributions of the relative errors are plotted on the left and the distributions of the empirical relative errors on the right. The horizontal lines on the graphics indicate the level 0.05.

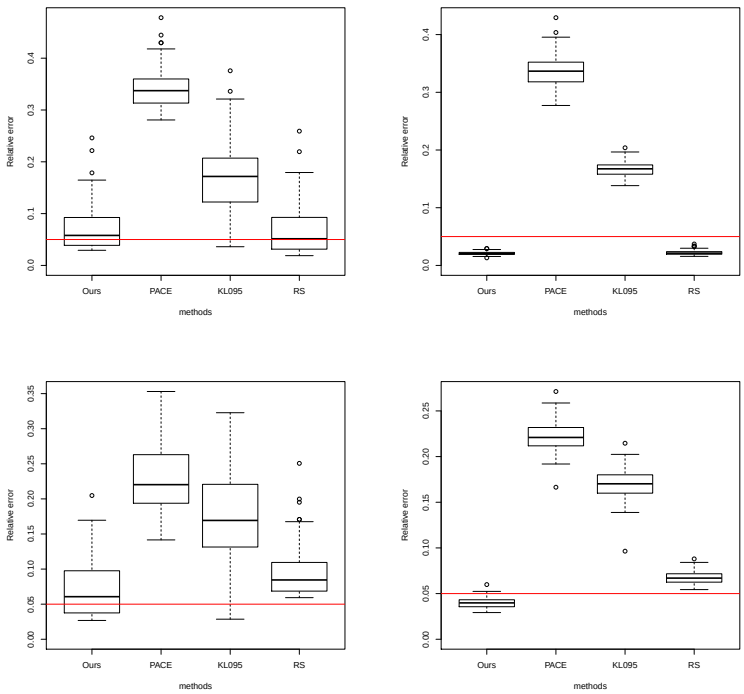

Figure A.1 – Scenario A, combinations 1 and 2 (from top to bottom) of regime 1.

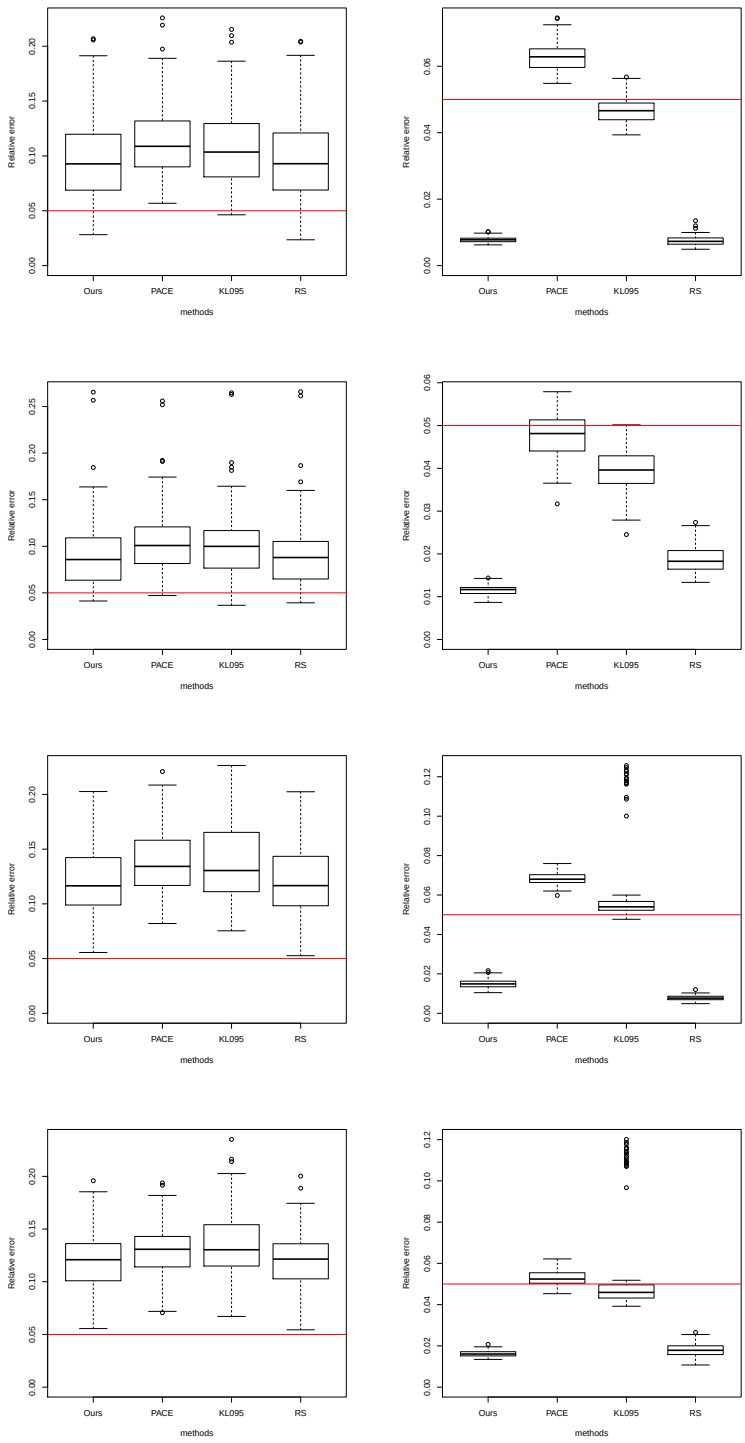

Figure A.2 – Scenario A, combinations 3 to 6 (from top to bottom) of regime 1.

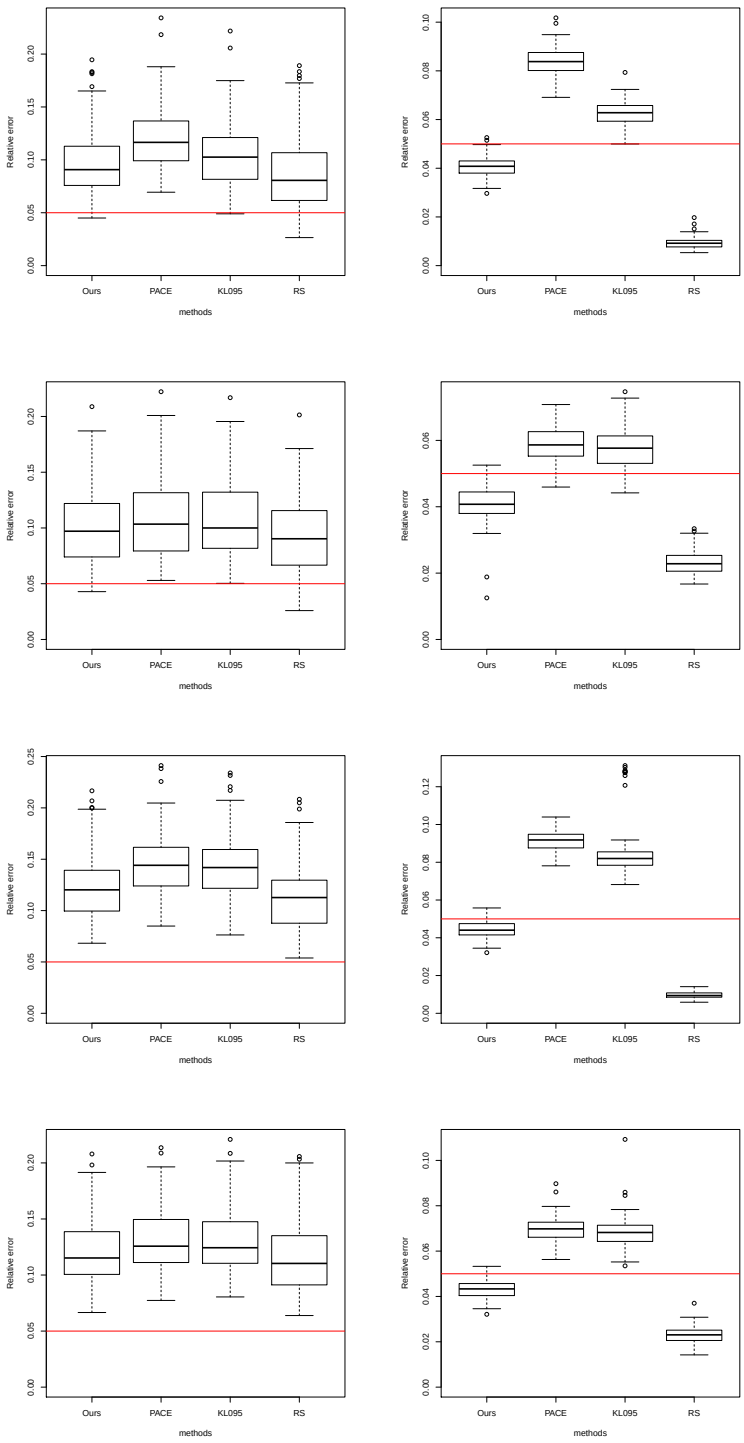

Figure A.3 – Scenario A, combinations 3 to 6 (from top to bottom) of regime 2.

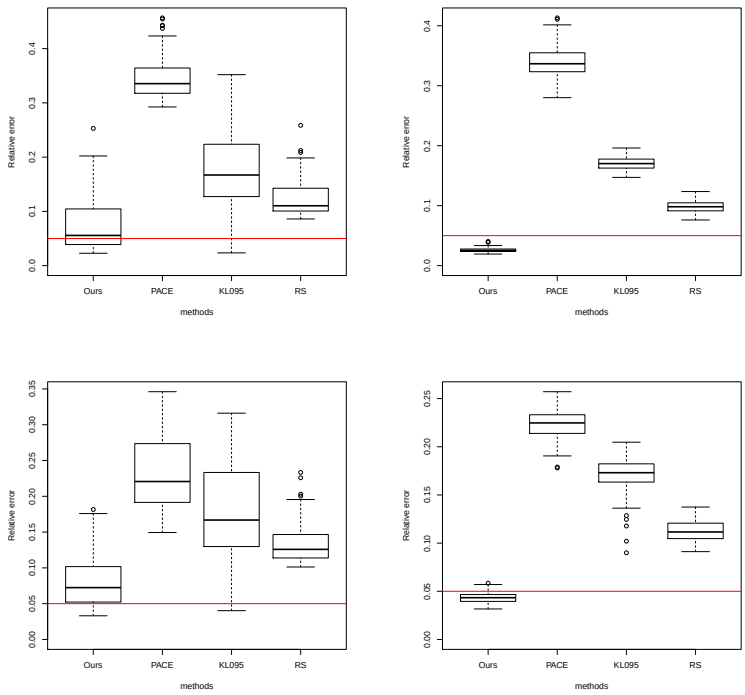

Figure A.4 – Scenario B, combinations 1 and 2 (from top to bottom) of regime 1.

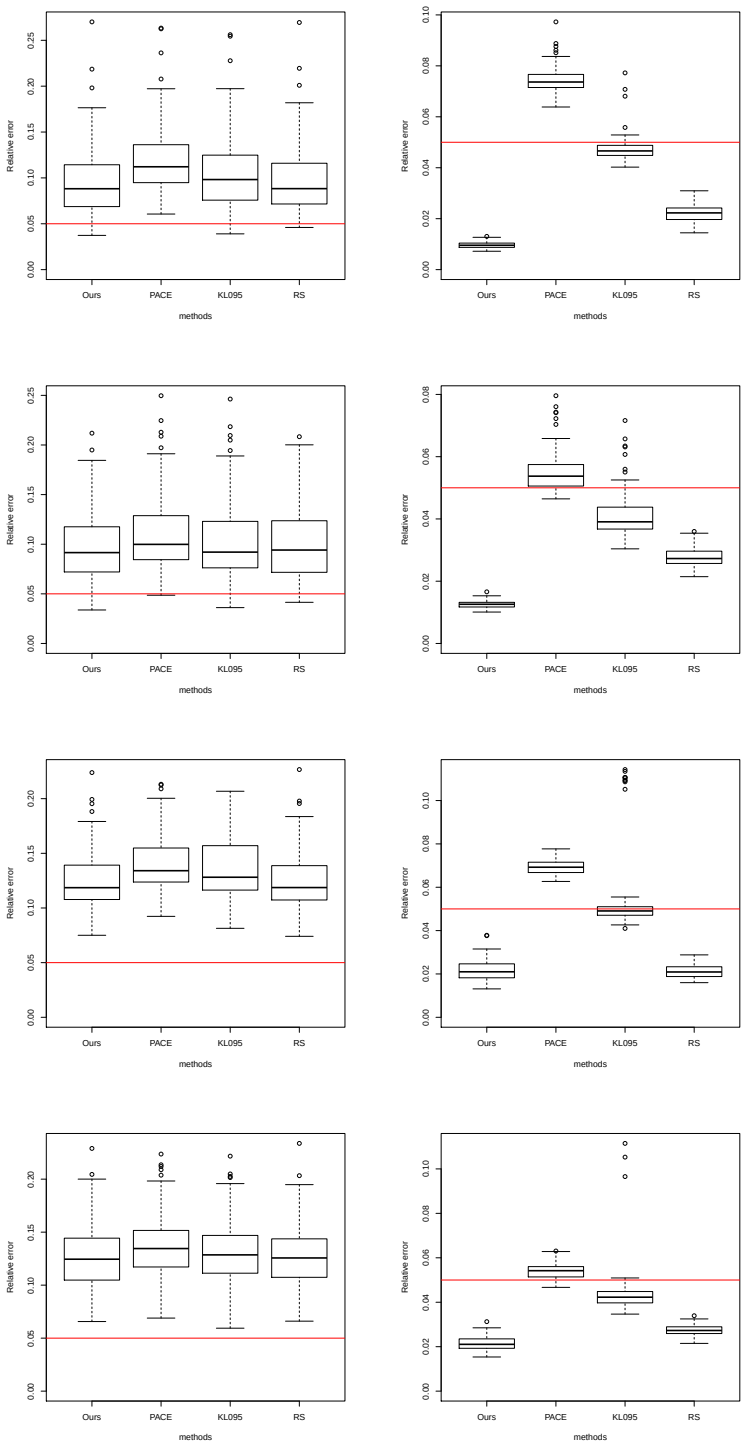

Figure A.5 – Scenario B, combinations 3 to 6 (from top to bottom) of regime 1.

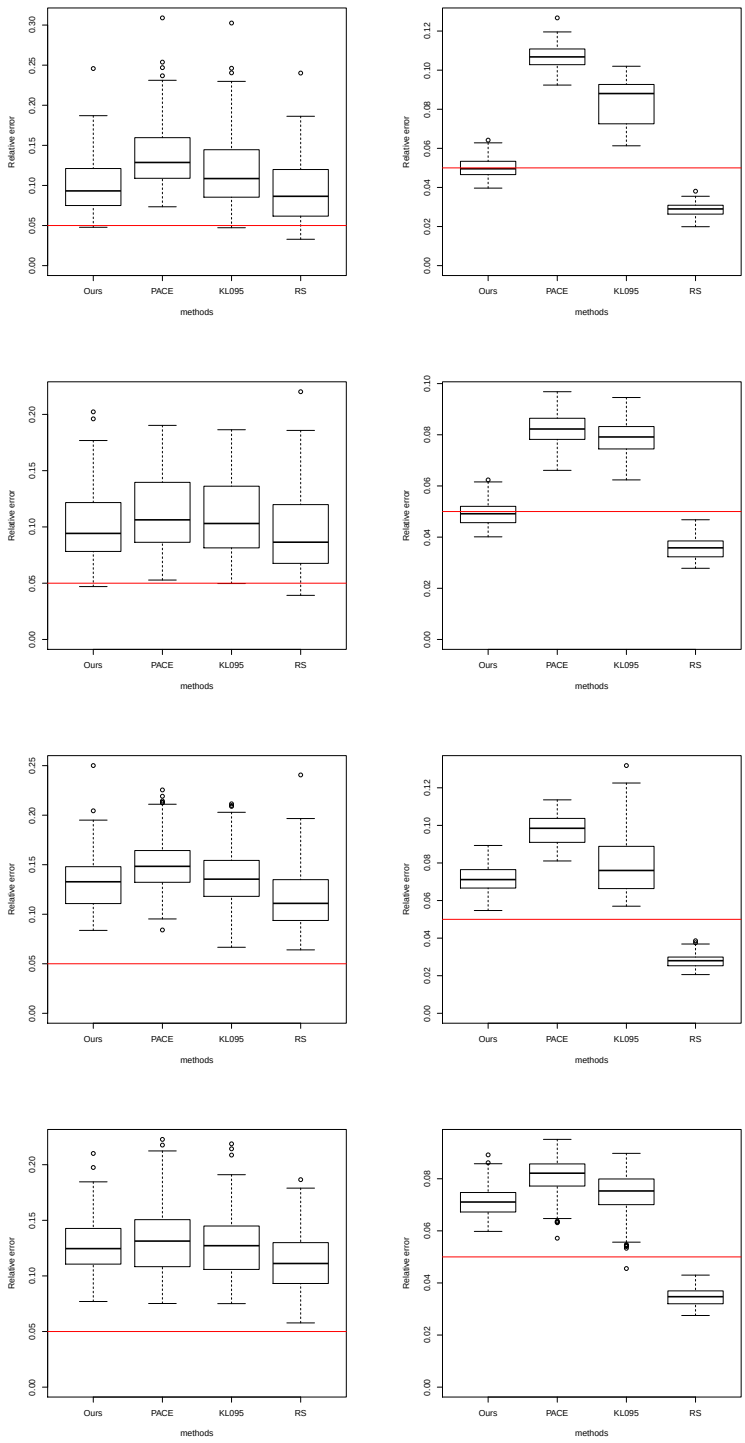

Figure A.6 – Scenario B, combinations 3 to 6 (from top to bottom) of regime 2.

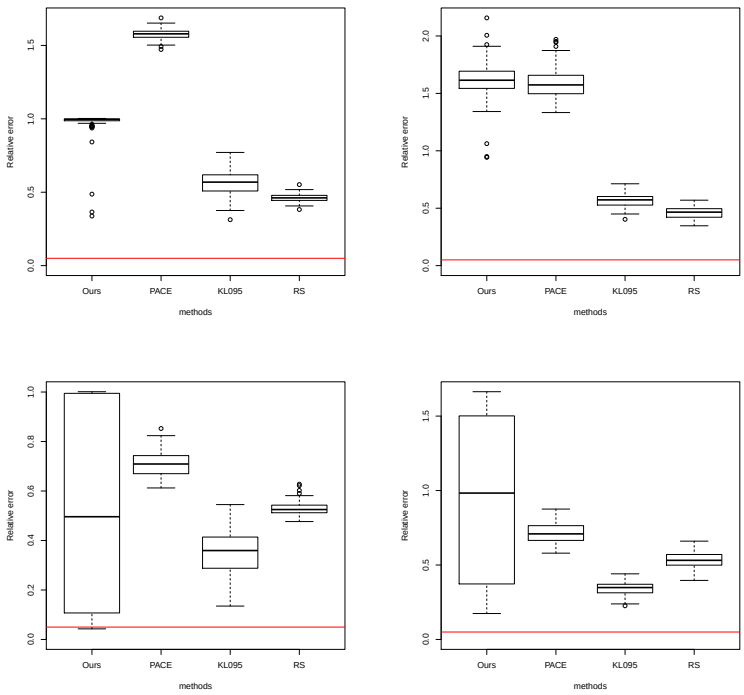

Figure A.7 – Scenario C, combinations 1 and 2 (from top to bottom) of regime 1.

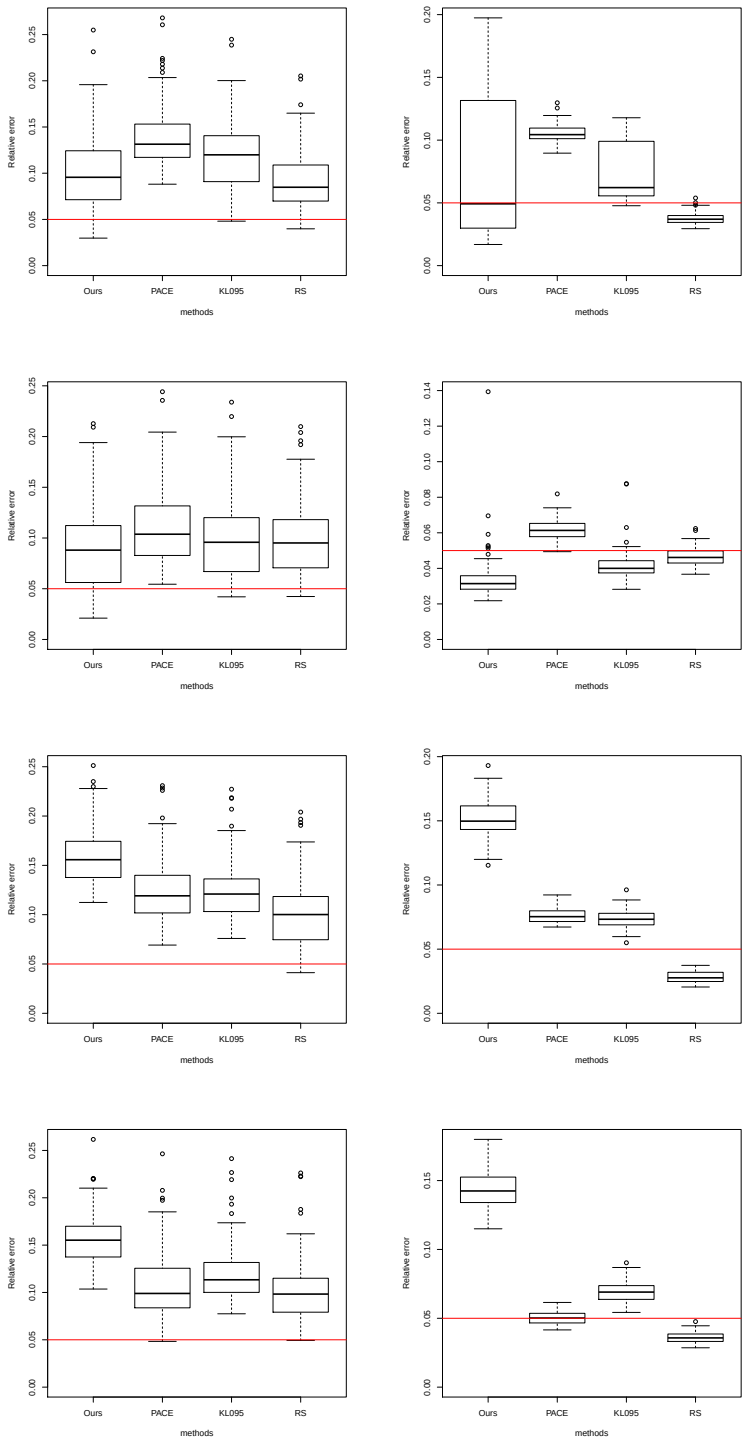

Figure A.8 – Scenario C, combinations 3 to 6 (from top to bottom) of regime 1.

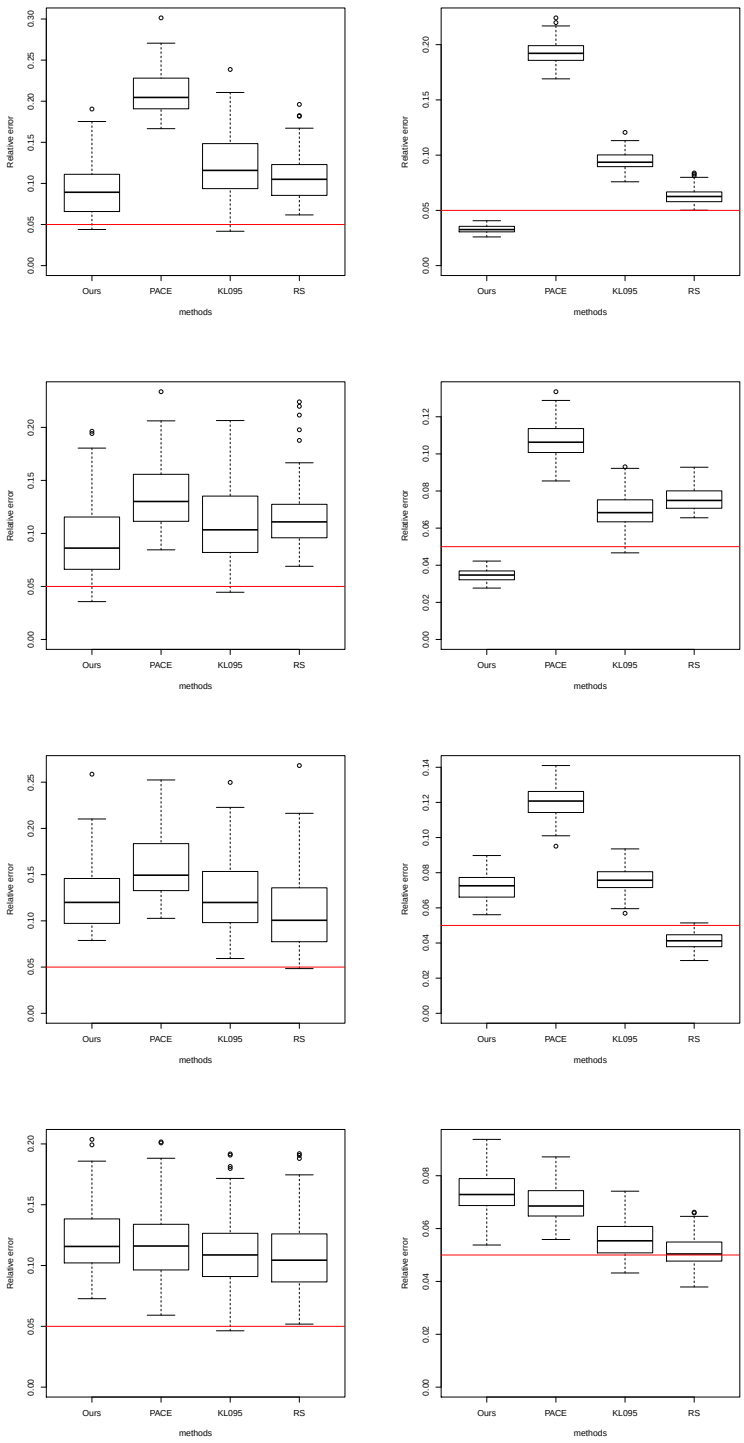

Figure A.9 – Scenario C, combinations 3 to 6 (from top to bottom) of regime 2.

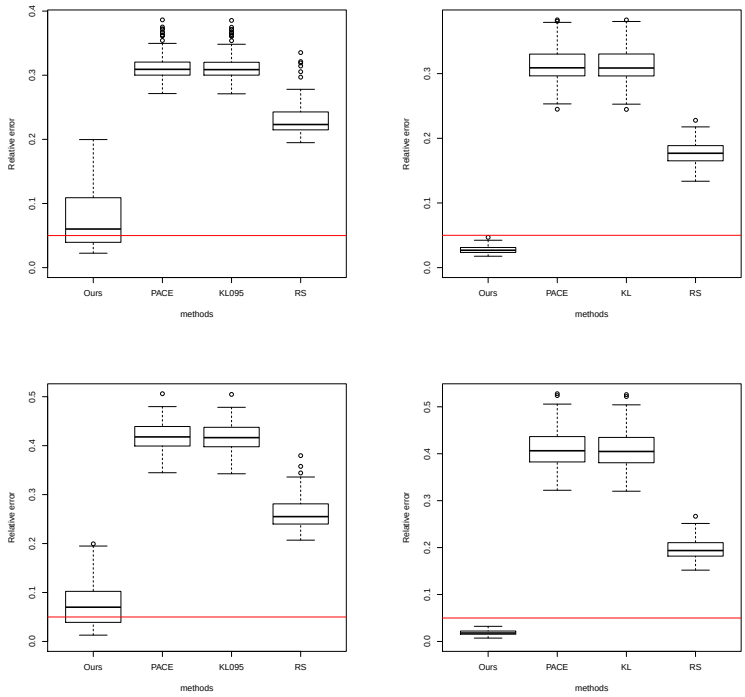

Figure A.10 – Scenario D, combinations 1 and 2 (from top to bottom) of regime 1.

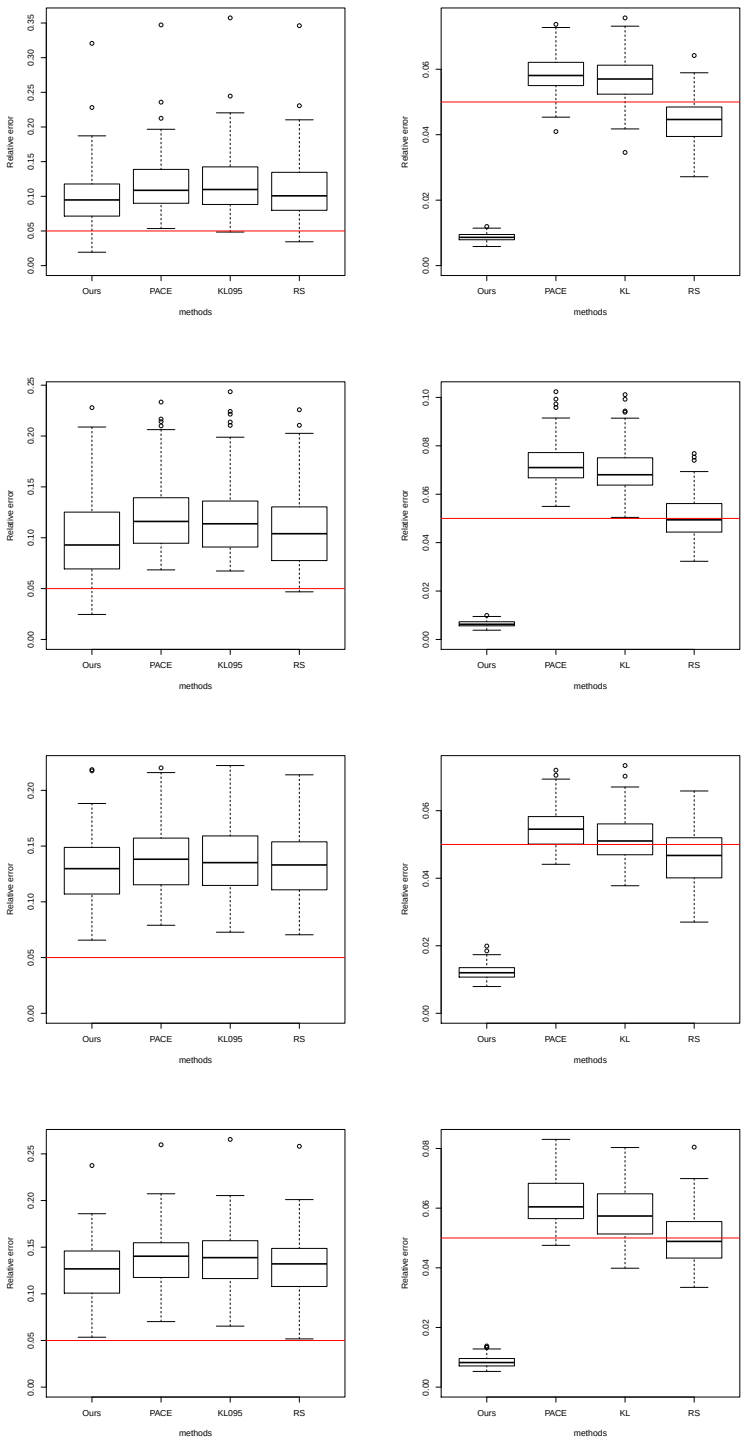

Figure A.11 – Scenario D, combinations 3 to 6 (from top to bottom) of regime 1.

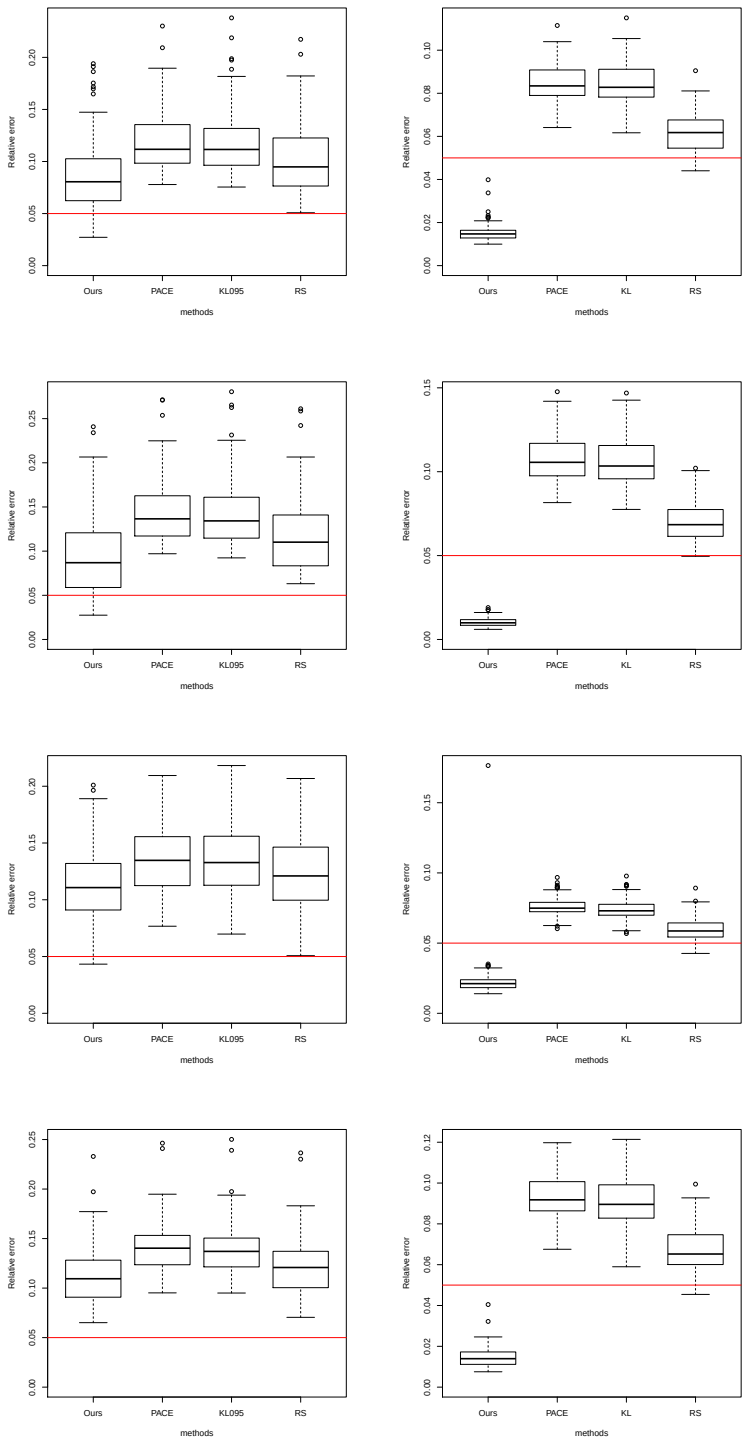

Figure A.12 – Scenario D, combinations 3 to 6 (from top to bottom) of regime 2.

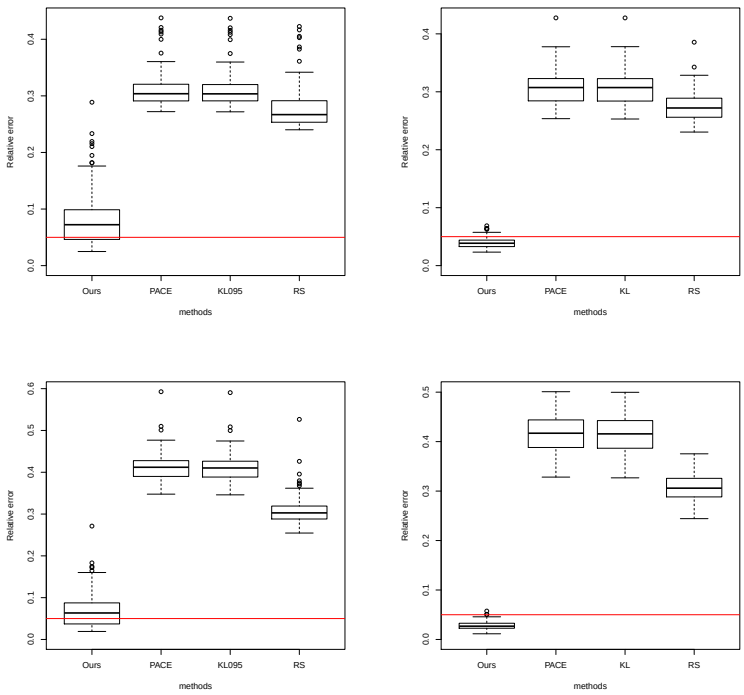

Figure A.13 – Scenario E, combinations 1 and 2 (from top to bottom) of regime 1.

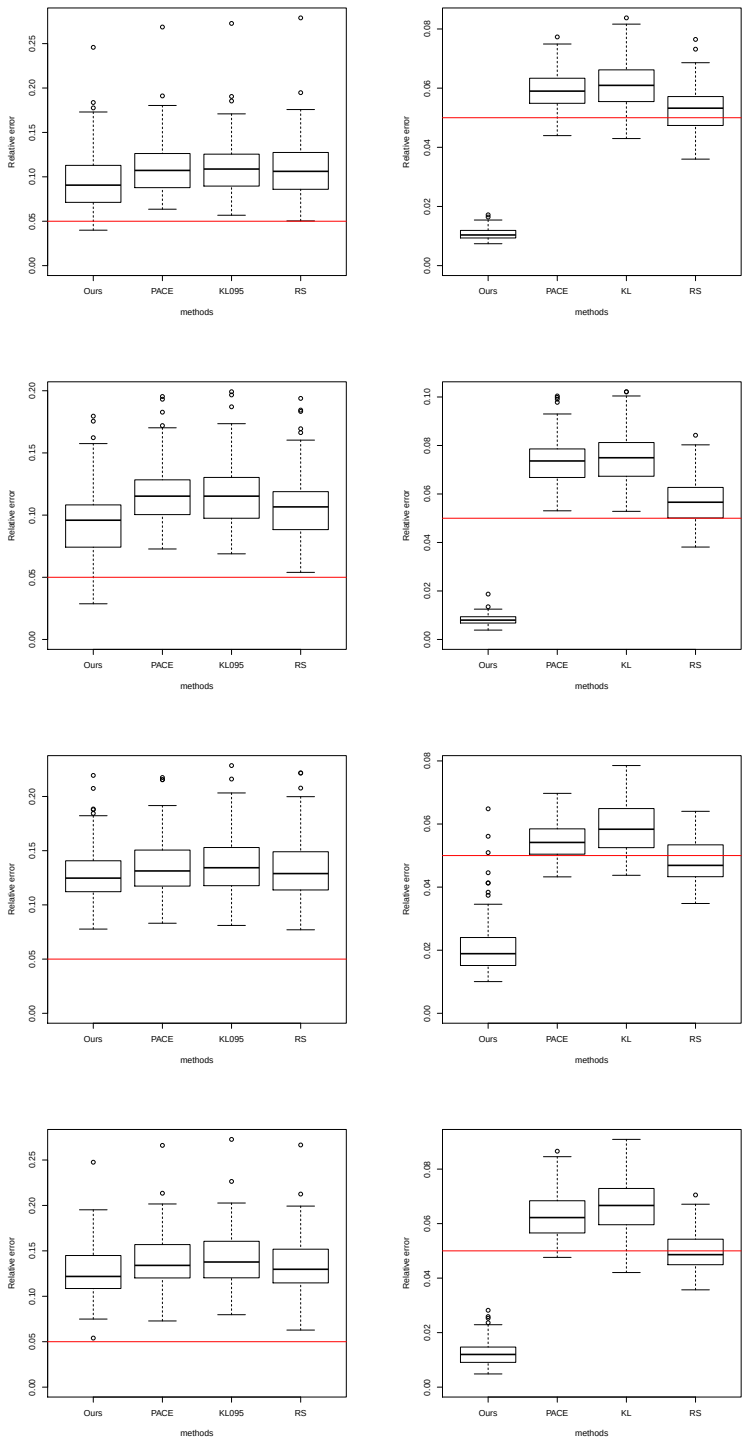

Figure A.14 – Scenario E, combinations 3 to 6 (from top to bottom) of regime 1.

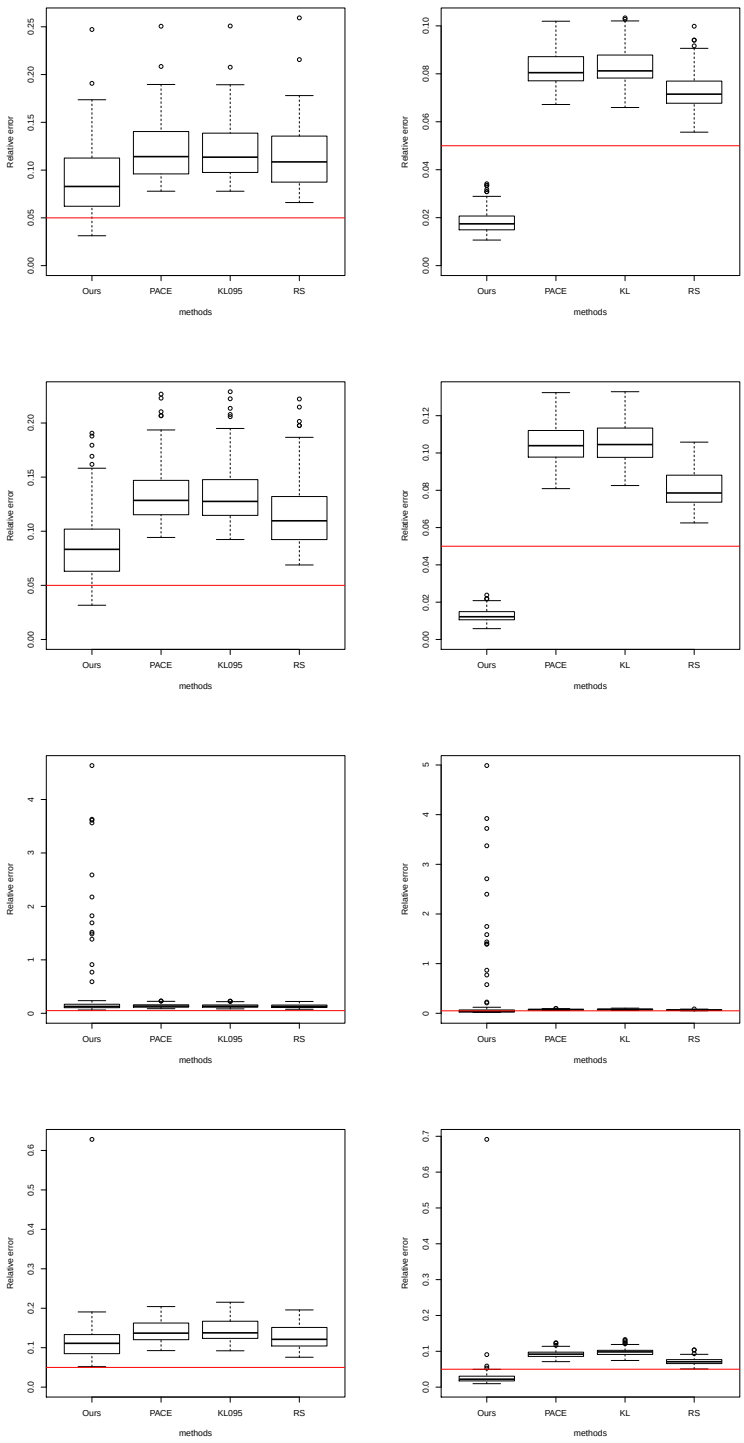

Figure A.15 – Scenario E, combinations 3 to 6 (from top to bottom) of regime 2.

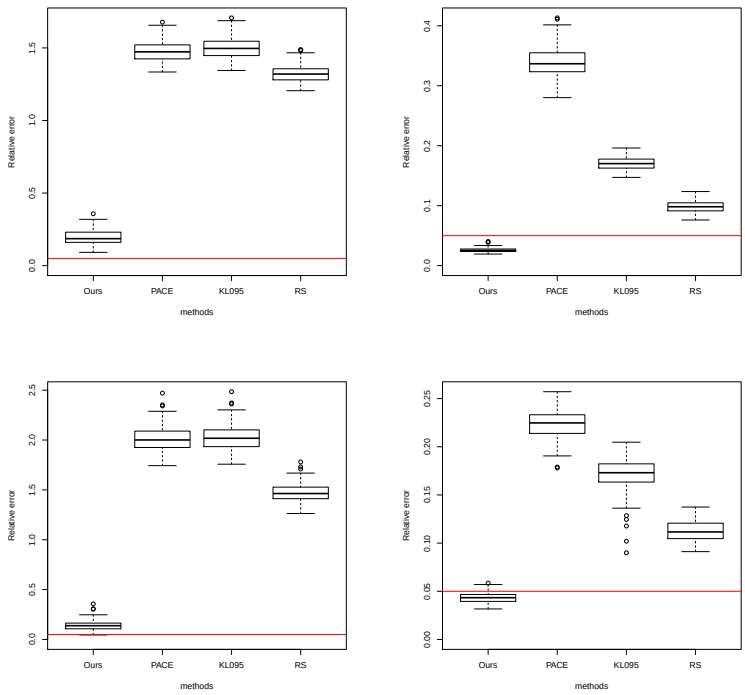

Figure A.16 – Scenario F, combinations 1 and 2 (from top to bottom) of regime 1.

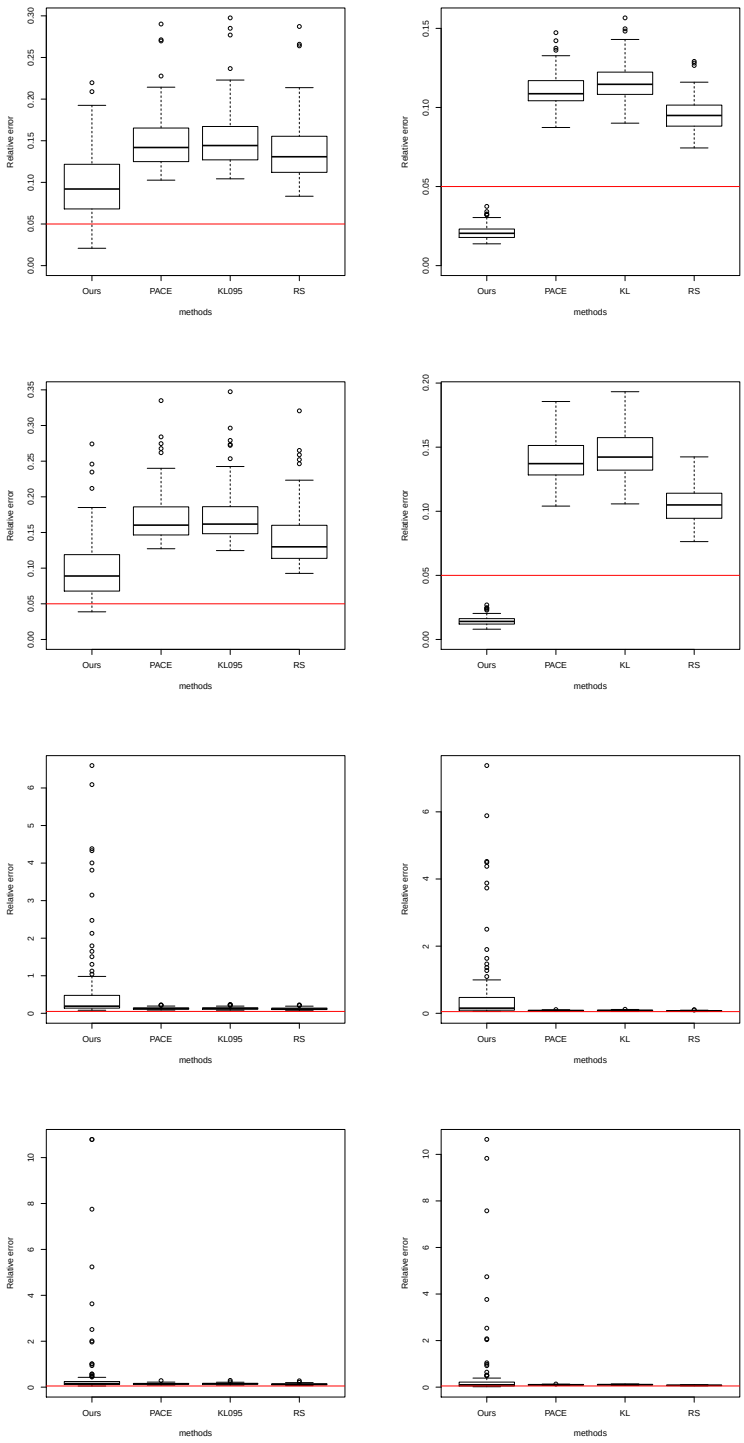

Figure A.17 – Scenario F, combinations 3 to 6 (from top to bottom) of regime 1.

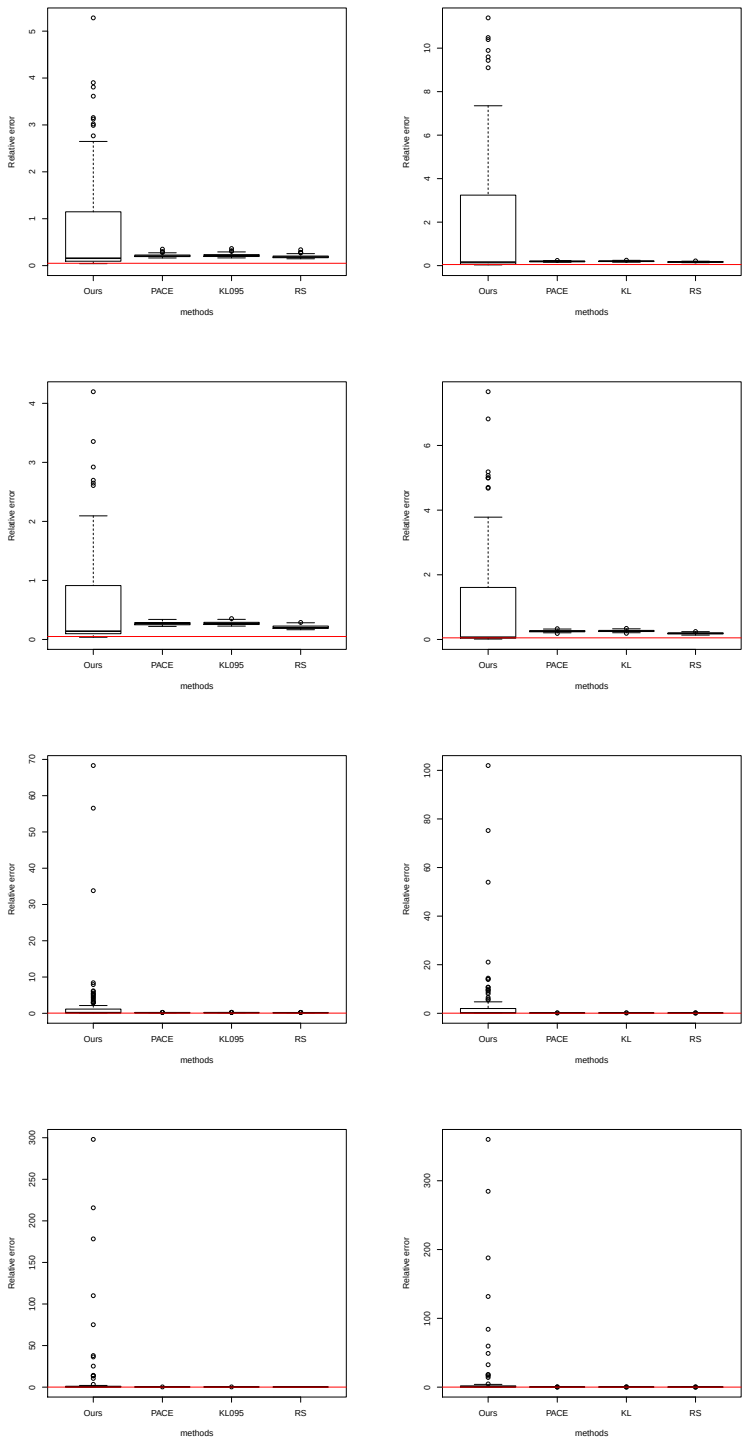

Figure A.18 – Scenario F, combinations 3 to 6 (from top to bottom) of regime 2.

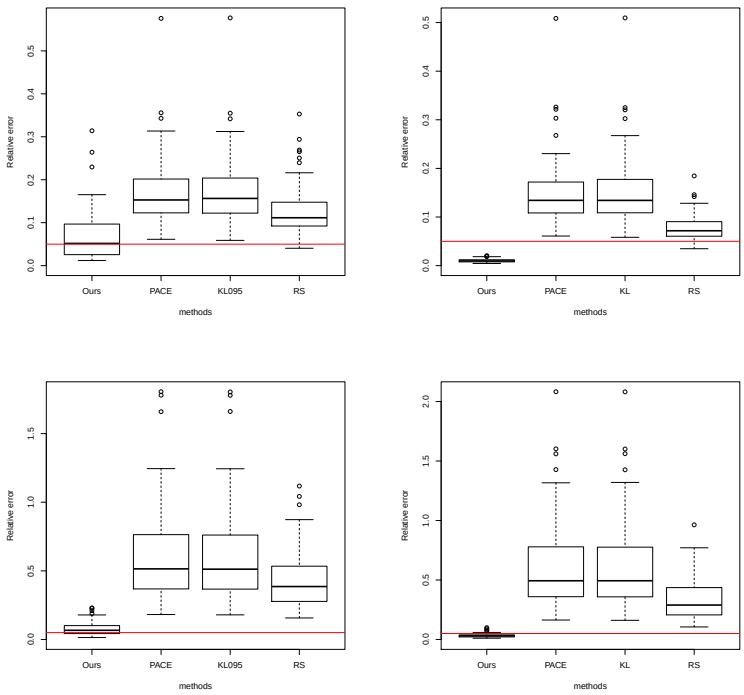

Figure A.19 – Scenario G, combinations 1 and 2 (from top to bottom) of regime 1.

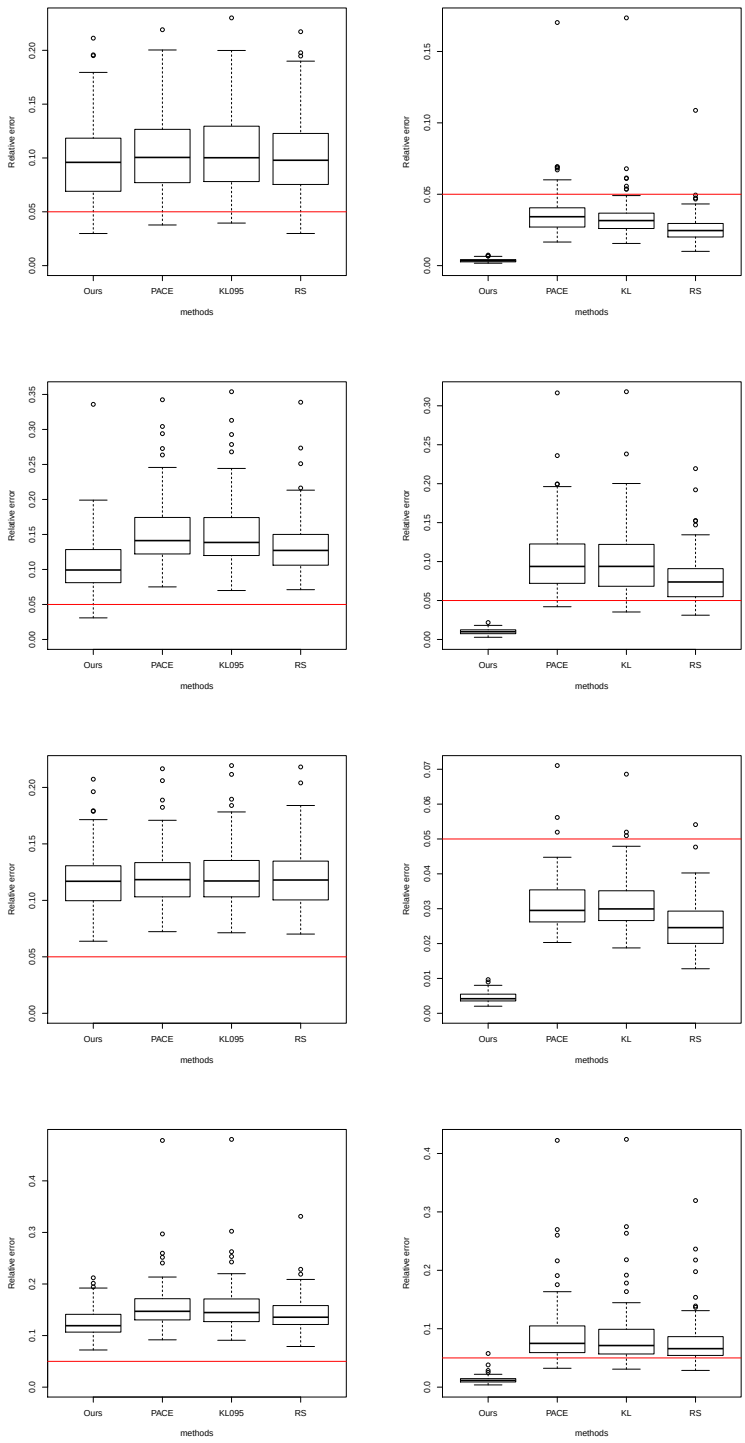

Figure A.20 – Scenario G, combinations 3 to 6 (from top to bottom) of regime 1.

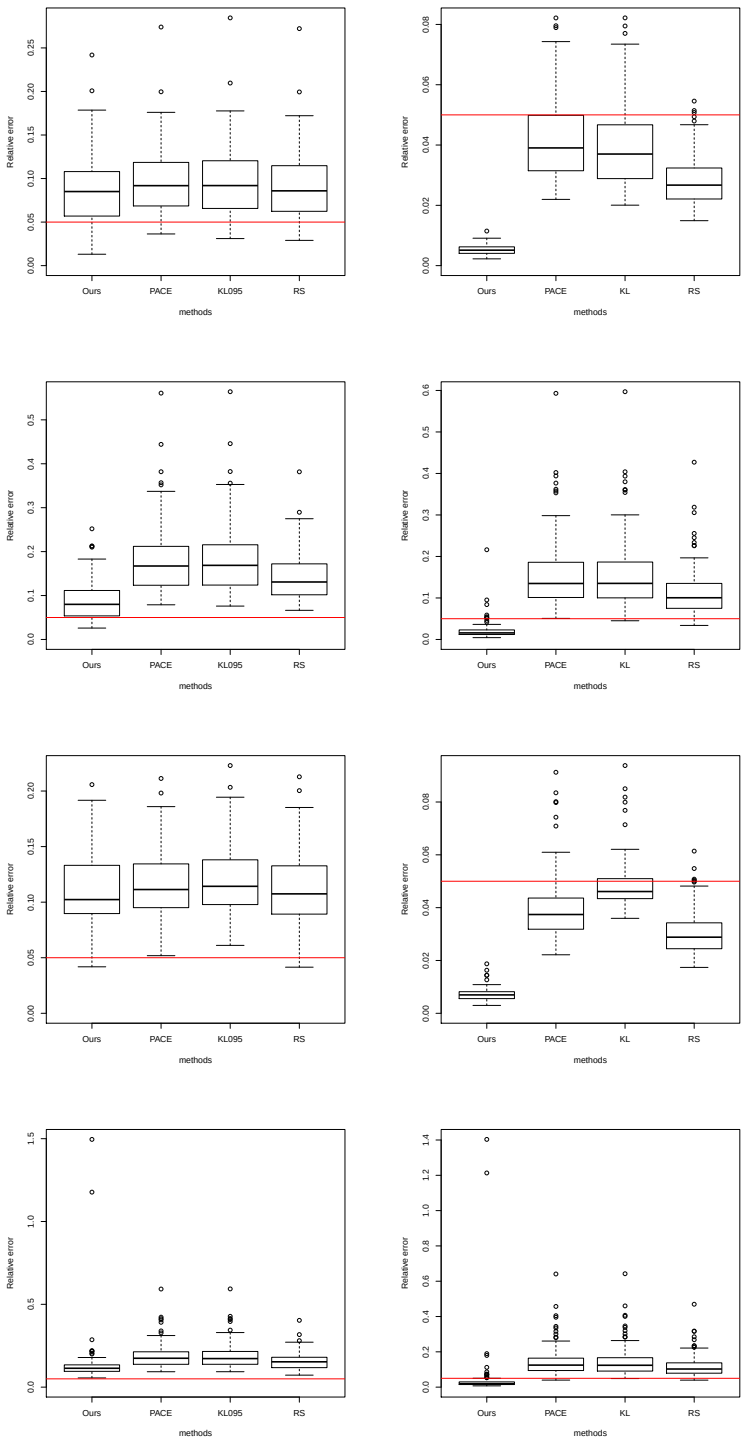

Figure A.21 – Scenario G, combinations 3 to 6 (from top to bottom) of regime 2.

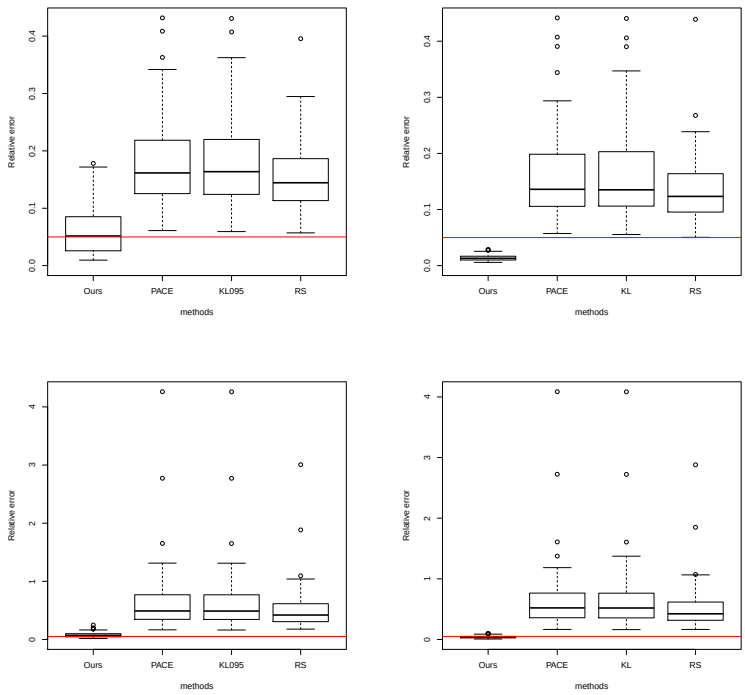

Figure A.22 – Scenario H, combinations 1 and 2 (from top to bottom) of regime 1.

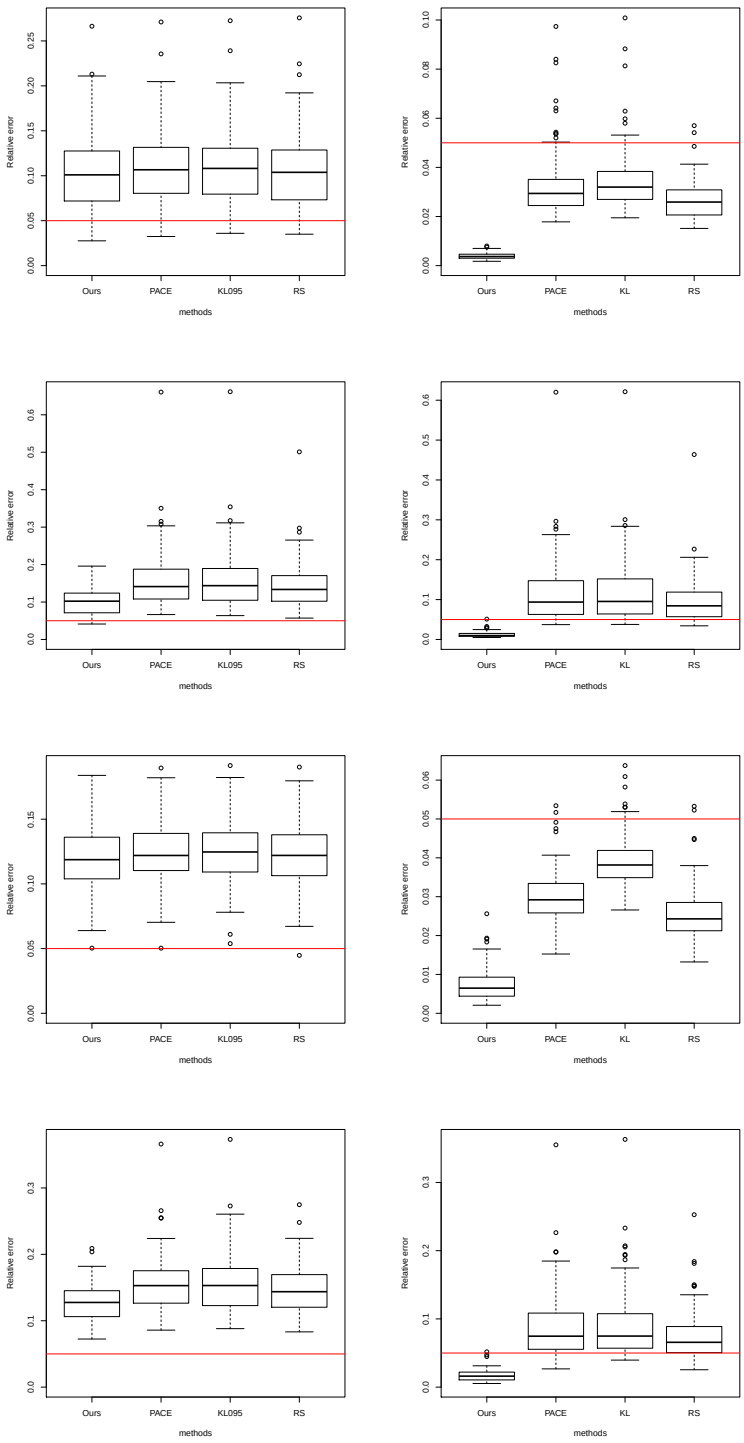

Figure A.23 – Scenario H, combinations 3 to 6 (from top to bottom) of regime 1.

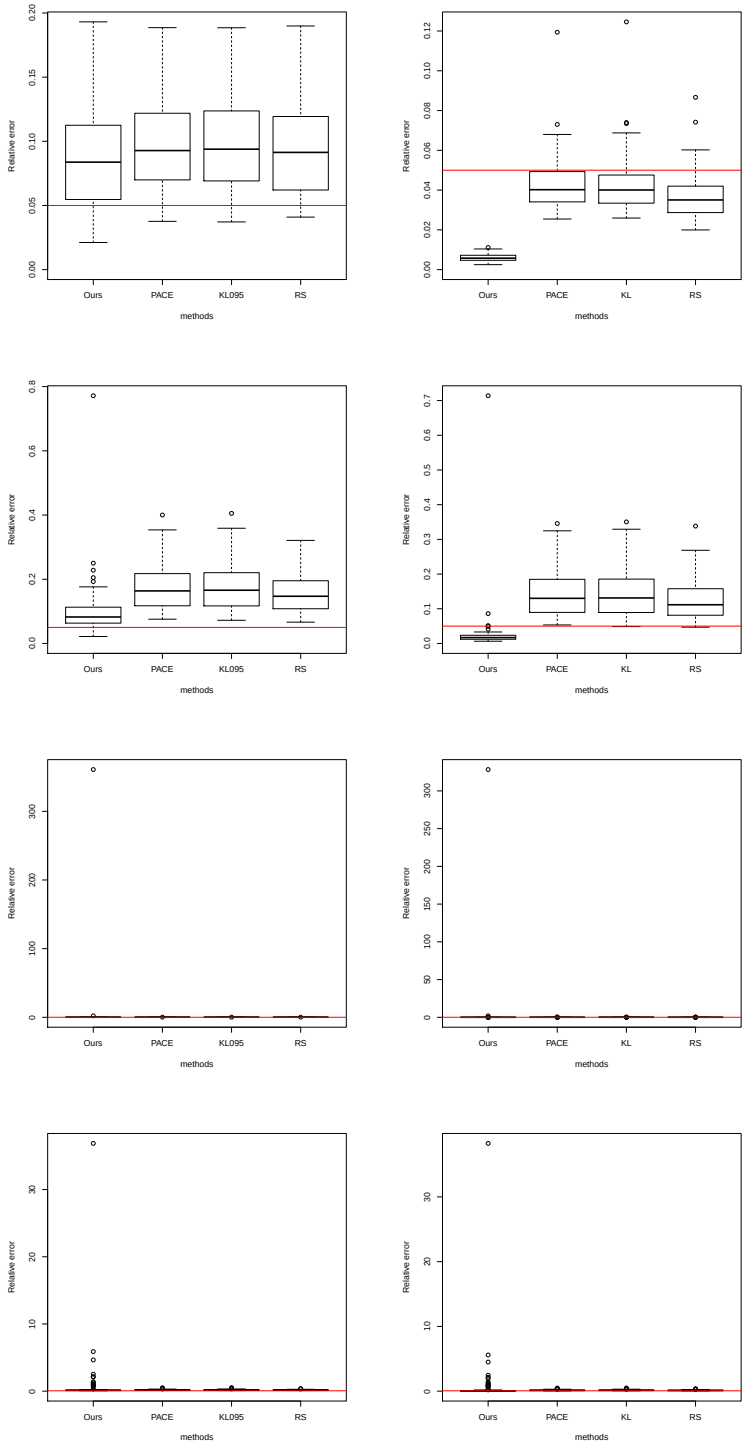

Figure A.24 – Scenario H, combinations 3 to 6 (from top to bottom) of regime 2.

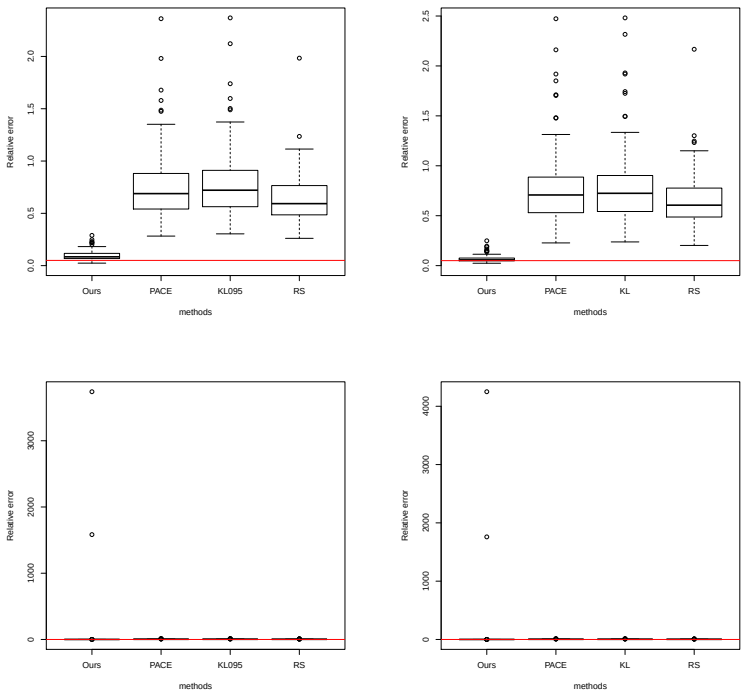

Figure A.25 – Scenario I, combinations 1 and 2 (from top to bottom) of regime 1.

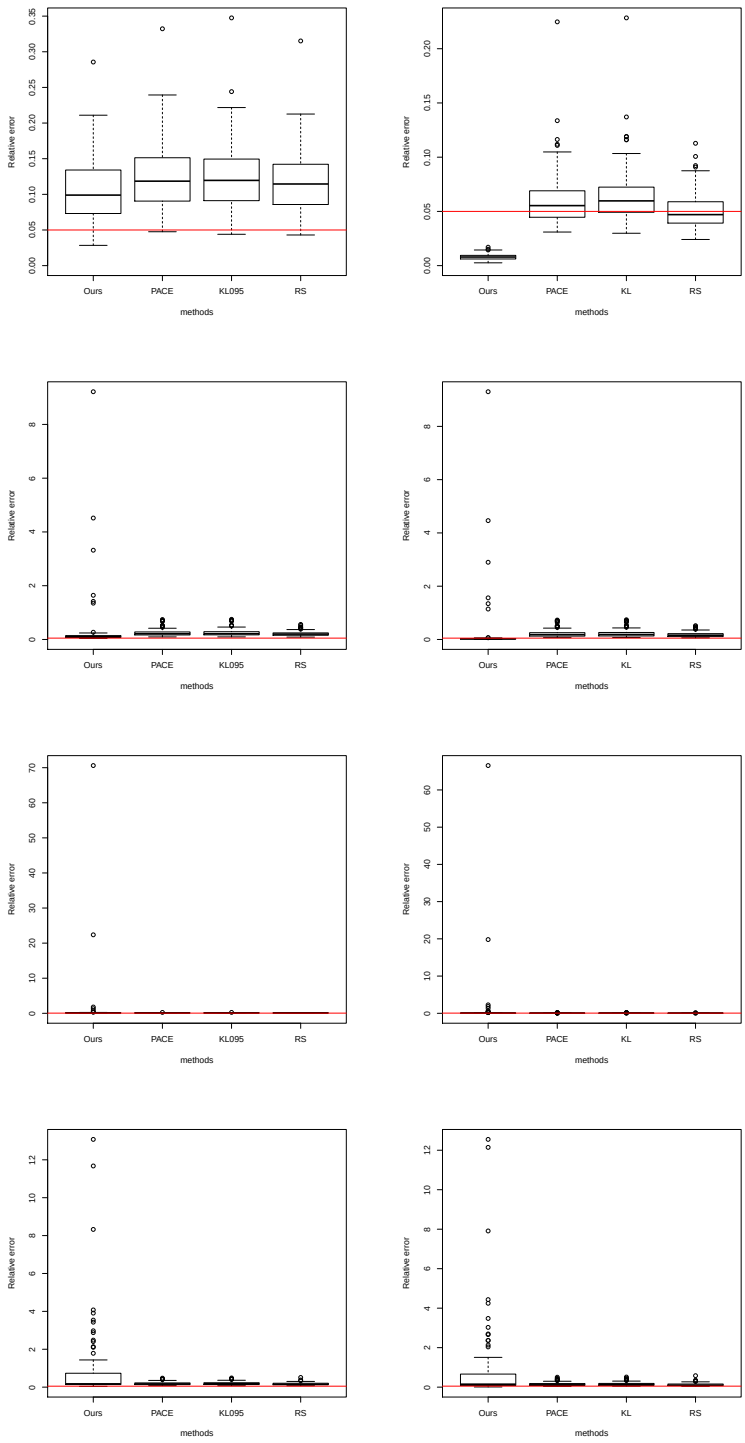

Figure A.26 – Scenario I, combinations 3 to 6 (from top to bottom) of regime 1.

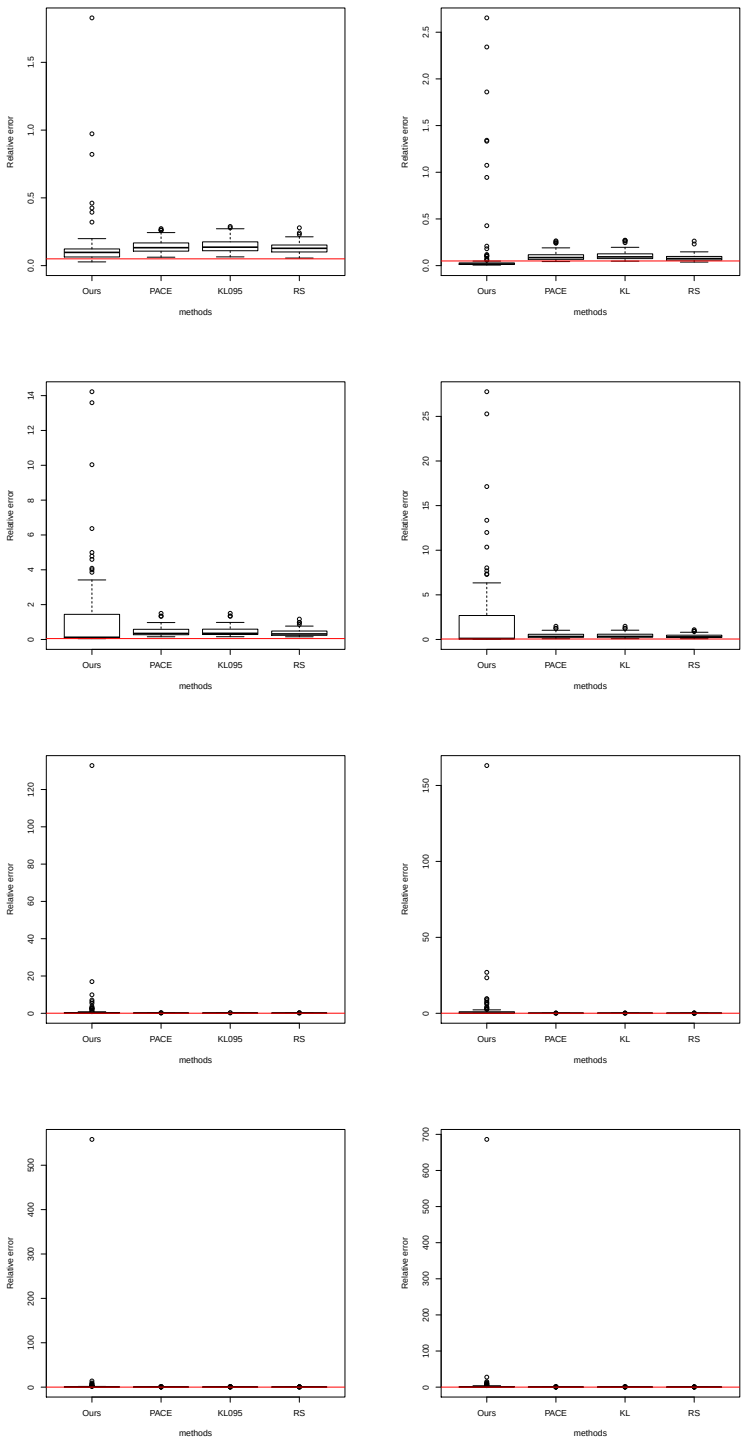

Figure A.27 – Scenario I, combinations 3 to 6 (from top to bottom) of regime 2.

Bibliography

[AW12] A.A. Amini and M. J. Wainwright. Sampled forms of functional pca in reproducing kernel hilbert spaces. *The Annals of Statistics*, 40(5), 2012.

[BB96] H. Bauschke and J. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM review*, 38(3):367–426, 1996.

[BCF97] P. Besse, H. Cardot, and F. Ferraty. Simultaneous nonparametric regressions of unbalanced longitudinal data. *Computational Statistics and Data Analysis*, 24, 1997.

[BD86] J.P. Boyle and R.L. Dykstra. A method for finding projections onto the intersection of convex sets in hilbert spaces. In *Advances in order restricted statistical inference*, pages 28–47. Springer, 1986.

[BJ03] R. Basri and D. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(2), 2003.

[Bos00] D. Bosq. *Linear processes in function space*. Springer, 2000.

[Bos14] D. Bosq. Computing the best linear predictor in a hilbert space; applications to general armah processes. *Journal of Multivariate Analysis*, 124:436–450, 2014.

[BR86] P. Besse and J.O. Ramsay. Principal components analysis of sampled functions. *Psychometrika*, 51, 1986.

[BTA11] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Sciences & Business Media, 2011.

[CaSPW11] V. Chandrasekaran, a. Sanghavi, P.A. Parrilo, and A.S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[CBSW14] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward. Coherent matrix completion. *Proceedings of the 31th International Conference on Machine Learning*, 32, 2014.

[Che15] Y. Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5), 2015.

[CJSC13] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis. Low-rank matrix recovery from errors and erasures. *IEEE Transactions on Information Theory*, 59(7), 2013.

[CL94] T. Coleman and Y. Li. On the convergence of reflective newton methods for large-scale nonlinear minimization subject to bounds. *Mathematical Programming*, 67(2):189–224, 1994.

[CL96] T. Coleman and Y. Li. An interior, trust region approach for nonlinear minimization subjects to bounds. *SIAM Journal on Optimization*, 6:418–445, 1996.

[CL08] J.-M. Chiou and P.-L. Li. Correlation-based functional clustering via subspace projection. *Journal of the American Statistical Association*, 103, 2008.

[CLMW11] E.J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3), 2011.

[CP10] E.J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6), 2010.

[CR09] E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 717(9), 2009.

[CT10] E.J. Candès and T. Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5), 2010.

[CW15] Y. Chen and M.-J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *ArXiv e-prints*, September 2015.

[CY11] T. Cai and M. Yuan. Optimal estimation of the mean function based on discretely sampled functional data: phase transition. *The Annals of Statistics*, 39(5), 2011.

[dB01] C. de Boor. *A practical guide to splines*, volume 27. Springer verlag, 2001.

[DHLZ02] P. Diggle, P. Heagerty, K. Liang, and S. Zeger. *Principal component analysis*. Oxford University Press, 2nd edition, 2002.

[DP16] M.-H. Descary and V.M. Panaretos. Functional data analysis by matrix completion. *ArXiv e-prints*, September 2016.

[DPR82] J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of Multivariate Analysis*, 12(1):136–154, 1982.

[Faz02] M. Fazel. *Matrix rank minimization with applications*. PhD thesis, Stanford University, 2002.

[FG96] J. Fan and I Gijbels. *Local polynomial modelling and its applications*. London: Chapman and Hall, 1996.

[FV06] F. Ferraty and P. Vieu. *Nonparametric functional data analysis*. Springer Series in Statistics, 2006.

[Gre50] U. Grenander. Stochastic processes and statistical inference. *Arkiv för Matematik 1*, 1950.

[Gre81] U. Grenander. *Abstract inference*. Wiley New-York, 1981.

[Gro11] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3), 2011.

[GS94] P.J. Green and B.W. Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. London: Chapman and Hall, 1994.

[HE15] T. Hsing and R. Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*. John Wiley & Sons, 2015.

[HHN06] P. Hall and M. Hosseini-Nasab. On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 2006.

[HK12] L. Horváth and P. Kokoszka. *Inference for functional data with applications*, volume 200. Springer Science & Business Media, 2012.

[HMW06] P. Hall, H.-G. Müller, and J.-L Wang. Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, pages 1493–1517, 2006.

[JH01] G.M. James and T.J. Hastie. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Methodological)*, 63, 2001.

[Jol02] I.T. Jolliffe. *Principal component analysis*. Springer Series in Statistics, 2002.

[Kar46] K. Karhunen. Zur spektraltheorie stochastischer prozesse. *Suomalainen tiedeakatemia*, 1946.

[Kat80] T. Kato. *Perturbation theory for linear operators*. Springer, 1980.

[KP02] S.G. Krantz and H.R. Parks. *A primer of real analytic functions*. Birkhäuser, 2002.

[KT12] F. Király and R. Tomioka. A combinatorial algebraic approach for the identifiability of low-rank matrix completion. *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[LH10] Y. Li and T. Hsing. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, pages 3321–3351, 2010.

[LM06] X. Leng and H.-G. Müller. Time ordering of gene co-expression. *Biostatistics*, 7, 2006.

[Loe46] M. Loeve. Fonctions aléatoire à décomposition orthogonale exponentielle. *La revue scientifique*, 84, 1946.

[Mal09] S. Mallat. *A wavelet tour of signal processing, third edition*. Elsevier, 2009.

[MAT12] MATLAB. *version 8.0.0.783 (R2012b)*. The MathWorks Inc., Natick, Massachusetts, 2012.

[Mor15] J.S. Morris. Functional regression. *Annual Review of Statistics and its Application*, 2, 2015.

[Mul05] H.-G. Muller. Functional modelling and classification of longitudinal data. *Scandinavian Journal of statistics*, 32, 2005.

[Net] Inc Netflix. The netflix prize. *http://www.netflixprize.com/*.

[OWY01] J. Opsomer, Y. Wang, and Y. Yang. Nonparametric regression with correlated errors. *Statistical Science*, pages 134–153, 2001.

[PKM10] V.M. Panaretos, D. Kraus, and J.H. Maddocks. Second-order comparison of gaussian random functions and the geometry of dna minicircles. *Journal of the American Statistical Association*, 105(490):670–682, 2010.

[PM08] J. Peng and H.-G. Müller. Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *Annals of Applied Statistics*, 2, 2008.

[PP11] D. Paul and J. Peng. Principal components analysis for sparsely observed correlated functional data using a kernel smoothing approach. *Electronic Journal of Statistics*, 5, 2011.

[PS93] S.D. Pezzulli and B.W. Silverman. Some properties of smoothed principal components analysis for functional data. *Computational Statistics*, 8, 1993.

[Ram82] J.O. Ramsay. When the data are functions. *Psychometrika*, 47, 1982.

[Rao58] C.R. Rao. Some statistical methods for comparison of growth curves. *Biometrics*, 14, 1958.

[RD91] J.O. Ramsay and C. Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 51, 1991.

[Rec11] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

[RS91] J. Rice and B.W. Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53, 1991.

[RS05] J.O. Ramsay and B.W. Silverman. *Functional Data Analysis.* Springer, New York, 2005.

[Sil84] B.W. Silverman. Spline smoothing: the equivalent variable kernel method. *The Annals of Statistics*, pages 898–916, 1984.

[Sil96] B.W. Silverman. Smoothed functional principal components analysis by choice of norm. *Annals of statistics*, 24, 1996.

[SL98] J.G. Staniswalis and J.J. Lee. Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 93, 1998.

[SWT96] M. Shi, R.E. Weiss, and J.M. Taylor. An analysis of paediatric cd4 counts for acquired immune deficiency syndrome using flexible random curves. *Applied Statistics*, 45, 1996.

[vdVW96] A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes with applications to statistics.* Springer series in statistics, 1996.

[WCM16] J.-L. Wang, J.-M. Chiou, and H.-G. Müller. Functional data analysis. *Annual Review of Statistics and its Application*, 3, 2016.

[YM+03] F. Yao, H.-G. Müller, et al. Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics*, 59, 2003.

[YMW05] F. Yao, H.-G. Müller, and J.-L. Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.

[ZLW+10] Z. Zhou, X. Li, J. Wright, E.J. Candès, and Y. Ma. Stable principal component pursuit. *Proceedings of 2010 IEEE International Symposium on Information Theory*, 2010.

[ZW16] X. Zhang and J.-L. Wang. From sparse to dense functional data and beyond. *The Annals of Statistics*, 44(5), 2016.