Université de Lille 1 SCIENCES ET TECHNOLOGIES

## Université de Lille

Doctoral School **ED Régionale SPI 72**

University Department **Laboratoire de Mathématiques Paul Painlevé**

Thesis defended by **Jérémie Kellner**

Defended on **1st December, 2016**

In order to become Doctor from Université de Lille

Academic Field **Mathematics**

Speciality **Probability and Statistics**

Thesis Title

# Gaussian Models and Kernel Methods

| | | |
|---|---|---|
| **Thesis supervised by** | Christophe Biernacki | Supervisor |
| | Alain Celisse | Co-Supervisor |

**Committee members**

| | | |
|---|---|---|
| *Referees* | Gilles Blanchard | Professor at Universität Potsdam |
| | Yannick Baraud | Professor at Université de Nice Sophia-Antipolis |
| *Examiners* | Cristina Butucea | Professor at Université Paris-Est Marne-la-Vallée |
| | Vincent Rivoirard | Professor at Université Paris-Dauphine |
| | Charles Bouveyron | Professor at Université Paris Descartes |
| *Supervisors* | Christophe Biernacki | Professor at Université Lille I |
| | Alain Celisse | Lecturer at Université Lille I |

Université de Lille 1 SCIENCES ET TECHNOLOGIES

## Université de Lille

Doctoral School **ED Régionale SPI 72**

University Department **Laboratoire de Mathématiques Paul Painlevé**

Thesis defended by **Jérémie Kellner**

Defended on **1st December, 2016**

In order to become Doctor from Université de Lille

Academic Field **Mathematics**

Speciality **Probability and Statistics**

Thesis Title

# Gaussian Models and Kernel Methods

| | | |
|---|---|---|
| **Thesis supervised by** | Christophe Biernacki | Supervisor |
| | Alain Celisse | Co-Supervisor |

**Committee members**

| | | |
|---|---|---|
| *Referees* | Gilles Blanchard | Professor at Universität Potsdam |
| | Yannick Baraud | Professor at Université de Nice Sophia-Antipolis |
| *Examiners* | Cristina Butucea | Professor at Université Paris-Est Marne-la-Vallée |
| | Vincent Rivoirard | Professor at Université Paris-Dauphine |
| | Charles Bouveyron | Professor at Université Paris Descartes |
| *Supervisors* | Christophe Biernacki | Professor at Université Lille I |
| | Alain Celisse | Lecturer at Université Lille I |

Université de Lille 1 SCIENCES ET TECHNOLOGIES

## Université de Lille

École doctorale **ED Régionale SPI 72**

Unité de recherche **Laboratoire de Mathématiques Paul Painlevé**

Thèse présentée par **Jérémie Kellner**

Soutenue le **1er décembre 2016**

En vue de l'obtention du grade de docteur de l'Université de Lille

Discipline **Mathématiques**

Spécialité **Probabilités et Statistiques**

Titre de la thèse

# Modèles Gaussiens et Méthodes à Noyaux

| | | |
|---|---|---|
| **Thèse dirigée par** | Christophe Biernacki | directeur |
| | Alain Celisse | co-directeur |

**Composition du jury**

| | | |
|---|---|---|
| *Rapporteurs* | Gilles Blanchard | professeur à l'Universität Potsdam |
| | Yannick Baraud | professeur à l'Université de Nice Sophia-Antipolis |
| *Examinateurs* | Cristina Butucea | professeure à l'Université Paris-Est Marne-la-Vallée |
| | Vincent Rivoirard | professeur à l'Université Paris-Dauphine |
| | Charles Bouveyron | professeur à l'Université Paris Descartes |
| *Directeurs de thèse* | Christophe Biernacki | professeur à l'Université Lille I |
| | Alain Celisse | mcf à l'Université Lille I |

This thesis has been prepared at

**Laboratoire de Mathématiques Paul Painlevé**
UMR 8524 CNRS - Université Lille 1
59655, Villeneuve d'Ascq Cedex
France

Web Site `http://math.univ-lille1.fr/`

Rule 34 of Statistics: if it exists, there is a kernel version of it. No exceptions.

Anonymous

When you get new rules that work, you're changing the physiology of your brain. And then your brain has to reconfigure itself in order to deal with it.

Richard D. James (Aphex Twin)

There are other worlds (they have not told you of).

Sun Ra

### Gaussian Models and Kernel Methods

**Abstract**

Kernel methods have been extensively used to transform initial datasets by mapping them into a so-called kernel space or RKHS, before applying some statistical procedure onto transformed data. In particular, this kind of approach has been explored in the literature to try and make some prescribed probabilistic model more accurate in the RKHS, for instance Gaussian mixtures for classification or mere Gaussians for outlier detection. Therefore this thesis studies the relevancy of such models in kernel spaces. In a first time, we focus on a family of parameterized kernels - Gaussian RBF kernels - and study theoretically the distribution of an embedded random variable in a corresponding RKHS. We managed to prove that most marginals of such a distribution converge weakly to a so-called "scale-mixture" of Gaussians - basically a Gaussian with a random variance - when the parameter of the kernel tends to infinity. This result is used in practice to device a new method for outlier detection. In a second time, we present a one-sample test for normality in an RKHS based on the Maximum Mean Discrepancy. In particular, our test uses a fast parametric bootstrap procedure which circumvents the need for re-estimating Gaussian parameters for each bootstrap replication.

**Keywords:** kernel methods, rkhs, normality test, outlier detection

### Modèles Gaussiens et Méthodes à Noyaux

**Résumé**

Les méthodes à noyaux ont été beaucoup utilisées pour transformer un jeu de données initial en les envoyant dans un espace dit « à noyau » ou RKHS, pour ensuite appliquer une procédure statistique sur les données transformées. En particulier, cette approche a été envisagée dans la littérature pour tenter de rendre un modèle probabiliste donné plus juste dans l'espace à noyaux, qu'il s'agisse de mélanges de gaussiennes pour faire de la classification ou d'une simple gaussienne pour de la détection d'anomalie. Ainsi, cette thèse s'intéresse à la pertinence de tels modèles probabilistes dans ces espaces à noyaux. Dans un premier temps, nous nous concentrons sur une famille de noyaux paramétrée - la famille des noyaux radiaux gaussiens - et étudions d'un point de vue théorique la distribution d'une variable aléatoire projetée vers un RKHS correspondant. Nous établissons que la plupart des marginales d'une telle distribution est asymptotiquement proche d'un « scale-mixture » de gaussiennes - autrement dit une gaussienne avec une variance aléatoire - lorsque le paramètre du noyau tend vers l'infini. Une nouvelle méthode de détection d'anomalie utilisant ce résultat théorique est introduite. Dans un second temps, nous introduisons un test d'adéquation basé sur la Maximum Mean Discrepancy pour tester des modèles gaussiens dans un RKHS. En particulier, notre test utilise une procédure de bootstrap paramétrique rapide qui permet d'éviter de ré-estimer les paramètres de la distribution gaussienne à chaque réplication bootstrap.

**Mots clés :** méthodes à noyaux, rkhs, test de normalité, détection d'anomalie

**Laboratoire de Mathématiques Paul Painlevé**
UMR 8524 CNRS - Université Lille 1 – 59655, Villeneuve d'Ascq Cedex – France

# Remerciements

Il faut être honnête deux secondes: mener une thèse de A á Z est moins motivé par l'envie d'être une fourmi ouvrière parmi la myriade d'autres fourmis ouvrières qui chacune apporte ou ont apporté leur humble contribution à cet imposant édifice qu'on appelle la science, que par le besoin d'accéder à la gloire éternelle que peut apporter le titre de docteur (tant qu'on se garde de préciser dans les soirées mondaines qu'il s'agit en fait de docteur *en mathématiques* et non en médecine). Et pourtant, malgré le fait que mon nom figure en tête de proue du présent manuscrit, je serais de mauvaise foi si je m'en attribuerais le mérite à cent pour cent. Outre les travaux antécédents menés par d'autres sur lesquelles se base une thèse, les quelques nouveautés qu'apporte cette dernière sont les résultante d'un environnement dans lequel le doctorant baigne pendant trois années, où la moindre collision fortuite d'idées peut faire éclore une direction complètement neuve. Pour cela, je me dois de rendre hommage aux acteurs de cet environnement qui ont forcément chacun à leur manière — plus ou moins directement — influencé l'issue de la présente thèse.

En premier lieu, j'aimerais naturellement remercier mon directeur de thèse Alain — sans qui il n'y aurait simplement pas eu de thèse! — pour avoir proposer le sujet de cette thèse original et d'avoir accepté de me suivre pendant ces trois dernières années. Sans lui, je serais probablement encore maintenant en train de m'enliser dans des chemins de traverse tortueux et le présent manuscrit aurait été plus compliqué que nécessaire.

Un grand merci à Gilles Blanchard et Yannick Baraud d'avoir accepté de rapporter cette thèse, ainsi qu'à Cristina Butucea, Vincent Rivoirard et Charles Bouveyron pour avoir bien voulu compléter le jury de thèse en tant qu'examinateurs.

Pendant ces trois ans, les membres de l'équipe MODAL ont constitué une sorte de seconde famille, en particulier les autres doctorants et ingénieurs avec qui j'ai partagé la fameuse salle A106, qui a vu naître une symbiose quasi-mystique entre ses occupants. Plus qu'un simple bureau, l'"esprit A106" représente une émulsion explosive d'idées folles, de celles qui nécessitent de longues décennies avant que l'*establishment* scientifique ne daigne enfin en accepter les fruits. (Et je m'abstiendrai de mentionner le package-dont-on-a-assez-prononcé-le-nom.)

Je voudrais également saluer les différentes personnes avec qui j'ai pu discuter en conférence ou en école d'été, que ce soit à Lübeck, Fréjus ou aux RJS de Bordeaux (je serais bien resté plus longtemps!).

Puisqu'il y a une vie en dehors de la sphère de la recherche, je voudrais *tip my fedora* aux */mu/tants* avec qui j'ai le plaisir de partager toutes ces séances musicales de *listenalong* dominicales. *Keep on memeing in a meme world, senpais!*

Et bien sûr, un grand merci aux fans de A Loathsome Smile qui sont nombreux à travers le monde à propager les ondes vénéneuses de la *dark electro* tellurique.

Le meilleur pour la fin, je voudrais remercier ma famille de m'avoir soutenu pendant ces trois années en ayant évité de me poser des questions quant à l'avancement de ma thèse.

Et enfin, merci à toi lecteur, d'abord d'avoir eu la patience de lire ce texte de remerciements

pompeux, mais surtout de lire la suite...

# Contents

Chapter 1

# General Introduction

## 1.1 Issues adressed by this thesis

The world of statistics is traditionally divided into two halves: non-parametric statistics and parametric statistics. Parametric statistics — as the name suggests — are based on a parametric model, that is a collection $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ of distributions that may have generated the observed dataset. There are several advantages of using a parametric model: firstly the parameter $\theta$ stands as a summary of the underlying mechanism governing the data which enables interpretation, and secondly narrowing the possible representation of this mechanism to a simple model allows a better inference on $\theta$ — which is well-known by statisticians as the bias-variance trade-off. On the other hand, an inaccurate model may lead to bad performances of the learning task at play. Besides, a model must be defined with respect to the type of observed data, that is one must design specific models for vectorial data, qualitative data, structured objects, heterogeneous data, and so on.

By contrast, non-parametric statistics do not impose such distributional assumptions on data. One family of such non-parametric procedures are *kernel methods*. Kernel methods rely on a transformation $\phi : \mathcal{X} \to \mathcal{H}$ (the *feature map*) where $\mathcal{X}$ (the *input space*) is the set where our data take values and $\mathcal{H}$ is a typically high-dimensional space also called the *feature space* or *kernel space*. The idea is to map the initial dataset into $\mathcal{H}$ through $\phi$ and to work on the transformed data lying in the kernel space instead of the initial data. The interest of such mapping is to get properties on transformed data that initial data may not satisfy. As an illustration, let us take a look at Support Vector Machines (SVM) [BGV92; CV95] which is the archetype of kernel methods. SVM is an algorithm of supervised binary classification which separates two classes of observations by means of a linear boundary. Clearly using a linear classifier directly on the initial data may be problematic as the two classes may not be linearly separable. Hence SVM aims at linearly separating the two classes in the kernel space instead of the input space, which

is a feasible task because of the high dimensionality of the kernel space. Since the introduction of SVM, many other kernel methods have been conceived throughout the last two decades and cover a wide field of applications: regression [SGV98], denoising [SSM97], hypothesis testing [Gre+07a; Gre+07b], outlier detection [Sch+01; Hof07], ...

A remarkable feature of kernel methods is that it is possible to do non-parametric statistics *and* parametric statistics at the same time, by applying a parametric statistical procedure in the kernel space. Such an approach is non-parametric in the sense that no distributional assumption is made in the input space. Instead, we consider a family of mappings and if this family is large enough, chances are that an adequately chosen mapping will "shape" the data to make it fit a given parametric model in the kernel space. So far, this approach has only been exploited by a few papers in the litterature, including [BFG15] for classification and [Rot06] for batch outlier detection.

Albeit promising, this parametric approach to kernel methods raises several questions:

- **Is a given parametric model well-defined in the kernel space?** The kernel space is often represented as a functional space (reproducing kernel Hilbert space or RKHS) which is typically infinite dimensional. Since the number of dimensions is infinite, special care must be taken to define properly a distribution in the kernel space, in particular such a distribution must be guaranteed to lie almost surely in the kernel space.
- **Is a given parametric model realistic in the kernel space?** Considering a family of feature maps, the question is whether this family is "rich" enough so that it contains a map that yields a distribution in the kernel space which is accurately described by a given parametric model. Another concern is whether such a map preserves the information carried by the initial data so that the parametric procedure in the kernel space still allows inference.
- **How to test competing parametric models in the kernel space?** Two or more parametric models may be in competition in the kernel space. To decide between those different models, a goodness-of-fit test may be used. The objective is to devise such a test that handles samples taking values in a possibly infinite-dimensional Hilbert space, such as a kernel space.

Therefore the aim of the present thesis is to bring answers to the questions above.

## 1.2 Outline of this thesis

This thesis is outlined as follows:

Chapter 2 consists in an introduction to kernel methods. The aim of the chapter is to explain the interest of kernel methods through motivating examples, to introduce essential notions such as kernels, positive definite functions and reproducing kernel Hilbert spaces, and also to provide several instances of kernels. Besides, we also mention elements that will be useful throughout

the remainder of the thesis: we introduce in details three examples of kernel methods that are Support Vector Machines (SVM) and Kernel Principal Component Analysis (kernel PCA) — that will be mentioned as rival methods in Chapter 5 — and the Maximum Mean Discrepancy (MMD) — which is essential to understand Chapter 6. Besides we also present methods for low-rank approximations of the so-called Gram matrices that are Nyström method and Random Kitchen Sinks — which will be useful in Chapter 5.

Chapter 3 is devoted to properly defining distributions in a kernel space represented as a reproducing kernel Hilbert space (RKHS). Such a distribution can be defined in two ways, either through a random element when seeing the RKHS as a vector space or through a stochastic process when seeing the RKHS as a function space. We see how these two definitions are linked to one another. We focus in particular on the definition of Gaussian distributions in an RKHS. We also cover the notion of *nuclear dominance* that provide criteria for a random element/stochastic process to be well defined in an RKHS and to take values almost surely in the RKHS. Furthermore, we check by means of nuclear dominance that in practice, Gaussian processes whose parameters are estimated on the basis of a fintie sample are well defined in an RKHS. Finally, we review three existing methods that make use of Gaussian processes in kernel spaces.

Chapter 4 focuses on a family of Gaussian RBF kernels and studies theoretically the distribution of a random variable embedded into corresponding RKHS. The distribution of such embedded variable is characterized through the distribution of its projection onto a randomly chosen low-dimensional subspace of the RKHS. We prove that when the parameter of the kernel tends to infinity, this projection converges weakly to a *scale-mixture of Gaussians* (SMG), which corresponds basically to a Gaussian distribution with a random variance. Furthermore, we extend this result to the empirical setting where the distribution embedded into the RKHS is only known through a finite sample of size $n$. In this case, we establish that the result in the non-empirical setting still holds as long as $n$ tends to infinity and the kernel parameter grows to infinity slowly enough compared to $n$. Finally, we show that with an adequate renormalization of the kernel, the previous results still hold but with an asymptotic Gaussian distribution instead of a SMG.

Chapter 5 introduces a new algorithm for outlier detection (OD) in an online setting. This algorithm relies on random projection in an RKHS induced by a Gaussian RBF kernel, and makes use of the results evidenced in Chapter 4 to get a control of the false alarm rate — that is the probability of incorrectly tagging a "normal" observation as an outlier. This new kernel-based OD method is compared with two existing kernel methods for OD that are one-class SVM and kernel PCA. It is shown that our proposed procedure manages to circumvent shortcomings of the two rival algorithms, that are a lack of false alarm rate control for one-class SVM and the lack of scalability for kernel PCA. Furthermore, we theoretically guarantee the consistency of our OD method, in other words the missed detection rate (the probability of missing an outlier) tends to 0 when the number of past observations grows to infinity.

Chapter 6 presents a goodness-of-fit test for normality dedicated to samples taking values in a general Hilbert space (which may be an RKHS). This test is based on the Maximum Mean Discrepancy (MMD) which is known to be used for homogeneity and independence testing. Our procedure allows to test a large variety of Gaussian models (zero-mean Gaussian, Gaussian with covariance of rank $r$, ...). Furthermore, a fast parametric bootstrap procedure adapted from [KY12] is successfully applied to our test and allows to reduce its computational cost. Our test displays good empirical performances and mild sensibility to high dimensions (unlike common multivariate normality tests), outperforming goodness-of-fit tests based on random projections [Cue+06]. Finally, we propose an application of our test to covariance rank selection.

Finally, Chapter 7 concludes by providing some possible future works which could stem from the results of this thesis.

Chapter 2

# Background: an Overview of Kernel Methods

This chapter presents an introduction to kernel methods, a set of methods used in the machine learning community for the past two decades and relying on the same principle — the "kernel trick". The aim of the present chapter is not to be exhaustive, but to provide a broad overview of this topic.

## 2.1 Motivating examples

Broadly speaking, kernel methods are based on a transformation called *feature map* that maps available data taking values in a set $\mathcal{X}$ — called the *input space* — into a (typically) high-dimensional space called *feature space* or *kernel space*. This introductory section aims at presenting three different cases that motivate the use of such a mapping: supervised binary classification, Fisher discriminant analysis and structured objects. Each of these examples highlights one useful aspect of kernelization, which are respectively:

- Benefit from the high-dimension nature of the feature space (Section 2.1.1),
- Applying a linear method in the kernel space to solve a non-linear problem in the input space (Section 2.1.2),
- Applying methods made for vectorial data in the kernel space to deal with non-vectorial data in the input space (Section 2.1.3).

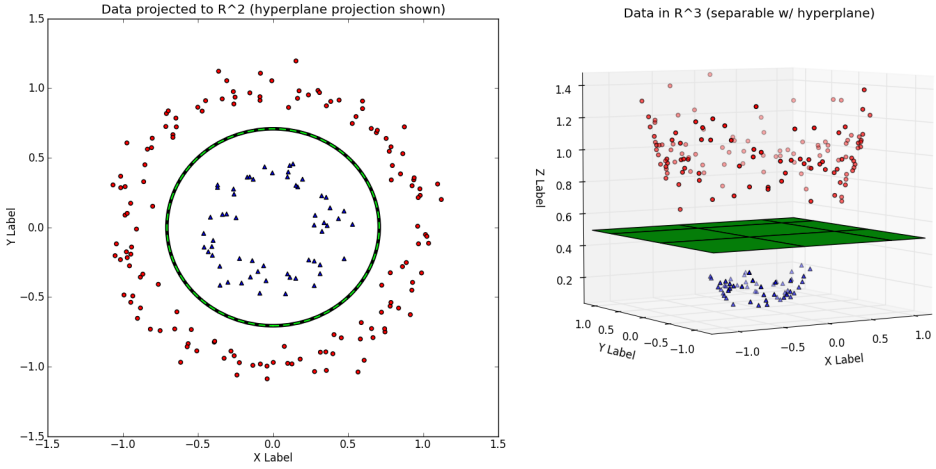


Figure 2.1 – **Left plot:** Two groups of points not linearly separable in $\mathbb{R}^2$; **right plot:** Two linearly separable classes after embedding into $\mathbb{R}^3$.

### 2.1.1 Supervised binary classification

Let us consider the problem of supervised binary classification. Assume that we are given $n$ observations $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{-1, 1\}$ where $\mathcal{X}$ is some set. Given a test point $x \in \mathcal{X}$, the goal is to guess the corresponding $y \in \{-1, 1\}$ based on the $n$ observations.

Consider for instance $\mathcal{X} = \mathbb{R}^d$ with $d \in \mathbb{N}^*$. A way of fulfilling this goal is to separate the two sets of points $\{x_i \mid y_i = -1\}$ and $\{x_i \mid y_i = 1\}$ by an hyperplane $\mathscr{H}$ of $\mathbb{R}^d$ defined by a corresponding normal vector $u \in \mathbb{R}^d$ and an offset $b \in \mathbb{R}$, that is $\mathscr{H} = \{x \in \mathbb{R}^d \mid x^T u + b = 0\}$. Therefore a new point $x \in \mathbb{R}^d$ is assigned to the class $\hat{y}$ defined by

$$\hat{y} \stackrel{\Delta}{=} \operatorname{sgn}\left(x^T u + b\right) \ ,$$

where sgn(.) denotes the sign of its argument.

The advantage of such a linear classifier is its simplicity. However, the drawback of such approach is that the existence of an hyperplane that separates the two classes is not always guaranteed. For instance, the left plot in Figure 2.1 [1] displays two classes of points in $\mathbb{R}^2$, one class (blue, triangle) being enclosed by the other one (red, circle). It is clear that there exists no straight line in $\mathbb{R}^2$ to separate these two groups of points, as a natural boundary would be instead the green circle displayed in the left plot of Figure 2.1. However, one can apply the following map to our observations:

$$\mathbb{R}^2 \to \mathbb{R}^3, \ \ (x_1, x_2) \mapsto (x_1, x_2, x_1^2 + x_2^2) \ ,$$

which yields the transformed dataset represented in three dimensions in the right plot of

[1] The plots in Figure 2.1 were taken from `http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html`

Figure 2.1. In this new setting, it is possible to find a plane (e.g. the plane of equation $z = 0.5$) in $\mathbb{R}^3$ which separates the two classes.

This toy example illustrates the usefulness of embedding a dataset into a space of higher dimension. More generally, linear classifiers are known to be powerful enough in high-dimensional settings. This is related to the notion of the Vapnik-Chernovenkis dimension (or VC-dimension) of a set of classifiers. Considering a set of classifiers $\mathscr{C} \subset \{(A, A^c) \mid A \subset \mathcal{X}\}$ as an ensemble of two-part partitions of a set $\mathcal{X}$, the VC-dimension of $\mathscr{C}$ is defined as the maximum number $h(\mathscr{C})$ of points in $\mathcal{X}$ such that, for each of the $2^{h(\mathscr{C})}$ ways of splitting these $h(\mathscr{C})$ points into two classes, there exists a classifier in $\mathscr{C}$ that classifies correctly all of the $h(\mathscr{C})$ points. In more formal terms,

$$h(\mathscr{C}) \triangleq \max\left\{ h \in \mathbb{N}^* \;\middle|\; \forall (x_1,\dots,x_h) \in \mathcal{X}^h,\ \ \forall I \subseteq \{1,2,\dots,h\},\ \ \exists (A,A^c) \in \mathscr{C},\ \ \sum_{i=1}^{h} \mathbb{1}_{x_i \in A}\, \mathbb{1}_{i \in I} \in \{0,h\} \right\} .$$

In the particular case of linear classifiers in $\mathbb{R}^d$, it is known that the corresponding VC-dimension is $d+1$ (see [Vap00], Section 3.6), that is a set of $n$ points — split into two groups – lying in a vector space of dimension $d$ with $d \geqslant n-1$ is always linearly separable.

### 2.1.2 Fisher discriminant analysis

Let $x_1,\dots,x_n \in \mathbb{R}^D$ be a sample of $n$ $D$-variate vectors split into $M$ classes, and let $y_1,\dots,y_n \in \{1,2,\dots,M\}$ be the corresponding labels, that is $y_i = j$ means that $x_i$ belongs to the class $j$ for every $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant M$. When $d$ is large, it is desirable to search for a (linear) transformation $\mathcal{T} : \mathbb{R}^D \to \mathbb{R}^d$ such that $d$ is much smaller than $D$ and $\mathcal{T}(x_1),\dots,\mathcal{T}(x_n)$ allows to visualize data (when $p \leqslant 3$) and identify the structure of the classes.

Fisher discriminant analysis (FDA) [Fis36] is a classical method of dimensionality reduction that consists in finding a direction that best separates those $M$ clusters. Therefore it corresponds to a transformation $\mathcal{T} : \mathbb{R}^D \to \mathbb{R}^d$ that is linear and such that $d = 1$. Introducing the pooled mean $\mu = (1/n)\sum_{i=1}^n x_i$ and the inner class means $\mu_j = \sum_{i=1}^n x_i \mathbb{1}_{y_i=j} \,/\, \sum_{i=1}^n \mathbb{1}_{y_i=j}$ for $1 \leqslant j \leqslant M$, the goal is to find a direction $u^* \in \mathbb{R}^D$ which maximizes the following quantity

$$u^* \in \underset{u \in \mathbb{R}^D, |u|_2 = 1}{\operatorname{argmax}} \frac{u^T \mathbf{S}_B u}{u^T \mathbf{S}_W u} \ , \tag{2.1.1}$$

where $|u|_2 = \sqrt{u^T u}$ denotes the Euclidean norm in $\mathbb{R}^D$ and

$$\mathbf{S}_B = \sum_{j=1}^{M} (\mu_j - \mu)(\mu_j - \mu)^T$$
$$\mathbf{S}_W = \sum_{j=1}^{M} \sum_{i=1}^{n} (x_i - \mu_j)(x_i - \mu_j)^T \mathbb{1}_{y_i=j} \ ,$$

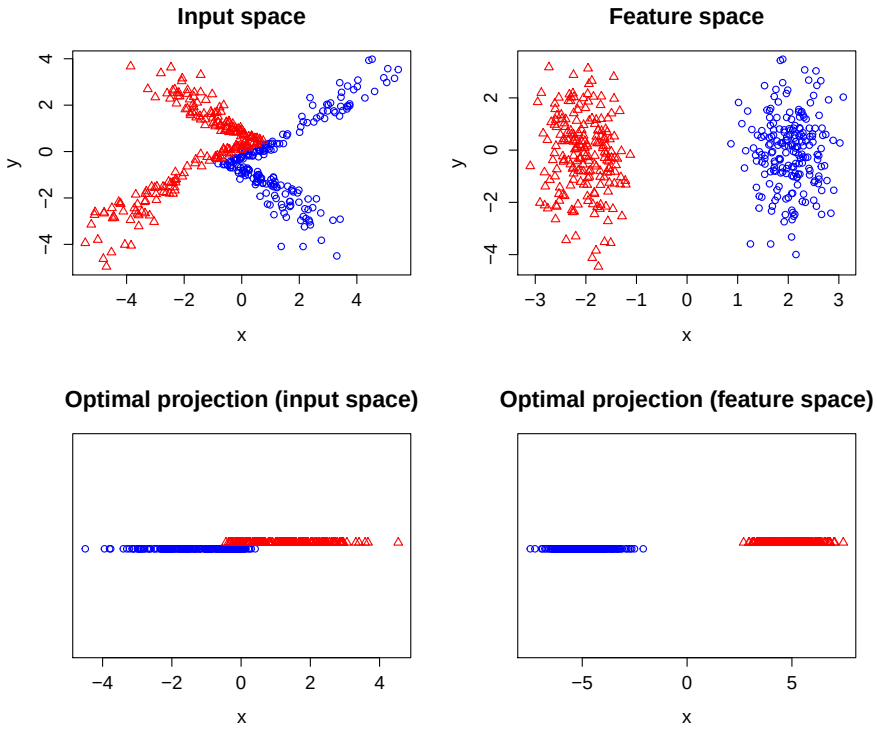


Figure 2.2 – **Top-left:** points $X_1,\dots,X_n$ split into two classes in $\mathbb{R}^2$ (red triangles and blue circles); **bottom-left:** overlapping projected points found by FDA; **top-right:** transformed points $\phi(X_1),\dots,\phi(X_n)$ in the feature space; **bottom-right:** well separated projected points in the feature space found by FDA.

denote respectively the *between class scatter matrix* and *within class scatter matrix*.

The idea is that maximizing (2.1.1) is equivalent to maximizing the numerator $u^T\mathbf{S}_B u$ and minimizing the denominator $u^T\mathbf{S}_W u$. A large $u^T\mathbf{S}_B u$ means that the within class means of the projected sample $u^T x_1,\dots,u^T x_n$ are distant from each other, while a small $u^T\mathbf{S}_W u$ implies that the within class variances of the projected sample are all small. Therefore a larger value for the objective (2.1.1) indicates a better separability between the $M$ clusters.

However, an optimal direction $u^*$ does not necessarily provide satisfactory separation for the projected sample, even though the clusters are not overlapping in $\mathbb{R}^D$ before projection. For instance, the top left plot of Figure 2.2 presents two classes of points in $\mathbb{R}^2$ (red triangles, blue circles) which lie on distinct subsets of $\mathbb{R}^2$. Despite this, the projected points onto the optimal direction found by FDA (Figure 2.2, bottom left plot) form two overlapping classes.

In the example of Figure 2.2, the use of an appropriate feature map may be helpful. For

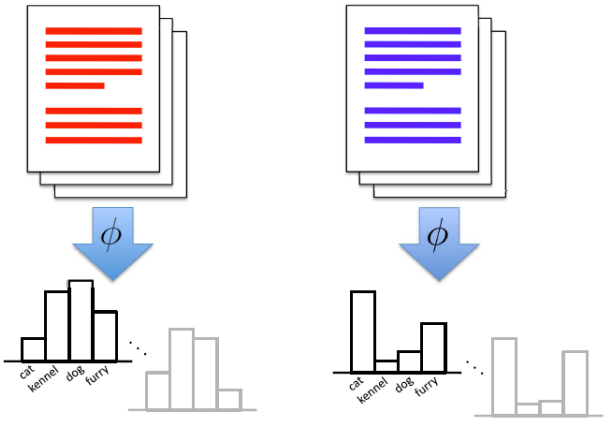


Figure 2.3 – Illustration of the *bag-of-words* approach: given a set of relevant words, transform a text into a histogram with occurrence of each of the relevant words, and compare texts by comparing corresponding histograms.

instance, if one introduces the feature map $\phi$ as follows

$$\phi : \mathbb{R}^2 \to \mathbb{R}^2, \quad (x,y) \mapsto \left( \frac{x + \text{sgn}(0.5x - y)}{|y|^{1/2}}, y + 0.5\text{sgn}(0.5x - y) \right) ,$$

then the transformed data $\phi(x_1), \ldots, \phi(x_n) \in \mathbb{R}^2$ as displayed in the top right plot of Figure 2.2 admit a linear projection that perfectly separates the classes (Figure 2.2, bottom right plot).

This example shows that through the use of a non-linear feature map $\phi$, a problem in the input space for which linear methods such as FDA fail turns to a problem in the feature space that can be solved by such methods.

### 2.1.3 Text data and *bag-of-words*

In the era of *Big Data*, data are often available in a raw form, that is observations are not always represented as multivariate vectors with real entries, but as a batch of heterogeneous types of data: vectors but also graphs, trees, categorical data, text, images, audio streams,... The main problem is that most data analysis tools are designed for a specific type data – typically multivariate data. Instead of creating *ad hoc* methods corresponding to each conceivable type of data, one may instead transform the initial dataset to get vectorial data.

We illustrate this approach with the *bag-of-words* technique [Joa02] used to represent a text as a multivariate vector. The *bag-of-words* method is simple (Figure 2.3 [2]). Given a finite set of $m$ relevant words (e.g. "dog", "cat", "economy", "kafkaesque" but not "the", "where", "but")

[2] The picture in Figure 2.3 was taken from `http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/Slides4A.pdf`

denoted by $w_1, \dots, w_m$, an observed text $x$ is mapped to a histogram $\phi(x) \in \mathbb{R}^m$ where $[\phi(x)]_i$ is the number of occurrence of the word $w_i$ in the text $x$, for every $i \in \{1, \dots, m\}$. This way, any algorithm designed for multivariate data (linear classification, principal component analysis,...) can be applied to $\phi(x)$ to treat $x$.

A straightforward way of choosing the relevant words is to take a dictionary and, based on a corpus of texts, to weed out the most observed words (*stopword elimination*) — typically words like "the" or "if" — and the least observed ones (*document frequency thresholding*).

## 2.2 Kernels: definition and basic properties

Section 2.1 showed through examples the usefulness of embedding data into a feature space $\mathcal{H}$ through a feature map $\phi$. In practice, this embedding is represented by a function called *kernel*. The objective of this section is to define formally the notion of kernel, how it is linked to positive definite functions and to illustrate the so-called *"kernel trick"*, the latter allowing to perform all practical computations only in the input space. The notion of *reproducing kernel Hilbert spaces* (RKHS) [Aro50] is covered as well, in order to provide an explicit construction of the feature map and the feature space.

### 2.2.1 Definition

In the following, we introduce the formal definition of a kernel.

**Definition 2.2.1** (Kernel)**.** *Let $\mathcal{X}$ be a set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.*
*$k$ is called a (real valued) kernel if there exist a real Hilbert space $\mathcal{H}$ endowed with an inner product $\langle .,. \rangle_{\mathcal{H}}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that:*

$$\forall x, y \in \mathcal{X}, \quad k(x,y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} \ .$$

In particular, a kernel $k$ can be defined by an expansion of the form

$$k(x,y) = \sum_{r=1}^{D} \phi_r(x)\phi_r(y) \ , \tag{2.2.2}$$

where $(\phi_r)_{1 \leqslant r \leqslant D}$ are real-valued functions defined on $\mathcal{X}$ and $D \in \mathbb{N}^* \cup \{+\infty\}$, when considering a feature map of the form

$$\phi : \mathcal{X} \to \mathbb{R}^D, \quad x \mapsto (\phi_1(x) \dots \phi_D(x))^T \ .$$

A kernel $k$ defined by an expansion as in (2.2.2) is called a *Mercer kernel*. Furthermore, *Mercer's theorem* [Mer09] states that conversely, a kernel $k$ admit the representation (2.2.2) under the assumption that $\mathcal{X}$ is a compact topological space and $k$ is continuous. However this result still

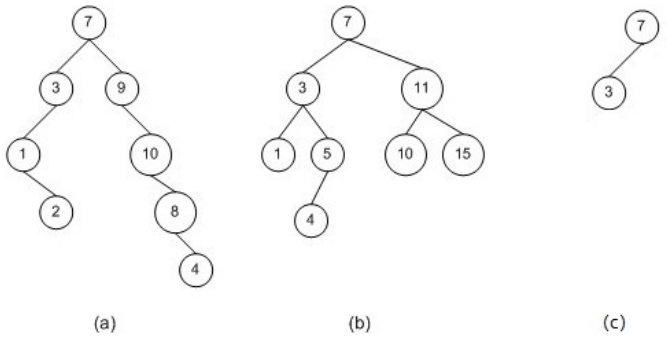


Figure 2.4 – Trees (a) and (b) are compared by counting the number of their common subtrees. For instance, Tree (c) is a common subtree of (a) and (b).

holds true if the compactness assumption for $\mathcal{X}$ is removed [SS12].

In Examples 2.2.1 to 2.2.3 thereafter, we provide examples of kernels defined through an implicit feature map as in Definition 2.2.1.

**Example 2.2.1.** (Linear kernel)
The most simple kernel on $\mathcal{X} = \mathbb{R}^d$ ($d \geqslant 1$) is the canonical scalar product

$$\forall x, y \in \mathbb{R}^d, \;\; k(x,y) = y^T x \;\; .$$

The corresponding feature map/space are

$$\mathcal{H} = \mathcal{X} = \mathbb{R}^d, \;\; \phi = \mathrm{Id}_{\mathbb{R}^d} \;\; .$$

**Example 2.2.2.** (Tree kernel)
In this example, we will define a tree kernel as proposed in [CD01]. To compare two given trees $T_1$ and $T_2$, we build a kernel following the idea that two trees are similar if they share many common subtrees. Let $\{S_i\}_{1 \leq i \leq M}$ be the set of all the $M$ possible subtrees.

Consider the feature map: $h(T) = (h_1(T), \ldots, h_M(T))$ where $h_i(T)$ counts the number of occurrences of $S_i$ in $T$.

The tree kernel is defined by the canonical dot product in $\mathbb{R}^M$:

$$k(T_1, T_2) = \sum_{i=1}^{M} h_i(T_1) h_i(T_2) \;\; .$$

Remark that $M$ is exponentially much larger than the number of nodes in $T_1$ and $T_2$. To be able to compute $k$, the tree kernel is written in another form. Let $I_i(n)$ the indicator equal to 1 if $S_i$ is rooted at the node $n$ and 0 otherwise. Then $h_i(T) = \sum_{n \in T} I_i(n)$ and $K$ can be formulated in a

different way

$$k(T_1, T_2) = \sum_{i=1}^{M} \left( \sum_{n_1 \in T_1} I_i(n_1) \right) \left( \sum_{n_2 \in T_2} I_i(n_2) \right)$$
$$= \sum_{n_1 \in T_1} \sum_{n_2 \in T_2} \underbrace{\sum_{i=1}^{M} I_i(n_1) I_i(n_2)}_{=C(n_1, n_2)} \ .$$

On the first hand, $k$ becomes a sum over the pairs of nodes $(n_1, n_2)$, and on the other hand, $C(n_1, n_2)$ can be computed quickly with a recursive procedure. Therefore the computation of the tree kernel does not depend on $M$.

**Example 2.2.3.** (Fisher kernel)
An *a priori* about the behaviour of the dataset can be expressed by a generative model. Such a model involves a family of distributions $\mathbb{P}(.|\theta)$ on $\mathcal{X}$ parameterized by $\theta \in \Omega \subseteq \mathbb{R}^D$ where $\Omega$ is the parameter space. Haussler and Jaakkola (1999, [JDH99]) introduced a type of kernels called *Fisher kernels* which take into account such a model.

Fisher kernels are built upon the feature map

$$\phi_{\theta_0} : \mathcal{X} \to \mathbb{R}^D, \ \ \phi_{\theta_0}(x) = \left( \frac{\partial \log \mathbb{P}(x|\theta)}{\partial \theta_i}(\theta_0) \right)_{1 \leq i \leq D} \ ,$$

where $\theta_0$ is a given point in $\Omega$. $\phi_{\theta_0}(x)$ represents how to modify the current setting (with parameter set at $\theta_0$) to maximize $\log(\mathbb{P}(x|\theta))$ and make the model fit $x$ better.

Two points $x$ and $y$ are similar if they "draw" the model towards similar directions, thus a Fisher kernel is defined as an inner product in $\mathbb{R}^D$ between $\phi_{\theta_0}(x)$ and $\phi_{\theta_0}(y)$. Namely, let $I$ be the Fisher information matrix

$$I = \left[ \mathbb{E}_X \phi_{\theta_0}(X)_i \phi_{\theta_0}(X)_j \right]_{1 \leq i,j \leq D} \ ,$$

where $X$ follows the distribution $\mathbb{P}(.|\theta_0)$. For any $x, y \in \mathcal{X}$, the Fisher kernel $k$ related to the model $\{\mathbb{P}(.|\theta)\}_{\theta \in \Omega}$ and a given parameter value $\theta_0$ is defined by

$$k(x, y) = \phi_{\theta_0}(x)^T I^{-1} \phi_{\theta_0}(y) \ .$$

Note that given a kernel $k$, the corresponding pair of feature map/space $(\phi, \mathcal{H})$ is not unique. For instance, consider the linear kernel of Example 2.2.1, $k(x, y) = x^T y$ for every $x, y \in \mathbb{R}^d$ $(d \geqslant 1)$. Example 2.2.1 defined $\phi = \mathrm{Id}_{\mathbb{R}^d}$ and $\mathcal{H} = \mathbb{R}^d$ as the corresponding feature map/space. However,

one could have also considered the following feature space/map $\tilde{\mathcal{H}}$ and $\tilde{\phi}$ as

$$\tilde{\mathcal{H}} = \mathbb{R}^{2d}, \qquad \tilde{\phi} : \mathbb{R}^d \to \mathbb{R}^{2d}, \ x \mapsto \left( \frac{1}{\sqrt{2}} x, \frac{1}{\sqrt{2}} x \right) ,$$

where for any $u, v \in \mathbb{R}^d$, $(u, v) \in \mathbb{R}^{2d}$ denotes the vector obtained by concatenating $u$ and $v$. It is then trivial to check that

$$\forall x, y \in \mathcal{X}, \ k(x, y) = \phi(x)^T \phi(y) = \tilde{\phi}(x)^T \tilde{\phi}(y) .$$

### 2.2.2 Positive definite functions and reproducing kernel Hilbert spaces

The kernels shown in the examples of Section 2.2.1 were constructed by explicitly defining a feature map $\phi$ and a feature space $\mathcal{H}$. However the function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is sometimes defined in the first place without any preliminary definition of $\phi$ and $\mathcal{H}$. In fact, $k(x, y)$ is often understood as a measure of similarity between $x, y \in \mathcal{X}$ and is directly constructed following this rationale. For instance if $\mathcal{X}$ is endowed with a distance function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, the function $k$ can be written as

$$\forall x, y \in \mathbb{R}^d, \ k(x, y) = \kappa(d(x, y)) ,$$

for some function $\kappa : \mathbb{R}^+ \to \mathbb{R}$ with $\kappa(0) = \sup_{t \in \mathbb{R}} \kappa(t)$.

This raises the following question: how to tell whether a given $k$ is a kernel or not? It turns out that answering this question does not require to explicitly build a feature space and a feature map and is related to the notion of *symmetric, positive definite functions*.

**Definition 2.2.2** (Symmetric, positive definite function)**.** *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if for every $a_1, \ldots, a_n \in \mathbb{R}$ and $x_1, \ldots, x_n \in \mathcal{X}$:*

$$\sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) \geq 0 .$$

*Moreover if $k(x, y) = k(y, x)$ for every $x, y \in \mathcal{X}$, $k$ is symmetric.*

*Remark: Often in the literature, the expression* positive definite function *is replaced by* positive semi-definite function. *With the additional condition that $\sum_{i,j} a_i a_j k(x_i, x_j) = 0$ if and only if $a_1 = \ldots = a_n = 0$, "positive definite" becomes "strictly positive definite" in the first naming, and "positive semi-definite" becomes "positive definite" in the second one, which may be confusing. In the following, we stick to the first terminology.*

The interest of this notion is that $k$ is a symmetric, positive definite function if and only if $k$ is a kernel, as we show in the following.

Straightforwardly, a kernel $k$ is necessarily symmetric and definite positive. Given a corresponding feature space $\mathcal{H}$ — equipped with the inner product $\langle .,. \rangle_{\mathcal{H}}$ and associated norm

$\|.\|_{\mathcal{H}}$ with $\|h\|_{\mathcal{H}} = \sqrt{\langle h,h\rangle_{\mathcal{H}}}$ for every $h \in \mathcal{H}$ — and feature map $\phi$, then for any $a_1,\dots,a_n \in \mathbb{R}$ and $x_1,\dots,x_n \in \mathcal{X}$

$$\sum_{i,j=1}^{n} a_i a_j k(x_i,x_j) = \sum_{i,j=1}^{n} a_i a_j \langle \phi(x_i),\phi(x_j)\rangle_{\mathcal{H}} = \left\| \sum_{i=1}^{n} a_i \phi(x_i) \right\|_{\mathcal{H}}^2 \geq 0 \ ,$$

and for every $x,y \in \mathcal{X}$

$$k(x,y) = \langle \phi(x),\phi(y)\rangle_{\mathcal{H}} = \langle \phi(y),\phi(x)\rangle_{\mathcal{H}} = k(y,x) \ ,$$

by symmetry of the inner product of $\mathcal{H}$. Therefore, a kernel $k$ is always positive definite and symmetric.

Showing the converse requires to build a canonical feature map $\phi$ and feature space $\mathcal{H}$. The construction of $\phi$ and $\mathcal{H}$ thereafter follows the proof of Theorem 4.16 from [SC08]. Given a symmetric, positive definite function $k$, consider the following pre-hilbertian space of functions $\mathcal{H}_0$ consisting of linear combinations of evaluations functions $k(x,.),\ \ x \in \mathcal{X}$

$$\mathcal{H}_0 = \left\{ \sum_{i=1}^{n} a_i k(x_i,.) \ \middle|\ n \in \mathbb{N}^*,\ \ a_1,\dots,a_n \in \mathbb{R},\ \ x_1,\dots,x_n \in \mathcal{X} \right\} \ .$$

which is equipped with the inner product $\langle .,. \rangle_{\mathcal{H}_0}$ defined for any $h = \sum_{i=1}^{n} a_i k(x_i,.) \in \mathcal{H}_0$ and $h' = \sum_{j=1}^{m} a'_j k(x'_j,.) \in \mathcal{H}_0$ as

$$\langle h,h'\rangle_{\mathcal{H}_0} = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i a'_j k(x_i,x'_j) \ .$$

Let us check that $\langle .,. \rangle_{\mathcal{H}_0}$ actually defines an inner product.

First of all, $\langle .,. \rangle_{\mathcal{H}_0}$ is well defined since $\langle h,h'\rangle_{\mathcal{H}_0}$ can be written $\langle h,h'\rangle = \sum_{j=1}^{m} a'_j h(x'_j) = \sum_{i=1}^{m} a_i h'(x_i)$, which shows that $\langle h,h'\rangle_{\mathcal{H}_0}$ does not depend on the representation of $h$ and $h'$.

The positivity condition is fulfilled by the positive definiteness of $k$, since for every $h = \sum_{i=1}^{n} a_i k(x_i,.) \in \mathcal{H}_0$

$$\langle h,h\rangle_{\mathcal{H}_0} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i,x_j) \geqslant 0 \ .$$

It is straightforward that $\langle .,. \rangle_{\mathcal{H}_0}$ is symmetric and bilinear. Therefore it satisfies the Cauchy-Schwarz's inequality $\langle h,h'\rangle_{\mathcal{H}_0} \leqslant \sqrt{\langle h,h\rangle_{\mathcal{H}_0}\langle h',h'\rangle_{\mathcal{H}_0}}$ which implies that $\langle \cdot,\cdot \rangle_{\mathcal{H}_0}$ is definite as follows: for any $h \in \mathcal{H}_0$ such that $\langle h,h\rangle_{\mathcal{H}_0} = 0$,

$$\forall x \in \mathcal{X},\ \ |h(x)| = |\langle h,k(x,\cdot)\rangle_{\mathcal{H}_0}| \leqslant |\langle h,h\rangle_{\mathcal{H}_0}||k(x,x)| = 0 \ ,$$

hence $h = 0$.

Now we are in a position to define a feature space $\mathcal{H}$ as the completion of $\mathcal{H}_0$ with respect to $\langle .,. \rangle_{\mathcal{H}_0}$. $\mathcal{H}$ is therefore a Hilbert space whose inner product $\langle .,. \rangle_{\mathcal{H}}$ satisfies

$$\forall h, h' \in \mathcal{H}_0, \ \langle h, h' \rangle_{\mathcal{H}} = \langle h, h' \rangle_{\mathcal{H}_0} \ .$$

In particular, this implies the so-called *reproducing property*

$$\forall x, y \in \mathcal{X}, \ \langle k(x,.), k(y,.) \rangle_{\mathcal{H}} = k(x,y) \ .$$

The reproducing property shows that choosing $\mathcal{H}$ as a feature space and

$$\phi : \mathcal{X} \to \mathcal{H}, \ x \mapsto k(x,.) \ ,$$

as a feature map yields

$$\forall x, y \in \mathcal{X}, \ k(x,y) = \langle k(x,.), k(y,.) \rangle_{\mathcal{H}} = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} \ ,$$

proving that $k$ is a kernel.

$\mathcal{H}$ is called the *reproducing kernel Hilbert space* (RKHS) of $k$ [Aro50] and is denoted $H(k)$. Furthermore, every kernel $k$ is related to a unique RKHS ([SC08], Theorem 4.21).

**Definition 2.2.3** (Reproducing kernel Hilbert space)**.** *For any symmetric, positive definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists a unique Hilbert space $H(k)$ equipped with an inner product $\langle .,. \rangle_{H(k)}$ which satisfies:*

- $\forall x \in \mathcal{X}, \ k(x,.) \in H(k)$ ,
- $\forall f \in H(k), \ \forall x \in \mathcal{X}, \ \langle f, k(x,.) \rangle_{H(k)} = f(x)$ .

*Such an $H(k)$ is called* reproducing kernel Hilbert space *with reproducing kernel $k$.*

Apart from providing a canonical form for the feature space and the feature map, $H(k)$ has the remarkable property of being the simplest" possible feature space corresponding to $k$, that is $H(k)$ is the smallest possible feature space in the following sense: Theorem 4.21 from [SC08] states that given any feature space/map pair $(\tilde{\mathcal{H}}, \tilde{\phi})$ related to $k$, the following operator

$$\tilde{\mathcal{H}} \to H(k), \ w \mapsto \langle w, \tilde{\phi}(.) \rangle_{\tilde{\mathcal{H}}} \ ,$$

is surjective. As Section 2.1.2 suggested, one of the motivations for the use of a kernel is to apply a linear algorithm in a feature space such as $\tilde{\mathcal{H}}$. Such an algorithm deals with inner product such as $\langle w, \tilde{\phi}(x) \rangle_{\tilde{\mathcal{H}}}$ where $w \in \tilde{\mathcal{H}}$ and $x \in \mathcal{X}$, which shows that viewing the feature space as an RKHS suffices.

Finally, we illustrate this section by introducing a family of symmetric, positive definite functions (hence kernels): *translation-invariant kernels*.

**Example 2.2.4.** (Translation-invariant kernels)
Let us set $\mathcal{X} = \mathbb{R}^d$ for some $d \geqslant 1$ and define $k$ as

$$\forall x, y \in \mathbb{R}^d, \;\; k(x,y) \overset{\Delta}{=} K(x-y) \;\; ,$$

where $K : \mathbb{R}^d \to \mathbb{R}$.

The function $K$ is said to be positive definite if the corresponding kernel $k$ is positive definite. Furthermore, assuming that $K$ is continuous, Bochner's theorem states that $K$ is positive definite if and only if $K$ is the Fourier transform of a bounded positive measure on $\mathbb{R}^d$ ([BT04], Theorem 20), that is

$$K(z) = \int e^{iw^T z} \Lambda(\mathrm{d}w) \;\; ,$$

for some bounded positive measure $\Lambda$ on $\mathbb{R}^d$.

Therefore, the family of translation-invariant kernels encompasses several instances of commonly used kernels on $\mathbb{R}^d$ such as:

- **Gaussian kernel:** $k(x,y) = \exp(-\gamma \|x-y\|^2), \;\; \gamma > 0 \;\; .$

  This kind of kernel corresponds to $K(z) = \exp(-\gamma\|z\|^2)$ which is positive definite since $K$ is the characteristic function of a $\mathcal{N}(0, 2\gamma I_d)$ Gaussian distribution.

- **Laplace kernel:** $k(x,y) = \exp\left(-\gamma\|x-y\|\right), \;\; \gamma > 0 \;\; .$

  Here $K(z) = \exp(-\gamma\|z\|)$ corresponds to the characteristic function of a random vector $rU$, where $U$ is uniform on the unit sphere of $\mathbb{R}^d$ and $r$ is independent of $U$ and follows a Cauchy distribution with density function $f(t) = \left[\pi\gamma\left(1 + (t/\gamma)^2\right)\right]^{-1}$.

- **Cauchy kernel:** The Cauchy kernel is a parametric kernel (with parameter $\sigma > 0$) with formula

  $$k(x,y) = \frac{1}{1 + \sigma^{-2}\|x-y\|^2} \;\; ,$$

  for every $x, y \in \mathbb{R}^d$ Compared to the Gaussian and Laplace kernels, the Cauchy kernel emphasized more possible influences between distant points [Bas08].

- **B-spline kernels:** Define $B_0 = \mathbb{1}_{\mathcal{B}_{1,d}}$ the indicator function of the unit ball $\mathcal{B}_{1,d}$ of $\mathbb{R}^d$. For every function $f, g : \mathbb{R}^d \to \mathbb{R}$, let $f \otimes g$ denote the convolution of $f$ and $g$ that is $(f \otimes g)(x) = \int_{\mathbb{R}^d} f(x')g(x'-x)\mathrm{d}x'$. Then define iteratively each function $B_i : \mathbb{R}^d \to \mathbb{R}$ by $B_{i+1} = B_i \otimes B_0$ for each $i \geqslant 0$. The function $B_{2p+1}$ where $p \geqslant 0$ are positive definite and

therefore the kernel $k_p$

$$k_p(x,y) = B_{2p+1}(x-y) \quad \text{for all } x,y \in \mathbb{R}^d \ ,$$

defines a translation-invariant kernel [V+97].

Some translation-invariant kernels such as the Gaussian kernel, the Laplace kernel and the Cauchy kernel admit actually in the more specific form $k(x,y) = \kappa(\|x-y\|)$ where $\kappa : \mathbb{R}^+ \to \mathbb{R}$. These particular kernels are called *radial kernels* or *radial basis function kernels* (RBF kernels).

### 2.2.3 The "kernel trick"

As shown in Section 2.2.1, a kernel $k$ is designed to match the inner product of some feature space. The feature space is typically high-dimensional — a property that is desired within some frameworks as the example of Section 2.1.1. However, manipulating high-dimensional vectors may be more of a curse than a blessing — especially in regard of computational costs.

In the following, we expose how using a kernel allows to eventually perform computations in the input space, hence avoiding the computational burden of high dimension. This practical usefulness of kernels is due to two aspects: the *"kernel trick"* and the *representer theorem*.

#### "Kernel trick"

Here is a simple illustration of the "kernel trick". Given $x$ and $y$ two elements in the input space and a kernel $k$, the distance between the embedded points $\phi(x)$ and $\phi(y)$ in $H(k)$ can be calculated as follows

$$\begin{aligned}\|\phi(x)-\phi(y)\|^2_{H(k)} &= \langle \phi(x)-\phi(y), \phi(x)-\phi(y)\rangle_{H(k)} \\ &= \langle \phi(x),\phi(x)\rangle_{H(k)} + \langle \phi(y),\phi(y)\rangle_{H(k)} - 2\langle \phi(x),\phi(y)\rangle_{H(k)} \\ &= k(x,x) + k(y,y) - 2k(x,y) \ .\end{aligned}$$

Therefore computing a distance in the RKHS only requires to evaluate the kernel $k$ on elements of the input space. There is no need to manipulate directly any vectors in the feature space: this is the principle of the "kernel trick".

More generally, given a set of observations $x_1,\dots,x_n \in \mathcal{X}$, if computations in the RKHS only consist of pairwise inner products $\langle \phi(x_i),\phi(x_j)\rangle_{H(k)}$ for $1 \leqslant i,j \leqslant n$, then calculating the *Gram matrix* $K \in \mathcal{M}_n(\mathbb{R})$ defined by

$$K = \Big[k(x_i,x_j)\Big]_{1\leqslant i,j\leqslant n} \ ,$$

suffices. It turns out that most learning algorithms only require the Gram matrix, as implied by the *representer theorem* stated thereafter.

**Representer theorem**

Most learning algorithms can be cast as an optimization problem. Given observations $(x_1, y_1), \dots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$, the goal is to seek out a prediction function $f^*$ such that $Y_i = f^*(x_i) + \epsilon_i$ for every $1 \leqslant i \leqslant n$ where $\epsilon_1, \dots, \epsilon_n$ are error terms. Such a function is determined by minimizing (or maximizing) a functional of the general form

$$\Omega(f) = c\left((x_1, y_1, f(x_1)), \dots, (x_n, y_n, f(x_n))\right) + \mathcal{R}(f) \ , \tag{2.2.3}$$

where $f \in \mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ with $\mathcal{F}$ a prescribed set of functions, $c : (\mathcal{X} \times \mathbb{R}^2)^m \to \mathbb{R}^+$ is a loss function and $\mathcal{R}(f)$ is a regularizer term which typically controls the "smoothness" of $f$ in some sense to avoid overfitting. For instance in the case of ridge regression, an optimal regression function $f^* : \mathbb{R}^d \to \mathbb{R}$ satisfies

$$f^* \in \underset{f(x)=a^T x,\ a \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} (a^T x_i - y_i)^2 + \lambda \|a\|^2 \ ,$$

where $\lambda > 0$. Here $\mathcal{F} = \{x \mapsto a^T x \mid a \in \mathbb{R}^d\}$ is the set of linear regression functions, the loss function $c$ is the squared loss $\sum_{i=1}^{n} (a^T x_i - y_i)^2$ and $\mathcal{R}(f) = \lambda \|a\|^2$.

The *representer theorem* provides an explicit form for the optimizers of functionals as (2.2.3) when $\mathcal{F}$ is an RKHS. Originally introduced in [KW71] in the special case of the squared loss, it has been generalized by [SHS01]. The latter version is stated in Proposition 2.2.4.

**Proposition 2.2.4** (Representer theorem, [SHS01])**.** *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel where $\mathcal{X}$ is a non-empty set and let $(x_1, y_1), \dots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$.*
*Assume that the regularizer $\mathcal{R}(f)$ in (2.2.3) is a strictly increasing (resp. decreasing) function of $\|f\|_{H(k)}$.*
*Then, any $f^* \in H(k)$ that minimizes (resp. maximizes) $\Omega(f^*)$ lies in the span of $k(x_1, .), \dots, k(x_n, .)$, that is*

$$\exists a_1, \dots, a_n \in \mathbb{R}, \quad f^* = \sum_{i=1}^{n} a_i k(x_i, .) \ . \tag{2.2.4}$$

*Proof of Proposition 2.2.4.* We focus on the case where $\Omega(f)$ is strictly increasing w.r.t. $\|f\|$, the other case being proved likewise.

Let $V = \operatorname{span}(k(x_1, .), \dots, k(x_n, .))$, and $f \in H(k)$. Let us split $f$ into the sum of two orthogonal functions:

$$f = \underbrace{P_V f}_{\in V} + \underbrace{f^{\perp}}_{\in V^{\perp}} \ ,$$

which is possible since $V$ is finite dimensional and $H(k)$ is a Hilbert space (hence complete).

Using the reproducing property and the fact that $f^{\perp}$ is orthogonal to $V$ and $k(x_i, .) \in V$ for

every $1 \leqslant i \leqslant n$,

$$\forall 1 \leqslant i \leqslant n, \;\; f(x_i) = P_V f(x_i) + f^{\perp}(x_i) = P_V f(x_i) \;\; ,$$

since $f^{\perp}(x_i) = \langle f^{\perp}, k(x_i, .)\rangle_{H(k)} = 0$.

It follows

$$\begin{aligned}\Omega(f) &= c((x_1, y_1, f(x_1)), \ldots, (x_n, y_n, f(x_n))) + \mathcal{R}(f) \\ &= c((x_1, y_1, P_V f(x_1)), \ldots, (x_n, y_n, P_V f(x_n))) + \mathcal{R}(f) \;\; .\end{aligned}$$

Finally, since $P_V f$ and $f^T$ are orthogonal

$$\|f\|^2 = \|P_V f\|^2 + \|f^{\perp}\|^2 \geq \|P_V f\|^2 \;\; .$$

Therefore since $\mathcal{R}$ is strictly increasing with respect to $\|f\|_{H(k)}$,

$$\begin{aligned}\Omega(f) &= c((x_1, y_1, P_V f(x_1)), \ldots, (x_n, y_n, P_V f(x_n))) + \mathcal{R}(f) \\ &> c((x_1, y_1, P_V f(x_1)), \ldots, (x_n, y_n, P_V f(x_n))) + \mathcal{R}(P_V f) = \Omega(P_V f) \;\; .\end{aligned}$$

It proves that any minimizer of $\Omega(f)$ necessarily belongs to $V$. □

Thus determining $f^*$ reduces to determining the scalars $a_1, \ldots, a_n \in \mathbb{R}$. We can therefore reformulate the problem in terms of matrices and vectors (called the *dual problem*) which is more convenient for resolution. Furthermore, the dual writing of the optimization problem does not involve the number of dimensions of the feature space but only the number of observations $n$.

## 2.3 Some examples of kernel methods

Now that we have properly introduced the definition of a kernel and some of their key properties, we are in a position to review some of the most encountered statistical and/or learning methods that make use of kernels. Many of them turn out to be generalized version of pre-existing procedures: *Kernel Principal Component Analysis* [SSM97] which is a non-linear extension of PCA, *Kernel Ridge Regression* [SGV98], *Kernel Fisher discriminant analysis* [Mik+99], *Kernel Canonical Correlation Analysis* [LF00], among others.

Due to the vast variety of kernel methods, this section has not the ambition to present an exhaustive taxonomy of such methods. We restrict ourselves to three examples that will be mentioned or useful throughout this thesis: *Support Vector Machine* for classification, *Kernel PCA* for dimension reduction and *Maximum Mean Discrepancy* for two-sample and independence testing.

### 2.3.1 Support Vector Machine

Introduced by [BGV92] and then [CV95], *Support Vector Machine* (SVM) is a learning algorithm designed for binary classification. We recall the binary classification framework — that we have already encountered in Section 2.1.1: we are given $n$ pairs of observations $(x_1, y_1), \dots, (x_n, y_n) \in \mathcal{X} \times \{-1, 1\}$ where $\mathcal{X}$ is an arbitrary set. The objective is to guess the class $y \in \{-1, 1\}$ corresponding to a new point $x \in \mathcal{X}$ based on $(x_i, y_i)_{1 \leqslant i \leqslant n}$.

The main idea of SVM is to construct a linear decision function that separates the two classes and to determine $y$ according to the side of the decision hyperplane where the tested point $x$ lies. Furthermore, this decision boundary is not set in the input space $\mathcal{X}$ but in a high dimensional feature space $\mathcal{H}$ where $x_1, \dots, x_n$ are embedded into, in order to make the two classes linearly separable. More precisely, assuming that $\mathcal{X}$ is endowed with a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with corresponding feature space/map $(\mathcal{H}, \phi)$, the SVM algorithm aims at finding a vector $f \in \mathcal{H}$ that solves the following constrained optimization scheme

$$\begin{array}{lll} \text{minimize} & \frac{1}{2}\|f\|_{\mathcal{H}}^2 + C\sum_{i=1}^n \xi_i & \text{over } f \in \mathcal{H}, b \in \mathbb{R}, \xi_1, \dots, \xi_n \in \mathbb{R}^+ \\ \text{such that} & y_i(\langle f, \phi(x_i)\rangle_{\mathcal{H}} + b) \geqslant 1 - \xi_i & \text{for all } 1 \leqslant i \leqslant n \ , \end{array} \tag{2.3.5}$$

where $C > 0$.

Let us denote optimal separating hyperplane as $\{h \in \mathcal{H} \mid \langle f^*, h\rangle_{\mathcal{H}} + b^* = 0\}$ where $(f^*, b^*)$ is a solution of (2.3.5). Thus a tested point $x \in \mathcal{X}$ will be assigned the class $y \in \{-1, 1\}$ defined by

$$y = \operatorname{sgn}(\langle f^*, \phi(x)\rangle_{\mathcal{H}} + b^*) \ .$$

Let us discuss the inequality constraints $y_i(\langle f, \phi(x_i)\rangle_{\mathcal{H}} + b) \geqslant 1 - \xi_i$ in (2.3.5). In a first time, we omit the variables $\xi_1, \dots, \xi_n$ and focus on the simplified optimization scheme

$$\begin{array}{lll} \text{minimize} & \frac{1}{2}\|w\|_{\mathcal{H}}^2 & \text{over } f \in \mathcal{H}, \ b \in \mathbb{R} \\ \text{such that} & y_i(\langle f, \phi(x_i)\rangle_{\mathcal{H}} + b) \geqslant 1 & \text{for all } 1 \leqslant i \leqslant n \ . \end{array} \tag{2.3.6}$$

The effect of optimizing (2.3.6) is twofold. On the first hand, the constraints $y_i(\langle f, \phi(x_i)\rangle_{\mathcal{H}} + b) \geqslant 1$ require that $y_i(\langle f, \phi(x_i)\rangle_{\mathcal{H}} + b)$ is positive for every $1 \leqslant i \leqslant n$, which means that every training point $\phi(x_1), \dots, \phi(x_n)$ is correctly classified by the separating hyperplane. On the other hand, minimizing $\|f\|_{\mathcal{H}}^2$ makes the distance between the training points and the decision hyperplane as large as possible. To see this, rewrite (2.3.6) as follows

$$\begin{array}{lll} \text{minimize} & N & \text{over } N \in \mathbb{R}_+, \ \tilde{b} \in \mathbb{R}_+, \ \|\tilde{f}\|_{\mathcal{H}} = 1 \\ \text{such that} & y_i(\langle \tilde{f}, \phi(x_i)\rangle_{\mathcal{H}} + \tilde{b}) \geqslant \frac{1}{N} & \text{for all } 1 \leqslant i \leqslant n \ . \end{array}$$

where $\tilde{f} = f/\|f\|_{H(k)}^2$, $\tilde{b} = b/\|f\|_{H(k)}^2$ and $N = \|f\|_{H(k)}^2$.

The drawback of the optimal decision boundary that stems from (2.3.6) is that it tolerates no classification error among the training points. It may be a problem if one of the training points is

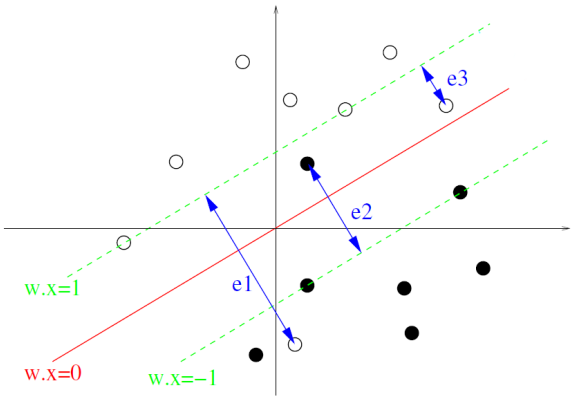


Figure 2.5 – Illustration of Support Vector Machine

an outlier. This lack of robustness may lead to a well-known problem in learning theory called *overfitting*, that is the learning algorithm sticks too much to the training set and causes a large *generalization error* — a tested point is more likely to be misclassified.

To tackle this overfitting issue, [CV95] introduced the slack variables $\xi_1,\dots,\xi_n$ in (2.3.5). These slack variables allow a few training points to be exceptionally close to the separating hyperplane or even to be misclassified. This more robust approach leads to a smaller generalization error.

Now that the objective of the SVM is clearly set, we show that solving (2.3.5) is feasible in practice, despite the fact that the optimization is done over a possibly infinite dimensional space $\mathcal{H}$. In particular, the representer theorem can be applied in this framework so that the optimization problem does not depend on the dimension of $\mathcal{H}$. Namely, (2.3.5) is equivalent to

$$\text{minimize} \quad \tfrac{1}{2}\|f\|_{\mathcal{H}}^2 + C\sum_{i=1}^n \max(0, 1 - y_i(\langle f, \phi(x_i)\rangle_{\mathcal{H}} + b)) \quad \text{over } f \in \mathcal{H},\ b \in \mathbb{R}\ . \tag{2.3.7}$$

If the feature space $\mathcal{H}$ is chosen as the RKHS $H(k)$, the reproducing property entails $\langle f, \phi(x_i)\rangle_{H(k)} = f(x_i)$ for every $1 \leqslant i \leqslant n$ and (2.3.7) becomes

$$\text{minimize} \quad \tfrac{1}{2}\|f\|_{\mathcal{H}}^2 + C\sum_{i=1}^n \max(0, 1 - y_i(f(x_i) + b)) \quad \text{over } f \in \mathcal{H},\ b \in \mathbb{R}\ ,$$

which is of the same form as the objective (2.2.3) in Section 2.2.3 with $c(x_i, y_i, f(x_i)) = C\max(0, 1 - y_i(f(x_i) + b))$ as the loss function — in this case, the *hinge loss* — and $\mathcal{R}(f) = (1/2)\|f\|_{H(k)}^2$. Hence the representer theorem states that the optimal $f^*$ admits the following expansion

$$f^* = \sum_{i=1}^n a_i k(x_i, .)\ .$$

Introducing the vector $\mathbf{a} = (a_1 \dots a_n)^T \in \mathbb{R}^n$ and the Gram matrix $\mathbf{K} = [k(x_i, x_j)]_{1 leq si, j \leqslant n}$, (2.3.5) finally becomes

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\mathbf{a}^T\mathbf{K}\mathbf{a} + C\sum_{i=1}^n \xi_i \quad \text{over } f \in \mathcal{H},\ b \in \mathbb{R},\ \xi_1, \dots, \xi_n \in \mathbb{R}_+ \\ \text{such that} & y_i(\mathbf{K}\mathbf{a} + b) \geqslant 1 - \xi_i \quad \text{for all } 1 \leqslant i \leqslant n\ , \end{array}$$

which depends on $n$ but not on the dimension of $H(k)$.

### 2.3.2 Kernel PCA

Before introducing kernel principal component analysis (kernel PCA), let us recall the principle of linear PCA which is a classical tool in statistics for dimension reduction.

In the classical framework, we are given $n$ *i.i.d.* observations $\mathbf{X} = [X_1 \dots X_n]$ in $\mathbb{R}^m$ which are assumed zero-mean. The goal is to find the directions $u_1, \dots, u_d$ that capture the major part of the variance of the observations. To do so, we first have to find the unit vector $u_1$ that maximizes the variance $u_1^T \Sigma u_1$, where $\Sigma = \mathbb{E}_X XX^T$ is the covariance matrix of $X$. The associated Laplacian to be maximized is

$$\mathcal{L}(u_1, \lambda) = u_1^T \Sigma u_1 - \lambda(u_1^T u_1 - 1)\ ,$$

which is equivalent to solving the eigenvalue equation

$$\Sigma u_1 = \lambda_1 u_1\ , \tag{2.3.8}$$

where $\lambda_1$ the largest eigenvalues of $\Sigma$. The following vector $u_2$ is sought in the same way but such that it is orthogonal to $u_1$, and so on. Therefore, the principal components $u_1, \dots, u_d$ correspond to the first $d$ eigenvectors of $\Sigma$.

In practice, we consider the sample covariance estimator $\hat{\Sigma} = (1/n)\sum_{i=1}^n X_i X_i^T$, so that the equation (2.3.8) becomes

$$\frac{1}{n}\mathbf{X}\mathbf{X}^T u_1 = \lambda_1 u_1\ .$$

In [SSM97], the authors have proposed a kernelized version of the PCA, and applied it successfully to image denoising. Given a kernel $k$, kernel PCA applies usual PCA in the feature space $\mathcal{H}$. For the sake of clarity, kernel PCA is presented in the case where $\mathcal{H}$ is a finite dimensional linear space but it can be extended as well to infinite dimensional settings by means of Hilbert-Schmidt operators algebra.

Assume $\mathcal{H} = \mathbb{R}^D$ where $D$ is typically very large. Writing the feature map as $\phi(x) = (\phi_1(x), \dots, \phi_D(x))^T \in \mathbb{R}^D$, the images of our observations in $\mathcal{H}$ are stored in the matrix $\mathbf{\Phi} \in$

$\mathcal{M}_{D,n}(\mathbb{R})$

$$\mathbf{\Phi} = [\phi(X_1), \dots, \phi(X_n)] \ .$$

As in classical PCA, we have to solve the following eigenvalue equation

$$\frac{1}{n}\mathbf{\Phi}\mathbf{\Phi}^T u_1 = \lambda_1 u_1 \ . \tag{2.3.9}$$

We can not solve it directly, since the matrix $\mathbf{\Phi}\mathbf{\Phi}^T \in \mathcal{M}_D(\mathbb{R})$ is too large. Instead we transform (2.3.9) and multiply it from the left by $\mathbf{\Phi}^T$

$$\frac{1}{n}\underbrace{\mathbf{\Phi}^T\mathbf{\Phi}}_{=\mathbf{K}}\underbrace{\mathbf{\Phi}^T u_1}_{=v_1 \in \mathbb{R}^n} = \lambda \underbrace{\mathbf{\Phi}^T u_1}_{=v_1} \ ,$$

where $\mathbf{K} = [\langle \phi(X_i), \phi(X_j)\rangle_{\mathcal{H}}]_{1\leqslant i,j\leqslant n} = [k(X_i, X_j)]_{1\leqslant i,j\leqslant n}$ is the Gram matrix. Therefore, it becomes an eigenvalue problem in $\mathbb{R}^n$

$$\mathbf{K}v_1 = n\lambda_1 v_1 \ .$$

Having found a wanted $v_1$, we can express the eigenvector $u_1$ as follows

$$\begin{aligned} \mathbf{\Phi}^T u_1 = v_1 \ &\Leftrightarrow \ \underbrace{\mathbf{\Phi}\mathbf{\Phi}^T u_1}_{=n\lambda_1 u_1} = \mathbf{\Phi} v_1 \\ &\Leftrightarrow \ u_1 = \frac{1}{n\lambda_1}\mathbf{\Phi} v_1 \ . \end{aligned}$$

Finally we normalize $u_1$ by $\|u_1\|_{\mathcal{H}}$ to obtain $\tilde{u}_1 \in \mathbb{R}^D$. Since

$$\|u\|_{\mathcal{H}}^2 = \frac{1}{n^2\lambda_1^2} v_1^T \mathbf{\Phi}^T \mathbf{\Phi} v_1 = \frac{1}{n^2\lambda_1^2} v_1^T \mathbf{K} v_1 = \frac{1}{n\lambda_1} \ ,$$

we get

$$\tilde{u}_1 = \frac{1}{\sqrt{n\lambda_1}}\mathbf{\Phi} v_1 \ .$$

Projecting a new embedded point $\phi(X)$ with $X \in \mathcal{X}$ onto the normalized principal component $\tilde{u}_1$ yields

$$\phi(X)^T \tilde{u}_1 = \frac{1}{\sqrt{n\lambda_1}}\phi(X)^T \mathbf{\Phi} v_1 \ ,$$

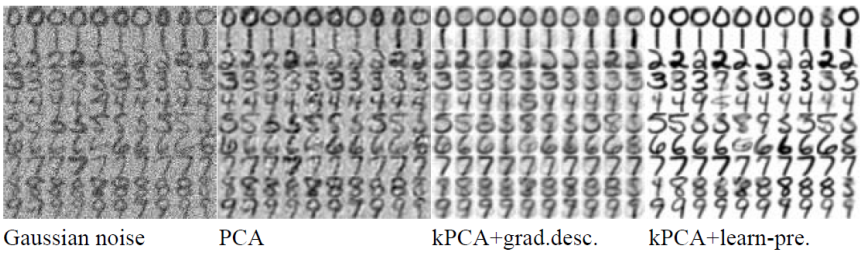


Figure 2.6 – Application of kernel PCA to image de-noising. **From left to right:** original noisy image; de-noised image with linear PCA; de-noised image with kernel PCA and gradient descent to learn the pre-image; de-noised image with kernel PCA and regularized regression to learn the pre-image.

with

$$\phi(X)^T \mathbf{\Phi} = (k(X_1, X), \dots, k(X_n, X)) \ .$$

Note that every actual computation has been made all along in the input space through $k$, which is an instance of the kernel trick. In the seminal paper of [SSM97], an application of kernel PCA consists in "cleaning out" noisy data by projecting an embedded point $\phi(X) \in \mathcal{H}$ onto principal components in the feature space to obtain $P\phi(X) \in \mathcal{H}$, then by going back into the input space. The latter step is not trivial since most vector in $\mathcal{H}$ do not admit a pre-image through the feature map $\phi$. In particular, there exists in general no element $X^* \in \mathcal{X}$ satisfying $\phi(X^*) = P\phi(X)$. Instead, $X^*$ is determined as a solution of the following optimization problem

$$X^* = \underset{x \in \mathcal{X}}{\operatorname{argmin}} \ \|\phi(x) - P\phi(X)\|_{\mathcal{H}}^2 \ ,$$

which is solved by gradient descent.

Another way to define $X^*$ is to consider the problem of going back to the input space as a regression problem [BWS04]. The regularization in the regression procedure allows to get a "smoother" approximate pre-image, which is shown to be efficient in practice for image de-noising (see Figure 2.6 taken from [BWS04]).

Other instances of applications of kernel PCA include regularized binary classification [Zwa05] and outlier detection [Hof07].

### 2.3.3 Maximum Mean Discrepancy

Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$ be two samples of *i.i.d.* $\mathcal{X}$-valued random variables, where $(X_i)_{1 \leqslant i \leqslant n}$ were drawn from a distribution $P_X$ and $(Y_i)_{1 \leqslant i \leqslant m}$ from a distribution $P_Q$, where $P_X$ and $P_Q$ are both defined on the measurable space $(\mathcal{X}, \mathcal{A})$. The purpose of a two-sample test is to test the null-hypothesis $\mathbf{H}_0 : P_X = P_Y$ against the alternative hypothesis $\mathbf{H}_1 : P_X \neq P_Y$. In order to set such

a test, one needs a test statistic that measures the discrepancy between $P_X$ and $P_Y$ based on the two available samples. This test statistic is typically an estimator of some quantity $\Delta(P_X, P_Y)$ that quantifies the gap between the unknown distributions $P_X$ and $P_Y$.

A series of papers [Bor+06; Gre+07a; Gre+09] have proposed to use kernels for this task. Their method rely on an *Hilbert space embedding of distributions*.

**Definition 2.3.1** (Hilbert space embedding of distributions, Lemma 3 in [Gre+12a])**.** *Let $\mathcal{M}_1^+(\mathcal{X})$ be the set of probability distributions on $\mathcal{X}$ and $k$ a (measurable) kernel on $\mathcal{X}$ that satisfies $\sup_{x\in\mathcal{X}} k(x,x) < +\infty$.*
*Then the following mapping $\mu$ is well-defined*

$$\mu : \mathcal{M}_1^+(\mathcal{X}) \to H(k), \quad P \mapsto \mu[P] \ ,$$

*where for each $P \in \mathcal{M}_1^+(\mathcal{X})$, $\mu[P] \in H(k)$ is the element of $H(k)$ that satisfies*

$$\forall f \in H(k), \quad \langle \mu[P], f \rangle_{H(k)} = \mathbb{E}_{X\sim P}[\langle k(X,\cdot), f \rangle_{H(k)}] \ .$$

*For every $P \in \mathcal{M}_1^+(\mathcal{X})$, $\mu[P]$ is called the Hilbert space embedding of $P$ in $H(k)$.*

This can be seen as a generalization of the notion of moment-generating function — *i.e.* a function of the type $\varphi(w) = \mathbb{E}_X e^{w^T X}$ related to some random vector $X$. It is clear that a moment-generating function corresponds to a Hilbert space embedding with an exponential kernel $k(x,y) = \exp(x^T y)$.

Through the embedding $\mu$, each distribution $P$ on $\mathcal{X}$ is represented by a mean element $\mu[P]$ in $H(k)$. This mapping is injective for a class of kernels called *characteristic kernels*.

**Definition 2.3.2** (Characteristic kernel)**.** *Let $k$ a kernel on $\mathcal{X}$ and $\mu$ the corresponding Hilbert space embedding of distributions.*
*$k$ is said to be characteristic if and only if $\mu$ is injective, that is*

$$\forall P, Q \in \mathcal{M}_1^+(\mathcal{X}), \quad P = Q \iff \mu[P] = \mu[Q] \ .$$

Criteria for a kernel to be characteristic have been widely investigated in [Fuk+09; Sri+10; CS10; SFL11]. Examples of characteristic kernels include commonly used kernels such as the Gaussian kernel and the Laplacian kernel on $\mathbb{R}^d$.

Through the Hilbert space embedding of distribution $\mu$, the discrepancy between $P_X$ and $P_Y$ can be gauged by the following quantity

$$\mathrm{MMD}(P_X, P_Y) \stackrel{\Delta}{=} \left\| \mu[P_X] - \mu[P_Y] \right\|_{H(k)}^2 \ ,$$

that is equal to 0 if and only if $P_X = P_Y$ whenever $k$ is a characteristic kernel. $\mathrm{MMD}(P_X, P_Y)$ is called the *maximum mean discrepancy* (MMD) between $P_X$ and $P_Y$.

To set a two-sample test, $\mathrm{MMD}(P_X, P_Y)$ is estimated based on $(X_i)_{1\leqslant i\leqslant n}$ and $(Y_i)_{1\leqslant i\leqslant m}$. Introducing $\hat{P}_X = (1/n)\sum_{i=1}^n \delta_{X_i}$ (resp. $\hat{P}_Y = (1/n)\sum_{i=1}^m \delta_{Y_i}$) the empirical measure of the sample $(X_i)_{1\leqslant i\leqslant n}$ (resp. $(Y_i)_{1\leqslant i\leqslant m}$) leads to the following test statistic

$$\widehat{\mathrm{MMD}}(P_X, P_Y) \triangleq \mathrm{MMD}(\hat{P}_X, \hat{P}_Y) = \left\| \frac{1}{n}\sum_{i=1}^n k(X_i,\cdot) - \frac{1}{m}\sum_{j=1}^m k(Y_j,\cdot) \right\|^2_{H(k)} .$$

Expanding the squared norm and the kernel trick lead to a fully computable statistic

$$\widehat{\mathrm{MMD}}(P_X, P_Y) = \frac{1}{n^2}\sum_{i,j=1}^n k(X_i, X_j) + \frac{1}{m^2}\sum_{i,j=1}^m k(Y_i, Y_j) - \frac{2}{mn}\sum_{i,j=1}^{n,m} k(X_i, Y_j) .$$

The same rationale can be applied to design an independence test. Let $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X}\times\mathcal{Y}$ be $n$ independent copies of an $\mathcal{X}\times\mathcal{Y}$-valued random variable $(X, Y)$ following the joint distribution $P_{XY}$. Let $P_X$ and $P_Y$ denote respectively the marginal distributions of $X$ and $Y$. An independence test aims at testing the null hypothesis $\mathbf{H}_0$ that $X$ and $Y$ are independent against the alternative $\mathbf{H}_1$ that they are dependent. By definition, $X$ and $Y$ are independent if and only if the joint distribution $P_{XY}$ and the product distribution $P_X \otimes P_Y$ coincide [3].

Therefore the independence testing framework is reformulated as follows

$$\mathbf{H}_0 : P_{XY} = P_X \otimes P_Y \quad \text{versus} \quad \mathbf{H}_1 : P_{XY} \neq P_X \otimes P_Y ,$$

and an adequate test statistic is an the MMD between the empirical measures of $P_{XY}$ and $P_X \otimes P_Y$ [Gre+07b]. This statistic is called the *Hilbert space independence criterion* (HSIC). Namely, given a kernel $K$ defined on $\mathcal{X}\times\mathcal{Y}$, the HSIC writes

$$\begin{aligned}
\mathrm{HSIC}(X, Y) &= \widehat{\mathrm{MMD}}(P_{XY}, P_X \otimes P_Y) \\
&= \left\| \frac{1}{n}\sum_{i,j=1}^n K\left((x_i, y_i), (.,.)\right) - \frac{1}{n^2}\sum_{i,j=1}^n K\Big((x_i, y_j), (.,.)\Big) \right\|^2_{H(K)} \\
&= \frac{1}{n^2}\sum_{i,j=1}^n K\Big((x_i, y_i), (x_j, y_j)\Big) + \frac{1}{n^4}\sum_{i,j,k,l=1}^n K\Big((x_i, y_j), (x_k, y_l)\Big) \\
&\qquad - \frac{2}{n^3}\sum_{i,j,k=1}^n K\Big((x_i, y_i), (x_j, y_k)\Big) .
\end{aligned}$$

In [Gre+07b], kernels $K$ of the form $K((x, y), (x', y') = k(x, x')l(y, y')$ are considered where $k$ is a

[3] $P_X \otimes P_Y$ is defined as the measure on $\mathcal{X}\times\mathcal{Y}$ such that for every measurable $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$, $P_X \otimes P_Y(A \times B) = P_X(A)P_Y(B)$.

kernel on $\mathcal{X}$ and $l$ is a kernel on $\mathcal{Y}$ which leads to the following expression of the HSIC

$$\mathrm{HSIC}(X,Y) = \frac{1}{n^2}\mathrm{Tr}\left(\mathbf{KHLH}\right) \ ,$$

where Tr(.) denotes the trace of a matrix, $\mathbf{K} = [k(X_i,X_j)]_{1\leqslant i,j,n}$, $\mathbf{L} = [l(Y_i,Y_j)]_{1\leqslant i,j,n}$ and $\mathbf{H} = \mathrm{I}_n - (1/n)\mathbb{1}_n\mathbb{1}_n^T$ is a centering matrix. In this case, [Gre15] shows that if $k$ and $l$ are both characteristic, then the HSIC is consistent — that is $\mathrm{HSIC}(X,Y) \to 0$ almost surely as $n \to +\infty$ under $\mathbf{H}_1$. This is a useful result since one does not need to check whether the product kernel $K$ is characteristic.

## 2.4 Efficient computation of the kernel matrix

Through the kernel trick, we are able to run algorithm in some high dimensional feature space while performing actual calculations in the input space through evaluations of the kernel $k$ on pairs of observations $X_1,\dots,X_n \in \mathcal{X}$ summarized by the Gram matrix (or kernel matrix)

$$\mathbf{K} = \left[k(X_i,X_j)\right]_{1\leqslant i,j\leqslant n} \ ,$$

so that computational costs are not affected by the number of dimensions of the feature space but only on the sample size $n$. However, kernel methods may still be computationally expensive when $n$ is large. In particular, the obtainment of $\mathbf{K}$ requires $n^2$ entries to compute and store, while algorithms in linear time with respect to $n$ are more desirable nowadays.

This section is devoted to presenting methods used to tackle this problem by approximating $\mathbf{K}$ with low-rank representations, thus setting the computation/memory costs down to the order of $\mathcal{O}(n)$. More exactly, we are interested in factorizations of the Gram matrix as follows

$$\mathbf{K} = \mathbf{U}\mathbf{V}^T \ ,$$

such that $\mathbf{U},\mathbf{V} \in \mathcal{M}_{n,r}(\mathbb{R})$ and $r$ is much smaller than $n$. Such approaches include the *Nyström method* [WS01] and *Random Kitchen Sinks* with its variation *Fastfood* [RR07; LSS13].

Note that the goal is not to find the best low-rank factorization of $\mathbf{K}$ but to find a satisfactory low-rank representation of the Gram matrix at a reduced computational cost, typically in linear time. Actually the best $r$-rank approximation $\mathbf{K}_r$ of $\mathbf{K}$ can be defined as the matrix $\mathbf{K}_r$ of rank $r < n$ that minimizes $\|\mathbf{K} - \mathbf{K_r}\|_{\mathcal{F}}$ where $\|A\|_{\mathcal{F}}^2 = \mathrm{Tr}(AA^T)$ defines the Froebenius norm for every $A \in \mathcal{M}_n(\mathbb{R})$. The expression of $\mathbf{K_r}$ depends on the eigenexpansion of the Gram matrix since it reads

$$\mathbf{K_r} = \mathbf{U_r}\mathbf{D_r}\mathbf{U_r} \ ,$$

where $\mathbf{D_r} = \mathrm{diag}(d_1,\dots,d_r)$ contains the $r$ largest eigenvalues of $\mathbf{K}$ on its diagonal and where the

columns of $\mathbf{U_r}$ are the eigenvectors of $\mathbf{K}$ (with unit norm) corresponding to $d_1, \dots, d_r$. However obtaining $\mathbf{K_r}$ requires to compute — at least partially — the eigenexpansion of $\mathbf{K}$ which costs a too demanding execution time of order $\mathcal{O}(n^2 r)$.

### 2.4.1 Nyström method

The Nyström method consists in subsampling the columns of $\mathbf{K}$ to build a low-rank approximation of $\mathbf{K}$. Namely, consider a subset $\mathcal{I} \subseteq \{1, \dots, n\}$ of cardinality $r < n$ and define the approximate Gram matrix $\tilde{\mathbf{K}}$ as

$$\tilde{\mathbf{K}} \stackrel{\Delta}{=} \mathbf{K}_{\cdot,\mathcal{I}} \mathbf{K}_{\mathcal{I},\mathcal{I}}^{\dagger} \mathbf{K}_{\mathcal{I},\cdot} \ , \tag{2.4.10}$$

where $\mathbf{K}_{\mathcal{I},\mathcal{I}} = [k(X_i, X_j)]_{i,j \in \mathcal{I}}$, $\mathbf{K}_{\mathcal{I},\cdot} = [k(X_i, X_j)]_{i \in \mathcal{I}, 1 \leqslant j \leqslant n}$, $\mathbf{K}_{\cdot,\mathcal{I}} = \mathbf{K}_{\mathcal{I},\cdot}^T$ and for any matrix $A \in \mathcal{M}_r(\mathbb{R})$, $A^{\dagger}$ denotes the pseudo-inverse [4] of $A$.

The computation of (2.4.10) is in linear time with respect to $n$ since $\mathbf{K}_{\mathcal{I},\cdot}$ takes $nr$ entries to calculate, $\mathbf{K}_{\mathcal{I},\mathcal{I}}$ has $r^2$ entries to calculate and its inversion costs $r^3$ operations hence a total of $nr + r^2 + r^3$ operations, which is of order $\mathcal{O}(n)$ if $r$ is negligible compared to $n$.

Note that $\tilde{\mathbf{K}}_{\mathcal{I},\cdot} = \mathbf{K}_{\mathcal{I},\mathcal{I}} \mathbf{K}_{\mathcal{I},\mathcal{I}}^{\dagger} \mathbf{K}_{\mathcal{I},\cdot} = \mathbf{K}_{\mathcal{I},\cdot}$, that is the rows (and likewise the columns) of $\mathbf{K}$ and $\tilde{\mathbf{K}}$ indexed by $\mathcal{I}$ coincide. This is due to the fact that the Nyström method amounts to project the embedded points $k(X_1,.), \dots, k(X_n,.)$ onto the subspace $V$ of $H(k)$ spanned by the subsample $\{k(X_i,.)\}_{i \in \mathcal{I}}$. To see this, let us introduce the orthogonal projector $\Pi : H(k) \to H(k)$ projecting onto $V$ and $\mathbf{L} = \langle \Pi k(X_i,.), \Pi k(X_j,.) \rangle_{H(k)}$ the Gram matrix corresponding to projected points in the RKHS. It follows for every $1 \leqslant i, j \leqslant n$ with $i \in \mathcal{I}$,

$$\begin{aligned} \mathbf{L}_{i,j} &= \langle \Pi k(X_i,.), \Pi k(X_j,.) \rangle_{H(k)} = \langle \Pi^* \Pi k(X_i,.), k(X_j,.) \rangle_{H(k)} \\ &= \langle \Pi k(X_i,.), k(X_j,.) \rangle_{H(k)} = \langle k(X_i,.), k(X_j,.) \rangle_{H(k)} \\ &= \mathbf{K}_{i,j} \ , \end{aligned}$$

where $\Pi^* = \Pi$ is the adjoint operator of $\Pi$. Furthermore, there exists only one matrix $\tilde{\mathbf{K}}$ of rank $r < n$ satisfying $\tilde{\mathbf{K}}_{\mathcal{I},\cdot} = \mathbf{K}_{\mathcal{I},\cdot}$ ([BJ05], Proposition 1), hence $\tilde{\mathbf{K}} = \mathbf{L}$.

There exist several strategies for the choice of the subset $\mathcal{I}$. In [WS01], $\mathcal{I}$ is picked at random uniformly from the set of all subsets of $\{1, \dots, n\}$ of cardinality $p$. [SS00] uses *sparse greedy approximation* that consists in selecting the indexes in $\mathcal{I}$ one after the other — namely at each iteration, the selected index optimizes some criterion. However, to make this greedy selection efficient, the new index is chosen at each iteration from a randomly picked subset of the non-chosen indexes instead of trying all the remaining indexes. The proposal of [DM05] is close to that of [WS01] except that they assign different probabilities $p_i = \mathbf{K}_{ii} / \sum_{i=1}^n \mathbf{K}_{ii}$ to each $i \in \{1, \dots, n\}$

[4] In general, the pseudo-inverse $M^{\dagger}$ of a matrix $M \in \mathcal{M}_{n,m}(\mathbb{R})$ is defined by $M^{\dagger} = \lim_{\delta \to 0} (M^T M + \delta I_m)^{-1} M^T = \lim_{\delta \to 0} M (M M^T + \delta I_n)^{-1}$.

and construct $\mathcal{I}$ by picking from $\{1,\dots,n\}$ $p$ subsequent times with replacement according to these probabilities.

It remains to determine the minimal value of $r$ that makes the approximation of $\mathbf{K}$ satisfactory. This problem is studied in [Bac13] in the case of a uniform sampling of $\mathcal{I}$. [Bac13] points out that the quality of approximation of the Gram matrix should not be assessed by the matrix error $\|\mathbf{K}-\tilde{\mathbf{K}}\|_{\mathcal{F}}$ but by the performance of the statistical task at stake when using the approximate Gram matrix. In particular, [Bac13] focuses on the case of regression problems — such as SVM — and shows that the generalization error obtained with $\tilde{\mathbf{K}}$ is of the same order as that with $\mathbf{K}$ if $p$ is larger than a certain quantity that depends on the eigenspectrum decay of $\mathbf{K}$ and the regression function.

### 2.4.2 Random Fourier features

Another way to approximate $k$ is to build a low-rank kernel $\tilde{k}$ defined explicitly through a feature map $\tilde{\phi}:\mathcal{X}\to\mathbb{R}^p$ where the number of features $p$ is small. *Random Kitchen Sinks* [RR07] proposes such an approach for translation-invariant kernels $k:\mathbb{R}^d\times\mathbb{R}^d\to\mathbb{R}$. We recall the definition of translation-invariant kernels mentioned previously in Section **??**. A kernel $k$ is translation-invariant if there exists a function $K:\mathbb{R}^d\to\mathbb{R}$ such that

$$\forall x,y\in\mathbb{R}^d,\ \ k(x,y)=K(x-y)\ .$$

Bochner's theorem states that $K$ is the Fourier transform of a positive measure $\Lambda$ on $\mathbb{R}^d$, that is

$$\forall z\in\mathbb{R}^d,\ \ K(z)=\int_{\mathbb{R}^d} e^{iz^Tw}\Lambda(\mathrm{d}w)\ . \tag{2.4.11}$$

Note that since $K$ is real-valued, (2.4.11) can be written in the following alternate way

$$\forall z\in\mathbb{R}^d,\ \ K(z)=\mathrm{Re}\left(\int_{\mathbb{R}^d} e^{iz^Tw}\Lambda(\mathrm{d}w)\right)=\int_{\mathbb{R}^d}\mathrm{Re}(e^{iz^Tw})\Lambda(\mathrm{d}w)=\int_{\mathbb{R}^d}\cos(z^Tw)\Lambda(\mathrm{d}w)\ , \tag{2.4.12}$$

where Re(.) denotes the real part of a complex number.

Let us assume that $K$ is properly normalized so that $\Lambda$ is a probability measure — that is $\int_{\mathbb{R}^d}\Lambda(\mathrm{d}w)=1$. The basic idea of Random Kitchen Sinks is to randomly draw a few points $w_1,\dots,w_p\in\mathbb{R}^d$ from the distribution $\Lambda$ and consider a Monte-Carlo estimation of the integral in (2.4.12). Namely, $k$ is approximated by a low-rank kernel $\tilde{k}$ as follows

$$\begin{aligned}\forall x,y\in\mathbb{R},\ \ k(x,y)=K(x-y)&=\int_{\mathbb{R}^d}\cos(\{x-y\}^Tw)\Lambda(\mathrm{d}w)\\&\simeq\frac{1}{r}\sum_{i=1}^{r}\cos(\{x-y\}^Tw_i)\end{aligned}$$

$$= \frac{1}{r}\sum_{i=1}^{r}\Big\{\cos(x^T w_i)\cos(y^T w_i) + \sin(x^T w_i)\sin(y^T w_i)\Big\}$$
$$= \tilde{\phi}(x)^T\tilde{\phi}(y) = \tilde{k}(x,y) \ ,$$

where

$$\forall x \in \mathbb{R}^d, \ \ \tilde{\phi}(x) = r^{-1/2}\begin{pmatrix} \cos(x^T w_1) \\ \dots \\ \cos(x^T w_r) \\ \sin(x^T w_1) \\ \dots \\ \sin(x^T w_r) \end{pmatrix} .$$

All in all, computing $\tilde{\phi}(x)$ for a single $x \in \mathcal{X}$ mostly involves the calculation of a matrix $W = [w_1 \dots w_r]^T \in \mathcal{M}_{2r,d}(\mathbb{R})$ and of the matrix product $Wx$, which amounts to a computational complexity of order $\mathcal{O}(dr)$. Repeating this computation for every $x = X_i$ with $1 \leqslant i \leqslant n$ leads to a running time of order $\mathcal{O}(ndr)$.

In the case of radial kernels where $K(z)$ is a function of $\|z\|$, the *fastfood* method proposed by [LSS13] improves the running time of Random Kitchen Sinks with respect to $d$ so that the complexity changes from linear to logarithmic dependence on $d$. Assuming in a first time that $r = d$, *Fastfood* bypasses the direct sampling of $W$ and the computation of $Wx$ by replacing $W$ with the matrix $\tilde{W}$

$$\tilde{W} \overset{\Delta}{=} \sqrt{\frac{2\gamma}{d}} S H_d G \Pi H_d B \in \mathcal{M}_d(\mathbb{R}) \ , \tag{2.4.13}$$

where $S, G, B \in \mathcal{M}_d(\mathbb{R})$ are diagonal matrices, $\Pi \in \mathcal{M}_d(\mathbb{R})$ is a permutation matrix and $H_d \in \mathcal{M}_d(\mathbb{R})$ is the so-called *Walsh-Hadamard matrix* defined recursively by

$$H_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \text{ and for every } l \in \mathbb{N}^* H_{2l} = H_2 \otimes H_l \ ,$$

where $\otimes$ stands for the Kronecker matrix product. The diagonal entries of $B$ are *i.i.d.* Rademacher variables [5] , those of $G$ are *i.i.d.* $\mathcal{N}(0,1)$ Gaussians and those of $S$ are *i.i.d.* variables whose distribution depends on the kernel $k$.

The iterative definition of $H_d$ allows to compute any matrix product $H_d v$ where $v \in \mathbb{R}^d$ recursively to get a $\mathcal{O}(d\log(d))$ computational cost instead of $\mathcal{O}(d^2)$. This way computing $\tilde{W}x$ costs $\mathcal{O}(d\log(d))$ in time in the case $r = d$. If $r > d$ is chosen as a multiple of $d$, $\tilde{W}$ can be written as the result of stacking $r/d$ square matrices $\tilde{W}_1, \dots, \tilde{W}_{r/d}$ of size $d$ sampled as in (2.4.13) which finally leads to a complexity of order $\mathcal{O}(r\log(d))$.

[5] A Rademacher random variable $\xi$ is a $\{-1,1\}$-valued r.v. such that $\mathbb{P}(\xi = 1) = \mathbb{P}(\xi = -1) = 1/2$.

To understand how $\tilde{W}$ emulates $W$, one must see that the rows of $H_d G \Pi H_d B$ follow a $\mathcal{N}(0, I_d)$ Gaussian distribution and have the same norm equal to $\|G\|_{\mathcal{F}}^2 d$. In other words the rows of $(1/\|G\|_{\mathcal{F}}^2 d) H_d G \Pi H_d B$ are uniformly distributed on the unit sphere $\mathcal{S}^{d-1}$ of $\mathbb{R}^d$. Since $k(\cdot,\cdot) = K(\cdot - \cdot)$ is assumed to be a radial kernel, $K$ is the Fourier transform of a spherical distribution $\Lambda$, that is the distribution of a random variable of the form $rU$ where $U$ is uniform on $\mathcal{S}^{d-1}$ and $r$ is a positive random variable independent of $U$. Therefore setting the diagonal entries of $S$ as *i.i.d.* copies of $r/(\|G\|_{\mathcal{F}}^2 d)$ implies that the rows of $\tilde{W}$ follow the distribution $\Lambda$. In the end, the actual difference between $\tilde{W}$ and $W$ is that the rows of $\tilde{W}$ are not independent. Despite this dependency, [LSS13] proved that the approximate kernel $\tilde{k}$ converges pointwise to $k$ almost surely as $r$ tends to infinity.

Chapter 3

# Gaussian Models in Reproducing Kernel Hilbert Spaces

This chapter is devoted to defining probability distributions in an RKHS, especially Gaussian distributions. Conditions that guarantee that such distributions lie almost surely in a given RKHS will be of interest. In particular in the case of a Gaussian process parameterized by mean and covariance functions, we establish that such a process is well defined in the empirical setting where those parameters are estimated on the basis of a finite sample. In a second time, we present a few existing methods using Gaussian models in kernel spaces that can be found in the literature.

## 3.1 Gaussian distributions in RKHS

Let $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be a symmetric, positive definite function defined on a set $\mathcal{X}$ and let $H(k)$ denote the corresponding RKHS. $H(k)$ can be seen either as a function space —whose elements are seen as real functions — or as a mere Hilbert space — whose elements are seen as vectors. Each of these two points of view corresponds to a different way of defining a probability distribution on an RKHS: either through a *stochastic process* or through a *random element*.

**Definition 3.1.1** (Stochastic process)**.** *Let $\mathcal{X}$ be a non-empty set and $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space. A stochastic process $Y$ is a collection $Y = \{Y(x)\}_{x\in\mathcal{X}}$ of real-valued random variables defined on $(\Omega, \mathcal{A}, \mathbb{P})$.*

**Definition 3.1.2** (Random element)**.** *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(\mathcal{H}, \mathcal{A}')$ a measurable space. A random element $Z$ in $(\mathcal{H}, \mathcal{A}')$ is a $(\mathcal{A}, \mathcal{A}')$-measurable function $Z : \Omega \to \mathcal{H}$, that is for every $B \in \mathcal{A}'$, the pre-image $Z^{-1}(B) = \{\omega \in \Omega \mid Z(\omega) \in B\}$ belongs to $\mathcal{A}$.*

Let us focus on the case where $\mathcal{H} = H(k)$ is an RKHS. Since $H(k)$ is endowed with an inner product $\langle\cdot,\cdot\rangle_{H(k)}$, the $\sigma$-algebra $\mathcal{A}'$ can be defined as the cylindrical $\sigma$-algebra $\mathcal{C}_{H(k)}$, which is the

coarsest $\sigma$-algebra that contains all the cylinder sets $C_{A,h_1,\dots,h_m} \subseteq H(k)$ with

$$C_{A,h_1,\dots,h_m} \triangleq \left\{ h \in H(k) \,\middle|\, \left( \langle h, h_1 \rangle_{H(k)} \dots \langle h, h_m \rangle_{H(k)} \right) \in A \right\} \ ,$$

where $m \in \mathbb{N}^*$, $h_1, \dots, h_m \in H(k)$ and $A \subseteq \mathbb{R}^m$ is a Borel subset of $\mathbb{R}^m$. In other words, an RKHS-valued random element $Z$ can be characterized by the distribution of its real-valued marginals $\langle Z, h \rangle_{H(k)}$, $h \in H(k)$.

A Gaussian distribution in $H(k)$ corresponds to either a *Gaussian stochastic process* or a *Gaussian random element*.

**Definition 3.1.3** (Gaussian stochastic process)**.** *Let $Y = \{Y(x)\}_{x \in \mathcal{X}}$ be a stochastic process where $\mathcal{X}$ is a non-empty set. $Y$ is a Gaussian stochastic process if for every $a_1, \dots, a_m \in \mathbb{R}$ and $x_1, \dots, x_m \in \mathcal{X}$ with $m \in \mathbb{N}^*$, $\sum_{i=1}^m a_i Y(x_i)$ is a (univariate) Gaussian random variable.*

**Definition 3.1.4** (Gaussian random element)**.** *Let $Z$ be a random element in $(H(k), \mathcal{C}_{H(k)})$. $Z$ is a Gaussian random element if for every $h \in H(k)$, $\langle Z, h \rangle_{H(k)}$ is a (univariate) Gaussian random variable.*

A multivariate Gaussian distribution is parametrized by its mean and its variance, which are respectively a vector and a square matrix. In the following, we show how the notions of mean and covariance are transposed to stochastic processes and RKHS-valued random elements.

The *mean of a stochastic process* $\{Y(x)\}_{x \in \mathcal{X}}$ is a map $m : \mathcal{X} \to \mathbb{R}$ defined by

$$\forall x \in \mathcal{X}, \ \ m(x) \triangleq \mathbb{E}_Y[Y(x)] \ , \tag{3.1.1}$$

where we assume that the expectation in (3.1.1) exists for every $x \in \mathcal{X}$.

The *covariance function* of a stochastic process $\{Y(x)\}_{x \in \mathcal{X}}$ is a map $R : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by

$$\forall x, x' \in \mathcal{X}, \ \ R(x, x') \triangleq \mathbb{E}_Y[\{Y(x) - m(x)\}\{Y(x') - m(x')\}] \ , \tag{3.1.2}$$

assuming that the expectation in (3.1.2) exists for every $x, x' \in \mathcal{X}$.

In the case of a random element $Z$, the mean and the covariance are defined differently. Assume that $Z$ is a random element with weak first order, that is

$$\forall h \in H(k), \ \ \mathbb{E}_Z[\langle Z, h \rangle_{H(k)}] < +\infty \ . \tag{3.1.3}$$

Then the *mean of the random element* $Z$ is the element $\mu \in H(k)$ that satisfies

$$\forall h \in H(k), \ \ \mathbb{E}_Z[\langle Z, h \rangle_{H(k)}] = \langle \mu, h \rangle_{H(k)} \ . \tag{3.1.4}$$

The existence of $\mu$ is guaranteed by Riesz's Representation Theorem. Indeed, (3.1.3) implies that the linear map $h \mapsto \mathbb{E}_Z[\langle Z, h \rangle_{H(k)}]$ is bounded hence continuous so that Riesz's Representation Theorem states the existence of $\mu \in H(k)$ so that (3.1.4) holds.

Assume that the random element $Z$ is of weak second order, that is

$$\forall h \in H(k),\ \ \mathbb{E}_Z[\langle Z, h\rangle_{H(k)}^2] < +\infty \ . \tag{3.1.5}$$

If (3.1.5) holds then the following bilinear form $S : H(k) \times H(k) \to \mathbb{R}$ is well defined

$$\forall h, h',\ \ S(h,h') \overset{\Delta}{=} \mathbb{E}_Z[\langle Z-\mu, h\rangle_{H(k)} \langle Z-\mu, h'\rangle_{H(k)}] \ , \tag{3.1.6}$$

The *covariance operator* of the random element $Z$ is the operator $\Sigma : H(k) \to H(k)$ that satisfies

$$\forall h, h' \in H(k),\ \ \langle \Sigma \cdot h, h'\rangle_{H(k)} = S(h,h') \ . \tag{3.1.7}$$

To prove the existence of the covariance operator, one can build $\Sigma$ as follows. Firstly, consider the linear map $h \mapsto (h' \mapsto S(h,h'))$. By the Riesz's Representation Theorem and the assumption (3.1.5), each linear map $h' \mapsto S(h,h')$ with $h \in H(k)$ fixed corresponds to an element $\Sigma \cdot h \in H(k)$ with $\langle \Sigma \cdot h, h'\rangle_{H(k)} = S(h,h')$ for every $h' \in H(k)$. Clearly $h \mapsto \Sigma \cdot h$ is linear so that the property in (3.1.7) holds.

Because of the RKHS nature of $H(k)$, the notions of Gaussian stochastic process and Gaussian random element turn out to be the two sides of the same coin. Consider a Gaussian random element $Z$ in $H(k)$ and define the stochastic process $Y = \{Y(x)\}_{x \in H(k)}$ as

$$\forall x \in H(k),\ \ Y(x) \overset{\Delta}{=} \langle Z, k(x,\cdot)\rangle_{H(k)} \ . \tag{3.1.8}$$

Clearly $Y$ defines a Gaussian stochastic process. Indeed for every $a_1, \ldots, a_m \in \mathbb{R}$ and $x_1, \ldots, x_m \in \mathcal{X}$,

$$\sum_{i=1}^m a_i Y(x_i) = \sum_{i=1}^m a_i \langle Z, k(x_i,\cdot)\rangle_{H(k)} = \left\langle Z, \sum_{i=1}^m a_i k(x_i,\cdot)\right\rangle_{H(k)} \ ,$$

is a Gaussian random variable because $k(x_i,\cdot) \in H(k)$ for every $1 \leqslant i \leqslant m$ since $H(k)$ is an RKHS.

The respective mean and covariances of $Z$ and $Y$ as defined in (3.1.8) can be linked as follows. Because of the reproducing property of $H(k)$, the mean function $m : \mathcal{X} \to \mathbb{R}$ of $Y$ is expressed with respect to the mean $\mu \in H(k)$ of $Z$ as follows

$$\forall x \in \mathcal{X},\ \ m(x) = \langle \mu, k(x,\cdot)\rangle_{H(k)} = \mu(x) \ ,$$

so that $m$ and $\mu$ actually coincide.

Straightforwardly, the covariance function $R(x,x')$ of $Y$ satisfies for every $x, x' \in \mathcal{X}$

$$R(x,x') = \langle \Sigma \cdot k(x,\cdot), k(x',\cdot)\rangle_{H(k)} \ .$$

Defining conversely a random element $Z$ from a stochastic process $Y$ is less straightforward.

To do so, one has to assume that there exists a subset of trajectories — or sample paths — $\mathcal{T} \subseteq \{Y(x)(\omega) \mid \omega \in \Omega\}$ of $Y$ that lie in $H(k)$ with $\mathbb{P}(Y \in \mathcal{T}) = 1$. When this is the case, one can define the random element $Z$ in $H(k)$ as

$$Z(\omega) \overset{\Delta}{=} Y(\cdot)(\omega) \ ,$$

for every $\omega \in \Omega$ such that $Y \in \mathcal{T}$ and and $Z(\omega) = 0$ for instance for every $\omega \in \Omega$ such that $Y \notin \mathcal{T}$. However, such $\mathcal{T}$ does not necessarily exist. Here is a simple example that illustrates this. Let $k = \langle \cdot, \cdot \rangle$ be the linear kernel on $\mathbb{R}^m$. In this case the RKHS $H(k)$ consists in the set of every linear map $u \mapsto u^T v$ with $v \in \mathbb{R}^m$. Let $Y = \{Y(u)\}_{u \in \mathbb{R}^m}$ be a Gaussian stochastic process indexed by $\mathbb{R}^m$ such that the $Y(u)$ are *i.i.d.* $\mathcal{N}(0,1)$ Gaussian variables. Clearly $Y$ is not a linear map with high probability. This raises the question about conditions for a Gaussian process to lie in a given RKHS with probability one. This issue is treated in Section 3.2 thereafter which introduces the notion of *nuclear dominance*.

## 3.2 Nuclear dominance

We saw that a $H(k)$-valued random element $Z$ always defines a corresponding stochastic process $Y(x) = \langle Z, k(x, \cdot) \rangle_{H(k)}$ but that the converse is possible only if the paths of a given process $Y$ belong to $H(k)$ almost surely. In this section, we are interested in sufficient conditions that ensure this requirement.

Let us assume that the process $Y = \{Y(x)\}_{x \in \mathcal{X}}$ admits a mean function $m : \mathcal{X} \to \mathbb{R}$ and a covariance function $R : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. In the case of a Gaussian process $Y$, [Dri73] studied the relationships between $R$ and the kernel $k$ that entail $\mathbb{P}(Y \in H(k)) = 1$.

**Proposition 3.2.1** ([Dri73])**.** *Assume $\mathcal{X}$ is a separable metric space and $Y = \{Y(x)\}_{x \in \mathcal{X}}$ is a Gaussian process with mean $m(\cdot) \in H(k)$, covariance function $R(\cdot, \cdot)$ continuous on $\mathcal{X} \times \mathcal{X}$ and whose paths are almost surely continuous on $\mathcal{X}$. Let $\{x_1, x_2, \ldots\}$ be a countable dense subset of $\mathcal{X}$ and define the matrices $R_n = [R(x_i, x_j)]_{1 \leqslant i,j \leqslant n}$ and $K_n = [k(x_i, x_j)]_{1 \leqslant i,j \leqslant n}$ for every $n \geqslant 1$. If the quantity $\tau$ defined as*

$$\tau = \lim_{n \to +\infty} \mathrm{Tr}(R_n K_n^{-1}) \ , \tag{3.2.9}$$

*is finite, then $\mathbb{P}(Y \in H(k)) = 1$.*

Even though it provides some answers, the result of [Dri73] relies on some restrictive assumptions, such as the almost sure continuity of $Y$ or the continuity of the covariance function $R(\cdot, \cdot)$. [LB01] improved Driscoll's result by relaxing those assumptions. In particular, [LB01] simplified the definition of $\tau$ in (3.2.9) by relating $\tau$ to the trace of some operator called *dominance operator*.

**Definition 3.2.2** (Dominance operator)**.** *Let $R$ and $k$ be two symmetric, positive definite functions and $H(R), H(k)$ denote their respective RKHS.*

*Assume $H(R) \subseteq H(k)$. Then there exists a bounded operator $L : H(k) \to H(R)$ such that*

$$\forall h \in H(k),\ \ \forall f \in H(R),\ \ \langle L \cdot h, f \rangle_{H(R)} = \langle h, f \rangle_{H(k)} \ .$$

*In particular, $L \cdot k(x, \cdot) = R(x, \cdot)$ for every $x \in \mathcal{X}$.*
*The operator $L$ is called dominance operator of $H(k)$ over $H(R)$.*
*Moreover, if $L : H(k) \to H(k)$ is a positive, continuous and self-adjoint operator, then $R(x, x') = \langle L \cdot k(x, \cdot), k(x', \cdot) \rangle_{H(k)}$ is a positive definite kernel such that $H(R) \subseteq H(k)$.*

The existence of $L$ as asserted by definition 3.2.2 comes from Theorem 1.1 in [LB01]. Remark that the operator $L$ as defined in 3.2.2 coincide with the covariance operator $\Sigma$ of the Gaussian random element $Z$ defined from $Y$ when $Y \in H(k)$ almost surely. Indeed for every $x, x' \in \mathcal{X}$

$$\langle \Sigma \cdot k(x, \cdot), k(x', \cdot) \rangle_{H(k)} = R(x, x') = \langle R(x, \cdot), k(x', \cdot) \rangle_{H(k)} = \langle L \cdot k(x, \cdot), k(x', \cdot) \rangle_{H(k)} \ ,$$

hence by bi-linearity and continuity of $\langle \cdot, \cdot \rangle_{H(k)}$, $\langle \Sigma \cdot f, g \rangle_{H(k)} = \langle L \cdot f, g \rangle_{H(k)}$ for every $f, g \in H(k)$ and $\Sigma = L$.

According to Proposition 4.5 in [LB01], the quantity $\tau$ as defined in (3.2.9) corresponds to the trace of the dominance operator $L$. To define the trace of $L$, we need to assume that $H(k)$ is separable so that an orthonormal Hilbertian basis $\{e_i\}_{i \in \mathbb{N}^*}$ exists and

$$\tau = \mathrm{Tr}(L) \overset{\Delta}{=} \sum_{i \geqslant 1} \langle L \cdot e_i, e_i \rangle_{H(k)} \ .$$

Note that this definition of the trace does not depend on the choice of the orthonormal basis $\{e_i\}_{i \geqslant 1}$ (see Theorem VI.18 in [RS80]). Also remark that a sufficient condition for $H(k)$ to be separable is that $\mathcal{X}$ is a separable topological space and $k$ is continuous on $\mathcal{X} \times \mathcal{X}$ (see Lemma 4.33 in [SC08]).

Therefore Driscoll's condition about the finiteness of $\tau$ is replaced by the finiteness of $\mathrm{Tr}(L)$ and the condition $H(R) \subseteq H(k)$ that allows the existence of the dominance operator $L$. These two conditions defines the *nuclear dominance* of $H(k)$ over $H(R)$.

**Definition 3.2.3** (Nuclear dominance)**.** *Let $R$ and $k$ be two symmetric, positive definite functions and $H(R), H(k)$ denote their respective RKHS.*
*One says that there is nuclear dominance of $H(k)$ over $H(R)$ if the following assertions hold true:*

- $H(R) \subseteq H(k)$,
- $\mathrm{Tr}(L) < \infty$,

*where $L$ is the dominance operator as defined in Definition 3.2.2. This relationship of nuclear dominance is denoted by $k >> R$.*

Proposition 3.2.4 thereafter follows Proposition 7.2 from [LB01].

**Proposition 3.2.4** ([LB01]). *Let $R$ and $k$ be two symmetric, positive definite functions and $H(R), H(k)$ denote their respective RKHS. Assume that $H(k)$ is separable.*
*Then there exists a Gaussian process $Y = \{Y(x)\}_{x\in\mathcal{X}}$ with mean $m \in H(k)$ and covariance function $R(\cdot,\cdot)$ such that $\mathbb{P}(Y \in H(k)) = 1$ if and only if $k >> R$.*

In particular, a consequence of Proposition 3.2.4 is that there cannot be any isotropic Gaussian process in an infinite-dimensional RKHS. In an RKHS $H(k)$, an isotropic stochastic process is a stochastic process with covariance function $R(x,x') = k(x,x')$ for every $x, x' \in \mathcal{X}$. In other words, the dominance operator of $H(k)$ over $H(R)$ is the identity operator $L = \mathrm{Id}_{H(k)}$. Given an orthonormal Hilbertian basis $\{e_i\}_{i\in\mathbb{N}^*}$ of $H(k)$,

$$\mathrm{Tr}(L) = \mathrm{Tr}(\mathrm{Id}_{H(k)}) = \sum_{i\geqslant 1} \langle \mathrm{Id}_{H(k)} \cdot e_i, e_i \rangle_{H(k)} = \sum_{i\geqslant 1} \|e_i\|^2_{H(k)} = \sum_{i\geqslant 1} 1 = +\infty \ ,$$

hence nuclear dominance does not hold and Proposition 3.2.4 entails the non-existence of an isotropic Gaussian process in $H(k)$.

## 3.3 Validity of Gaussian processes in the empirical case

In this section, we consider the empirical case where a sample $x_1, \ldots, x_n \in \mathcal{X}$ of $\mathcal{X}$-valued observations is available and where the embedded data $k(x_1,\cdot), \ldots, k(x_n,\cdot) \in H(k)$ are modeled as a Gaussian process. The parameters of this Gaussian process need to be chosen with respect to the dataset. Hence two things are required: firstly, define empirical estimators for the mean and the covariance function of the Gaussian process, and secondly check whether the corresponding Gaussian process is well-defined in $H(k)$ for this choice of parameters.

A natural estimator for the mean and the covariance function is to consider the mean/covariance of the empirical measure $\sum_{i=1}^n \delta_{k(x_i,\cdot)}$, which gives rise to the *empirical mean* and the *empirical covariance function*.

**Definition 3.3.1** (Empirical mean). *Given an embedded sample $k(x_1,\cdot), \ldots, k(x_n,\cdot)$, the corresponding empirical mean $\hat{m} : \mathcal{X} \to \mathbb{R}$ is defined by*

$$\forall x \in \mathcal{X}, \ \ \hat{m}(x) \stackrel{\Delta}{=} \frac{1}{n} \sum_{i=1}^n k(x_i, x) \ . \tag{3.3.10}$$

**Definition 3.3.2** (Empirical covariance function). *Given an embedded sample $k(x_1,\cdot), \ldots, k(x_n,\cdot)$, the corresponding empirical covariance function $\hat{R} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined by*

$$\forall x, x' \in \mathcal{X}, \ \ \hat{R}(x,x') \stackrel{\Delta}{=} \frac{1}{n} \sum_{i=1}^n [k(x_i,x) - m(x)][k(x_i,x') - m(x')] \ . \tag{3.3.11}$$

Let $Y = \{Y(x)\}_{x\in\mathcal{X}}$ be a Gaussian process with mean $\hat{m}(\cdot)$ and covariance function $\hat{R}(\cdot,\cdot)$. The

next step consists in ensuring that $Y$ is well-defined in $H(k)$ by means of the results exposed in Section 3.2, which is the case as stated by Proposition 3.3.3. Note that the nuclear dominance holds even if the actual covariance function $R(\cdot,\cdot)$ related to the true underlying distribution of the sample $x_1,\dots,x_n$ does not satisfy $R << k$.

**Proposition 3.3.3.** *Assume $H(k)$ is separable. Then the empirical mean $\hat{m}(\cdot)$ and covariance function $\hat{R}(\cdot,\cdot)$ as defined in* (3.3.10) *and* (3.3.11) *satisfy $\hat{m} \in H(k)$ and $\hat{R} << k$. Therefore, there exists a Gaussian process with mean $\hat{m}$ and covariance $\hat{R}$ whose sample paths belong almost surely to $H(k)$.*

*Proof of Proposition 3.3.3.* It is straightforward that $\hat{m} \in H(k)$ since $H(k)$ is a vector space.

We show that $\hat{R} << k$ by exhibiting the corresponding dominance operator $\hat{L}$:

$$\hat{L}\cdot f \triangleq \frac{1}{n}\sum_{i=1}^{n}\langle f, k(x_i,\cdot)\rangle_{H(k)}k(x_i,\cdot) - \frac{\hat{m}(\cdot)}{n}\sum_{i=1}^{n}\langle f, k(x_i,\cdot)\rangle_{H(k)} \ , \tag{3.3.12}$$

for every $f \in H(k)$.

First of all, it is clear that $\hat{L}\cdot k(x,\cdot) = \hat{R}(x,\cdot)$ for all $x \in \mathcal{X}$. Let us check that $\hat{L}$ is a bounded operator. Since $|\langle f, k(x_i,\cdot)\rangle|_{H(k)} \leqslant \|f\|_{H(k)} k^{\frac{1}{2}}(x_i,x_i)$ by Cauchy-Schwarz's inequality, one obtains that for $\|f\|_{H(k)} = 1$,

$$\|\hat{L}\cdot f\|_{H(k)} \leqslant \frac{1}{n}\sum_{i=1}^{n}k(x_i,x_i) + \frac{1}{n}\left(\frac{1}{n^2}\sum_{j,l=1}^{n}k(x_j,x_l)\right)^{1/2}\sum_{i=1}^{n}k(x_i,x_i)^{1/2} < +\infty \ ,$$

hence the boundedness of $\hat{L}$.

Finally since the operator $\hat{L}$ is of finite rank (up to $n$), it has a finite trace. Therefore we can conclude that $\hat{R} << k$. □

Due to the high-dimensional nature of an RKHS, the empirical covariance function as defined in (3.1.2) may be unwieldy for some applications. For instance, in the methods described later in Sections 3.4.1 and 3.4.2, the covariance function needs to be "inverted" in the following sense. Consider the eigen-expansion of the empirical covariance function $\hat{R}$

$$\forall x,x' \in \mathcal{X},\ \ \hat{R}(x,x') = \sum_{j=1}^{r}\hat{\lambda}_j\hat{\varphi}_j(x)\hat{\varphi}_j(x') \ , \tag{3.3.13}$$

where the pairs $(\hat{\lambda}_j,\hat{\varphi}_j)_{1\leqslant j\leqslant r}$ are the eigenvalues/eigenvectors of the corresponding dominance operator $\hat{L}$ — defined in (3.3.12) — that satisfy

$$\forall 1 \leqslant j \leqslant r, \qquad \hat{L}\cdot\hat{\varphi}_j = \hat{\lambda}_j\hat{\varphi}_j \ ,$$

and $\langle\hat{\varphi}_j,\hat{\varphi}_l\rangle_{H(k)} = \delta_{jl}$ for every $1 \leqslant j,l \leqslant r$ and $\delta_{jl}$ denotes the Kronecker delta. Here $r \leqslant n$ denotes the rank of $\hat{L}$. Therefore "inverting" $\hat{R}(\cdot,\cdot)$ means inverting the operator $\hat{L}$ which is done by

inverting the eigenvalue in the eigen-expansion (3.3.13). This operation becomes problematic when $r$ is very large as the smallest eigenvalues tend to 0 when $r \to +\infty$ — because of the finiteness of $\mathrm{Tr}(\hat{L})$ required by the nuclear dominance conditions. Hence inverting those small eigenvalues may result in computational instability in practice.

To overcome this issue, a class of parsimonious Gaussian models have been proposed [MM08; BFG15] where the $r-d$ smallest (non-zero) eigenvalues are assumed to be equal to the same value (for some $1 \leqslant d \leqslant r$). The parameter $d$ of the model is called *intrinsic* dimension and has to be interpreted as the number of directions that carry most of the information of the dataset whereas the $r-d$ other directions consists of noise. Following this model, the empirical covariance function is "tweaked" in the following way: whereas the $d$ largest eigenvalues of $\hat{R}$ remain unchanged , the other eigenvalues are replaced by their average value. The resulting parsimonious covariance estimator is formally defined thereafter.

**Definition 3.3.4** (Parsimonious covariance estimator)**.** *Write the eigen-expansion of $\hat{R}$ as in* (3.3.13).

*The parsimonious sample covariance $\hat{R}_d(x,x')$ is defined for every $x, x'$ by*

$$\hat{R}_d(x,x') \overset{\Delta}{=} \sum_{j=1}^{d} \hat{\lambda}_j \hat{\varphi}_j(x)\hat{\varphi}_j(x') + \frac{\sum_{l=d+1}^{r} \hat{\gamma}_l}{r-d} \sum_{j=d+1}^{r} \hat{\varphi}_j(x)\hat{\varphi}_j(x') \ . \tag{3.3.14}$$

It remains to show that a Gaussian process with $\hat{R}_d$ as covariance function is well-defined in an RKHS. This is done by first proving that nuclear dominance of $\hat{R}$ over $\hat{R}_d$ (Lemma 3.3.5) which combined with the already known relationship $\hat{R} << k$ entails $\hat{R}_d << k$ (Proposition 3.3.6).

**Lemma 3.3.5.** *Nuclear dominance of the sample covariance estimator $\hat{R}$ over the parsimonious estimator $\hat{R}_d$ holds.*

*Proof of Lemma 3.3.5.* Considering the eigenexpansion of $\hat{R}$ as written in (3.3.13), the dominance operator between $\hat{R}$ and $\hat{R}_d$ can be defined by

$$\hat{L}_d \cdot f = \sum_{j=1}^{d} \langle f, \hat{\varphi}_j \rangle_{H(k)} \hat{\varphi}_j + \left( \frac{1}{r-d} \sum_{l=d+1}^{r} \hat{\lambda}_l \right) \sum_{j=d+1}^{r} \hat{\lambda}_j^{-1} \langle f, \hat{\varphi}_j \rangle_{H(k)} \hat{\varphi}_j \ .$$

Straightforwardly $\hat{L}_d(\hat{R}(x,.)) = \hat{R}_d(x,.)$ for every $x \in \mathcal{X}$. Let us check that $\hat{L}_d$ is bounded. For any $\|f\|_{H(\hat{R})} \leq 1$,

$$\begin{aligned} \|\hat{L}_d \cdot f\|_{H(\hat{R})} &\leqslant \sum_{j=1}^{d} |\langle f, \hat{\varphi}_j \rangle|_{H(k)} \|\hat{\varphi}_j\|_{H(\hat{R})} + \left( \frac{1}{r-d} \sum_{l=d+1}^{r} \hat{\lambda}_l \right) \sum_{j=d+1}^{r} \hat{\lambda}_j^{-1} |\langle f, \hat{\varphi}_j \rangle|_{H(k)} \|\hat{\varphi}_j\|_{H(\hat{R})} \\ &\leqslant \sum_{j=1}^{d} \|f\|_{H(k)} \|\hat{\varphi}_j\|_{H(\hat{R})} + \left( \frac{1}{r-d} \sum_{l=d+1}^{r} \hat{\lambda}_l \right) \sum_{j=d+1}^{r} \hat{\lambda}_j^{-1} \|f\|_{H(k)} \|\hat{\varphi}_j\|_{H(\hat{R})} \end{aligned}$$

$$\leqslant \sum_{j=1}^{d} \|f\|_{H(\hat{R})} \|\hat{\varphi}_j\|_{H(\hat{R})} + \left( \frac{1}{r-d} \sum_{l=d+1}^{r} \hat{\lambda}_l \right) \sum_{j=d+1}^{r} \hat{\lambda}_j^{-1} \|f\|_{H(\hat{R})} \|\hat{\varphi}_j\|_{H(\hat{R})}$$
$$\leqslant \sum_{j=1}^{d} \|\hat{\varphi}_j\|_{H(\hat{R})} + \left( \frac{1}{r-d} \sum_{l=d+1}^{r} \hat{\lambda}_l \right) \sum_{j=d+1}^{r} \hat{\lambda}_j^{-1} \|\hat{\varphi}_j\|_{H(\hat{R})} < +\infty \ ,$$

where we used the inequality $\|f\|_{H(k)} \leqslant \|f\|_{H(\hat{R})}$ due to $H(\hat{R}) \subseteq H(k)$ (see Theorem 1.1 in [LB01]).

Furthermore $\hat{L}_d$ has a finite trace because of the finiteness of its rank. □

**Proposition 3.3.6.** *There is nuclear dominance of $k$ over the parsimonious sample covariance estimator $\hat{R}_d$, that is $\hat{R}_d << k$.*

*Proof of Proposition 3.3.6.* Having proved that $k >> \hat{R}$ and $\hat{R} >> \hat{R}_d$, one would like to deduce that $k >> \hat{R}_d$. It is straightforward that $H(\hat{R}_d) \subseteq H(k)$. A candidate for being the dominance operator of $k$ over $\hat{R}_d$ is $\hat{L}_d \hat{L}$. Clearly $\hat{L}_d \hat{L}$ is bounded since $\hat{L}_d$ and $\hat{L}$ are bounded. It remains to check that $\mathrm{Tr}(\hat{L}_d \hat{L}) < \infty$:

$$\begin{aligned} \mathrm{Tr}(\hat{L}_d \hat{L}) &= \sum_{j \geq 1} \langle \hat{L}_d \hat{L} \hat{\varphi}_j, \hat{\varphi}_j \rangle_{H(k)} \\ &= \sum_{j,l \geq 1} \langle \hat{L} \hat{\varphi}_j, \hat{\varphi}_l \rangle_{H(k)} \langle \hat{L}_d \hat{\varphi}_j, \hat{\varphi}_l \rangle_{H(k)} \ , \end{aligned}$$

since $\hat{L}_d$ is symmetric that is $\langle \hat{L}_d f, g \rangle_{H(k)} = \langle f, \hat{L}_d g \rangle_{H(k)}$ for all $f, g \in H(k)$.

Note that

$$\langle \hat{L}_d \hat{\varphi}_j, \hat{\varphi}_l \rangle_{H(k)} = \begin{cases} 0 & \text{if } j \neq l \text{ or } j > d \text{ or } l > d \\ 1 & \text{if } j = l \leqslant d \end{cases}$$

which leads to

$$\mathrm{Tr}(\hat{L}_d \hat{L}) = \sum_{j,l \geq 1} \langle \hat{L} \hat{\varphi}_j, \hat{\varphi}_l \rangle_{H(k)} = \sum_{j=1}^{d} \hat{\lambda}_k < +\infty \ .$$

□

## 3.4 Gaussian models and RKHS: examples of application

This section presents how a Gaussian model in an RKHS can be used in practice for some learning tasks. Three existing methods that make use of such models are reviewed: one for multigroup classification through a mixture of Gaussian processes in the kernel space [BFG15],a method for outlier detection [Rot06] and Gaussian process regression [RW06].

### 3.4.1 Multigroup classification with a mixture of Gaussian processes

Consider an *i.i.d.* sample $(X_1, Z_1), \dots, (X_n, Z_n) \in \mathcal{X} \times \{1, \dots, C\}$. The $n$ observations $X_1, \dots, X_n$ are assumed to come from $M$ different classes, and for each $1 \leqslant i \leqslant n$ the variable $Z_i$ represents the class membership of $X_i$ — that is $Z_i = j$ if $X_i$ comes from the $j$-th class. Given a new point $X$ in $\mathcal{X}$, the goal is to assign correctly $X$ to one of the $C$ classes.

One way of doing this is to compute the *maximum a posteriori* (MAP) probabilities $\mathbb{P}(Z = j \mid X = x)$ for each $1 \leqslant j \leqslant C$, where $Z$ denotes the class label of $X$ and $x$ is the observed value of $X$. The new point is assigned to the class that maximizes the MAP.

In a first time, let us assume that $\mathcal{X} = \mathbb{R}^d$. The MAP is calculated through the Bayes rule as

$$\mathbb{P}(Z = j \mid X = x) = \frac{\pi_j f_j(x)}{\sum_{l=1}^{C} \pi_l f_l(x)} \ , \tag{3.4.15}$$

where $\pi_j = \mathbb{P}(Z = j)$ and $f_j$ denotes the density function of $X$ conditionally to $Z = j$ for every $1 \leqslant j \leqslant C$ and $x$ is the observed value of $X$.

A method to get a tractable expression for the MAP is to model the dataset $X_1, \dots, X_n$ as a *mixture of Gaussians*, that is each $f_j$ is modeled as a Gaussian density

$$f_j(x) = (2\pi)^{-d/2} |\Sigma_j|^{-1/2} \exp\left(-\frac{1}{2}(x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j)\right) \ ,$$

where $\mu_1, \dots, \mu_C \in \mathbb{R}^d$ are the respective mean vectors of each class, $\Sigma_1, \dots, \Sigma_C \in \mathcal{M}_d(\mathbb{R})$ their respective covariance matrices and $|\cdot|$ denotes the determinant of a matrix.

Now we are in a position to assign a score function to each class based on the MAP. Note that the denominator in (3.4.15) does not depend on the class label $j$, hence the score function $D_j(x)$ of the $j$-th class is defined solely via the numerator in (3.4.15). Namely, $D_j(x)$ is written as a decreasing function of $\pi_j f_j(x)$ that is

$$\begin{aligned} D_j(x) &= -2\log(\pi_j f_j(x)) - d\log(2\pi) \\ &= (x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j) + \log(|\Sigma_j|) - 2\log(\pi_j) \ , \end{aligned} \tag{3.4.16}$$

so that the new point $X$ is assigned to the class that *minimizes* $D_j(x)$. Let us write the eigen-expansion of each matrix $\Sigma_j$ as $\Sigma_j = \sum_{l=1}^{r_l} \lambda_{j,l} u_{j,l} u_{j,l}^T$ where $r_j \in \{1, \dots, d\}$ is the rank of $\Sigma_j$, $\lambda_{j,1} \geqslant \lambda_{j,2} \geqslant \dots \geqslant \lambda_{j,r_j} > 0$ are the non-zero eigenvalues of $\Sigma_j$ and $u_{j,1}, \dots, u_{j,r_j} \in \mathbb{R}^d$ the corresponding eigenvectors. Then (3.4.16) can be written alternatively as

$$D_j(x) = \sum_{l=1}^{r_j} \frac{1}{\lambda_{j,l}} [(x - \mu_j)^T u_{j,l}]^2 + \sum_{l=1}^{r_j} \log(\lambda_{j,l}) - 2\log(\pi_j) \ . \tag{3.4.17}$$

However this approach may suffer some shortcomings. The actual distribution of $X$ conditionally to $Z = j$ may be far from a Gaussian distribution which may lead to bad classification

performances. Moreover, this method does not cover the case of non-vectorial data, that is when $\mathcal{X} \neq \mathbb{R}^d$.

[BFG15] proposed a variation of this classification method by using a kernel. Let $k$ be a kernel on $\mathcal{X}$ with corresponding feature map $\phi : \mathcal{X} \to H(k)$. The idea of [BFG15] is to model the embedded data $Y_i = \phi(X_i)$, $i = 1, \dots, n$ — instead of the initial data $X_1, \dots, X_n$ — as a mixture of Gaussian processes. Namely, $Y \mid Z = j$ is modeled as a Gaussian process with mean $m_j : \mathcal{X} \to \mathbb{R}$ and covariance function $R_j : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Following Section 3.2, we assume that $m_j \in H(k)$ and $R_j << k$ so that these Gaussian processes are well-defined in $H(k)$. The dominance operator of $H(k)$ over $H(R_j)$ is denoted $\Sigma_j$. Remember that as remarked in Section 3.2, $\Sigma_j$ coincides with the covariance operator of the random element defined from $Y$ conditionally to $Z = j$.

This kernelization allows to tackle the two issues mentioned above: there is no requirement about the nature of the set $\mathcal{X}$ where the initial dataset lies, and the kernel $k$ can be chosen from a family of kernels that is rich enough so that the Gaussian model in the RKHS is reliable.

On the other hand, the transposition to the RKHS version is not straightforward. Indeed, the score functions $D_j(x)$ defined in the $\mathbb{R}^d$ case were based on Gaussian densities, which was defined with respect to the Lebesgue measure on $\mathbb{R}^d$. In the case of an infinite-dimensional RKHS, there exists no canonical reference measure such as the Lebesgue measure on $\mathbb{R}^d$ from which a Gaussian density could be defined. For this reason, we have to assume that $H(k)$ is of finite dimension $d$ even if it means that $d$ is very large. The score function in (3.4.17) therefore becomes

$$D_j(x) = \sum_{l=1}^{d} \frac{1}{\lambda_{j,l}} \langle \phi(x) - m_j, \varphi_{j,l} \rangle_{H(k)}^2 + \sum_{l=1}^{d} \log(\lambda_{j,l}) - 2\log(\pi_j) \ , \tag{3.4.18}$$

where the pairs $(\lambda_{j,1}, \varphi_{j,1}), \dots, (\lambda_{j,d}, \varphi_{j,d}) \in \mathbb{R}_+ \times H(k)$ are the eigenvalue/eigenvectors of the covariance operator $\Sigma_j$, that is

$$\Sigma_j \cdot \varphi_{j,l} = \lambda_{j,l} \varphi_{j,l} \ ,$$

for every $1 \leqslant j \leqslant C$ and $1 \leqslant l \leqslant d$.

Since $d$ is potentially very large, a parsimony assumption is made on the eigenvalues of $\Sigma_1, \dots, \Sigma_C$. Namely, we assume the existence of so-called *intrinsic* dimensions $d_1, \dots, d_C \in \{1, \dots, d\}$ such that for every $1 \leqslant j \leqslant C$ and $l > d_j$, $\lambda_{j,l} = \lambda$ with $\lambda > 0$. This way (3.4.18) becomes

$$\begin{aligned} D_j(x) = \sum_{l=1}^{d_j} \left( \frac{1}{\lambda_{j,l}} - \frac{1}{\lambda} \right) \langle \phi(x) - m_j, \varphi_{j,l} \rangle_{H(k)}^2 + \frac{1}{\lambda} \| \phi(x) - m_j \|_{H(k)}^2 \\ + \sum_{l=1}^{d_j} \log(\lambda_{j,l}) + (d - d_j)\log(\lambda) - 2\log(\pi_j) \ . \end{aligned} \tag{3.4.19}$$

Note that if $d = +\infty$, this parsimony assumption entails that $\mathrm{Tr}(\Sigma_j) = +\infty$ for every $1 \leqslant j \leqslant C$

which violates one of the conditions for the relationships $R_j << k$ to hold.

In practice, (3.4.19) is expressed in a different form that does not involve $\phi$ or any direct calculations in $H(k)$ (the "kernel trick"). Beforehand, we need to estimate some parameters on the basis of the sample $(X_1, Z_1), \dots, (X_n, Z_n)$. The within class means $m_1, \dots, m_C$ and covariance functions $R_1, \dots, R_C$ are replaced by their empirical counterparts $\hat{m}_1, \dots, \hat{m}_C$ and $\hat{R}_1, \dots, \hat{R}_C$ as defined in Section 3.3. The empirical eigenvalues $\{\hat{\lambda}_{j,l}\}_l$ and eigenfunctions $\{\hat{\varphi}_{j,l}\}_l$ corresponding to $\hat{R}_j$ are estimated as in kernel PCA (see Section 2.3.2) from the eigendecomposition of the matrix $[k(X_i, X_{i'})]_{i,i' \in \mathcal{I}_j}$ where $\mathcal{I}_j = \{i \in \{1, \dots, n\} \mid Z_i = j\}$. In particular the eigenfunctions admits an expansion of the form $\hat{\varphi}_{j,l} = \sum_{i \in \mathcal{I}_j} a_{j,l}^{(i)} \phi(X_i)$. Besides, the class proportions $\pi_j$ are simply estimated by $\hat{\pi}_j = n_j / n$ where $n_j$ is the cardinal of $\mathcal{I}_j$. Plugging those empirical estimates into the score function (3.4.19) leads to the empirical score functions $\hat{D}_j(x)$

$$\begin{aligned}\hat{D}_j(x) = \sum_{i=1}^{d_j} \left( \frac{1}{\hat{\lambda}_{j,l}} - \frac{1}{\hat{\lambda}} \right) \left( \sum_{i \in \mathcal{I}_j} a_{j,l}^{(i)} k(x, x_i) - \frac{1}{n_j} \sum_{i,i' \in \mathcal{I}_j} k(x_i, x_{i'}) \right)^2 \\ + \left( k(x,x) + \frac{1}{n_j^2} \sum_{i,i' \in \mathcal{I}_j} k(x_i, x_{i'}) - \frac{2}{n_j} \sum_{i \in \mathcal{I}_j} k(x, x_i) \right)^2 \\ + \sum_{l=1}^{d_j} \log(\hat{\lambda}_{j,l}) + (d - d_j) \log(\hat{\lambda}) - 2 \log(\hat{\pi}_j) \ .\end{aligned}$$

### 3.4.2 Batch outlier detection

The problem of outlier detection can be described as follows: given a sample of $n$ observations $X_1, \dots, X_n \in \mathcal{X}$, one assumes that there exists a subsample with indexes $\mathcal{I}_{out} \subset \{1, \dots, n\}$ such that $\{X_i\}_{i \in \mathcal{I}_{out}}$ — called the *outliers* — were generated by an underlying distribution $Q$ different from the distribution $P$ the $\{X_i\}_{i \notin \mathcal{I}_{out}}$ stem from. The goal is to identify those outliers. The difficulty lies in the fact that the number of outliers is typically very small compared to $n$. This issue is solved by assuming that outliers fall outside the support of $P$. Therefore, the problem amounts to define an acceptance region that contains non-outliers with high probability.

[Rot06] proposes to solve this problem by setting a Gaussian model in a feature space. Let us assume that the $X_1, \dots, X_n$ take values in $\mathcal{X} = \mathbb{R}^d$ and consider a feature map $\phi : \mathbb{R}^d \to \mathbb{R}^p$ with related kernel $k(x, x') = \phi(x)^T \phi(x')$ for every $x, x' \in \mathbb{R}^d$. The proposed model assumes that $\phi(X)$ follows a $\mathcal{N}(\mu, \Sigma)$ Gaussian distribution with mean vector $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathcal{M}_p(\mathbb{R})$ when $X \sim P$.

The useful quantity to detect outliers is the *Mahalanobis distance* $D(\phi(x), \mu)$ between an embedded observation $\phi(x) \in \mathbb{R}^p$ and the mean vector $\mu$

$$D(\phi(x), \mu) = (\phi(x) - \mu)^T \Sigma^{-1} (\phi(x) - \mu) \ . \tag{3.4.20}$$

When $X$ is a non-outlier, the distance $D(\phi(X),\mu)$ is expected to follow a $\chi_2(p)$ chi-square distribution. To detect outliers, the empirical distribution of the $D(\phi(x_1),\mu),\dots,D(\phi(x_n),\mu)$ is compared with the $\chi_2(p)$ distribution by means of a quantile-quantile plot or QQ-plot. The QQ-plot consists in the pairs of points $(F^{-1}(i/n),\hat{F}^{-1}(i/n))_{1\leqslant i\leqslant n}$ where $F^{-1}$ denotes the quantile function of the $\chi_2(p)$ distribution and $\hat{F}^{-1}$ the empirical quantile function of $D(\phi(X),\mu)$. In the case when no outlier is present, the QQ-plot should ideally be close to a straight line of equation $y=x$. Thus a linear model is applied to the QQ-plot and a confidence interval around the fitted line is calculated. The observations declared as outliers are those that fall outside of the confidence interval.

Note that both $\mu$ and $\Sigma$ are unknown and should be estimated. Moreover, the inversion of the covariance matrix $\Sigma$ may be cumbersome when the number of dimensions $p$ in the feature space becomes very large. However, [Rot06] uses an alternative approach that bypasses this estimation phase by relating the Mahalanobis distance in (3.4.20) to a *kernel ridge regression* formulation of the outlier problem. Namely, they consider the regression problem

$$w^* \in \operatorname{argmin}_{w\in\mathbb{R}^p} \ \|\mathbf{y}-\mathbf{\Phi}^T w\|^2+\delta\|w\|^2 \quad \text{with} \quad \begin{array}{l}\mathbf{y}=(1\dots 1)^T\in\mathbb{R}^n\\ \mathbf{\Phi}=[\phi(x_1)\dots\phi(x_n)]\in\mathbb{R}^{p\times n}\\ \delta>0\ ,\end{array}$$

where $\mathbf{y}$ represents the realizations of a latent variable $Y$ that indicates whether $X$ is an outlier or not. The solution of this regression problem is known to be $w^*=(\mathbf{K}+\delta I)^{-1}\mathbf{y}$ where $\mathbf{K}=\mathbf{\Phi}^T\mathbf{\Phi}$ is the Gram matrix. Therefore the Mahalanobis distance (3.4.20) can be split into two parts by projecting $\phi(x)$ onto some vector $\tilde{w}=\alpha w^*$ where $\alpha\in\mathbb{R}$ is adequately chosen to get an expression of the form

$$D(\phi(x),\mu)=(\phi(x)^T\tilde{w}-m_+)^2+D_\perp \quad \text{where } m_+,D_\perp\in\mathbb{R}\ . \tag{3.4.21}$$

The detailed expression of (3.4.21) writes for every $1\leqslant i\leqslant n$

$$D(\phi(x_i),\mu)=\frac{s^{-2}}{1-s^2}\left[y^T(\mathbf{K}+\delta I_n)^{-1}\mathbf{k_i}-s^2\right]^2-s^{-1}\left[\mathbf{y^T}(\mathbf{K}+\delta I_n)^{-1}\mathbf{k_i}\right]^2+n\mathbf{k_i^T}(\mathbf{K}+\delta I_n)^{-1}e_i\ ,$$

where $s^2=n^{-1}\mathbf{y^T}\mathbf{K}(\mathbf{K}+\delta I)^{-1}\mathbf{y}$, $\mathbf{k_i}$ is the $i$-th column of $K$ and $e_i\in\mathbb{R}^n$ is the vector whose entries are all equal to 0 except for the $i$-th entry that is equal to 1. The advantage of this formulation is two-fold. Firstly, the parameters $\mu$ and $\Sigma$ do not need to be estimated and in particular $\Sigma$ does not need to be inverted. Secondly introducing the solution $w^*$ of the kernel ridge regression problem above leads to invert only the matrix $(\mathbf{K}+\delta I_n)$ instead of the Gram matrix $\mathbf{K}$. Apart from avoiding computational instability, this also introduces a quantity that [Rot06] calls *effective degrees of freedom* $df=\operatorname{Tr}(\mathbf{K}(\mathbf{K}+\delta I_n)^{-1})$. This quantity is useful when the number of dimensions $p$ of the feature space is infinite so that $D(\phi(x),\mu)$ is modeled as a $\chi_2(df)$ random variable instead

of a $\chi_2(p)$ variable.

As for theoretical guarantees, [Rot06] focuses on the case of a Gaussian RBF kernel $k(x,x') = \phi(x)^T\phi(x') = \exp(-\gamma\|x-x'\|^2)$ with parameter $\gamma > 0$ and justifies that in some sense this class of kernels is rich enough so that the Gaussian model in the feature space is relevant. Firstly in order to select an optimal $\gamma$, they consider a *likelihood cross-validation* procedure which necessitates a proper Gaussian density in the feature space. Since the feature space is infinite-dimensional for a Gaussian RBF kernel, a proper Gaussian density cannot be defined. However the problem is circumvented by setting a density $p_n(x)$ in the input space that "mimics" the contour lines of a Gaussian density in the feature space if the latter was finite dimensional, that is

$$p_n(x) \overset{\Delta}{=} C^{-1}\exp\left(-\frac{1}{2}D(\phi(x),\mu)\right) \quad \text{with} \quad C = \mathbb{E}_X\left[\exp\left(-\frac{1}{2}D(\phi(X),\mu)\right)\right] \ .$$

[Rot06] proves that when $\gamma \to +\infty$, $p_n(x)$ tends pointwise to a Parzen window density estimator $(1/n)\sum_{i=1}^n \delta(x-x_i)$ where $\delta(\cdot)$ denotes the Dirac function centered at 0. From this observation, [Rot06] deduces that $p_n(x)$ is close to the actual distribution of $X$ when $\gamma, n \to +\infty$ and therefore the Gaussian model in the feature space is accurate. However, this conclusion is questionable since the notion of density is the RKHS of $k$ is not well-defined. Actually, it is proved later in this thesis (see Chapter 4) that for Gaussian RBF kernels, most of the marginals of an embedded variable $\phi(X)$ are close to a *scale-mixture of isotropic Gaussians* when $\gamma \to +\infty$ — that is a random variable of the form $\eta G$ where $G$ is Gaussian and $\eta$ is a positive variable independent of $G$ — which contradicts the Gaussian model of [Rot06].

### 3.4.3 Gaussian process regression

A regression model can be cast in the following general form

$$y = f(x) + \epsilon \ ,$$

where $x$ is an $\mathcal{X}$-valued variable called the *input*, $y$ is a real valued variable called the *output* and $\epsilon$ is an error term. Given a sample $(x_1,y_1),\dots,(x_n,y_n) \in \mathcal{X}\times\mathbb{R}$, the goal is to estimate the regression function $f$, or at least to predict the value $f(x^*)$ for a test point $x^* \in \mathcal{X}$.

[RW06] have introduced *Gaussian process regression* as a method of regression by using Gaussian processes in a Bayesian manner.

To illustrate their method, we consider in a first time the standard linear model

$$\mathcal{X} = \mathbb{R}^d, \qquad y = f(x) + \epsilon \quad \text{with} \quad f(x) = x^T w \ , \tag{3.4.22}$$

for some $w \in \mathbb{R}^d$ and where $\epsilon \sim \mathcal{N}(0,\sigma^2)$, $\sigma^2 > 0$. Instead of pursuing an estimator of $w$ like other regression methods such as ridge regression or Lasso, the approach of [RW06] consists in

setting a prior distribution on $w$ which is

$$w \sim \mathcal{N}(0, \Sigma) \ ,$$

for some covariance matrix $\Sigma \in \mathcal{M}_d(\mathbb{R})$. This prior distribution induces a distribution on the stacked vector $(\mathbf{y} \ \ f(x^*))^T$ where $\mathbf{y} = (y_1 \dots, y_n)^T$ that is

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \\ f(x^*) \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} X^T\Sigma X + \sigma^2 I_n & X^T\Sigma x^* \\ (x^*)^T\Sigma X & (x^*)^T\Sigma x^* \end{bmatrix}\right) \ ,$$

where $X = [x_1 \dots x_n] \in \mathcal{M}_{d,n}(\mathbb{R})$.

Well-known formulas for the conditional distribution of covariates of Gaussian vectors yield the posterior distribution of $f(x^*)$

$$f(x^*) \mid X, \mathbf{y}, x^* \sim \mathcal{N}\Big((x^*)^T\Sigma X(X^T\Sigma X + \sigma I_n)^{-1}\mathbf{y}, \\ (x^*)^T\Sigma x^* - (x^*)^T\Sigma X(X^T\Sigma X + \sigma I_n)^{-1}X^T\Sigma x^*\Big) \ . \tag{3.4.23}$$

An extension of (3.4.22) to a non-linear regression framework can be done by considering a (non-linear) feature map $\phi : \mathcal{X} \to \mathbb{R}^p$ and applying the linear regression method above to the embedded points $\phi(x_1), \dots, \phi(x_n)$, which leads to the regression model

$$y = f(x) + \epsilon \quad \text{with} \quad f(x) = \phi(x)^T w \quad \text{and} \quad w \in \mathbb{R}^p \ . \tag{3.4.24}$$

Introducing the kernel $k$ defined by

$$\forall x, x' \in \mathcal{X}, \ \ k(x, x') \overset{\Delta}{=} \phi(x)^T\Sigma\phi(x') \ ,$$

the kernel trick allows to express the posterior distribution of $f(x^*)$ in this case as

$$f(x^*) \mid X, \mathbf{y}, x^* \sim \mathcal{N}\left(\mathbf{k}_{\mathbf{x}^*}^{\mathbf{T}}(\mathbf{K} + \sigma I_n)^{-1}\mathbf{y}, \ \ k(x^*, x^*) - \mathbf{k}_{\mathbf{x}^*}^{\mathbf{T}}(\mathbf{K} + \sigma I_n)^{-1}\mathbf{k}_{\mathbf{x}^*}\right) \ , \tag{3.4.25}$$

where $\mathbf{k}_{\mathbf{x}^*} = (k(x_1, x^*) \dots k(x_n, x^*))^T$ and $\mathbf{K} = [k(x_i, x_j)]_{1 \leqslant i,j \leqslant n}$.

Note that in the procedure above works as well for more general feature maps $\phi : \mathcal{X} \to \mathcal{H}$ with $\mathcal{H} \neq \mathbb{R}^p$, and in practice we can define directly the kernel $k$ and bypass the explicit definition of $\Sigma$ and $\phi$.

Chapter 4

# Asymptotic Distribution of Random Projections in RKHS

This chapter presents theoretical results concerning embedded distributions in an RKHS induced by a Gaussian RBF kernel. The content of this chapter will be useful for the new outlier detection method introduced later in Chapter 5.

## 4.1 Introduction

Many statistical methods rely on distributional assumptions which often involve normality. This kind of Gaussian assumption can be found in classification [RW06; MM08; BFG15], dimension reduction [Bla+06; Die+10; DJS13] or novelty detection [Rot06]. However in most practical cases, this Gaussian assumption may be questionable. One possible way of dealing with this issue is to transform the dataset to make the normality assumption more reliable.

In the case of real-valued or multivariate data, a few approaches have been considered to get a nearly Gaussian distribution by using Box-Cox transformations [HJ80] or more general power transformations [YJ00]. To handle more general types of data, this transformation can be done through a positive definite kernel. Only a few papers have considered kernels to obtain Gaussian distributions. Among them, [BFG15] proposes to perform supervised and unsupervised classification through a Gaussian mixture model in the kernel space. However, this work does not provide any theoretical guarantee about the normality of embedded data. In [Rot06], novelty detection is conducted by setting a Gaussian density in the kernel space. The validity of their method to detect outliers is justified by relating this method to density estimation of the distribution in the initial space. However, they do not assert straightforwardly that kernelized data are actually close to a Gaussian in general. All in all, there is a lack of understanding about the probabilistic behaviour of random variables embedded into an RKHS.

The purpose of the present chapter is to fill this gap and describe embedded distributions in a parametric class of RKHS. Namely, we study the asymptotic distribution of low-dimensional random projections in such RKHS when the "size" of the RKHS (which is controlled by the parameter of the kernel) increases.

Since RKHS — which are typically high-dimensional — are difficult to handle at first sight, let us resort to the case of $\mathbb{R}^D$ when $D$ is large. A series of several papers, ranging from [DF84] to [DHV06] have studied the distribution of low-dimensional projections of random vectors X in $\mathbb{R}^D$ as $D$ grows to infinity. In the seminal paper of [DF84], unidimensional marginal probability distributions of $X$ are proven to converge weakly to a $\mathcal{N}(0, \sigma^2)$ Gaussian distribution (for some $\sigma^2 > 0$) as $D \to +\infty$, with high probability over the direction of the projection (whose measure is given by the uniform distribution on the unit sphere $\mathbb{S}^{D-1}$). However, this result is based on the strong assumption that $X$ behaves as an isotropic random vector, that is with a covariance matrix of the form $v^2 I$ where $v^2 > 0$ and $I$ denotes the identity matrix. They also provide counter-examples where this isotropic assumption is not satisfied and where normality does not hold. A series of subsequent papers such as [Wei97], [BK03] or more recently [DHV06] have generalized this result in broader frameworks. In particular, [DHV06] considers the case where $X$ is projected on a $p$-dimensional subspace with $p \geqslant 1$ and for more general covariance structures of $X$. In this general setting, the distribution of the marginal is not necessarily Gaussian but an isotropic scale-mixture of Gaussians (SMG), that is a random variable $\mathfrak{s}G$ where $G \sim \mathcal{N}(0, I_p)$ and $\mathfrak{s}$ is a real-valued random variable independent of $G$.

However, most of these findings hold in the finite-dimensional case and thus are not straightforwardly applicable to RKHS-valued variables since RKHS are often infinite-dimensional. The only exception is [Wei97] which deals with random vectors in Hilbert spaces, but their result assumes that the direction of the projection is picked randomly from an isotropic Gaussian process, that is whose covariance operator is the identity operator. This makes their result unusable in the RKHS framework for the following reason. The covariance function of a Gaussian process must satisfy some conditions of so-called *nuclear dominance* [LB01] for its trajectories to lie in an RKHS almost surely (see Section 3.2). These conditions require in particular that the covariance operator of the process must have a finite trace, which does not hold for an isotropic Gaussian process in an infinite-dimensional RKHS. Therefore [Wei97] is not helpful in our case.

In the present work, we propose an extension of the results mentioned above from $\mathbb{R}^D$ to RKHS. We restrict ourselves to the family of Gaussian RBF kernels $k_\gamma(x, y) = \exp(-\gamma d^2(x, y))$ with $\gamma > 0$ where $d(\cdot, \cdot)$ is a metric on the input space $\mathcal{X}$. The advantage of such kernels is that the covariance eigenstructure of an embedded distribution tends to be "flat" when $\gamma$ is large, which is an ideal situation to get convergence to SMG according to results in $\mathbb{R}^D$. For reasons that will be explained later, the parameter $\gamma$ controls the size of the RKHS in the same way $D$ controls that of $\mathbb{R}^D$.

Our contribution is the following. First, we prove the weak convergence of random projections in the RKHS of $k_\gamma$ to a Gaussian scale-mixture as $\gamma \to +\infty$. Secondly, we examine the empirical

framework where the distribution in the input space is represented by an *i.i.d.* sample $X_1, \dots, X_n$ and show that convergence still holds as long as $\gamma$ grows to infinity slowly enough compared to $n$. Finally, we show that when renormalizing properly $k_\gamma$, most random projections in the RKHS converge weakly to a $\mathcal{N}(0, I_p)$ Gaussian distribution instead of a SMG. This may be an advantage in practice since it is no longer required to know the distribution of $\mathfrak{s}$ in the SMG $\mathfrak{s}G$. The latter result holds when considering the empirical distribution and the true underlying distribution as well.

This chapter is outlined as follows. In Section 4.2, previous results on marginal distributions in the multivariate case are detailed and the rationale to extend these results to the RKHS case is given. In Section 4.3, two theorems describing the distribution of random projections in $H(k_\gamma)$ — respectively in non-empirical and empirical cases — are stated. In Section 4.4, the results of Section 4.3 are extended to the case where $k_\gamma$ is adequately renormalized, which gives rise to a Gaussian limiting distribution instead of a SMG. In Section 4.5, we discuss how our results improve on existing results in $\mathbb{R}^D$ and how the loss of information carried by random projections in $H(k_\gamma)$ entails the existence of a "trade-off" value of $\gamma$ in practice. Finally, the proofs of our main theorems are provided in Appendix 4.A.

## 4.2 From $\mathbb{R}^D$ to reproducing kernel Hilbert spaces

In this section, we introduce existing results established in the multivariate case and present some properties of RBF kernels and their RKHS in order to link these two frameworks.

### 4.2.1 Concentration of random projections in $\mathbb{R}^D$

In the following, we state the main result of [DHV06] since it is one of the most recent and general result of this kind to our knowledge.

Let $X$ be some random variable with values in $\mathbb{R}^D$. $X$ is projected onto a subspace generated by $p \geqslant 1$ vectors of $\mathbb{R}^D$ given by the rows of a random matrix $\Theta \in \mathcal{M}_{p,D}(\mathbb{R})$. To simplify, let us assume that the rows of $\Theta$ are independent Gaussian $\mathcal{N}(0, D^{-1} I_D)$ random variables, which are asymptotically orthonormal as $D \to +\infty$. In the following, the probability distribution of $\Theta X$ (conditionally on $\Theta$) is denoted by $f_\Theta$.

Consider the probability distribution $\mu$ corresponding to the random variable $\|X\|_D / \sqrt{D}$ (where $\|.\|_D$ denotes the Euclidean norm in $\mathbb{R}^D$) and define the probability measure $\overline{f}$ which assigns to any Borelian subset $S \subseteq \mathbb{R}^D$

$$\overline{f}(S) \triangleq \int \nu_\sigma(S) \mu(\mathrm{d}\sigma) \ ,$$

where $\nu_\sigma$ denotes the measure of an isotropic Gaussian $\mathcal{N}(0, \sigma^2 I_p)$. $\overline{f}$ corresponds to the distribution of a random variable $\sigma G$, where $G \sim \mathcal{N}(0, I_p)$ and $\sigma$ are independent with $\sigma$ being a copy

of $\|X\|_D/\sqrt{D}$. Such a distribution $\overline{f}$ is called a scale-mixture of (isotropic) Gaussians (SMG).

Let $d(\Theta) = \sup_B |f_\Theta(B) - \overline{f}(B)|$ define the total variation distance between the probability measure $f_\Theta$ and $\overline{f}$ where the supremum is taken over every ball $B$ of $\mathbb{R}^D$. Theorem 4.2.1 proved by [DHV06] gives an upper bound for $d(\Theta)$ which holds true with high probability over the projection matrix $\Theta$.

**Theorem 4.2.1** ([DHV06])**.** *Assume $X$ is centered and admits finite second moments. Let $\lambda_{\max}$ (resp. $\lambda_{avg}$) denote the largest (resp. average) eigenvalue of the covariance matrix of $X$ and let $\sigma_\epsilon$ be the $\epsilon$-quantile of the distribution of $\|X\|_D/\sqrt{D}$. Then for any $\epsilon \in ]0, 1/2[$,*

$$\mathbb{P}_\Theta\left(d(\Theta) > \epsilon\right) \leqslant \left(\frac{C_1 p^2 \lambda_{avg} \ln(1/\epsilon)}{\epsilon^3 \sigma_\epsilon^2}\right)^p \exp\left(-\frac{C_2 \epsilon^4 D \sigma_\epsilon^2}{p \lambda_{\max} \ln(1/\epsilon)}\right) , \tag{4.2.1}$$

*where $\mathbb{P}_\Theta$ is taken with respect to $\Theta$ and $C_1, C_2$ are numerical constants.*

The upper bound in (4.2.1) implies that the weak convergence of finite dimensional Gaussian random projections occurs when $\lambda_{\max}/\sigma_\epsilon^2 = o(D)$ as $D \to +\infty$ (and for any fixed $\epsilon \in ]0, 1/2[$). In [DHV06], the term $\lambda_{\max}/\sigma_\epsilon^2$ is denoted $\mathrm{ecc}(X)$ and is presented as a measure of eccentricity of $X$. In a nutshell $\mathrm{ecc}(X)$ is small when the covariance eigenspectrum of $X$ is flat and large otherwise. For instance when $X$ is isotropic with independent entries (*ie* its covariance is of the form $v^2 I_D$ with $v^2 > 0$), $\lambda_{\max}$ is equal to $v^2$ and $\|X\|_D^2/D$ converges almost surely to $v^2$ as $D \to +\infty$ so that $\mathrm{ecc}(X)$ is close to 1 and the above convergence arises. On the contrary if the distribution of $X$ is supported on a line (for example, $X = Z.e_1$ where $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^D$ and $Z$ is some real-valued r.v.) , $\mathrm{ecc}(X)$ is of order $D$ and the bound of Theorem 4.2.1 does not guarantee the weak convergence toward a SMG.

Note that the result of [DF84] stated in the case where $X$ is isotropic and $p = 1$ is a special case of Theorem 4.2.1. [DF84] asserts that most unidimensional marginals of $X$ converges weakly as $D \to +\infty$ to a Gaussian distribution instead of a SMG. This holds true in the setting of Theorem 4.2.1 since $\|X\|_D^2/D$ converges to a deterministic quantity almost surely under the assumptions of [DF84].

### 4.2.2 Connections with RBF kernel spaces

Our purpose is to extend the result of Theorem 4.2.1 from $\mathbb{R}^D$ to kernel spaces, which will be done in Section 4.3. Since kernel spaces are typically infinite-dimensional, Theorem 4.2.1 is not straightforwardly applicable to RKHS. However, Section 4.2.1 provides two useful ingredients governing the weak convergence of projections in the multivariate case:

1 - the asymptotic behaviour of projections has to be studied with respect to $D$, that is the parameter driving the "size" of the ambient space,

2 - the covariance eigenspectrum of the projected vector has to remain "flat", which corresponds to a small eccentricity $\mathrm{ecc}(X)$.

Let us now consider the class of RBF kernels $k_\gamma$ with $\gamma > 0$ defined as follows

$$\forall x, y \in \mathcal{X},\ \ k_\gamma(x,y) = \exp\left(-\gamma d^2(x,y)\right)\ ,$$

where $\mathcal{X}$ is assumed to be a metric space endowed with a metric $d(.,.)$ [1] . For each $\gamma > 0$, $H(k_\gamma)$ denotes the RKHS corresponding to $k_\gamma$ and $\langle \cdot, \cdot \rangle_\gamma$ denotes the associated inner product.

In the following, we recall two features of this family of kernels (in the case where $d$ is the usual Euclidean metric) that satisfies the two criteria mentioned above and thus hint at an extension of Theorem 4.2.1 to RBF kernel spaces.

First, [Tan+11] states that the RKHSs of RBF kernels are nested, that is

$$\forall \gamma_1, \gamma_2 \in \mathbb{R}_+^*,\ \ \gamma_1 < \gamma_2 \Longrightarrow H(k_{\gamma_1}) \subseteq H(k_{\gamma_2})\ .$$

In this sense, the parameter $\gamma$ controls the size of the RKHS $H(k_\gamma)$. Thus increasing $\gamma$ - that is enlarging $H(k_\gamma)$ - may have the same effect as increasing $D$ in Theorem 4.2.1.

Secondly, we show that the second criterion is met. Let $\lambda_1 \geqslant \lambda_2 \geqslant \dots$ denote the eigenvalues of the covariance operator [2] $\mathbb{E}k_\gamma(X,.) \otimes k_\gamma(X,.)$. When $X$ is a univariate $\mathcal{N}(0, \nu^2)$ variable ($\nu^2 > 0$), it is known (see [RW06], p. 97) that for any $k \geqslant 1$

$$\lambda_k = \frac{\gamma^k / \sqrt{2\nu^2}}{\left((4\nu^2)^{-1} + \gamma + \sqrt{(16\nu^4)^{-1} + \gamma(2\nu^2)^{-1}}\right)^{k+1/2}} \underset{\gamma \to +\infty}{\sim} \frac{\gamma^{-1/2}}{\sqrt{2\nu^2}}\ .$$

Note that the asymptotic expression of $\lambda_k$ as $\gamma \to +\infty$ does not depend on $k$. Therefore the covariance eigenspectrum of $k_\gamma(X,.)$ becomes flat as $\gamma$ grows to infinity, hence a minimum eccentricity. This fact can be extended for more general distributions of $X$ as shows Proposition 4.2.2 that is proved in Appendix 4.B.1.

**Proposition 4.2.2.** *Let $(\lambda_r)_r$ be the eigenvalues of the covariance operator $\Sigma_\gamma = \mathbb{E}_X k_\gamma(X,.)^{\otimes 2}$. Let $\mathrm{supp}(X) = \{x \in \mathcal{X} \mid \forall \epsilon > 0,\ \ \mathbb{P}(d^2(x,X) < \epsilon) > 0\}$. Assume that the distribution of $X$ admits no point mass and that there exists a function $A : \mathrm{supp}(X) \to \mathbb{R}_+^*$ and $s > 0$ such that*

$$\forall x \in \mathrm{supp}(X)\ ,\ \ \mathbb{P}\left(d^2(x,X) \leqslant t\right) \sim A(x) t^s\ \ ,\ \ \textit{when } t \to 0\ .$$

*Then for any integer $r$,*

$$\lambda_r \sim \frac{\mathbb{E}_X[A(X)]\Gamma(s+1)}{\gamma^s}\ \ ,\ \ \textit{when } \gamma \to +\infty\ ,$$

[1] Note that the metric $d$ must be chosen such that the resulting kernel $k_\gamma$ is definite positive. This condition is fulfilled in the general case where $d(x,x') = \|\varphi(x) - \varphi(x')\|_{\mathbb{H}}$ for any $x, x' \in \mathcal{X}$, where $\varphi : \mathcal{X} \to \mathbb{H}$ and $\mathbb{H}$ is a separable Hilbert space.

[2] Given two vectors $u, v \in \mathcal{H}$ where $\mathcal{H}$ denotes some prehilbertian space with dot product $\langle .,. \rangle$, the tensor product $u \otimes v$ is defined as the operator $u \otimes v : \mathcal{H} \to \mathcal{H}$ such that for any $w \in \mathcal{H}$, $(u \otimes v) \cdot w = \langle v, w \rangle \cdot u$.

*where* $\Gamma(s) = \int_0^{+\infty} x^{s-1} e^{-x} \mathrm{d}x$ *denotes the Gamma function.*

Therefore under the milder assumptions on $X$ required by Proposition 4.2.2, $(\lambda_r)_{r\geqslant 1}$ admits a same asymptotic expression hence a "flat" eigenspectrum.

## 4.3 Asymptotic distribution of random projections

### 4.3.1 Notation

Let us further introduce additional notation needed for our main statements. Let $X$ be a $\mathcal{X}$-valued random variable with distribution $P$ where $\mathcal{X}$ denotes some set. Using the notation introduced in Section 4.2.2, we consider the Radial Basis Function (RBF) kernel $k_\gamma$ parameterized by $\gamma > 0$, its RKHS $H(k_\gamma)$ with corresponding inner product $\langle .,. \rangle_\gamma$ and norm $\|.\|_\gamma$. As a shorthand, we denote the embedded variable $k_\gamma(X,.)$ taking values in $H(k_\gamma)$ as $k_\gamma(X,.) = k_X$.

Following [DHV06], we examine the distribution of projections of $k_X$ onto $p$-linear subspaces of $H(k_\gamma)$ denoted by $V = \mathrm{Span}(h_1,\dots,h_p)$ where $h_1,\dots,h_p \in H(k_\gamma)$ are random vectors. Ideally $h_1,\dots,h_p$ should form an orthonormal family of $H(k_\gamma)$ almost surely. For the sake of simplicity, we follow an approach similar as [DHV06] and consider $h_1,\dots,h_p$ $p$ independent zero-mean Gaussian processes of covariance $\Sigma_\gamma$, where $\Sigma_\gamma = \mathbb{E}_X k_X \otimes k_X$ stands for the (non-centered) covariance operator of $k_X$ (to avoid additional notation, we also note $\Sigma_\gamma(x,x') = \langle \Sigma_\gamma k_\gamma(x,.), k_\gamma(x',.) \rangle_\gamma$ for any $x,x' \in \mathcal{X}$)[3]. In this case, orthonormality holds true asymptotically as $\gamma \to +\infty$. This is guaranteed by Lemma 4.3.1 whose proof is provided in Appendix 4.B.2.

**Lemma 4.3.1.** *Consider $p$ independent zero-mean Gaussian processes $h_1,\dots,h_p$ with covariance $\Sigma_\gamma = \mathbb{E}_{X\sim P} k_\gamma(X,.)^{\otimes 2}$ for some probability measure $P$. Then these $p$ variables are asymptotically orthonormal when $\gamma \to +\infty$, that is*

$$\forall 1 \leqslant i \leqslant p, \ \ \|h_i\|_\gamma^2 \underset{\gamma\to+\infty}{\longrightarrow} 1 \quad a.s. \quad ,$$

*and*

$$\forall 1 \leqslant i,j \leqslant p, \ \ i \neq j, \ \ \langle h_i, h_j \rangle_\gamma \underset{\gamma\to+\infty}{\longrightarrow} 0 \quad a.s. \ .$$

The "projection" of $k_X$ onto $V$ is denoted

$$p_V(X) = \begin{pmatrix} \langle k_X, h_1 \rangle_\gamma \\ \vdots \\ \langle k_X, h_p \rangle_\gamma \end{pmatrix} \in \mathbb{R}^p \quad ,$$

[3] According to Section 3.2, those Gaussian processes are well defined in $H(k_\gamma)$ if $H(\Sigma_\gamma) \subseteq H(k_\gamma)$ and $\mathrm{Tr}(\Sigma_\gamma) < +\infty$. The latter condition is easily checked as follows: $\mathrm{Tr}(\Sigma_\gamma) = \mathrm{Tr}(\mathbb{E}_X k_X \otimes k_X) = \mathbb{E}_X \mathrm{Tr}(k_X \otimes k_X) = \mathbb{E}_X \|k_X\|^2 = \mathbb{E} k_\gamma(X,X) = 1 < +\infty$. As for the former condition, it suffices to apply the last assertion of Definition 3.2.2 to the operator $\Sigma_\gamma$.

We also consider the empirical case when $P$ is not known and an *i.i.d.* sample $X_1,\dots,X_n \sim P$ is available. In this case, the covariance operator $\Sigma_\gamma$ is not known thus we resort to the empirical covariance $\Sigma_{\gamma,n} = (1/n)\sum_{i=1}^n k_\gamma(X_i,.)^{\otimes 2}$. Therefore we generate instead $p$ independent (conditionally to the sample) Gaussian processes $h_{1,n},\dots,h_{p,n}$ with mean zero and covariance $\Sigma_{\gamma,n}$. In practice this is done by setting $h_{j,n} = (1/n)\sum_{i=1}^n u_i^{(j)} k_\gamma(X_i,.)$ for each $1 \leqslant j \leqslant p$, where $(u_i^{(j)})_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}$ are *i.i.d.* $\mathcal{N}(0,1)$. The linear subspace spanned by $h_{1,n},\dots,h_{p,n}$ is denoted $V_n$ and the projection of $k_X$ onto $V_n$ (where $X$ is independent of the sample $X_1,\dots,X_n$) is denoted

$$p_{V_n}(X) = \begin{pmatrix} \langle k_X, h_{1,n}\rangle_\gamma \\ \vdots \\ \langle k_X, h_{p,n}\rangle_\gamma \end{pmatrix} \in \mathbb{R}^p \ .$$

Let $\mathrm{Tr}(\Sigma_\gamma^2) = \mathbb{E}_X \Sigma_\gamma(X,X)$. Our main results in next sections focus on the discrepancy between the distributions of $[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2} p_V(X)$ (resp. $[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2} p_{V_n}(X)$) and that of some scale-mixture of Gaussians $\mathfrak{s}G$ where $G \sim \mathcal{N}(0, I_p)$ and $\mathfrak{s}$ is some random vector independent of $G$ (The normalization of $p_V(X)$ and $p_{V_n}(X)$ is meant to prevent the variance from decreasing to 0).

Let $\omega$ be some random variable on $\mathbb{R}^p$ and $\phi_W(\omega) = \mathbb{E}_W \exp(i\langle W, \omega\rangle_p)$ denote the characteristic function of $W$ for any r.v. $W$ [4]. If the distribution of $\omega$ is fully supported on $\mathbb{R}^p$, then we can consider the following $L_2$-distance between distributions on $\mathbb{R}^p$ [Sri+10]: for any pair of $p$-variate random variables $W, W'$,

$$\Delta(W,W') \overset{\Delta}{=} \left\{ \mathbb{E}_\omega \left| \phi_W(\omega) - \phi_{W'}(\omega) \right|^2 \right\}^{1/2} \ .$$

### 4.3.2 Main assumptions

The results stated in Section 4.3.3 and Section 4.3.4 assume that the following assumptions hold:

(A1) $\mathcal{X}$ is a separable metric space endowed with a metric $d(.,.)$,

(A2) There exists a continuous, bounded function $A : \mathcal{X} \to \mathbb{R}_+$ and $s > 0$ such that

$$\forall x \in \mathrm{supp}(P) \ , \ \mathbb{P}\left(d^2(x,X) \leqslant t\right) \sim A(x) t^s \ , \quad \text{when } t \to 0 \ ,$$

where $\mathrm{supp}(P) = \{x \in \mathcal{X} \mid \forall \epsilon > 0, \ \mathbb{P}(d^2(x,X) < \epsilon) > 0\}$ defines the support of the distribution of $X$, and $A(x) = 0$ otherwise,

(A3) $\mathrm{supp}(P)$ is a compact subset of $\mathcal{X}$,

(A4) $(t,x) \mapsto \mathbb{P}_X(d^2(x,X) \leqslant t)$ is continuous.

The separability condition required by Assumption (A1) is made so that $\Sigma_\gamma$ admits a discrete eigenexpansion, that is there exists a sequence of eigenvalues $(\lambda_r)_{r\geqslant 1} \in (\mathbb{R}_+)^{\mathbb{N}^*}$ and eigenvectors

[4] Here $\langle .,. \rangle_p$ denotes the usual inner product of $\mathbb{R}^p$

$(\psi_r)_{r\geqslant 1} \in (H(k_\gamma))^{\mathbb{N}^*}$ such that $\Sigma_\gamma = \sum_{r\geqslant 1} \lambda_r \psi_r \otimes \psi_r$ and $(\psi_r)_{r\geqslant 1}$ form an orthonormal family of vectors of $H(k_\gamma)$. It is known (see e.g. Lemma 4.33 in [SC08]) that the RKHS $H(k_\gamma)$ is separable if $k_\gamma$ is continuous (straightforwardly this is true) and if the input space $\mathcal{X}$ is a separable topological space. Therefore under Assumption (A1) $H(k_\gamma)$ is separable and since the operator $\Sigma_\gamma : H(k_\gamma) \to H(k_\gamma)$ is compact and self-adjoint, it admits a discrete eigenspectrum.

Assumption (A2) may seem strong, especially because of the non-dependence between the exponent $s$ and $x$. But this assumption turns out to be mostly true when $\mathcal{X} = \mathbb{R}^D$ endowed with the Euclidean metric $d(.,.) = \|. - .\|$, as stated by Proposition 4.3.2.

**Proposition 4.3.2.** *Assume $\mathcal{X} = \mathbb{R}^D$ and $d(\cdot,\cdot) = \|\cdot - \cdot\|$ is the usual Euclidean distance. Let $f$ denote the density function of $X$ with respect to the Lebesgue measure $\mu$ of $\mathbb{R}^D$ and assume $f$ is continuous and bounded on its support. Then Assumption (A2) holds with $A(x) = \pi^{D/2} f(x)/\Gamma(D/2+1)$ and $s = D/2$.*

*Proof of Proposition 4.3.2.* Let $x \in \mathcal{X}$ and $B(x,\sqrt{t})$ denote the ball of radius $\sqrt{t}$ and center $x$. Since $f$ is a density function, $f \in L_1(\mu)$ where $\mu$ is the Lebesgue measure of $\mathbb{R}^D$ and Lebesgue's differentiation theorem entails for $\mu$-almost every $x \in \mathrm{supp}(f)$

$$\frac{\mathbb{P}(\|x-X\|^2 \leqslant t)}{\mu\Big(B(x,\sqrt{t})\Big)} = \frac{\int_{B(x,\sqrt{t})} f(y)\mu(dy)}{\mu\Big(B(x,\sqrt{t})\Big)} \underset{t\to 0}{\longrightarrow} f(x) \ ,$$

hence

$$\mathbb{P}(\|x-X\|^2 \leqslant t) \underset{t\to 0}{\sim} \mu\Big(B(x,\sqrt{t})\Big) f(x) = \frac{\pi^{D/2} t^{D/2} f(x)}{\Gamma(D/2+1)} \ ,$$

which correspond to Assumption (A2) with $A(x) = \pi^{D/2} f(x)/\Gamma(D/2+1)$ and $s = D/2$. This holds for $x \in \mathcal{S} \subseteq \mathrm{supp}(f)$ for some $\mathcal{S}$ satisfying $\mu(\mathrm{supp}(f) \setminus \mathcal{S}) = 0$. Thus the continuity of $f$ on its support implies that the result also holds for every $x \in \mathrm{supp}(f)$. □

Note that this extends to more general metrics in $\mathbb{R}^D$. For instance, consider Mahalanobis distances $d(x,y) = \sqrt{(x-y)^\top \mathbf{S}^{-1}(x-y)}, x, y \in \mathbb{R}^D$ where $\mathbf{S} \in \mathcal{M}_D(\mathbb{R})$ is typically a covariance matrix. Applying Proposition 4.3.2 to the vector $\tilde{X} = \mathbf{S}^{-1/2}X$ straightforwardly entails that Assumption (A2) holds true for $X$. More generally one may consider metrics of the form

$$d(x,y) = \|\varphi(x) - \varphi(y)\| \ , \tag{4.3.2}$$

for some $\varphi : \mathcal{X} \to \mathbb{R}^q$ and $q \in \mathbb{N}^*$. Applying Proposition 4.3.2 to $\varphi(X)$ implies that Assumption (A2) holds with $A(x) = \pi^{q/2} f_{\varphi(X)}(x)/\Gamma(q/2+1)$ and $s = q/2$, where $f_{\varphi(X)}$ denotes the density of $\phi(X)$. Conversely, if $d^2(\cdot,\cdot)$ is of negative type [5] , $d(\cdot,\cdot)$ admits the form (4.3.2) where $\varphi : \mathcal{X} \to \mathcal{H}$ is an injective map and $\mathcal{H}$ is some Hilbert space (Proposition 3 in [Sej+13]), which makes Assumption (A2) true if $\mathcal{H}$ is of finite dimension.

[5] A function $\ell : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is of negative type if for every $x_1,\dots,x_n \in \mathcal{X}$ and $a_1,\dots,a_n \in \mathbb{R}$ such that $\sum_{i=1}^n a_i = 0$, $\sum_{i,j=1}^n a_i a_j \ell(x_i,x_j) \leqslant 0$.

On the other hand, Assumption (A2) may be false especially when $\mathcal{X}$ is infinite dimensional. For example, consider $\mathcal{X} = \mathcal{C}([0,1],\mathbb{R})$ the space of continuous functions $g : [0,1] \to \mathbb{R}$ endowed with the supremum norm $|g|_\infty = \sup_{t\in[0,1]}|g(t)|$. Assume $X = (X(t))_{0\leqslant t\leqslant 1}$ is a fractional Brownian motion, that is $X(0) = 0$ a.s. and for any $s,t \in [0,1]$, $\mathbb{E}|X(s) - X(t)|^2 = |s-t|^\alpha$ for some $\alpha \in (0,2)$. According to Theorem 4.6 from [LS01], there exists $0 < K_1 \leqslant K_2 < +\infty$ only depending on $\alpha$ such that for any $\epsilon \in (0,1]$

$$\exp(-K_2\epsilon^{-2/\alpha}) \leqslant \mathbb{P}(|X|_\infty \leqslant \epsilon) \leqslant \exp(-K_1\epsilon^{-2/\alpha}) \ , \tag{4.3.3}$$

so that small ball probabilities in Assumption (A2) decrease to 0 at exponential rate instead of polynomial rate. This hints at the idea that the polynomial rate of decay required by Assumption (A2) only corresponds to the case where $\mathcal{X}$ is of finite dimensional nature. Note that this does not imply that the new convergence results presented in Section 4.3.3 and Section 4.3.4 do not hold true when $X$ is infinite-dimensional, even though the rates of decrease of the upper bounds might change. Besides the small ball probability in (4.3.3) only corresponds to a specific case (*i.e.* fractional Brownian motions). To the best of our knowledge, it seems difficult to provide a general expression that holds for most infinite-dimensional distributions, unlike that of Assumption (A2) which holds in all generality in the finite dimensional case. For this reason we restrict our analysis to finite dimensional input spaces, even if it means to preprocess infinite-dimensional data in practice through dimension reduction techniques (*e.g.* functional principal component analysis [JR92]).

### 4.3.3 Asymptotic distribution when $P$ is known

We are now in a position to provide our main Theorem 4.3.3, which describes the asymptotic distribution of $[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2}p_V(X)$ as $\gamma \to +\infty$ when $P$ is known. The proof of Theorem 4.3.3 is provided in Appendix 4.A.1.

**Theorem 4.3.3.** *Let us assume (A1), (A2), (A3), (A4) and $\mathbb{E}_\omega\|\omega\| < +\infty$ where $\|\cdot\|$ stands for the Euclidean norm of $\mathbb{R}^p$, and define a real-valued random variable $\mathfrak{s}^2$ by*

$$\mathfrak{s}^2 \overset{law}{=} \frac{\Sigma_\gamma(X,X)}{\mathrm{Tr}(\Sigma_\gamma^2)} \ .$$

*Then for every $\delta \in (0,1)$, there exists an event with probability larger than $1-\delta$ on which*

$$\Delta^2\left(\frac{p_V(X)}{\sqrt{\mathrm{Tr}(\Sigma_\gamma^2)}}, \mathfrak{s}G\right) \leqslant 8\mathbb{E}_\omega[\|\omega\|]\sqrt{p\mathbb{E}_X[A(X)]\Gamma(s+1)\log\left(\frac{2(p+1)}{\delta}\right)}\left(\frac{\gamma}{2}\right)^{-s/2}[1+o_\gamma(1)] \ , \tag{4.3.4}$$

*where $G \sim \mathcal{N}(0, I_p)$ and where $A : \mathcal{X} \to \mathbb{R}_+$ and $s \in \mathbb{R}_+^*$ are defined as in Assumption (A2) and $\Gamma(.)$ denotes the Gamma function $\Gamma(u) = \int_0^{+\infty} x^{u-1}e^{-x}\mathrm{d}x$, $\forall u \geqslant 0$.*

In the special case $\mathcal{X} = \mathbb{R}^D$ with $d(.,.)$ as the Euclidean distance, Proposition 4.3.2 provides $s = D/2$ and $A(x) = \pi^{D/2} f(x)/\Gamma(D/2+1)$, hence the upper bound in (4.3.4) becomes

$$\Delta^2\left(\frac{p_V(X)}{\sqrt{\mathrm{Tr}(\Sigma_\gamma^2)}}, \mathfrak{s}G\right) \leqslant 8\mathbb{E}_\omega[\|\omega\|]\sqrt{p\log\left(\frac{2(p+1)}{\delta}\right)}|f|_2\left(\frac{2\pi}{\gamma}\right)^{D/4}[1+o_\gamma(1)] \ , \tag{4.3.5}$$

with high probability over $h_1, \ldots, h_p$, where $f$ denotes the density function of the distribution $P$ of $X$ and $|f|_2 = (\int f^2)^{1/2}$ denotes the $\ell_2$-norm.

The upper bound in (4.3.5) hints at the influence of some key quantities on the distributional approximation of $p_V(X)$.

A large value of $|f|_2$ yields a loose bound which suggests that in this case weak convergence does not occur. This observation may be justified by the following rationale: a large $|f|_2$ corresponds intuitively to a distribution $P$ concentrated around a few "peaks". Consider the limit case where $|f|_2$ tends to infinity and $P$ gets close to a discrete distribution or at least admits a point mass. If $P$ admits a point mass $x_0 \in \mathcal{X}$ such that $\mathbb{P}(X = x_0) = p > 0$, then $\mathbb{P}(p_V(X) = p_V(x_0)) \geqslant p > 0$ and the limit distribution of $p_V(X)$ would also admit a point mass which is not compatible with the asymptotic distribution foretold by Theorem 4.3.3. Therefore the influence of $|f|_2$ displayed by the bound in (4.3.5) is consistent.

Another way to explain the role of $|f|_2$ is to relate $|f|_2$ with the covariance eigenspectrum in $H(k_\gamma)$. Remember that a small eccentricity of the covariance eigenspectrum fosters the convergence to a SMG, as Theorem 4.2.1 suggests in $\mathbb{R}^D$. When $|f|_2$ is large and $P$ close to a discrete distribution (let us assume that the support is of cardinality $m \in \mathbb{N}^*$), the support of the distribution of $k_\gamma(X,.)$ is contained in a subspace of $H(k_\gamma)$ of finite dimension at most $m$. Therefore the covariance eigenvalues of $k_\gamma(X,.)$ of order larger than $m$ vanish to 0 and weak convergence to a SMG may not happen.

Besides since the upper bound displays a term of order $\gamma^{-D/4}$, it is tempting to conclude that a high-dimensional input space fosters the weak convergence of $p_V(X)$. However, one has to be careful: when $D$ varies, the input space - and consequently $P$ and $|f|_2$ - do also change. Thus a decreasing value of $\gamma^{-D/4}$ might be canceled out by a high value of $|f|_2$. To see this, let us consider a simple example. Assume $X$ follows a multivariate Gaussian $\mathcal{N}(0, \mathbf{S})$ distribution in the input space, where the covariance matrix $\mathbf{S} = \mathrm{diag}(l_1^{(D)}, \ldots, l_D^{(D)})$ is a diagonal matrix with $l_1^{(D)} \geqslant l_2^{(D)} \geqslant \ldots \geqslant l_D^{(D)} > 0$. The density function $f$ of $P$ is defined as

$$f(x) = (2\pi)^{-D/2}\det(\mathbf{S})^{-1/2}\exp(-(1/2)x^T\mathbf{S}^{-1}x)$$

for every $x \in \mathbb{R}^D$. Then straightforward algebra yields

$$|f|_2 = (4\pi)^{-D/4}\det^{-1/4}(\mathbf{S}) \geqslant (4\pi l_1^{(D)})^{-D/4} \ ,$$

and choosing $l_1^{(D)}$ such that $\sup_{D\geqslant 1} l_1^{(D)} < 1/(2\gamma)$ leads to a loose upper bound in (4.3.5) as $D$

grows to infinity.

A special case of Theorem 4.3.3 is when $P$ is a uniform measure on some subset $\mathcal{S} \subset \mathbb{R}^D$. In this setting, $\mathfrak{s}$ converges almost surely to 1 as $\gamma$ grows to infinity hence the asymptotic distribution of $[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2} p_V(X)$ is Gaussian with mean 0 and covariance matrix $I_p$. This is stated by Corollary 4.3.4 thereafter.

**Corollary 4.3.4.** *Let us use the same notation and assumptions as in Theorem 4.3.3. Assume $\mathcal{X} = \mathbb{R}^D$ and $P$ is the uniform measure on some measurable subset $\mathcal{S} \subset \mathbb{R}^D$ with finite Lebesgue measure $\mu(\mathcal{S}) > 0$. Then there exists an event with probability larger than $1 - \delta$ for any $\delta \in (0,1)$ on which $[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2} p_V(X)$ converges weakly to $\mathcal{N}(0, I_p)$ as $\gamma \to +\infty$.*

*Proof.* It suffices to show that $\mathfrak{s}^2 \underset{\gamma \to +\infty}{\longrightarrow} 1$ almost surely. Lemma A.2.1 combined with Proposition 4.3.2 yield

$$\begin{aligned}\Sigma_\gamma(X,X) = \frac{A(X)\Gamma(s+1)}{(2\gamma)^s}[1 + o_\gamma(1)] = f(X)\left(\frac{\pi}{2\gamma}\right)^{D/2}[1 + o_\gamma(1)] \\ = \mu^{-1}(\mathcal{S})\left(\frac{\pi}{2\gamma}\right)^{D/2}[1 + o_\gamma(1)]\ ,\end{aligned}$$

and Lemma A.2.2 provides

$$\begin{aligned}\mathrm{Tr}(\Sigma_\gamma^2) = \frac{\mathbb{E}A(X)\Gamma(s+1)}{(2\gamma)^s}[1 + o_\gamma(1)] = \left(\int_{\mu(\mathcal{S})} f^2(x)\mathrm{d}x\right)\left(\frac{\pi}{2\gamma}\right)^{D/2}[1 + o_\gamma(1)] \\ = \mu^{-1}(\mathcal{S})\left(\frac{\pi}{2\gamma}\right)^{D/2}[1 + o_\gamma(1)]\ ,\end{aligned}$$

□

which proves that $\mathfrak{s}^2 = 1 + o_\gamma(1)$ as $\gamma \to +\infty$. Theorem 4.3.3 and Slutsky's lemma allow to conclude.

### 4.3.4 Asymptotic distribution when $P$ is unknown

In this section, we consider the empirical case where $P$ is unknown and only an *i.i.d.* sample $X_1, \ldots, X_n$ is available. The question is whether the convergence in distribution to the scale-mixture of Gaussian as stated in Theorem 4.3.3 still holds when $\gamma$ and $n$ both grow to infinity. Theorem 4.3.5 states that this is the case under the additional condition $\gamma = o(n^{1/s})$. The proof of Theorem 4.3.5 is provided in Appendix 4.A.2.

**Theorem 4.3.5.** *Let us assume (A1), (A2), (A3), (A4) and $\mathbb{E}_\omega\|\omega\| < +\infty$, and let $\mathfrak{s}^2$ be defined as in Theorem 4.3.3. Then for every $\delta \in (0,1)$, there exists an event with probability larger than $1 - 4\delta$ on*

*which*

$$\Delta^2\left(\frac{p_{V_n}(X)}{\sqrt{\mathrm{Tr}(\Sigma^2)}}, \mathfrak{s}G\right) \leqslant C_1\gamma^{-s/2} + C_2\gamma^{s/2}n^{-1/2} \ ,$$

*where* $G \sim \mathcal{N}(0, I_p)$ *and*

$$C_1 = 2^{(s+6)/2}\mathbb{E}\|\omega\|\sqrt{p\log\left(\frac{2(p+1)}{\delta}\right)\mathbb{E}A(X)\Gamma(s+1)}\{1+o_\gamma(1)\}$$

$$C_2 = \frac{1}{\sqrt{\mathbb{E}A(X)\Gamma(s+1)}}\left[\frac{2\mathbb{E}[\|\omega\|^2]\sqrt{\log(1/\delta)}[1+\epsilon_{\gamma,n}]}{1-\delta/2} + 4\mathbb{E}[\|\omega\|]\sqrt{p\log\left(\frac{2[p+1]}{\delta}\right)}\left(2^{s/2+1} + 2^{s/2}\cdot 3\cdot\sqrt{\log(3/\delta)}\right)\right] \ .$$

*with* $\epsilon_{\gamma,n} \to 0$ *as* $\gamma \to +\infty$ *and* $\gamma^s = o(n)$.

The fact that $\gamma$ must be small compared to $n$ for convergence to hold can be understood as follows. Since $h_1, \ldots, h_p$ are Gaussian processes with covariance $\Sigma_{\gamma,n}$, they take values almost surely in the $n$-dimensional subspace $S_n$ of $H(k_\gamma)$ generated by $k_\gamma(X_1,.), \ldots, k_\gamma(X_n,.)$ and $V_n$ is included in $S_n$ as a consequence. If $n$ is fixed and $\gamma$ grows to infinity, each embedded point $k_\gamma(x,.)$ with $x \in \mathcal{X} \setminus \{X_1, \ldots, X_n\}$ satisfies by the reproducing property $\langle k_\gamma(x,.), k_\gamma(X_i,.)\rangle_\gamma = k_\gamma(X_i, x) = \exp(-\gamma d^2(X_i, x)) \underset{\gamma\to+\infty}{\longrightarrow} 0$ for any $i = 1, \ldots, n$. Therefore in this case $k_\gamma(x,.)$ tends to be orthogonal to $S_n$ and thus to $V_n$, and the distribution of $p_{V_n}(X)$ is highly concentrated around the origin and is far from a scale-mixture of Gaussians. This explains why $n$ must be large enough compared to $\gamma$ to counter this effect.

Straightforwardly, Corollary 4.3.4 still holds in the empirical case and a uniform distribution in the input space yields a $\mathcal{N}(0, I_p)$ Gaussian asymptotic distribution for $p_{V_n}(X)$.

## 4.4 Asymptotic distribution when the kernel is renormalized

In Section 4.3, random projections in the RKHS $H(k_\gamma)$ have been proved to converge weakly to an SMG $\mathfrak{s}G$ where $\mathfrak{s}^2$ is equal to $\Sigma_\gamma(X, X)$ *in distribution*. However, if the latter equality holds *P-almost surely*, one gets a convergence in law to a $\mathcal{N}(0, I_p)$ Gaussian by normalizing the embedded variable $k_X$ by $\sqrt{\Sigma_\gamma(X, X)}$ — which amounts to consider the renormalized kernel $\tilde{k}_\gamma$

$$\tilde{k}_\gamma(x, y) \triangleq \frac{k_\gamma(x, y)}{\sqrt{\Sigma_\gamma(x, x)\Sigma_\gamma(y, y)}} \ , \tag{4.4.6}$$

where $\Sigma_\gamma(\cdot,\cdot)$ is replaced by $\Sigma_{\gamma,n}(\cdot,\cdot)$ in the empirical framework.

In this case, we get a simpler limiting distribution that is parameter-free, whereas for $k_\gamma$ the asymptotic distribution involves the distribution of $\Sigma_\gamma(X,X)$ that must be estimated in practice.

The present section is devoted to show that this Gaussian convergence holds when renormalizing $k_\gamma$ as in (4.4.6) — both in the non-empirical and in the empirical case. In the non-empirical case, we study the asymptotic distribution of projections $\tilde{p}_V(X)$

$$\tilde{p}_V(X) = \begin{pmatrix} \langle \tilde{k}_X, h_1 \rangle_\gamma \\ \vdots \\ \langle \tilde{k}_X, h_p \rangle_\gamma \end{pmatrix} \quad \text{with} \quad \tilde{k}_X = \left\{ \Sigma_\gamma(X,X) \right\}^{-1/2} k_X \ ,$$

and in the empirical case we study the projections $\tilde{p}_{V_n}(X)$

$$\tilde{p}_{V_n}(X) = \begin{pmatrix} \langle \tilde{k}_X^{(n)}, h_{1,n} \rangle_\gamma \\ \vdots \\ \langle \tilde{k}_X^{(n)}, h_{p,n} \rangle_\gamma \end{pmatrix} \quad \text{with} \quad \tilde{k}_X^{(n)} = \left\{ \Sigma_{\gamma,n}(X,X) \right\}^{-1/2} k_X \ .$$

Most of the notation and assumptions from Section 4.3 are continued in this section.

### 4.4.1 Asymptotic distribution when $P$ is known

Theorem 4.4.1 thereafter provides an upper bound for the discrepancy between the distribution of $\tilde{p}_V(X)$ as defined above and the $\mathcal{N}(0, I_p)$ Gaussian distribution. This bound decreases to 0 when $\gamma \to +\infty$ — assessing the claimed asymptotic distribution of $\tilde{p}_V(X)$. The proof of this theorem is detailed in Appendix 4.A.3.

**Theorem 4.4.1.** *Let us assume (A1), (A2), (A3), (A4) where $A : \mathcal{X} \to \mathbb{R}_+$ as defined in (A2) satisfies $0 < a \leqslant A(x) \leqslant A < +\infty$ for every $x \in \mathrm{supp}(P)$. Also assume that $\mathbb{E}_\omega \|\omega\| < +\infty$.*

*Then for every $\delta \in (0,1)$, there exists an event with probability larger than $1-\delta$ on which*

$$\Delta^2\left(\tilde{p}_V(X), G\right) \leqslant 8\mathbb{E}\|\omega\| \sqrt{pA\Gamma(s+1)\log\left(\frac{2(p+1)}{\delta}\right)} \left(\frac{\gamma}{2}\right)^{-s/2} \left[1 + o_\gamma(1)\right] \ , \tag{4.4.7}$$

*where $G \sim \mathcal{N}(0, I_p)$, $s \in \mathbb{R}_+^*$ is the same quantity introduced in Assumption (A2) and $\Gamma(.)$ denotes the Gamma function $\Gamma(u) = \int_0^{+\infty} x^{u-1} e^{-x} \mathrm{d}x$, $u \geqslant 0$.*

Note the additional assumption that the function $A(\cdot)$ is bounded from below by a positive quantity $a$. This assumption is due to the renormalization of the projection $p_V(X)$ by $\Sigma_\gamma^{1/2}(X,X)$ that needs to be lower bounded. Namely we use the equivalent term $\Sigma_\gamma(X,X) \underset{\gamma \to +\infty}{\sim} A(X)\Gamma(s+1)(2\gamma)^{-s}$ (provided by Lemma A.2.1), hence a natural lower bound equivalent to $a\Gamma(s+1)\gamma^{-s}$.

### 4.4.2 Asymptotic distribution when $P$ is unknown

Theorem 4.4.2 extends Theorem 4.4.1 to the empirical case. The proof of Theorem 4.4.2 can be found in Appendix 4.A.4.

**Theorem 4.4.2.** *Let us assume (A1), (A2), (A3), (A4) where $A : \mathcal{X} \to \mathbb{R}_+$ as defined in (A2) satisfies $0 < a \leqslant A(x) \leqslant A < +\infty$ for every $x \in \mathrm{supp}(P)$. Also assume that $\mathbb{E}_\omega\|\omega\| < +\infty$.*

*Then for every $\delta \in (0,1)$, there exists an event with probability larger than $1-\delta$ on which*

$$\Delta^2\left(\tilde{p}_{V_n}(X), G\right) \leqslant \mathbb{E}\|\omega\|^2 \left\{ \frac{A}{a}\frac{\Gamma(s+1)}{(2\gamma)^s} + \sqrt{\frac{A^2\Gamma(s+1)\log(1/\delta)}{2^{1+2s}a^3\gamma^s n}} + \frac{2^{(1-s)/2}\kappa A \log(1/\delta)}{a^2 n} \right\} \min\left(\frac{1}{\delta}, 1+\xi_{\gamma,n}\right)^{[1+o_\gamma(1)]} , \tag{4.4.8}$$

*where $s \in \mathbb{R}_+^*$ is the same quantity introduced in Assumption (A2), $\kappa$ is a numerical constant and $\xi_{\gamma,n}$ is a quantity that tends to 0 when $\gamma, n \to +\infty$ and $\gamma = o(n^{3/2s})$.*

Unlike the case where $k_\gamma$ is not renormalized, the remarkable novelty is that the upper bound in (4.4.8) suggests that weak convergence holds when $\gamma, n \to +\infty$ even without any constraint on the growth of $\gamma$ with respect to $n$. Empirical investigation in Section 4.5.3 below confirms that this is actually the case. However this raises a practical problem since $\gamma$ can be chosen as large as wanted and there is no apparent "counter-weight" that forbids too large values of $\gamma$ and leads to the existence of a "trade-off" optimal value of the hyperparameter. Section 4.5.3 below suggests that a possible counter-weight could be the loss of information carried by $\tilde{p}_{V_n}(X)$ when $\gamma \to +\infty$, which is quantified by the dependence between $\tilde{p}_{V_n}(X)$ and $X$.

## 4.5 Discussion

### 4.5.1 Advantages compared to previous results

We proved most of finite-dimensional projections of an embedded variable $k_X$ in a Gaussian RBF kernel space are close to a scale-mixture of Gaussians when the kernel hyperparameter $\gamma$ is large, or to a $\mathcal{N}(0, I_p)$ Gaussian if the distribution in the input space is a uniform on a subset of $\mathbb{R}^D$ (for some $D \geqslant 1$) or if the kernel is renormalized adequately. This matches existing results about projections of a random vector $X$ in $\mathbb{R}^D$ for large $D$.

The main advantage of our result over previous results in $\mathbb{R}^D$ is the lack of strong assumptions on $X$. As noted in Section 4.2.2, the closeness of projections in $\mathbb{R}^D$ to a scale-mixture of Gaussians is due to two factors: the dimension $D$ of the ambient space and a relatively flat covariance eigenspectrum of $X$. However, $D$ is usually an immutable parameter whereas in the kernel case $\gamma$ can be freely chosen and adapted to the dataset. As for the latter factor, it is tempting to flatten the eigenspectrum of $X$, for instance by making multiplying $X$ at left by the square root of its inverse covariance matrix to get an isotropic vector. However, inverting a covariance matrix is a difficult task in the high-dimensional framework. It is usually done with methods involving shrinkage

[LW04], matrix tapering [C+10] or sparsity [LV12], and most of these methods counterbalance the large number of covariates of $X$ with parsimony assumptions. Because of these underlying assumptions, it is difficult to foretell whether a vector $X$ renormalized by such a parsimonious estimator still admits marginals close to a SMG. On the other hand, Theorem 4.3.3 does not require any strong assumption on the distribution of $X$, thus avoiding any necessary pre-processing step on $X$. Moreover, our results include more general cases where the input space $\mathcal{X}$ is different from $\mathbb{R}^D$.

### 4.5.2 To renormalize or not to renormalize the kernel?

Whether one considers $k_\gamma$ or its renormalized version $\tilde{k}_\gamma$, one obtains either SMG as marginals in the former case or $\mathcal{N}(0, I_p)$ Gaussians in the latter case. Since the latter distribution is simpler than the former, it would be natural to conclude that $\tilde{k}_\gamma$ should always be used in place of $k_\gamma$. However there are some cases where $k_\gamma$ is preferable to $\tilde{k}_\gamma$. This holds for instance in the application to outlier detection presented later in Chapter 5. It turns out that when transforming data through $k_\gamma$, typical observations and outliers are naturally separated which allows to set a boundary between them and detect outliers (Section 5.2.1). However when using $\tilde{k}_\gamma$, the renormalizing factor $\Sigma_\gamma^{1/2}(X, X)$ tends faster to 0 when $X$ is an outlier and $\gamma \to +\infty$, so that the renormalization cancels out that separation property.

### 4.5.3 Loss of information

In practice, the distribution of $X$ is represented by a sample of $n$ observations $X_1, \dots, X_n$ which are independent copies of $X$. Consider the Gram matrix $\mathbf{K}_\gamma = [k_\gamma(X_i, X_j)]_{1 \leqslant i,j \leqslant n}$. When $\gamma \to +\infty$, $\mathbf{K}_\gamma$ converges entrywise to the identity matrix $I_n$, which means that the information carried by the dataset $(X_i)_{1 \leqslant i \leqslant n}$ has been entirely lost. This intuition is confirmed by the experiments whose results are summarized in Figure 4.1. The upper-left panel of Figure 4.1 displays a sample of simulated points in the input space $\mathbb{R}^2$ generated from a uniform distribution on a "donut"-shaped subset of $\mathbb{R}^2$. The upper-right plot shows an instance of a random projection of $k_X$ in $H(k_\gamma)$ (with $\gamma = 400$, $p = 2$ and $n = 5000$) which looks like a Gaussian (not a scale-mixture since $X$ is uniformly distributed) as expected. We examine several two-dimensional projections of $k_X$ for increasing values of $\gamma$ ranging from 50 to 5000 (100 trials for each value of $\gamma$). At each trial, we assess the normality of the projection with the Henze-Zirkler multivariate normality (MVN) test [HZ90] and we measure the dependence between $X$ and $\mathrm{Tr}^{-1/2}(\Sigma_\gamma^2) \cdot p_{V_n}(X)$ through a Hilbert Schmidt Independence Criterion (HSIC) test [Gre+07b]. The p-values for these two tests are displayed respectively in the lower-left and the lower-right boxplots of Figure 4.1. The p-values of the MVN test seem to almost follow the uniform distribution on $[0, 1]$ as $\gamma$ gets larger up to 400 then tend back to 0 when $\gamma$ gets too large, which means that most of the projections are close to a Gaussian when $\gamma$ is large enough but small compared to $n$. On the other hand, the p-values of the HSIC test converges weakly to a uniform law which shows that $X$ and projections

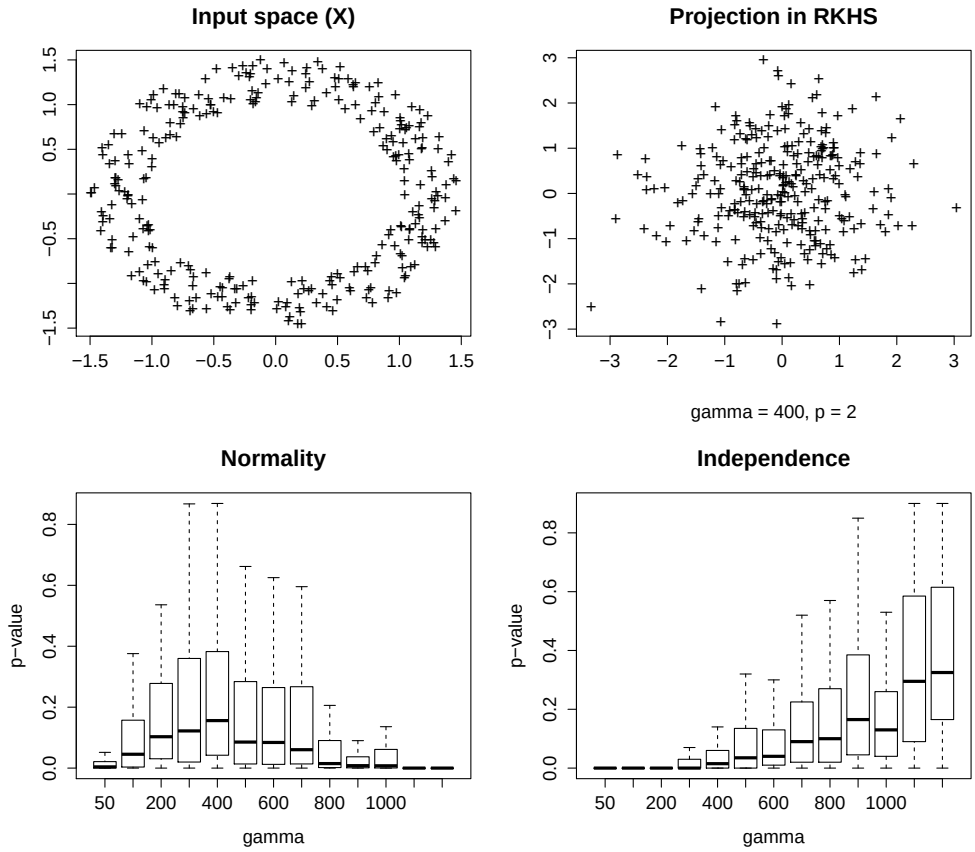


Figure 4.1 – **Upper-left:** Observations in the input space (uniform on a subset of $\mathbb{R}^2$); **upper-right:** Low-dimensional ($p = 2$) projection $p_{V_n}(X)$ of embedded points in the RKHS $H(k_\gamma)$ ($\gamma = 400$); **lower-left:** p-values given by Henze-Zirkler multivariate normality test on projections $p_{V_n}(X)$ (renormalized by $\mathrm{Tr}^{1/2}(\Sigma_\gamma^2)$) with varying values for $\gamma$; **lower-right:** p-values given by the HSIC test for independence between $X$ and $\mathrm{Tr}^{-1/2}(\Sigma_\gamma^2) \cdot p_{V_n}(X)$.

of $k_X$ tend to become independent as $\gamma \to +\infty$ and corroborates the expected loss of information.

We carry the same experiments in the case where $k$ is renormalized as in Section 4.4, and the corresponding results are represented in Figure 4.2. As foretold by Theorem 4.4.2, the convergence to a Gaussian distribution seems to hold when $\gamma \to +\infty$ even for a finite $n$ (lower-left plot in Figure 4.2). However $X$ and $\tilde{p}_{V_n}(X)$ tend to become independent when $\gamma$ grows to infinity which shows that a loss of information occurs as in the case where $k$ is not renormalized.

All in all this negative effect on data information can be seen as a counterweight that defines a range of optimal values of $\gamma$ for which a trade-off between distributional approximation and information is reached.

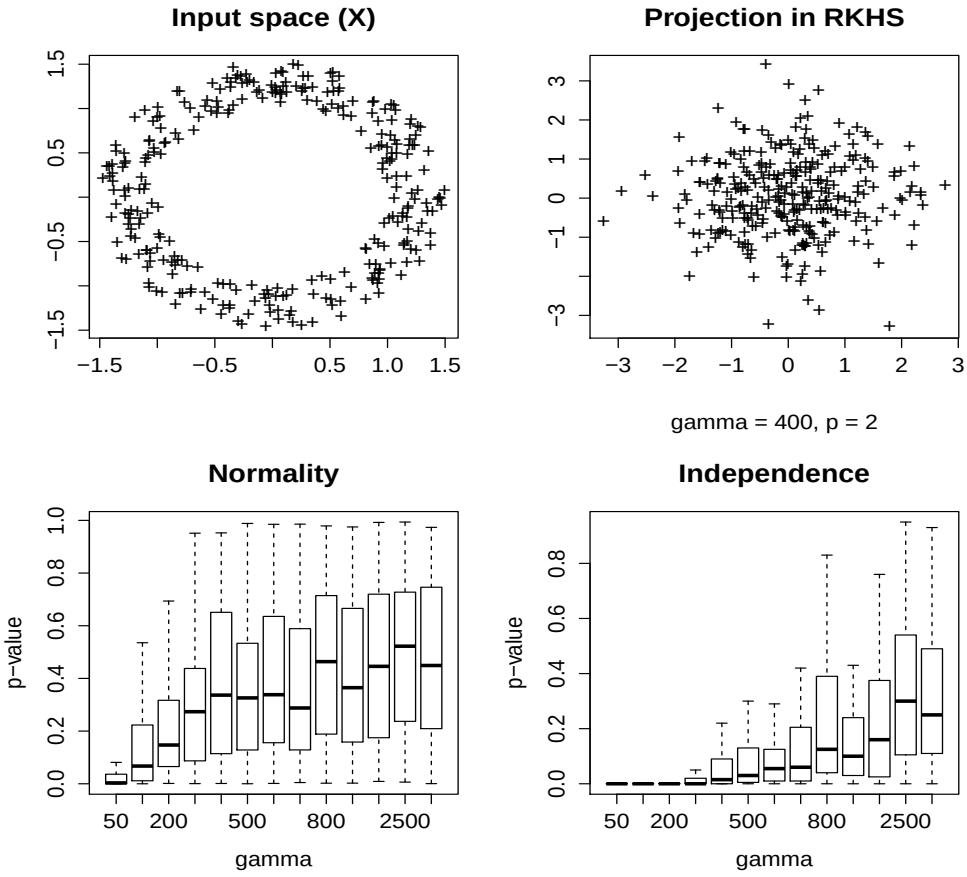


Figure 4.2 – **Upper-left:** Observations in the input space (uniform on a subset of $\mathbb{R}^2$); **upper-right:** Low-dimensional ($p = 2$) projection $\tilde{p}_{V_n}(X)$ of renormalized embedded points in the RKHS $H(k_\gamma)$ ($\gamma = 400$); **lower-left:** p-values given by Henze-Zirkler multivariate normality test on projections $\tilde{p}_{V_n}(X)$ with varying values for $\gamma$; **lower-right:** p-values given by the HSIC test for independence between $X$ and $\tilde{p}_{V_n}(X)$.

## 4.A Proofs of main results

This section presents in details the proofs of Theorem 4.3.3, 4.3.5, 4.4.1 and 4.4.2. All of these four proofs follow a similar outline consisting in three steps. Using the generic notation $L$ to denote the quantity of interest (the distance between the distribution of a random projection and the limit law), the first step aims at linking $L$ with its expectancy taken over the directions $h_1, \ldots, h_p$ (or $h_{1,n}, \ldots, h_{p,n}$ in the empirical case) through a McDiarmid-type concentration inequality. In the second step, a tight upper bound for the expectancy of $L$ is derived. Finally the results of two first steps are combined in the last step and yield the final upper bound of $L$.

### 4.A.1 Proof of Theorem 4.3.3

**First step: apply a concentration inequality**

In the following, the quantity of interest $\Delta^2\left(\frac{p_V(X)}{\sqrt{\mathrm{Tr}(\Sigma_\gamma^2)}}, \mathfrak{s}G\right)$ is rewritten as a function $L(\mathbf{h})$ of $\mathbf{h} = (h_1 \dots h_p)$

$$L(\mathbf{h}) = \Delta^2\left(\frac{p_V(X)}{\sqrt{\mathrm{Tr}(\Sigma_\gamma^2)}}, \mathfrak{s}G\right) = \mathbb{E}_\omega \left| \mathbb{E}_X e^{i[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2}\omega^T p_V(X)} - \mathbb{E}_{\mathfrak{s}G} e^{i\mathfrak{s}\omega^T G} \right|^2 .$$

The goal is to apply the McDiarmid inequality (Lemma A.1.1) to bound the probability $\mathbb{P}(L(h_1,\dots,h_p) > t)$ for some $t > 0$. This requires the existence of some deterministic quantity $c > 0$ such that for any $1 \leqslant j \leqslant p$ and $h_j, h_j' \in H(k_\gamma)$,

$$\left| L(\mathbf{h}^{-j}) - L(\mathbf{h}) \right| \leqslant c \ ,$$

where $\mathbf{h}^{-j} = (h_1, \dots, h_{j-1}, h_j', h_{j+1}, \dots, h_p)$. We show that this condition is fulfilled conditionally to the event $\mathcal{E}_M = \cap_{j=1}^p \{\|h_j\| \leqslant M\}$ for any $M > 0$. Assume $\mathcal{E}_M$ holds for some $M > 0$. Using Jensen's inequality (combined with the convexity of the modulus function) and the triangle inequality yielding $\left||z_1|^2 - |z_2|^2\right| = \left||z_1^2| - |z_2^2|\right| \leqslant \left|z_1^2 - z_2^2\right|$ for any $z_1, z_2 \in \mathbb{C}$,

$$\begin{aligned}
\left| L(\mathbf{h}^{-j}) - L(\mathbf{h}) \right| &\leqslant \mathbb{E}_\omega \left| \left( \mathbb{E}_X e^{i[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2} \sum_{k\neq j}^p \omega_k \langle h_k, k_X\rangle + \omega_j \langle h_j', k_X\rangle} - \mathbb{E}_{\mathfrak{s}G} e^{i\mathfrak{s}\omega^T G} \right)^2 \right. \\
&\qquad \left. - \left( \mathbb{E}_X e^{i[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2} \sum_{k\neq j}^p \omega_k \langle h_k, k_X\rangle + \omega_j \langle h_j, k_X\rangle} - \mathbb{E}_{\mathfrak{s}G} e^{i\mathfrak{s}\omega^T G} \right)^2 \right| \\
&\leqslant \mathbb{E}_\omega \left\{ \left| \mathbb{E}_X e^{i\sum_{k\neq j}^p \frac{\omega_k \langle h_k, k_X\rangle}{[\mathrm{Tr}(\Sigma_\gamma^2)]^{1/2}}} \left( e^{i\frac{\omega_j \langle h_j', k_X\rangle}{[\mathrm{Tr}(\Sigma_\gamma^2)]^{1/2}}} - e^{i\frac{\omega_j \langle h_j, k_X\rangle}{[\mathrm{Tr}(\Sigma_\gamma^2)]^{1/2}}} \right) \right| \right. \\
&\qquad \left. \cdot \left| \mathbb{E}_X e^{i\sum_{k\neq j}^p \frac{\omega_k \langle h_k, k_X\rangle}{[\mathrm{Tr}(\Sigma_\gamma^2)]^{1/2}}} \left( e^{i\frac{\omega_j \langle h_j', k_X\rangle}{[\mathrm{Tr}(\Sigma_\gamma^2)]^{1/2}}} + e^{i\frac{\omega_j \langle h_j, k_X\rangle}{[\mathrm{Tr}(\Sigma_\gamma^2)]^{1/2}}} \right) + 2\mathbb{E}_{\mathfrak{s}G} e^{i\mathfrak{s}\omega^T G} \right| \right\} \\
&\leqslant 4\mathbb{E}_\omega \left| \mathbb{E}_X \left( e^{i\frac{\omega_j \langle h_j', k_X\rangle}{[\mathrm{Tr}(\Sigma_\gamma^2)]^{1/2}}} - e^{i\frac{\omega_j \langle h_j, k_X\rangle}{[\mathrm{Tr}(\Sigma_\gamma^2)]^{1/2}}} \right) \right| \\
&\leqslant \frac{4\mathbb{E}_\omega |\omega_j| \ \ \mathbb{E}_X |\langle h_j' - h_j, k_X\rangle|}{[\mathrm{Tr}(\Sigma_\gamma^2)]^{1/2}} \\
&\leqslant 4\mathbb{E}_\omega \|\omega\| \ \sqrt{\frac{\langle \Sigma_\gamma (h_j' - h_j), h_j' - h_j\rangle}{\mathrm{Tr}(\Sigma_\gamma^2)}} \\
&\leqslant 4\mathbb{E}_\omega \|\omega\| \ \left( \frac{\lambda_1}{\mathrm{Tr}(\Sigma_\gamma^2)} \right)^{1/2} \|h_j' - h_j\|
\end{aligned}$$

$$\leqslant 8\mathbb{E}_\omega\|\omega\|\left(\frac{\lambda_1}{\mathrm{Tr}(\Sigma_\gamma^2)}\right)^{1/2} M := M(\gamma) \ ,$$

where we used the fact that $t \mapsto \exp(it)$ is 1-Lipschitz. Here $\lambda_1 = \sup_{\|u\|=1}\langle \Sigma_\gamma, u, u\rangle$ denotes the largest eigenvalue of $\Sigma_\gamma$.

We apply the McDiarmid inequality to $h_1, \dots, h_p$ conditionally to $\mathcal{E}_M$. Remark that $h_1, \dots, h_p$ are still *i.i.d.* when conditioned to $\mathcal{E}_M$. This yields for any $t > \mathbb{E}(L \mid \mathcal{E}_M)$

$$\begin{aligned}\mathbb{P}\left(L(\mathbf{h}) > t\right) &\leqslant \mathbb{P}\Big(L(\mathbf{h}) - \mathbb{E}_{\mathcal{E}_M} L(\mathbf{h}) > t - \mathbb{E}_{\mathcal{E}_M} L(\mathbf{h}) \mid \mathcal{E}_M\Big) + \mathbb{P}(\mathcal{E}_M^c) \\ &\leqslant \exp\left(-\frac{2(t - \mathbb{E}_{\mathcal{E}_M} L(\mathbf{h}))^2}{pM^2(\gamma)}\right) + \mathbb{P}(\mathcal{E}_M^c) \ .\end{aligned}$$

Using the exponential Markov inequality and given some $0 < \beta < (2\lambda_1)^{-1}$, $\mathbb{P}(\mathcal{E}_M^c)$ is bounded as follows

$$\begin{aligned}\mathbb{P}(\mathcal{E}_M^c) \leqslant p\mathbb{P}\left(\|h_1\| > M\right) \leqslant pe^{-\beta M^2}\mathbb{E}e^{\beta\|h_1\|^2} &= pe^{-\beta M^2}\prod_{j\geqslant 1}(1-2\beta\lambda_j)^{-1/2} \\ &= p\exp\left(-\beta M^2 + (1/2)\sum_{j\geqslant 1}\log\left(\frac{1}{1-2\beta\lambda_j}\right)\right) \\ &\leqslant p\exp\left(-\beta M^2 + (1/2)\sum_{j\geqslant 1}\frac{2\beta\lambda_j}{1-2\beta\lambda_j}\right) \\ &\leqslant p\exp\left(-\beta M^2 + \beta(1-2\beta\lambda_1)^{-1}\right) \ ,\end{aligned}$$

where we used the inequality $\log(x) \leqslant x - 1$ and the equality $\mathrm{Tr}(\Sigma_\gamma) = \mathrm{Tr}(\mathbb{E}_X k_X^{\otimes 2}) = \mathbb{E}_X \mathrm{Tr}(k_X^{\otimes 2}) = \mathbb{E}_X k_\gamma(X, X) = 1$.

Assuming $M < 1 - 1/\sqrt{2}$, one sets $\beta = (1-2M)[\sqrt{1 + (2M/(1-2M^2))^2} - 1]/(2\lambda_1)$ entails

$$\mathbb{P}(\mathcal{E}_M) \leqslant p\exp\left(-\frac{M^2}{2\lambda_1}\right) \ ,$$

hence

$$\mathbb{P}(L(\mathbf{h}) > t) \leqslant \exp\left(-\frac{2(t - \mathbb{E}_{\mathcal{E}_M} L(\mathbf{h}))^2}{pM^2(\gamma)}\right) + p\exp\left(-\frac{M^2}{2\lambda_1}\right) \ .$$

In other words, with probability larger than $1 - \delta$ (for any $M$ such that $\delta > pe^{-M^2/(2\lambda_1)}$)

$$\begin{aligned}L(\mathbf{h}) &\leqslant \mathbb{E}_{\mathbf{h}}[L(\mathbf{h}) \mid \mathcal{E}_M] + M(\gamma)\sqrt{\frac{p}{2}\log\left(\frac{1}{\delta - pe^{-M^2/(2\lambda_1)}}\right)} \\ &\leqslant \left[1 - pe^{-\frac{M^2}{2\lambda_1}}\right]^{-1} \mathbb{E}_{\mathbf{h}}[L(\mathbf{h})]\end{aligned}$$

$$+ 8\mathbb{E}\|\omega\|\left(\frac{\lambda_1}{\mathrm{Tr}(\Sigma_\gamma^2)}\right)^{1/2} M\sqrt{\frac{p}{2}\log\left(\frac{1}{\delta - pe^{-M^2/(2\lambda_1)}}\right)} \ . \qquad (4.A.9)$$

An upper bound for $\mathbb{E}_{\mathbf{h}}[L(\mathbf{h})]$ is given thereafter.

**Second step: find an upper bound of $\mathbb{E}_{\mathbf{h}}L(\mathbf{h})$**

In the following, $G$ and $G'$ denote two independent $\mathcal{N}(0, I_p)$ variables, and $\mathfrak{s}, \mathfrak{s}'$ are independent copies of $\sqrt{\Sigma_\gamma(X,X)/\mathrm{Tr}(\Sigma_\gamma^2)}$. Then $\mathbb{E}_{\mathbf{h}}[L(\mathbf{h})]$ can be written

$$\begin{aligned}\mathbb{E}_{\mathbf{h}}L(\mathbf{h}) &= \mathbb{E}_{\omega,\mathbf{h}}\left|\phi_{[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2}p_V(X)}(\omega) - \phi_{\mathfrak{s}G}(\omega)\right|^2 \\ &= \mathbb{E}_{\omega,\mathbf{h}}\Big(\mathbb{E}_{X,X'}e^{i[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2}\omega^T(p_V(X)-p_V(X'))} \\ &\qquad + \mathbb{E}_{G,G',\mathfrak{s},\mathfrak{s}'}e^{i\omega^T(\mathfrak{s}G-\mathfrak{s}'G')} - 2\mathbb{E}_{X,G,\mathfrak{s}}e^{i\omega^T([\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2}p_V(X)-\mathfrak{s}G)}\Big) \ . \qquad (4.A.10)\end{aligned}$$

Let $\psi(t) = \mathbb{E}_\omega e^{-t\|\omega\|^2/2}$ defined on $\mathbb{R}_+$ and write $\mathfrak{s}_X = (\mathrm{Tr}(\Sigma_\gamma^2))^{-1}\Sigma_\gamma(X,X)$ (likewise, $\mathfrak{s}_{X'} = (\mathrm{Tr}(\Sigma_\gamma^2))^{-1}\Sigma_\gamma(X',X')$) and $\delta_{XX'} = \Sigma_\gamma(X,X')$. On the first hand,

$$\begin{aligned}\mathbb{E}_{\omega,\mathbf{h},X,X'}e^{i[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2}\omega^T(p_V(X)-p_V(X'))} &= \mathbb{E}_{\omega,\mathbf{h},X,X'}e^{i[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2}\sum_{j=1}^p \omega_j\langle k_X - k_{X'}, h_j\rangle_\gamma} \\ &= \mathbb{E}_{\omega,X,X'}e^{-\frac{1}{2}[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1}\|\omega\|^2\langle\Sigma_\gamma(k_X-k_{X'}),k_X-k_{X'}\rangle_\gamma} \\ &= \mathbb{E}_{X,X'}\psi\Big([\mathfrak{s}_X + \mathfrak{s}_{X'} - 2[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1}\delta_{XX'}]\Big) \ .\end{aligned}$$

On the other hand,

$$\begin{aligned}\mathbb{E}_{\omega,G,G',\mathfrak{s},\mathfrak{s}'}e^{i\omega^T(\mathfrak{s}G-\mathfrak{s}'G')} = \mathbb{E}_{\omega,\mathfrak{s},\mathfrak{s}'}e^{-(\mathfrak{s}^2+[\mathfrak{s}']^2)\|\omega\|^2/2} &= \mathbb{E}_{\mathfrak{s},\mathfrak{s}'}\psi(\mathfrak{s}^2 + [\mathfrak{s}']^2) \\ &= \mathbb{E}_{X,X'}\psi(\mathfrak{s}_X^2 + \mathfrak{s}_{X'}^2) \ ,\end{aligned}$$

and finally, using $p_V(X) \sim \mathcal{N}(0, \Sigma_\gamma(X,X)I_p)$ with respect to $h_1, \ldots, h_p$ and conditionally to $X$,

$$\begin{aligned}\mathbb{E}_{\mathbf{h},X,G,\mathfrak{s},\omega}e^{i\omega^T([\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2}p_V(X)-\mathfrak{s}G)} &= \mathbb{E}_{\mathbf{h},X,\mathfrak{s},\omega}e^{i[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2}\omega^T p_V(X) - \mathfrak{s}^2\|\omega\|^2/2} \\ &= \mathbb{E}_{X,\mathfrak{s},\omega}e^{-(\mathfrak{s}_X^2+\mathfrak{s}^2)\|\omega\|^2/2} \\ &= \mathbb{E}_{X,X'}\psi(\mathfrak{s}_X^2 + \mathfrak{s}_{X'}^2) \ .\end{aligned}$$

Plugging these three equalities into (4.A.10),

$$\begin{aligned}\mathbb{E}_{\mathbf{h}}L(\mathbf{h}) &= \mathbb{E}_{X,X'}\psi\Big(\mathfrak{s}_X^2 + \mathfrak{s}_{X'}^2 - 2[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1}\delta_{XX'}\Big) - \psi(\mathfrak{s}_X^2 + \mathfrak{s}_{X'}^2) \\ &\leqslant 2\|\psi'\|_\infty\mathbb{E}_{X,X'}\left(\frac{\delta_{XX'}}{\mathrm{Tr}(\Sigma_\gamma^2)}\right) \ , \qquad (4.A.11)\end{aligned}$$

where $\|\psi'\|_\infty = \sup\{|\psi'(t)| : t \in \mathbb{R}^+\}$. Note that $\|\psi'\|_\infty$ is finite since $|\psi'(t)| = \left|\mathbb{E}_\omega\left(-(1/2)\|\omega\|^2 e^{-t\|\omega\|^2/2}\right)\right| \leqslant (1/2)\mathbb{E}\|\omega\|^2 < +\infty$. Here we are allowed to switch the mean operator and the derivation because of the Dominated Convergence Theorem and because $t \mapsto \exp(-t\|\omega\|^2/2)$ is 1-Lipschitz on $\mathbb{R}^+$.

Lemma A.2.3 states that

$$\mathbb{E}\delta_{XX'} = \frac{\mathbb{E}_X A^2(X)\Gamma^2(s+1)}{\gamma^{2s}}[1+o_\gamma(1)] \ ,$$

and according to Lemma A.2.2,

$$\mathrm{Tr}(\Sigma_\gamma^2) = \frac{\mathbb{E}_X A(X)\Gamma(s+1)}{2^s\gamma^s}[1+o_\gamma(1)] \ ,$$

so that

$$\mathbb{E}_{\mathbf{h}} L(\mathbf{h}) \leqslant \frac{2^s\mathbb{E}\|\omega\|^2\mathbb{E}A^2(X)\Gamma(s+1)}{\mathbb{E}A(X)\gamma^s}[1+o_\gamma(1)] \ . \tag{4.A.12}$$

**Conclusion of the proof**

Continuing from (4.A.9), we proved that with probability larger than $1-\delta$ for any $\delta > pe^{-M^2/(2\lambda_1)}$,

$$L(\mathbf{h}) \leqslant \left[1 - pe^{-\frac{M^2}{2\lambda_1}}\right]^{-1} \mathbb{E}L(\mathbf{h}) + 8\mathbb{E}\|\omega\|\left(\frac{\lambda_1}{\mathrm{Tr}(\Sigma_\gamma^2)}\right)^{1/2} M\sqrt{\frac{p}{2}\log\left(\frac{1}{\delta - pe^{-M^2/(2\lambda_1)}}\right)} \ .$$

Besides, (4.A.12) provided the following upper bound for $\mathbb{E}_{\mathbf{h}} L(\mathbf{h})$

$$\mathbb{E}_{\mathbf{h}} L(\mathbf{h}) \leqslant \frac{2^s\mathbb{E}\|\omega\|^2\mathbb{E}A^2(X)\Gamma(s+1)}{\mathbb{E}A(X)\gamma^s}[1+o_\gamma(1)] \ .$$

Lemma 4.B provides when $\gamma \to +\infty$

$$\lambda_1 = \frac{\mathbb{E}A(X)\Gamma(s+1)}{\gamma^s}[1+o_\gamma(1)] \ ,$$

and by Lemma A.2.2,

$$\mathrm{Tr}(\Sigma_\gamma^2) = \frac{\mathbb{E}A(X)\Gamma(s+1)}{(2\gamma)^s}[1+o_\gamma(1)] \ .$$

Hence if one sets $pe^{-M^2/(2\lambda_1)} = \delta/2$ that is $M = \sqrt{2\lambda_1\log(2p/\delta)}$ , then with probability larger than $1-\delta$ for any $0 < \delta < 1$ and $\gamma$ large enough so that $M = \sqrt{2\mathbb{E}A(X)\Gamma(s+1)\log(2p/\delta)}\gamma^{-s}[1+o_\gamma(1)] <$

$1 - 1/\sqrt{2}$,

$$
\begin{aligned}
L(\mathbf{h}) &= \Delta^2\left(\frac{p_V(X)}{\sqrt{\mathrm{Tr}(\Sigma_\gamma^2)}}, \mathfrak{s}G\right) \\
&\leqslant \left\{\mathbb{E}\|\omega\|^2 \frac{\mathbb{E}A^2(X)\Gamma(s+1)}{\mathbb{E}A(X)}\left(\frac{\gamma}{2}\right)^{-s} + 8\mathbb{E}\|\omega\|\sqrt{p\mathbb{E}A(X)\Gamma(s+1)\log\left(\frac{2(p+1)}{\delta}\right)}\left(\frac{\gamma}{2}\right)^{-s/2}\right\}[1+o_\gamma(1)] \\
&= 8\mathbb{E}\|\omega\|\sqrt{p\mathbb{E}A(X)\Gamma(s+1)\log\left(\frac{2(p+1)}{\delta}\right)}\left(\frac{\gamma}{2}\right)^{-s/2}[1+o_\gamma(1)] \ .
\end{aligned}
$$

## 4.A.2 Proof of Theorem 4.3.5

**First step: apply a concentration inequality**

Let $\mathbf{X} = (X_1, \ldots, X_n)$ denote an *i.i.d.* sample generated by $P$. In the following, we consider the function $L(\cdot; \mathbf{X}) : H(k_\gamma)^p \to \mathbb{R}$ defined by

$$
\Delta^2\left(\frac{p_{V_n}(X)}{\sqrt{\mathrm{Tr}(\Sigma_\gamma^2)}}, \mathfrak{s}G\right) = L(\mathbf{h}; \mathbf{X}) = \mathbb{E}_\omega\left|\mathbb{E}_X e^{i[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2}\omega^T p_V(X)} - \mathbb{E}_{\mathfrak{s}G} e^{i\mathfrak{s}\omega^T G}\right|^2 \ ,
$$

where $p_{V_n}(X) = \left(\langle h_{1,n}, k_X\rangle_\gamma \ldots \langle h_{p,n}, k_X\rangle_\gamma\right)^T$ with $\mathbf{h} = (h_{1,n}, \ldots, h_{p,n})$ following independently a Gaussian process of mean zero and of covariance $\Sigma_{\gamma,n} = (1/n)\sum_{i=1}^n k_\gamma^{\otimes 2}(X_i, .)$.

The goal is to apply the McDiarmid inequality (Lemma A.1.1) to bound the probability $\mathbb{P}_{|\mathbf{X}}(L(\mathbf{h}; \mathbf{X}) > t)$ for some $t > 0$ and conditionally to $\mathbf{X}$. This requires the existence of some deterministic quantity $c > 0$ such that for any $1 \leqslant j \leqslant p$ and $h_{j,n}, h'_{j,n} \in H(k_\gamma)$,

$$
\left|L(\mathbf{h}^{-j}; \mathbf{X}) - L(\mathbf{h}; \mathbf{X})\right| \leqslant c \ ,
$$

where $\mathbf{h}^{-j} = (h_{1,n}, \ldots, h_{j-1,n}, h'_{j,n}, h_{j+1,n}, \ldots, h_{p,n})$. We show that this condition is fulfilled conditionally to the event $\mathcal{E}_M = \cap_{j=1}^p \{\|h_{j,n}\| \leqslant M\}$ for any $M > 0$. Assume $\mathcal{E}_M$ holds for some $M > 0$. Using Jensen's inequality (combined with the convexity of the modulus function) and the triangle inequality yielding $\left||z_1|^2 - |z_2|^2\right| = \left||z_1^2| - |z_2^2|\right| \leqslant \left|z_1^2 - z_2^2\right|$ for any $z_1, z_2 \in \mathbb{C}$,

$$
\begin{aligned}
\left|L(\mathbf{h}^{-j}; \mathbf{X}) - L(\mathbf{h}; \mathbf{X})\right| &\leqslant \mathbb{E}_\omega\left|\left(\mathbb{E}_X e^{i[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2}\sum_{k\neq j}^p \omega_k\langle h_{k,n}, k_X\rangle + \omega_j\langle h'_{j,n}, k_X\rangle} - \mathbb{E}_{\mathfrak{s}G} e^{i\mathfrak{s}\omega^T G}\right)^2\right. \\
&\qquad \left. - \mathbb{E}_\omega\left(\mathbb{E}_X e^{i[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2}\sum_{k\neq j}^p \omega_k\langle h_{k,n}, k_X\rangle + \omega_j\langle h_{j,n}, k_X\rangle} - \mathbb{E}_{\mathfrak{s}G} e^{i\mathfrak{s}\omega^T G}\right)^2\right| \\
&\leqslant \mathbb{E}_\omega\left\{\left|\mathbb{E}_X e^{i\sum_{k\neq j}^p \frac{\omega_k\langle h_{k,n}, k_X\rangle}{[\mathrm{Tr}(\Sigma_\gamma^2)]^{1/2}}}\left(e^{i\frac{\omega_j\langle h'_{j,n}, k_X\rangle}{[\mathrm{Tr}(\Sigma_\gamma^2)]^{1/2}}} - e^{i\frac{\omega_j\langle h_{j,n}, k_X\rangle}{[\mathrm{Tr}(\Sigma_\gamma^2)]^{1/2}}}\right)\right|\right.
\end{aligned}
$$

$$\cdot \left| \mathbb{E}_X e^{i\sum_{k\neq j}^{p} \frac{\omega_k \langle h_{k,n}, k_X\rangle}{[\mathrm{Tr}(\Sigma_\gamma^2)]^{1/2}}} \left( e^{i\frac{\omega_j \langle h'_{j,n}, k_X\rangle}{[\mathrm{Tr}(\Sigma_\gamma^2)]^{1/2}}} + e^{i\frac{\omega_j \langle h_{j,n}, k_X\rangle}{[\mathrm{Tr}(\Sigma_\gamma^2)]^{1/2}}} \right) + 2\mathbb{E}_{\mathfrak{s}G} e^{i\mathfrak{s}\omega^T G} \right| \Bigg\}$$

$$\leqslant 4\mathbb{E}_\omega \left| \mathbb{E}_X \left( e^{i\frac{\omega_j \langle h'_{j,n}, k_X\rangle}{[\mathrm{Tr}(\Sigma_\gamma^2)]^{1/2}}} - e^{i\frac{\omega_j \langle h_{j,n}, k_X\rangle}{[\mathrm{Tr}(\Sigma_\gamma^2)]^{1/2}}} \right) \right|$$

$$\leqslant \frac{4\mathbb{E}_\omega |\omega_j| \ \mathbb{E}_X |\langle h'_{j,n} - h_{j,n}, k_X\rangle|}{[\mathrm{Tr}(\Sigma_\gamma^2)]^{1/2}}$$

$$\leqslant 4\mathbb{E}_\omega \|\omega\| \ \sqrt{\frac{\langle \Sigma_\gamma (h'_{j,n} - h_{j,n}), h'_{j,n} - h_{j,n}\rangle}{\mathrm{Tr}(\Sigma_\gamma^2)}}$$

$$\leqslant 4\mathbb{E}_\omega \|\omega\| \ \left( \frac{\lambda_1}{\mathrm{Tr}(\Sigma_\gamma^2)} \right)^{1/2} \|h'_{j,n} - h_{j,n}\|$$

$$\leqslant 8\mathbb{E}_\omega \|\omega\| \left( \frac{\lambda_1}{\mathrm{Tr}(\Sigma_\gamma^2)} \right)^{1/2} M := M(\gamma) \ ,$$

where we used the fact that $t \mapsto \exp(it)$ is 1-Lipschitz. Here $\lambda_1 = \sup_{\|u\|=1} \langle \Sigma_\gamma u, u\rangle$ denotes the largest eigenvalue of $\Sigma_\gamma$.

Therefore, the McDiarmid inequality yields conditionally to $\mathcal{E}_M$ and $\mathbf{X}$ for any $t > \mathbb{E}_{\mathbf{h}}(L(\mathbf{h};\mathbf{X}) \mid \mathcal{E}_M)$

$$\begin{aligned} \mathbb{P}_{\mathbf{h}}\left(L(\mathbf{h};\mathbf{X}) > t\right) &\leqslant \mathbb{P}_{\mathbf{h}}\left(L(\mathbf{h};\mathbf{X}) - \mathbb{E}_{\mathbf{h}}(L(\mathbf{h};\mathbf{X}) \mid \mathcal{E}_M) > t - \mathbb{E}_{\mathbf{h}}(L(\mathbf{h};\mathbf{X}) \mid \mathcal{E}_M) \mid \mathcal{E}_M\right) + \mathbb{P}_{\mathbf{h}}(\mathcal{E}_M^c) \\ &\leqslant \exp\left( -\frac{2(t - \mathbb{E}_{\mathbf{h}}[L(\mathbf{h};\mathbf{X}) \mid \mathcal{E}_M])^2}{pM^2(\gamma)} \right) + \mathbb{P}_{\mathbf{h}}(\mathcal{E}_M^c) \ . \end{aligned}$$

Let $\hat{\lambda}_1 \geqslant \hat{\lambda}_2 \geqslant \ldots \geqslant \hat{\lambda}_n$ denote the eigenvalues of $\Sigma_{\gamma,n}$. Using the exponential Markov inequality and given some $0 < \beta < (2\hat{\lambda}_1)^{-1}$, $\mathbb{P}_{\mathbf{h}}(\mathcal{E}_M^c)$ is bounded as follows

$$\begin{aligned} \mathbb{P}_{\mathbf{h}}(\mathcal{E}_M^c) \leqslant p\mathbb{P}_{\mathbf{h}}\left(\|h_{1,n}\| > M\right) \leqslant pe^{-\beta M^2} \mathbb{E}_{h_{1,n}} e^{\beta \|h_{1,n}\|^2} &= pe^{-\beta M^2} \prod_{j\geqslant 1} (1 - 2\beta\hat{\lambda}_j)^{-1/2} \\ &= p\exp\left( -\beta M^2 + (1/2) \sum_{j\geqslant 1} \log\left( \frac{1}{1 - 2\beta\hat{\lambda}_j} \right) \right) \\ &\leqslant p\exp\left( -\beta M^2 + (1/2) \sum_{j\geqslant 1} \frac{2\beta\hat{\lambda}_j}{1 - 2\beta\hat{\lambda}_j} \right) \\ &\leqslant p\exp\left( -\beta M^2 + \beta(1 - 2\beta\hat{\lambda}_1)^{-1} \right) \ , \end{aligned}$$

where we used the inequality $\log(x) \leqslant x - 1$ and the equality $\mathrm{Tr}(\Sigma_{\gamma,n}) = 1$.

Assuming $M < 1 - 1/\sqrt{2}$, one sets $\beta = (1-2M)[\sqrt{1+(2M/(1-2M^2))^2}-1]/(2\hat{\lambda}_1)$ entails

$$\mathbb{P}_{\mathbf{h}}(\mathcal{E}_M) \leqslant p \exp\left(-\frac{M^2}{2\hat{\lambda}_1}\right) ,$$

hence

$$\mathbb{P}_{\mathbf{h}}(L(\mathbf{h};\mathbf{X}) > t) \leqslant \exp\left(-\frac{2(t-\mathbb{E}_{\mathcal{E}_M} L)^2}{pM^2(\gamma)}\right) + p\exp\left(-\frac{M^2}{2\hat{\lambda}_1}\right) .$$

that is with probability larger than $1-\delta$ (with $\delta > pe^{-M^2/(2\hat{\lambda}_1)}$) over $\mathbf{h}$ and conditionally to $\mathbf{X}$

$$\begin{aligned} L(\mathbf{h};\mathbf{X}) &\leqslant \mathbb{E}_{\mathbf{h}|\mathcal{E}_M} L(\mathbf{h};\mathbf{X}) + M(\gamma)\sqrt{\frac{p}{2}\log\left(\frac{1}{\delta - pe^{-M^2/(2\hat{\lambda}_1)}}\right)} \\ &\leqslant \mathbb{P}^{-1}(\mathcal{E}_M)\mathbb{E}_{\mathbf{h}} L(\mathbf{h};\mathbf{X}) + M(\gamma)\sqrt{\frac{p}{2}\log\left(\frac{1}{\delta - pe^{-M^2/(2\hat{\lambda}_1)}}\right)} . \end{aligned} \tag{4.A.13}$$

The following section is devoted to finding an upper bound for $\mathbb{E}_{\mathbf{h}} L(\mathbf{h};\mathbf{X})$ that holds with large probability over the sample $\mathbf{X}$.

**Second step: find an upper bound for $\mathbb{E}_{\mathbf{h}} L(\mathbf{h};\mathbf{X})$**

$\mathbb{E}_{\mathbf{h}} L(\mathbf{h};\mathbf{X})$ can be expanded as follows

$$\begin{aligned} \mathbb{E}_{\mathbf{h}} L(\mathbf{h};\mathbf{X}) &= \mathbb{E}_{\omega,\mathbf{h}} \left|\phi_{\mathrm{Tr}(\Sigma_\gamma^2)^{-1/2} p_V(X)}(\omega) - \phi_{\mathfrak{s}G}(\omega)\right|^2 \\ &= \mathbb{E}_{\omega,\mathbf{h}} \Big\{ \mathbb{E}_{X,X'} e^{i[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2}\omega^T(p_V(X)-p_V(X'))} \\ &\qquad + \mathbb{E}_{G,G',\mathfrak{s},\mathfrak{s}'} e^{i\omega^T(\mathfrak{s}G - \mathfrak{s}'G')} - 2\mathbb{E}_{X,G,\mathfrak{s}} e^{i\omega^T([\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2} p_V(X) - \mathfrak{s}G)} \Big\} . \end{aligned} \tag{4.A.14}$$

Let $\psi(t) = \mathbb{E}_\omega e^{-t\|\omega\|^2/2}$ defined on $\mathbb{R}_+$ and write

$$\hat{\mathfrak{s}}_X = \frac{\Sigma_{\gamma,n}(X,X)}{\mathrm{Tr}(\Sigma_\gamma^2)}$$

$$\mathfrak{s}_X = \frac{\Sigma_\gamma(X,X)}{\mathrm{Tr}(\Sigma_\gamma^2)}$$

(likewise, $\hat{\mathfrak{s}}_{X'} = (\mathrm{Tr}(\Sigma_\gamma^2))^{-1}\Sigma_{\gamma,n}(X',X')$ and $\mathfrak{s}_{X'} = (\mathrm{Tr}(\Sigma_\gamma^2))^{-1}\Sigma_\gamma(X',X')$) and

$$\hat{\delta}_{XX'} = \frac{\Sigma_{\gamma,n}(X,X')}{\mathrm{Tr}(\Sigma_\gamma^2)} \ .$$

On the first hand,

$$\begin{aligned}\mathbb{E}_{\mathbf{h},\omega,X,X'} e^{i[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2}\omega^T(p_V(X)-p_V(X'))} &= \mathbb{E}_{\mathbf{h},\omega,X,X'} e^{i[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2}\sum_{j=1}^p \omega_j\langle k_X-k_{X'},h_{j,n}\rangle_\gamma} \\ &= \mathbb{E}_{\omega,X,X'} e^{-\frac{1}{2}[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1}\|\omega\|^2\langle\Sigma_{\gamma,n}(k_X-k_{X'}),k_X-k_{X'}\rangle_\gamma} \\ &= \mathbb{E}_{X,X'}\psi\left(\hat{\mathfrak{s}}_X+\hat{\mathfrak{s}}_{X'}-2\hat{\delta}_{XX'}\right) \ .\end{aligned}$$

On the other hand,

$$\mathbb{E}_{G,G',\mathfrak{s},\mathfrak{s}',\omega} e^{i\omega^T(\mathfrak{s}G-\mathfrak{s}'G')} = \mathbb{E}_{\mathfrak{s},\mathfrak{s},\omega} e^{-(\mathfrak{s}+\mathfrak{s}')\|\omega\|^2/2} = \mathbb{E}_{\mathfrak{s},\mathfrak{s}'}\psi(\mathfrak{s}+\mathfrak{s}') = \mathbb{E}_{X,X'}\psi(\mathfrak{s}_X+\mathfrak{s}_{X'}) \ ,$$

and finally, using $p_{V_n}(X) \sim \mathcal{N}(0,\langle\Sigma_{\gamma,n}k_X,k_X\rangle_\gamma I_p)$ with respect to $\mathbf{h}$ and conditionally to $X,X_1,\dots,X_n$,

$$\begin{aligned}\mathbb{E}_{\mathbf{h},X,G,\mathfrak{s}} e^{i\omega^T([\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2}p_V(X)-\mathfrak{s}G)} &= \mathbb{E}_{\mathbf{h},X,\mathfrak{s}} e^{i[\mathrm{Tr}(\Sigma_\gamma^2)]^{-1/2}\omega^T p_V(X)-\mathfrak{s}\|\omega\|^2/2} \\ &= \mathbb{E}_{X,\mathfrak{s}} e^{-(\hat{\mathfrak{s}}_X+\mathfrak{s})\|\omega\|^2/2} \\ &= \mathbb{E}_{X,X'}\psi(\hat{\mathfrak{s}}_X+\hat{\mathfrak{s}}_{X'}) \ .\end{aligned}$$

Plugging the three equalities into (4.A.14),

$$\begin{aligned}\mathbb{E}_{\mathbf{h}}L(\mathbf{h};\mathbf{X}) = \ & \mathbb{E}_{X,X'}[\psi(\mathfrak{s}_X+\mathfrak{s}_{X'}-2\hat{\delta}_{XX'})-\psi(\mathfrak{s}_X+\mathfrak{s}_{X'})] \\ & + \mathbb{E}_{X,X'}[\psi(\hat{\mathfrak{s}}_X+\hat{\mathfrak{s}}_{X'}-2\hat{\delta}_{XX'})-\psi(\mathfrak{s}_X+\mathfrak{s}_{X'}-2\hat{\delta}_{XX'})] \\ & + 2\mathbb{E}_{XX'}[\psi(\mathfrak{s}_X+\mathfrak{s}_{X'})-\psi(\hat{\mathfrak{s}}_X+\mathfrak{s}_{X'})] \\ \leqslant \ & 2\|\psi'\|_\infty\mathbb{E}_{X,X'}\hat{\delta}_{XX'} + 4\|\psi'\|_\infty\sqrt{\mathbb{E}_X\|\hat{\mathfrak{s}}_X-\mathfrak{s}_X\|^2} \qquad (4.A.15)\end{aligned}$$

where $\|\psi'\|_\infty = \sup\{|\psi'(t)| : t\in\mathbb{R}^+\}$. Note that $\|\psi'\|_\infty$ is finite since $|\psi'(t)| = \left|\mathbb{E}_\omega\left(-(1/2)\|\omega\|^2 e^{-t\|\omega\|^2/2}\right)\right| \leqslant (1/2)\mathbb{E}\|\omega\|^2 < +\infty$. Here we are allowed to switch the mean operator and the derivation because of the Dominated Convergence Theorem and because $t\mapsto\exp(-t\|\omega\|^2/2)$ is 1-Lipschitz on $\mathbb{R}^+$.

Since $\mathbb{E}_X k_\gamma(x,X) = \mathbb{E}_X k_{\gamma/2}^2(x,X) = \Sigma_{\gamma/2}(x,x)$, it follows

$$\mathbb{E}_{X,X'}\hat{\delta}_{XX'} = \mathbb{E}_{X,X'}\left(\frac{1}{n}\sum_{i=1}^n \frac{k_\gamma(X_i,X)k_\gamma(X_i,X')}{\mathrm{Tr}(\Sigma_\gamma^2)}\right) = \frac{1}{n}\sum_{i=1}^n \frac{\Sigma_{\gamma/2}^2(X_i,X_i)}{\mathrm{Tr}(\Sigma_\gamma^2)} \ .$$

By Bennett's concentration inequality (Lemma A.1.2), with probability larger than $1-\delta$ over $\mathbf{X}$

$$\mathbb{E}_{X,X'}\hat{\delta}_{XX'} \leqslant \mathbb{E}_{\mathbf{X}}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\Sigma^2_{\gamma/2}(X_i,X_i)}{\mathrm{Tr}(\Sigma^2_\gamma)}\right\} + \frac{\sqrt{2\nu\log(1/\delta)n^{-1}} + \kappa c\sqrt{2}\log(1/\delta)n^{-1}}{\mathrm{Tr}(\Sigma^2_\gamma)} \ , \tag{4.A.16}$$

where $\nu = \mathrm{Var}(\Sigma^2_{\gamma/2}(X,X))$ and $c = \sup_{x\in\mathcal{X}}\Sigma^2_{\gamma/2}(x,x)$.

First, Lemma A.2.2 provides

$$\mathrm{Tr}(\Sigma^2_\gamma) = \frac{\mathbb{E}A(X)\Gamma(s+1)}{2^s\gamma^s}[1+o_\gamma(1)] \ ,$$

and combined with Lemma A.2.3 entails

$$\begin{aligned}\mathbb{E}_{\mathbf{X}}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\Sigma^2_{\gamma/2}(X_i,X_i)}{\mathrm{Tr}(\Sigma^2_\gamma)}\right\} &= \frac{\mathbb{E}_X\Sigma^2_{\gamma/2}(X,X)}{\mathrm{Tr}(\Sigma^2_\gamma)} \\ &= \frac{\mathbb{E}_X\left\{A(X)\Gamma(s+1)\gamma^{-s}\right\}^2}{\mathbb{E}A(X)\Gamma(s+1)(2\gamma)^{-s}}[1+o_\gamma(1)] \\ &= \frac{\mathbb{E}A^2(X)}{\mathbb{E}A(X)}\frac{\Gamma(s+1)}{(\gamma/2)^s}[1+o_\gamma(1)] \ .\end{aligned}$$

Lemma A.2.3 also leads to

$$\begin{aligned}\nu = \mathrm{Var}(\Sigma^2_{\gamma/2}(X,X)) \leqslant \mathbb{E}\Sigma^4_{\gamma/2}(X,X) &= \mathbb{E}_X\left\{\left[\frac{A(X)\Gamma(s+1)}{\gamma^s}\right]^4[1+o_\gamma(1)]\right\} \\ &= \mathbb{E}A^4(X)\Gamma^4(s+1)\gamma^{-4s}[1+o_\gamma(1)] \ ,\end{aligned}$$

and by Lemma 4.B.1

$$c = \sup_{x\in\mathcal{X}}\Sigma^2_{\gamma/2}(x,x) \leqslant \lambda_1^2(\Sigma_{\gamma/2})\sup_{x\in\mathcal{X}}\|k_{\gamma/2}(x,.)\|^4_{\gamma/2} = \lambda_1^2(\Sigma_{\gamma/2}) = \left\{\frac{\mathbb{E}A(X)\Gamma(s+1)}{(\gamma/2)^s}\right\}^2[1+o_\gamma(1)] \ ,$$

where $\lambda_1(\Sigma_{\gamma/2})$ denotes the largest eigenvalue of $\Sigma_{\gamma/2}$ (the covariance operator).

Plugging all of the above into (4.A.16) provides with probability larger than $1-\delta$ over $\mathbf{X}$

$$\begin{aligned}\mathbb{E}_{X,X'}\hat{\delta}_{XX'} &\leqslant \left\{\frac{\mathbb{E}A^2(X)\Gamma(s+1)}{\mathbb{E}A(X)(\gamma/2)^s} + \frac{\sqrt{\frac{2\log(1/\delta)\Gamma^4(s+1)\mathbb{E}A^4(X)}{n\gamma^{4s}}} + \frac{\kappa\sqrt{2}\log(1/\delta)[\mathbb{E}A(X)]^2\Gamma^2(s+1)}{n(\gamma/2)^{2s}}}{\mathbb{E}A(X)\Gamma(s+1)(2\gamma)^{-s}}\right\}[1+o_\gamma(1)] \\ &= \frac{\mathbb{E}A^2(X)\Gamma(s+1)}{\mathbb{E}A(X)(\gamma/2)^s}[1+o_{\gamma,n}(1)] \ .\end{aligned}$$

Finally using Bennett's inequality again yields with probability larger than $1-\delta$ over $\mathbf{X}$

$$\mathbb{E}_X\left(\hat{\mathfrak{s}}_X-\mathfrak{s}_X\right)^2=\mathbb{E}_X\left[\frac{\Sigma_{\gamma,n}(X,X)-\Sigma_\gamma(X,X)}{\mathrm{Tr}(\Sigma_\gamma^2)}\right]^2=\mathbb{E}_X\left[\frac{\sum_{i=1}^n\left\{k_\gamma^2(X_i,X)-\mathbb{E}_{X'}k_\gamma^2(X',X)\right\}}{n\mathrm{Tr}(\Sigma_\gamma^2)}\right]^2$$
$$\leqslant\mathbb{E}_X\left[\frac{\sqrt{2\tilde{v}\log(1/\delta)/n}+\kappa\tilde{c}\sqrt{2}\log(1/\delta)/n}{\mathrm{Tr}(\Sigma_\gamma^2)}\right]^2\ ,$$

where $\kappa$ is a numerical constant and

$$\tilde{v}=\mathrm{Var}_{X'}(k_\gamma^2(x,X'))\leqslant\mathbb{E}_{X'}k_\gamma^4(X',x)=\Sigma_{2\gamma}(X,X)=\frac{A(X)\Gamma(s+1)}{(4\gamma)^s}[1+o_\gamma(1)]\ ,$$

and

$$\tilde{c}=\sup_{x\in\mathcal{X}}k_\gamma^2(X,x)\leqslant 1\ ,$$

hence introducing $a=\inf\{A(x):x\in\mathrm{supp}(P)\}>0$

$$\mathbb{E}_X\left(\hat{\mathfrak{s}}_X-\mathfrak{s}_X\right)^2\leqslant\mathbb{E}_X\left[\frac{\sqrt{\frac{2\log(1/\delta)A(X)\Gamma(s+1)}{n(4\gamma)}}+\frac{\kappa\sqrt{2}\log(1/\delta)}{n}}{\mathbb{E}A(X)\Gamma(s+1)(2\gamma)^{-s}}\right]^2[1+o_\gamma(1)]$$
$$\leqslant\mathbb{E}_X\left\{\frac{2\log(1/\delta)A(X)\Gamma(s+1)}{[\mathbb{E}A(X)\Gamma(s+1)]^2}\frac{\gamma^s}{n}\right\}\left[1+\frac{\kappa\sqrt{\log(1/\delta)}(2\gamma)^{s/2}}{\sqrt{na\Gamma(s+1)}}\right]^2[1+o_\gamma(1)]$$
$$=\frac{2\log(1/\delta)}{\mathbb{E}A(X)\Gamma(s+1)}\frac{\gamma^s}{n}[1+o_{\gamma,n}(1)]\ ,$$

where the '$o_{\gamma,n}(1)$' tends to 0 when $\gamma^s=o(n)$.

Therefore gathering all of the above into (4.A.15) yields with probability larger than $1-2\delta$ over $\mathbf{X}$

$$\mathbb{E}_{\mathbf{h}}L(\mathbf{h};\mathbf{X})\leqslant\mathbb{E}\|\omega\|^2\left\{\frac{\mathbb{E}A^2(X)\Gamma(s+1)}{\mathbb{E}A(X)(\gamma/2)^s}+2\sqrt{\frac{2\log(1/\delta)\gamma^s}{\mathbb{E}A(X)\Gamma(s+1)n}}\right\}[1+\epsilon_{\gamma,n}]\ ,\tag{4.A.17}$$

where $\epsilon_{\gamma,n}\to 0$ as $\gamma,n\to+\infty$ and $\gamma^s=o(n)$.

**Conclusion**

Continuing from (4.A.13), we proved that with probability larger than $1-\delta$ for any $\delta>pe^{-M^2/(2\lambda_1)}$,

$$L(\mathbf{h};\mathbf{X})\leqslant\mathbb{P}^{-1}(\mathcal{E}_M)\mathbb{E}_{\mathbf{h}}L+8\mathbb{E}\|\omega\|\left(\frac{\lambda_1}{\mathrm{Tr}(\Sigma_\gamma^2)}\right)^{1/2}M\sqrt{\frac{p}{2}\log\left(\frac{1}{\delta-pe^{-M^2/(2\hat{\lambda}_1)}}\right)}\ .$$

Besides, (4.A.17) provided the following upper bound that holds with probability larger than $1-2\delta$ over $\mathbf{X}$

$$\mathbb{E}_{\mathbf{h}} L(\mathbf{h};\mathbf{X}) \leqslant \mathbb{E}\|\omega\|^2 \left\{ \frac{\mathbb{E}A^2(X)\Gamma(s+1)}{\mathbb{E}A(X)(\gamma/2)^s} + 2\sqrt{\frac{2\log(1/\delta)\gamma^s}{\mathbb{E}A(X)\Gamma(s+1)n}} \right\} [1+\epsilon_{\gamma,n}] \ ,$$

where $\epsilon_{\gamma,n} \to 0$ as $\gamma, n \to +\infty$ and $\gamma^s = o(n)$.

Setting $\mathbb{P}(\mathcal{E}_M^c) = pe^{-M^2/(2\hat{\lambda}_1)} = \delta/2$ that is $M = \sqrt{2\hat{\lambda}_1 \log(2p/\delta)}$ yields with probability larger than $1-3\delta$ over $\mathbf{h}$ and $\mathbf{X}$

$$\begin{aligned} L(\mathbf{h};\mathbf{X}) \leqslant \frac{\mathbb{E}\|\omega\|^2}{1-\delta/2} \left\{ \frac{\mathbb{E}A^2(X)\Gamma(s+1)}{\mathbb{E}A(X)(\gamma/2)^s} + 2\sqrt{\frac{2\log(1/\delta)\gamma^s}{\mathbb{E}A(X)\Gamma(s+1)n}} \right\} [1+\epsilon_{\gamma,n}] \\ + 8\mathbb{E}\|\omega\| \left( \frac{\lambda_1}{\mathrm{Tr}(\Sigma_\gamma^2)} \right)^{1/2} \sqrt{p\hat{\lambda}_1 \log\left( \frac{2[p+1]}{\delta} \right)} \ . \end{aligned}$$

Theorem 3.3.2 in [Zwa05] provides with probability larger than $1-\delta$ over $\mathbf{X}$ for any $\delta > 0$ the inequality

$$|\hat{\lambda}_1 - \lambda_1| \leqslant \frac{2 + 3\sqrt{\log(3/\delta)}}{\sqrt{n}} \ ,$$

hence with probability larger than $1-4\delta$

$$\begin{aligned} L(\mathbf{h};\mathbf{X}) \leqslant \frac{\mathbb{E}\|\omega\|^2}{1-\delta/2} \left\{ \frac{\mathbb{E}A^2(X)\Gamma(s+1)}{\mathbb{E}A(X)(\gamma/2)^s} + 2\sqrt{\frac{2\log(1/\delta)\gamma^s}{\mathbb{E}A(X)\Gamma(s+1)n}} \right\} [1+\epsilon_{\gamma,n}] \\ + 8\mathbb{E}\|\omega\| \left( \frac{\lambda_1}{\mathrm{Tr}(\Sigma_\gamma^2)} \right)^{1/2} \sqrt{p\log\left( \frac{2[p+1]}{\delta} \right) \left[ \lambda_1 + \frac{2 + 3\sqrt{\log(3/\delta)}}{\sqrt{n}} \right]} \ . \end{aligned}$$

Lemma 4.B provides when $\gamma \to +\infty$

$$\lambda_1 = \frac{\mathbb{E}A(X)\Gamma(s+1)}{\gamma^s}[1+o_\gamma(1)] \ ,$$

and by Lemma A.2.2,

$$\mathrm{Tr}(\Sigma_\gamma^2) = \frac{\mathbb{E}A(X)\Gamma(s+1)}{(2\gamma)^s}[1+o_\gamma(1)] \ ,$$

therefore

$$L(\mathbf{h};\mathbf{X}) \leqslant \frac{\mathbb{E}\|\omega\|^2}{1-\delta/2}\left\{\frac{\mathbb{E}A^2(X)\Gamma(s+1)}{\mathbb{E}A(X)(\gamma/2)^s} + 2\sqrt{\frac{2\log(1/\delta)\gamma^s}{\mathbb{E}A(X)\Gamma(s+1)n}}\right\}[1+\epsilon_{\gamma,n}]$$
$$+ 8\mathbb{E}\|\omega\|\sqrt{p\log\left(\frac{2[p+1]}{\delta}\right)}\left[\frac{\mathbb{E}A(X)\Gamma(s+1)}{(\gamma/2)^s}\{1+o_\gamma(1)\} + \frac{2^{s+1}+2^s.3\sqrt{\log(3/\delta)}}{\sqrt{n}}\right] ,$$

which reads in a more concise way

$$L(\mathbf{h};\mathbf{X}) \leqslant C_1\gamma^{-s} + C_2\gamma^{s/2}n^{-1/2} + \sqrt{C_3\gamma^{-s} + C_4 n^{-1/2}} \ .$$

To simplify this upper bound, we use the concavity of the square root function to obtain $\sqrt{C_3\gamma^{-s} + C_4 n^{-1/2}} = C_3^{1/2}\gamma^{-s/2}\sqrt{1+(C_4/C_3)\gamma^s n^{-1/2}} \leqslant C_3^{1/2}\gamma^{-s/2}\left[1+(C_4/2C_3)\gamma^s n^{-1/2}\right]$, hence

$$\begin{aligned} L(\mathbf{h};\mathbf{X}) &\leqslant C_1\gamma^{-s} + C_2\gamma^{s/2}n^{-1/2} + C_3^{1/2}\gamma^{-s/2} + \frac{C_4}{2C_3^{1/2}}\gamma^{s/2}n^{-1/2} \\ &= \tilde{C}_1\gamma^{-s/2} + \tilde{C}_2\gamma^{s/2}n^{-1/2} \ . \end{aligned}$$

where

$$\tilde{C}_1 = 2^{(s+6)/2}\mathbb{E}\|\omega\|\sqrt{p\log\left(\frac{2(p+1)}{\delta}\right)\mathbb{E}A(X)\Gamma(s+1)}\{1+o_\gamma(1)\}$$
$$\tilde{C}_2 = \frac{1}{\sqrt{\mathbb{E}A(X)\Gamma(s+1)}}\left[\frac{2\mathbb{E}[\|\omega\|^2]\sqrt{\log(1/\delta)}[1+\epsilon_{\gamma,n}]}{1-\delta/2}\right.$$
$$\left. +4\mathbb{E}[\|\omega\|]\sqrt{p\log(\frac{2[p+1]}{\delta})}(2^{s/2+1} + 2^{s/2}\cdot 3\cdot\sqrt{\log(3/\delta)})\right] .$$

### 4.A.3 Proof of Theorem 4.4.1

**First step: apply a concentration inequality**

In the following, we consider the function $L : H(k_\gamma)^p \to \mathbb{R}$ defined by

$$\Delta^2(\tilde{p}_V(X), G) = L(\mathbf{h}) = \mathbb{E}_\omega\left|\mathbb{E}_X e^{i\omega^T\tilde{p}_V(X)} - \mathbb{E}_G e^{i\omega^T G}\right|^2 \ ,$$

where $\mathbf{h} = (h_1 \dots h_p)$. The goal is to apply the McDiarmid inequality (Lemma A.1.1) to bound the probability $\mathbb{P}(L(\mathbf{h}) > t)$ for some $t > 0$. This requires the existence of some deterministic quantity $c > 0$ such that for any $1 \leqslant j \leqslant p$ and $h_j, h_j' \in H(k_\gamma)$,

$$\left|L(\mathbf{h}^{-j}) - L(\mathbf{h})\right| \leqslant c \ ,$$

where $\mathbf{h}^{-j} = (h_1, \dots, h_{j-1}, h'_j, h_{j+1}, \dots, h_p)$.

We show that this condition is fulfilled conditionally to the event $\mathcal{E}_M = \cap_{j=1}^p \{\|h_j\| \leqslant M\}$ for any $M > 0$. Assume $\mathcal{E}_M$ holds for some $M > 0$. Using Jensen's inequality (combined with the convexity of the modulus function) and the triangle inequality yielding $\left||z_1|^2 - |z_2|^2\right| = \left||z_1^2| - |z_2^2|\right| \leqslant \left|z_1^2 - z_2^2\right|$ for any $z_1, z_2 \in \mathbb{C}$,

$$\begin{aligned}
\left|L(\mathbf{h}^{-j}) - L(\mathbf{h})\right| &\leqslant \mathbb{E}_\omega \Bigg| \left( \mathbb{E}_X e^{i \sum_{k\neq j}^p \omega_k \langle h_k, \tilde{k}_X\rangle + \omega_j \langle h'_j, \tilde{k}_X\rangle} - \mathbb{E}_G e^{i\omega^T G} \right)^2 \\
&\qquad - \mathbb{E}_\omega \left( \mathbb{E}_X e^{i \sum_{k\neq j}^p \omega_k \langle h_k, \tilde{k}_X\rangle + \omega_j \langle h_j, \tilde{k}_X\rangle} - \mathbb{E}_G e^{i\omega^T G} \right)^2 \Bigg| \\
&\leqslant \mathbb{E}_\omega \left\{ \left| \mathbb{E}_X e^{i \sum_{k\neq j}^p \omega_k \langle h_k, \tilde{k}_X\rangle} \left( e^{i\omega_j \langle h'_j, \tilde{k}_X\rangle} - e^{i\omega_j \langle h_j, \tilde{k}_X\rangle} \right) \right| \right. \\
&\qquad \left. \cdot \left| \mathbb{E}_X e^{i \sum_{k\neq j}^p \omega_k \langle h_k, \tilde{k}_X\rangle} \left( e^{i\omega_j \langle h'_j, \tilde{k}_X\rangle} + e^{i\omega_j \langle h_j, \tilde{k}_X\rangle} \right) + 2\mathbb{E}_{\mathfrak{s},G} e^{i\mathfrak{s}\omega^T G} \right| \right\} \\
&\leqslant 4\mathbb{E}_\omega \left| \mathbb{E}_X \left( e^{i\omega_j \langle h'_j, \tilde{k}_X\rangle} - e^{i\omega_j \langle h_j, \tilde{k}_X\rangle} \right) \right| \\
&\leqslant 4\mathbb{E}_\omega |\omega_j| \sqrt{\mathbb{E}_X \frac{\langle h'_j - h_j, k_X\rangle^2}{\Sigma_\gamma(X,X)}} \ .
\end{aligned}$$

where we used the fact that $t \mapsto \exp(it)$ is 1-Lipschitz.

Lemma A.2.1 combined with assumption $0 < a \leqslant A(X) \leqslant A$ entails

$$\Sigma_\gamma(X,X) = \frac{A(X)\Gamma(s+1)}{(2\gamma)^s}[1 + A^{-1}(X) o_\gamma(1)] \leqslant \frac{A\Gamma(s+1)}{(2\gamma)^s}[1 + a^{-1} o_\gamma(1)] \ ,$$

where the '$o_\gamma(1)$' term is uniformly bounded with respect to $X$. Therefore,

$$\begin{aligned}
\left|L(\mathbf{h}^{-j}) - L(\mathbf{h})\right| &\leqslant 4\mathbb{E}_\omega \|\omega\| \sqrt{\frac{\mathbb{E}_X \langle h'_j - h_j, k_X\rangle^2}{A\Gamma(s+1)(2\gamma)^{-s}}}[1 + o_\gamma(1)] \\
&\leqslant 4\mathbb{E}_\omega \|\omega\| \sqrt{\frac{\langle \Sigma_\gamma(h'_j - h_j), h'_j - h_j\rangle}{A\Gamma(s+1)(2\gamma)^{-s}}}[1 + o_\gamma(1)] \\
&\leqslant 4\mathbb{E}_\omega \|\omega\| \left( \frac{(2\gamma)^s \lambda_1}{A\Gamma(s+1)} \right)^{1/2} \|h'_j - h_j\| \ [1 + o_\gamma(1)] \\
&\leqslant 8\mathbb{E}_\omega \|\omega\| \left( \frac{(2\gamma)^s \lambda_1}{A\Gamma(s+1)} \right)^{1/2} M \ [1 + o_\gamma(1)] := M(\gamma) \ ,
\end{aligned}$$

Here $\lambda_1 = \sup_{\|u\|=1} \langle \Sigma_\gamma u, u\rangle$ denotes the largest eigenvalue of $\Sigma_\gamma$.

Therefore, the McDiarmid inequality yields conditionally to $\mathcal{E}_M$ and for any $t > \mathbb{E}_{\mathbf{h}}(L(\mathbf{h}) \mid \mathcal{E}_M)$

$$\mathbb{P}(L(\mathbf{h}) > t) \leqslant \mathbb{P}\left( L(\mathbf{h}) - \mathbb{E}_{\mathbf{h}}[L(\mathbf{h}) \mid \mathcal{E}_M] > t - \mathbb{E}_{\mathbf{h}}[L(\mathbf{h}) \mid \mathcal{E}_M] \,\middle|\, \mathcal{E}_M \right) + \mathbb{P}(\mathcal{E}_M^c)$$

$$\leqslant \exp\left(-\frac{2(t-\mathbb{E}_{\mathbf{h}}[L(\mathbf{h})\,|\,\mathcal{E}_M])^2}{pM^2(\gamma)}\right)+\mathbb{P}(\mathcal{E}_M^c)\ .$$

Using the exponential Markov inequality and given some $0<\beta<(2\lambda_1)^{-1}$, $\mathbb{P}(\mathcal{E}_M^c)$ is bounded as follows

$$\begin{aligned}\mathbb{P}(\mathcal{E}_M^c)\leqslant p\mathbb{P}(\|h_1\|>M)\leqslant pe^{-\beta M^2}\mathbb{E}e^{\beta\|h_1\|^2} &= pe^{-\beta M^2}\prod_{j\geqslant 1}(1-2\beta\lambda_j)^{-1/2}\\ &= p\exp\left(-\beta M^2+(1/2)\sum_{j\geqslant 1}\log\left(\frac{1}{1-2\beta\lambda_j}\right)\right)\\ &\leqslant p\exp\left(-\beta M^2+(1/2)\sum_{j\geqslant 1}\frac{2\beta\lambda_j}{1-2\beta\lambda_j}\right)\\ &\leqslant p\exp\left(-\beta M^2+\beta(1-2\beta\lambda_1)^{-1}\right)\ ,\end{aligned}$$

where we used the inequality $\log(x)\leqslant x-1$ and the equality $\mathrm{Tr}(\Sigma_\gamma)=1$.

Assuming $M<1-1/\sqrt{2}$, setting $\beta=(1-2M)[\sqrt{1+(2M/(1-2M^2))^2}-1]/(2\lambda_1)$ entails

$$\mathbb{P}(\mathcal{E}_M)\leqslant p\exp\left(-\frac{M^2}{2\lambda_1}\right)\ ,$$

hence

$$\mathbb{P}(L(\mathbf{h})>t)\leqslant\exp\left(-\frac{2(t-\mathbb{E}_{\mathbf{h}}[L(\mathbf{h})|\mathcal{E}_M])^2}{pM^2(\gamma)}\right)+p\exp\left(-\frac{M^2}{2\lambda_1}\right)\ .$$

In other words, with probability larger than $1-\delta$ for any $\delta>pe^{-M^2/(2\lambda_1)}$

$$\begin{aligned}L(\mathbf{h})&\leqslant\mathbb{E}_{\mathbf{h}}[L(\mathbf{h})\,|\,\mathcal{E}_M]+M(\gamma)\sqrt{\frac{p}{2}\log\left(\frac{1}{\delta-pe^{-M^2/(2\lambda_1)}}\right)}\\ &\leqslant\left[1-pe^{-\frac{M^2}{2\lambda_1}}\right]^{-1}\mathbb{E}_{\mathbf{h}}L(\mathbf{h})\\ &\qquad+8\mathbb{E}\|\omega\|\left(\frac{(2\gamma)^s\lambda_1}{A\Gamma(s+1)}\right)^{1/2}M\sqrt{\frac{p}{2}\log\left(\frac{1}{\delta-pe^{-M^2/(2\lambda_1)}}\right)}[1+o_\gamma(1)]\ .\end{aligned}\tag{4.A.18}$$

An upper bound for $\mathbb{E}_{\mathbf{h}}L(\mathbf{h})$ is derived in the second part thereafter.

**Second step: find an upper bound of $\mathbb{E}L(\mathbf{h})$**

$\mathbb{E}_{\mathbf{h}}[L(\mathbf{h})]$ can be expanded as

$$\mathbb{E}_{\mathbf{h}}L(\mathbf{h})=\mathbb{E}_{\omega,\mathbf{h}}\left|\phi_{\bar{p}_V(X)}(\omega)-\phi_G(\omega)\right|^2$$

$$= \mathbb{E}_{\omega,\mathbf{h}}\Big(\mathbb{E}_{X,X'}e^{i\omega^T(\tilde{p}_V(X)-\tilde{p}_V(X'))} + \mathbb{E}_{G,G'}e^{i\omega^T(G-G')} - 2\mathbb{E}_{X,G}e^{i\omega^T(\tilde{p}_V(X)-G)}\Big) \ . \tag{4.A.19}$$

Let $\psi(t) = \mathbb{E}_\omega e^{-t\|\omega\|^2/2}$ defined on $\mathbb{R}_+$ and write $\tilde{\delta}_X^2 = \langle \Sigma_\gamma \tilde{k}_X, \tilde{k}_X\rangle_\gamma = 1$ (likewise, $\tilde{\delta}_{X'}^2 = \langle \Sigma_\gamma \tilde{k}_{X'}, \tilde{k}_{X'}\rangle_\gamma = 1$) and $\tilde{\delta}_{XX'} = \langle \Sigma_\gamma \tilde{k}_X, \tilde{k}_{X'}\rangle_\gamma$. On the first hand,

$$\begin{aligned}
\mathbb{E}_{\mathbf{h},\omega}e^{i\omega^T(\tilde{p}_V(X)-\tilde{p}_V(X'))} &= \mathbb{E}_{\mathbf{h},\omega}e^{i\sum_{j=1}^p \omega_j\langle \tilde{k}_X-\tilde{k}_{X'},h_j\rangle_\gamma} \\
&= \mathbb{E}_\omega e^{-\frac{1}{2}\|\omega\|^2\langle \Sigma_\gamma(\tilde{k}_X-\tilde{k}_{X'}),\tilde{k}_X-\tilde{k}_{X'}\rangle_\gamma} \\
&= \psi\Big([\tilde{\delta}_X^2 + \tilde{\delta}_{X'}^2 - 2\tilde{\delta}_{XX'}]\Big) \\
&= \psi\Big(2 - 2\tilde{\delta}_{XX'}\Big) \ .
\end{aligned}$$

On the other hand,

$$\mathbb{E}_{G,G',\omega}e^{i\omega^T(G-G')} = \mathbb{E}_\omega e^{-\|\omega\|^2} = \psi(2) \ ,$$

and finally, using $\tilde{p}_V(X) \sim \mathcal{N}(0, I_p)$ with respect to $h_1, \dots, h_p$ and conditionally to $X$,

$$\begin{aligned}
\mathbb{E}_{\mathbf{h},\omega,G}e^{i\omega^T(\tilde{p}_V(X)-G)} &= \mathbb{E}_{\mathbf{h},\omega}e^{i\omega^T\tilde{p}_V(X)-(1/2)\|\omega\|^2} \\
&= \mathbb{E}_\omega e^{-\frac{1}{2}\|\omega\|^2\tilde{\delta}_X^2-\frac{1}{2}\|\omega\|^2} \\
&= \psi\Big(1+\tilde{\delta}_X^2\Big) = \psi(2) \ .
\end{aligned}$$

Plugging the three equalities into (4.A.19),

$$\mathbb{E}_{\mathbf{h}}L(\mathbf{h}) = \mathbb{E}_{X,X'}\psi\Big(2 - 2\tilde{\delta}_{XX'}\Big) - \psi(2) \ ,$$

and there exists $\xi$ satisfying $|\xi| \leqslant \left|\tilde{\delta}_{XX'}\right|$ such that

$$(I) = 2\mathbb{E}_{XX'}\left|\psi'(\xi)\tilde{\delta}_{XX'}\right| \leqslant 2\|\psi'\|_\infty \mathbb{E}_{X,X'}|\tilde{\delta}_{XX'}| = 2\|\psi'\|_\infty \mathbb{E}_{X,X'}\tilde{\delta}_{XX'} \ , \tag{4.A.20}$$

where $\|\psi'\|_\infty = \sup\{|\psi'(t)| : t \in \mathbb{R}^+\}$. Note that $\|\psi'\|_\infty$ is finite since $|\psi'(t)| = \left|\mathbb{E}_\omega\left(-(1/2)\|\omega\|^2 e^{-t\|\omega\|^2/2}\right)\right| \leqslant (1/2)\mathbb{E}\|\omega\|^2 < +\infty$. Here we are allowed to switch the mean operator and the derivation because of the Dominated Convergence Theorem and because $t \mapsto \exp(-t\|\omega\|^2/2)$ is 1-Lipschitz on $\mathbb{R}^+$.

It remains to find an equivalent term for $\mathbb{E}\tilde{\delta}_{XX'}$ when $\gamma \to +\infty$. Lemma A.2.1 and the assumption $A(x) \geqslant a > 0$ for every $x \in \mathrm{supp}(P)$ provides

$$\mathbb{E}_{X,X'}\tilde{\delta}_{XX'} = \mathbb{E}_{X,X'}\left[\frac{\Sigma_\gamma(X,X')}{\sqrt{\Sigma_\gamma(X,X)\Sigma_\gamma(X',X')}}\right] \leqslant \mathbb{E}_{X,X'}\left[\frac{\gamma^s\Sigma_\gamma(X,X')}{\sqrt{A(X)A(X')}\Gamma(s+1)}\{1+a^{-1}o_\gamma(1)\}\right]$$

$$\leqslant \mathbb{E}_{X,X'}\left[\frac{\gamma^s \Sigma_\gamma(X,X')}{a\Gamma(s+1)}\{1+a^{-1}o_\gamma(1)\}\right] ,$$

where the $o_\gamma(1)$ term is not random. It follows

$$\begin{aligned}
\mathbb{E}\tilde{\delta}_{XX'} &\leqslant \frac{\gamma^s \mathbb{E}_{X,X',X''}\left[e^{-\gamma[d^2(X,X'')+d^2(X',X'')]}\right]}{a\Gamma(s+1)}\{1+a^{-1}o_\gamma(1)\} \\
&= \frac{\gamma^s \mathbb{E}_{X''}\left[\Sigma^2_{\gamma/2}(X'',X'')\right]}{a\Gamma(s+1)}\{1+o_\gamma(1)\} \\
&\leqslant \frac{\gamma^s \mathbb{E}_{X''}\left[A^2(X'')\Gamma^2(s+1)(2\gamma)^{-2s}[1+a^{-1}o_\gamma(1)]\right]}{a\Gamma(s+1)}\{1+o_\gamma(1)\} \\
&= \frac{\mathbb{E}_X[A^2(X)]\Gamma^2(s+1)}{a(2\gamma)^{2s}}[1+o_\gamma(1)] ,
\end{aligned}$$

which finally leads to

$$\mathbb{E}_{\mathbf{h}}[L(\mathbf{h})] \leqslant \mathbb{E}\|\omega\|^2 \frac{\mathbb{E}_X[A^2(X)]\Gamma(s+1)}{a(4\gamma)^s}[1+o_\gamma(1)] . \tag{4.A.21}$$

**Conclusion of the proof**

Continuing from (4.A.18), we proved that with probability larger than $1-\delta$ for any $\delta > pe^{-M^2/(2\lambda_1)}$,

$$\begin{aligned}
L(\mathbf{h}) \leqslant \left[1-pe^{-\frac{M^2}{2\lambda_1}}\right]^{-1} &\mathbb{E}_{\mathbf{h}}L(\mathbf{h}) \\
&+ 8\mathbb{E}\|\omega\|\left(\frac{(2\gamma)^s\lambda_1}{A\Gamma(s+1)}\right)^{1/2} M\sqrt{\frac{p}{2}\log\left(\frac{1}{\delta-pe^{-M^2/(2\lambda_1)}}\right)}^{[1+o_\gamma(1)]} .
\end{aligned}$$

Besides, (4.A.21) provided the following upper bound for $\mathbb{E}_{\mathbf{h}}L(\mathbf{h})$

$$\mathbb{E}_{\mathbf{h}}L(\mathbf{h}) \leqslant \mathbb{E}\|\omega\|^2 \frac{\mathbb{E}_X[A^2(X)]\Gamma(s+1)}{a(4\gamma)^s}[1+o_\gamma(1)] ,$$

and Lemma 4.B.1 provides when $\gamma \to +\infty$

$$\lambda_1 = \frac{\mathbb{E}A(X)\Gamma(s+1)}{\gamma^s}[1+o_\gamma(1)] .$$

Hence if one sets $pe^{-M^2/(2\lambda_1)} = \delta/2$ that is $M = \sqrt{2\lambda_1 \log(2p/\delta)}$ , then with probability larger than $1-\delta$ for any $0 < \delta < 1$

$$L(\mathbf{h}) \leqslant \left\{ \mathbb{E}\|\omega\|^2 \mathbb{E}_X[A^2(X)]\Gamma(s+1)a^{-1}(4\gamma)^{-s} + 8\mathbb{E}\|\omega\| \sqrt{pA\Gamma(s+1)\log\left(\frac{2(p+1)}{\delta}\right)} \left(\frac{\gamma}{2}\right)^{-s/2} \right\} [1 + o_\gamma(1)]$$

$$= 8\mathbb{E}\|\omega\| \sqrt{pA\Gamma(s+1)\log\left(\frac{2(p+1)}{\delta}\right)} \left(\frac{\gamma}{2}\right)^{-s/2} [1 + o_\gamma(1)] \ .$$

### 4.A.4 Proof of Theorem 4.4.2

**First step: apply a concentration inequality**

Let $\mathbf{X} = (X_1, \ldots, X_n)$ denote an *i.i.d.* sample generated by $P$. In the following, we consider the function $L(., \mathbf{X}) : H(k_\gamma)^p \to \mathbb{R}$ defined by

$$\Delta^2\left(\tilde{p}_{V_n}(X), G\right) = L(\mathbf{h}; \mathbf{X}) = \mathbb{E}_\omega \left| \mathbb{E}_X e^{i\omega^T \tilde{p}_{V_n}(X)} - \mathbb{E}_G e^{i\omega^T G} \right|^2 \ ,$$

with $\mathbf{h} = (h_{1,n}, \ldots, h_{p,n})$ following independently a Gaussian process of mean zero and of covariance $\Sigma_{\gamma,n} = n^{-1} \sum_{i=1}^n k_\gamma^{\otimes 2}(X_i, .)$.

The goal is to apply the McDiarmid inequality (Lemma A.1.1) to bound the probability $\mathbb{P}_{\mathbf{h}}(L(\mathbf{h}; \mathbf{X}) > t)$ for some $t > 0$ and conditionally to $\mathbf{X}$. This requires the existence of some deterministic quantity $c > 0$ such that for any $1 \leqslant j \leqslant p$ and $h_j, h_j' \in H(k_\gamma)$,

$$\left| L(\mathbf{h}^{-j}; \mathbf{X}) - L(\mathbf{h}; \mathbf{X}) \right| \leqslant c \ ,$$

where $\mathbf{h}^{-j} = (h_{1,n}, \ldots, h_{j-1,n}, h'_{j,n}, h_{j+1,n}, \ldots, h_{p,n})$. We show that this condition is fulfilled conditionally to the event $\mathcal{E}_M = \cap_{j=1}^p \{\|h_{j,n}\| \leqslant M\}$ for any $M > 0$. Note that $h_{1,n}, \ldots, h_{p,n}$ are still independent under $\mathcal{E}_M$. Assume $\mathcal{E}_M$ holds for some $M > 0$. Using Jensen's inequality (combined with the convexity of the modulus function) and the triangle inequality yielding $\left| |z_1|^2 - |z_2|^2 \right| = \left| |z_1^2| - |z_2^2| \right| \leqslant \left| z_1^2 - z_2^2 \right|$ for any $z_1, z_2 \in \mathbb{C}$,

$$\begin{aligned}
\left| L(\mathbf{h}^{-j}; \mathbf{X}) - L(\mathbf{h}; \mathbf{X}) \right| \leqslant \mathbb{E}_\omega \Bigg| & \left( \mathbb{E}_X e^{i \sum_{k \neq j}^p \omega_k \langle h_k, \tilde{k}_X^{(n)} \rangle + \omega_j \langle h_j', \tilde{k}_X^{(n)} \rangle} - \mathbb{E}_G e^{i\omega^T G} \right)^2 \\
& - \mathbb{E}_\omega \left( \mathbb{E}_X e^{i \sum_{k \neq j}^p \omega_k \langle h_k, \tilde{k}_X^{(n)} \rangle + \omega_j \langle h_j, \tilde{k}_X^{(n)} \rangle} - \mathbb{E}_G e^{i\omega^T G} \right)^2 \Bigg| \\
\leqslant \mathbb{E}_\omega \Bigg\{ & \left| \mathbb{E}_X e^{i \sum_{k \neq j}^p \omega_k \langle h_k, \tilde{k}_X^{(n)} \rangle} \left( e^{i\omega_j \langle h_j', \tilde{k}_X^{(n)} \rangle} - e^{i\omega_j \langle h_j, \tilde{k}_X^{(n)} \rangle} \right) \right| \\
& \cdot \left| \mathbb{E}_X e^{i \sum_{k \neq j}^p \omega_k \langle h_k, \tilde{k}_X^{(n)} \rangle} \left( e^{i\omega_j \langle h_j', \tilde{k}_X^{(n)} \rangle} + e^{i\omega_j \langle h_j, \tilde{k}_X^{(n)} \rangle} \right) + 2\mathbb{E}_G e^{i\omega^T G} \right| \Bigg\}
\end{aligned}$$

$$\begin{aligned}
&\leqslant 4\mathbb{E}_\omega \left| \mathbb{E}_X \left( e^{i\omega_j \langle h_j', \tilde{k}_X^{(n)} \rangle} - e^{i\omega_j \langle h_j, \tilde{k}_X^{(n)} \rangle} \right) \right| \\
&\leqslant 4\mathbb{E}_\omega |\omega_j| \ \mathbb{E}_X \frac{|\langle h_j' - h_j, k_X \rangle|}{\Sigma_{\gamma,n}^{1/2}(X,X)} \\
&\leqslant 4\mathbb{E}_\omega \|\omega\| \ \sqrt{\mathbb{E}_X \frac{\langle h_j' - h_j, k_X \rangle^2}{\Sigma_{\gamma,n}(X,X)}} \ ,
\end{aligned}$$

where we used the fact that $t \mapsto \exp(it)$ is 1-Lipschitz.

By Lemma A.2.4, for any $\delta' \in (0,1)$

$$\left| \frac{\Sigma_{\gamma,n}(x,x)}{\Sigma_\gamma(x,x)} - 1 \right| \leqslant \sqrt{\frac{2\log(1/\delta')\gamma^s}{a\Gamma(s+1)n}} \left[ 1 + \kappa \sqrt{\frac{\log(1/\delta')(2\gamma)^s}{a\Gamma(s+1)n}} \{1 + o_\gamma(1)\} \right] \ ,$$

with probability larger than $1 - \delta'$ over $\mathbf{X}$. This event is denoted $\mathcal{B}$ in the following.

It follows that under the events $\mathcal{E}_M$ and $\mathcal{B}$

$$\begin{aligned}
\left| L(\mathbf{h}^{-j}; \mathbf{X}) - L(\mathbf{h}; \mathbf{X}) \right| &\leqslant 4\mathbb{E}_\omega \|\omega\| \ \sqrt{\mathbb{E}_X \frac{\langle h_j' - h_j, k_X \rangle^2}{\Sigma_{\gamma,n}(X,X)}} \\
&\leqslant 4\mathbb{E}_\omega \|\omega\| \ \sqrt{\mathbb{E}_X \frac{\langle h_j' - h_j, k_X \rangle^2}{\Sigma_\gamma(X,X)}} \left\{ 1 + \Delta_{\gamma,n} \right\}^{1/2} \ ,
\end{aligned}$$

where

$$\Delta_{\gamma,n} = \sqrt{\frac{2\log(1/\delta)\gamma^s}{a\Gamma(s+1)n}} + \frac{2^{(s+3)/2}\kappa \log(1/\delta)\gamma^s}{a\Gamma(s+1)n} [1 + o_\gamma(1)] \ .$$

Lemma A.2.1 combined with assumption $0 < a \leqslant A(X) \leqslant A$ entails

$$\Sigma_\gamma(X,X) = \frac{A(X)\Gamma(s+1)}{(2\gamma)^s} [1 + A^{-1}(X) o_\gamma(1)] \leqslant \frac{A\Gamma(s+1)}{(2\gamma)^s} [1 + a^{-1} o_\gamma(1)] \ ,$$

where the '$o_\gamma(1)$' term is uniformly bounded with respect to $X$. Therefore,

$$\begin{aligned}
\left| L(\mathbf{h}^{-j}; \mathbf{X}) - L(\mathbf{h}; \mathbf{X}) \right| &\leqslant 4\mathbb{E}_\omega \|\omega\| \ \sqrt{\frac{\mathbb{E}_X \langle h_j' - h_j, k_X \rangle^2}{A\Gamma(s+1)(2\gamma)^{-s}}} [1 + o_\gamma(1) + \Delta_{\gamma,n}] \\
&\leqslant 4\mathbb{E}_\omega \|\omega\| \ \sqrt{\frac{\langle \Sigma_\gamma(h_j' - h_j), h_j' - h_j \rangle}{A\Gamma(s+1)(2\gamma)^{-s}}} [1 + o_\gamma(1) + \Delta_{\gamma,n}] \\
&\leqslant 4\mathbb{E}_\omega \|\omega\| \ \left( \frac{(2\gamma)^s \lambda_1}{A\Gamma(s+1)} \right)^{1/2} \|h_j' - h_j\| \ [1 + o_\gamma(1) + \Delta_{\gamma,n}]
\end{aligned}$$

$$\leqslant 8\mathbb{E}_\omega\|\omega\|\left(\frac{(2\gamma)^s\lambda_1}{A\Gamma(s+1)}\right)^{1/2} M\ \left[1+o_\gamma(1)+\Delta_{\gamma,n}\right]\ ,$$

Here $\lambda_1 = \sup_{\|u\|=1}\langle\Sigma_\gamma, u, u\rangle$ denotes the largest eigenvalue of $\Sigma_\gamma$. Lemma 4.B.1 provides

$$\lambda_1 = \frac{\mathbb{E}A(X)\Gamma(s+1)}{\gamma^s}[1+o_\gamma(1)]\ ,$$

which entails

$$\left|L(\mathbf{h}^{-j};\mathbf{X}) - L(\mathbf{h};\mathbf{X})\right| \leqslant 2^{s/2+3}\mathbb{E}_\omega\|\omega\|M\ \left[1+o_\gamma(1)+\Delta_{\gamma,n}\right] \stackrel{\Delta}{=} M(\gamma)\ .$$

Therefore, the McDiarmid inequality yields conditionally to $\mathcal{E}_M\cap\mathcal{B}$ and $\mathbf{X}$ for any $t > \mathbb{E}(L \mid \mathcal{E}_M\cap\mathcal{B})$

$$\begin{aligned}\mathbb{P}\left(L(\mathbf{h};\mathbf{X}) > t\right) &\leqslant \mathbb{P}\left(L - \mathbb{E}[L \mid \mathcal{E}_M\cap\mathcal{B}] > t - \mathbb{E}[L \mid \mathcal{E}_M\cap\mathcal{B}]\ \middle|\ \mathcal{E}_M\cap\mathcal{B}\right) + \mathbb{P}(\mathcal{E}_M^c) + \mathbb{P}(\mathcal{B}^c)\\ &\leqslant \exp\left(-\frac{2(t-\mathbb{E}[L \mid \mathcal{E}_M\cap\mathcal{B}])^2}{pM^2(\gamma)}\right) + \mathbb{P}(\mathcal{E}_M^c) + \delta'\ .\end{aligned}$$

Let $\hat\lambda_1 \geqslant \hat\lambda_2 \geqslant \ldots \geqslant \hat\lambda_n$ denote the eigenvalues of $\Sigma_{\gamma,n}$. Using the exponential Markov inequality and given some $0 < \beta < (2\hat\lambda_1)^{-1}$, $\mathbb{P}_\mathbf{h}(\mathcal{E}_M^c)$ is bounded as follows

$$\begin{aligned}\mathbb{P}_\mathbf{h}(\mathcal{E}_M^c) \leqslant p\mathbb{P}_\mathbf{h}\left(\|h_{1,n}\| > M\right) \leqslant pe^{-\beta M^2}\mathbb{E}e^{\beta\|h_{1,n}\|^2} &= pe^{-\beta M^2}\prod_{j\geqslant 1}(1-2\beta\hat\lambda_j)^{-1/2}\\ &= p\exp\left(-\beta M^2 + (1/2)\sum_{j\geqslant 1}\log\left(\frac{1}{1-2\beta\hat\lambda_j}\right)\right)\\ &\leqslant p\exp\left(-\beta M^2 + (1/2)\sum_{j\geqslant 1}\frac{2\beta\hat\lambda_j}{1-2\beta\hat\lambda_j}\right)\\ &\leqslant p\exp\left(-\beta M^2 + \beta(1-2\beta\hat\lambda_1)^{-1}\right)\ ,\end{aligned}$$

where we used the inequality $\log(x) \leqslant x-1$ and the equality $\mathrm{Tr}(\Sigma_{\gamma,n}) = 1$.

Assuming $M < 1 - 1/\sqrt{2}$, one sets $\beta = (1-2M)[\sqrt{1+(2M/(1-2M^2))^2}-1]/(2\hat\lambda_1)$ entails

$$\mathbb{P}_\mathbf{h}(\mathcal{E}_M) \leqslant p\exp\left(-\frac{M^2}{2\hat\lambda_1}\right)\ ,$$

hence

$$\mathbb{P}_\mathbf{h}(L(\mathbf{h};\mathbf{X}) > t) \leqslant \exp\left(-\frac{2(t-\mathbb{E}[L \mid \mathcal{E}_M\cap\mathcal{B}])^2}{pM^2(\gamma)}\right) + p\exp\left(-\frac{M^2}{2\hat\lambda_1}\right) + \delta'\ . \tag{4.A.22}$$

In other words, McDiarmid's inequality entails with probability larger than $1-\delta$ for any $\delta >$

$pe^{-M^2/(2\hat{\lambda_1})} + \delta'$

$$
\begin{aligned}
L &\leqslant \mathbb{E}_{\mathbf{h}}[L \mid \mathcal{E}_M \cap \mathcal{B}] + M(\gamma)\sqrt{\frac{p}{2}\log\left(\frac{1}{\delta - \delta' - pe^{-M^2/(2\hat{\lambda}_1)}}\right)} \\
&\leqslant \mathbb{E}_{\mathbf{h}}[L \mid \mathcal{E}_M \cap \mathcal{B}] + 2^{s/2+3}\mathbb{E}_{\omega}\|\omega\| M\sqrt{\frac{p}{2}\log\left(\frac{1}{\delta - \delta' - pe^{-M^2/(2\hat{\lambda}_1)}}\right)}\{1 + o_\gamma(1) + \Delta_{\gamma,n}\} \\
&\leqslant \mathbb{P}^{-1}(\mathcal{E}_M \cap \mathcal{B})\mathbb{E}_{\mathbf{h}} L + 2^{s/2+3}\mathbb{E}_{\omega}\|\omega\| M\sqrt{\frac{p}{2}\log\left(\frac{1}{\delta - \delta' - pe^{-M^2/(2\hat{\lambda}_1)}}\right)}\{1 + o_\gamma(1) + \Delta_{\gamma,n}\} \ .
\end{aligned}
$$

**Second step: find an upper bound for $\mathbb{E}_{\mathbf{h}} L(\mathbf{h}; \mathbf{X})$**

The mean term in the rhs above can be written

$$
\begin{aligned}
\mathbb{E}_{\mathbf{h}} L(\mathbf{h}; \mathbf{X}) &= \mathbb{E}_{\omega,\mathbf{h}}\left|\phi_{\tilde{p}_{V_n}(X)}(\omega) - \phi_G(\omega)\right|^2 \\
&= \mathbb{E}_{\omega,\mathbf{h}}\Big\{\mathbb{E}_{X,X'} e^{i\omega^T(\tilde{p}_{V_n}(X) - \tilde{p}_{V_n}(X'))} \\
&\qquad + \mathbb{E}_{G,G'} e^{i\omega^T(G-G')} - 2\mathbb{E}_{X,G} e^{i\omega^T(\tilde{p}_{V_n}(X) - G)}\Big\} \ . \qquad (4.A.23)
\end{aligned}
$$

Let $\psi(t) = \mathbb{E}_{\omega} e^{-t\|\omega\|^2/2}$ defined on $\mathbb{R}_+$.On the first hand,

$$
\begin{aligned}
\mathbb{E}_{\mathbf{h},\omega,X,X'} e^{i\omega^T(\tilde{p}_V(X) - \tilde{p}_V(X'))} &= \mathbb{E}_{\mathbf{h},\omega,X,X'} e^{i\sum_{j=1}^p \omega_j\left\{\langle \tilde{k}_X^{(n)}, h_j\rangle_\gamma - \langle \tilde{k}_{X'}^{(n)}, h_j\rangle_\gamma\right\}} \\
&= \mathbb{E}_{\omega,X,X'} e^{-\frac{1}{2}\|\omega\|^2\{2 - 2\hat{\delta}_{XX'}\}} \\
&= \mathbb{E}_{X,X'}\psi\left(2 - 2\hat{\delta}_{XX'}\right) \ ,
\end{aligned}
$$

where

$$
\hat{\delta}_{XX'} = \frac{\langle \Sigma_{\gamma,n} k_X, k_{X'}\rangle_\gamma}{\sqrt{\Sigma_{\gamma,n}(X,X)\Sigma_{\gamma,n}(X',X')}} \ .
$$

On the other hand,

$$
\mathbb{E}_{G,G',\omega} e^{i\omega^T(G-G')} = \mathbb{E}_{\omega} e^{-\|\omega\|^2} = \psi(2) \ ,
$$

and finally, using $\tilde{p}_{V_n}(X) \sim \mathcal{N}(0, I_p)$ with respect to $\mathbf{h}$ and conditionally to $X, X_1, \ldots, X_n$,

$$
\begin{aligned}
\mathbb{E}_{\mathbf{h},X,G} e^{i\omega^T(\tilde{p}_{V_n}(X) - G)} &= \mathbb{E}_{\mathbf{h},X} e^{i\omega^T \tilde{p}_{V_n}(X) - \|\omega\|^2/2} \\
&= \mathbb{E}_{\omega} e^{-\|\omega\|^2} \\
&= \psi(2) \ .
\end{aligned}
$$

Plugging the three equalities into (4.A.23),

$$\begin{aligned}\mathbb{E}_{\mathbf{h}} L(\mathbf{h};\mathbf{X}) &= \mathbb{E}\psi(2-2\hat{\delta}_{XX'}) - \psi(2) \\ &\leqslant 2\|\psi'\|_\infty \mathbb{E}\hat{\delta}_{XX'} \ .\end{aligned}$$

where $\|\psi'\|_\infty = \sup\{|\psi'(t)| : t \in \mathbb{R}^+\}$. Note that $\|\psi'\|_\infty$ is finite since $|\psi'(t)| = \left|\mathbb{E}_\omega\left(-(1/2)\|\omega\|^2 e^{-t\|\omega\|^2/2}\right)\right| \leqslant (1/2)\mathbb{E}\|\omega\|^2 < +\infty$. Here we are allowed to switch the mean operator and the derivation because of the Dominated Convergence Theorem and because $t \mapsto \exp(-t\|\omega\|^2/2)$ is 1-Lipschitz on $\mathbb{R}^+$.

By Lemma A.2.4, the event

$$\mathcal{B} = \left\{ \left| \frac{\Sigma_{\gamma,n}(x,x)}{\Sigma_\gamma(x,x)} - 1 \right| \leqslant \sqrt{\frac{2\log(1/\delta)\gamma^s}{a\Gamma(s+1)n}} \left[ 1 + \kappa \sqrt{\frac{\log(1/\delta)(2\gamma)^s}{a\Gamma(s+1)n}} \{1 + o_\gamma(1)\} \right] \right\}$$

occurs with probability larger than $1-\delta$ for any $\delta \in (0,1]$ and any $x \in \mathcal{X}$. Moreover under $\mathcal{B}$

$$\begin{aligned}\sqrt{\frac{\Sigma_\gamma(X,X)}{\Sigma_{\gamma,n}(X,X)}} &= \left(1 + \left[\frac{\Sigma_{\gamma,n}(X,X)}{\Sigma_\gamma(X,X)} - 1\right]\right)^{-1/2} \\ &\leqslant \left(1 - \left|\frac{\Sigma_{\gamma,n}(X,X)}{\Sigma_\gamma(X,X)} - 1\right|\right)^{-1/2} \\ &\leqslant 1 + \frac{1}{2}\left|\frac{\Sigma_{\gamma,n}(X,X)}{\Sigma_\gamma(X,X)} - 1\right| \\ &\leqslant 1 + \sqrt{\frac{2\log(1/\delta)\gamma^s}{a\Gamma(s+1)n}} \left[ 1 + \kappa \sqrt{\frac{\log(1/\delta)(2\gamma)^s}{a\Gamma(s+1)n}} \{1 + o_\gamma(1)\} \right] \ .\end{aligned}$$

Introducing $A = \sup\{A(x) : x \in \mathcal{X}\} < +\infty$, it follows

$$\begin{aligned}\mathbb{E}\hat{\delta}_{XX'} &\leqslant \mathbb{E}_{X,X'}\left( \frac{1}{n}\sum_{i=1}^n \frac{k_\gamma(X,X_i)k_\gamma(X_i,X')}{\sqrt{\Sigma_{\gamma,n}(X,X)\Sigma_{\gamma,n}(X',X')}} \right) \\ &\leqslant \left[ 1 + \sqrt{\frac{\log(1/\delta)\gamma^s}{2a\Gamma(s+1)n}} + \frac{2^{\frac{s+1}{2}}\kappa\log(1/\delta)\gamma^s}{a\Gamma(s+1)n} \right] \mathbb{E}_{X,X'}\left( \frac{1}{n}\sum_{i=1}^n \frac{k_\gamma(X,X_i)k_\gamma(X_i,X')}{\sqrt{\Sigma_\gamma(X,X)\Sigma_\gamma(X',X')}} \right) \{1 + o_\gamma(1)\} \\ &\leqslant \left[ 1 + \sqrt{\frac{\log(1/\delta)\gamma^s}{2a\Gamma(s+1)n}} + \frac{2^{\frac{s+1}{2}}\kappa\log(1/\delta)\gamma^s}{a\Gamma(s+1)n} \right] \mathbb{E}_{X,X'}\left( \frac{1}{n}\sum_{i=1}^n \frac{k_\gamma(X,X_i)k_\gamma(X_i,X')}{a\Gamma(s+1)(2\gamma)^{-s}} \right) \{1 + o_\gamma(1)\} \\ &= \left[ 1 + \sqrt{\frac{\log(1/\delta)\gamma^s}{2a\Gamma(s+1)n}} + \frac{2^{\frac{s+1}{2}}\kappa\log(1/\delta)\gamma^s}{a\Gamma(s+1)n} \right] \frac{(1/n)\sum_{i=1}^n \Sigma^2_{\gamma/2}(X_i,X_i)}{a\Gamma(s+1)(2\gamma)^{-s}} \{1 + o_\gamma(1)\} \\ &\leqslant \left[ 1 + \sqrt{\frac{\log(1/\delta)\gamma^s}{2a\Gamma(s+1)n}} + \frac{2^{\frac{s+1}{2}}\kappa\log(1/\delta)\gamma^s}{a\Gamma(s+1)n} \right] \frac{A\Gamma^2(s+1)(2\gamma)^{-2s}}{a\Gamma(s+1)(2\gamma)^{-s}} \{1 + o_\gamma(1)\}\end{aligned}$$

$$= \left\{ \frac{A}{a} \frac{\Gamma(s+1)}{(2\gamma)^s} + \sqrt{\frac{A^2\Gamma(s+1)\log(1/\delta)}{2^{1+2s}a^3\gamma^s n}} + \frac{2^{(1-s)/2}\kappa A \log(1/\delta)}{a^2 n} \right\} \{1 + o_\gamma(1)\} \ ,$$

that is

$$\mathbb{E}_{\mathbf{h}} L(\mathbf{h};\mathbf{X}) \leqslant \mathbb{E}\|\omega\|^2 \left\{ \frac{A}{a} \frac{\Gamma(s+1)}{(2\gamma)^s} + \sqrt{\frac{A^2\Gamma(s+1)\log(1/\delta)}{2^{1+2s}a^3\gamma^s n}} + \frac{2^{(1-s)/2}\kappa A \log(1/\delta)}{a^2 n} \right\} \{1 + o_\gamma(1)\} \ , \tag{4.A.24}$$

with probability larger than $1-\delta$ over $\mathbf{X}$.

**Conclusion of the proof**

Continuing from (4.A.22), we proved that with probability larger than $1-\delta$ (over $\mathbf{h}$) for any $\delta > pe^{-M^2/(2\hat{\lambda}_1)} + \delta'$,

$$L(\mathbf{h};\mathbf{X}) \leqslant \frac{\mathbb{E}_{\mathbf{h}} L}{\mathbb{P}(\mathcal{E}_M \cap \mathcal{B})} + 2^{s/2+3}\mathbb{E}_\omega\|\omega\| M \sqrt{\frac{p}{2}\log\left(\frac{1}{\delta - \delta' - pe^{-M^2/(2\hat{\lambda}_1)}}\right)} \{1 + o_\gamma(1) + \Delta_{\gamma,n}\}^{1/2} \ ,$$

where $\Delta_{\gamma,n} = \sqrt{\frac{2\log(1/\delta)\gamma^s}{a\Gamma(s+1)n}} + \frac{2^{(s+3)/2}\kappa\log(1/\delta)\gamma^s}{a\Gamma(s+1)n}\{1 + o_\gamma(1)\}$.

Hence if one sets $\delta'$ and $M$ such that

$$\delta' = \frac{\delta}{4} \qquad \text{and} \qquad pe^{-M^2/[2\hat{\lambda}_1]} = \frac{\delta}{4} \geqslant \mathbb{P}(\mathcal{E}_M^c) \ ,$$

then $M = \sqrt{2\hat{\lambda}_1 \log(4p/\delta)}$ and with probability larger than $1-2\delta$ over $\mathbf{h}$ and $\mathbf{X}$

$$L(\mathbf{h};\mathbf{X}) \leqslant \frac{\mathbb{E}_{\mathbf{h}} L(\mathbf{h};\mathbf{X})}{1-\frac{\delta}{2}} + 2^{s/2+3}\mathbb{E}_\omega\|\omega\| \sqrt{p\hat{\lambda}_1 \log\left(\frac{2+4p}{\delta}\right)} \{1 + o_\gamma(1) + \Delta_{\gamma,n}\}^{1/2} \ .$$

Theorem 3.3.2 in [Zwa05] provides with probability larger than $1-\delta$ for any $\delta > 0$ the inequality

$$|\hat{\lambda}_1 - \lambda_1| \leqslant \frac{2 + 3\sqrt{\log(3/\delta)}}{\sqrt{n}} \ ,$$

which entails combined with Lemma 4.B.1

$$\hat{\lambda}_1 \leqslant \frac{\mathbb{E}A(X)\Gamma(s+1)}{\gamma^s}\{1 + o_\gamma(1)\} + \frac{2 + 3\sqrt{\log(3/\delta)}}{\sqrt{n}} \ .$$

Hence with probability larger than $1-2\delta$

$$L(\mathbf{h};\mathbf{X}) \leqslant \frac{\mathbb{E}_{\mathbf{h}} L(\mathbf{h};\mathbf{X})}{1-\frac{\delta}{2}}$$

$$+ 2^{s/2+3}\mathbb{E}_\omega\|\omega\|\sqrt{p\log\left(\frac{2+4p}{\delta}\right)\left[\frac{\mathbb{E}A(X)\Gamma(s+1)\{1+o_\gamma(1)\}}{\gamma^s}+\frac{2+3\sqrt{\log(3/\delta)}}{\sqrt{n}}\right]\left[1+o_\gamma(1)+\Delta_{\gamma,n}\right]}$$
$$\leqslant \frac{\mathbb{E}_{\mathbf{h}}L(\mathbf{h};\mathbf{X})}{1-\frac{\delta}{2}}$$
$$+ 2^{s/2+3}\mathbb{E}_\omega\|\omega\|\sqrt{p\log\left(\frac{2+4p}{\delta}\right)\left\{\frac{\mathbb{E}A(X)\Gamma(s+1)}{\gamma^s}+\frac{3\sqrt{\log(3/\delta)}}{\sqrt{n}}+\frac{3\cdot 2^{\frac{s+3}{2}}\kappa\sqrt{\log(3/\delta)\log(1/\delta)\gamma^s}}{a\Gamma(s+1)n^{3/2}}\right\}^{\{1+o_{\gamma,n}(1)\}}} \ , \tag{4.A.25}$$

where the '$o_{\gamma,n}(1)$' term converges to 0 when both $\gamma$ and $n$ tend to infinity.

This upper bound suggests that one must have $\gamma^s = o(n^{3/2})$ so that $L(\mathbf{h};\mathbf{X})$ converges to 0 with high probability. However it is easy to see that this condition is not necessary by deriving an alternative upper bound for $L(\mathbf{h};\mathbf{X})$ that is simpler but less tight. A basic Markov inequality entails with probability larger than $1-\delta$ over $\mathbf{h}$

$$L(\mathbf{h};\mathbf{X}) \leqslant \frac{1}{\delta}\mathbb{E}_{\mathbf{h}}L(\mathbf{h};\mathbf{X}) \ , \tag{4.A.26}$$

so that with probability larger than $1-3\delta$, $L(\mathbf{h};\mathbf{X})$ is smaller than

$$\frac{\mathbb{E}_{\mathbf{h}}L(\mathbf{h};\mathbf{X})}{1-\delta/2}\min\left(\frac{1}{\delta}, 1+\frac{2^{s/2+3}\mathbb{E}_\omega\|\omega\|\sqrt{p\log\left(\frac{2+4p}{\delta}\right)\left\{\frac{\mathbb{E}A(X)\Gamma(s+1)}{\gamma^s}+\frac{3\sqrt{\log(3/\delta)}}{\sqrt{n}}+\frac{3\cdot 2^{\frac{s+3}{2}}\kappa\sqrt{\log(3/\delta)\log(1/\delta)\gamma^s}}{a\Gamma(s+1)n^{3/2}}\right\}^{\{1+o_{\gamma,n}(1)\}}}}{\mathbb{E}_{\mathbf{h}}L(\mathbf{h};\mathbf{X})}\right) .$$

Finally, (4.A.24) provided the following upper bound with probability larger than $1-\delta$ over the sample $\mathbf{X}$

$$\mathbb{E}_{\mathbf{h}}L(\mathbf{h};\mathbf{X}) \leqslant \mathbb{E}\|\omega\|^2\left\{\frac{A}{a}\frac{\Gamma(s+1)}{(2\gamma)^s}+\sqrt{\frac{A^2\Gamma(s+1)\log(1/\delta)}{2^{1+2s}a^3\gamma^s n}}+\frac{2^{(1-s)/2}\kappa A\log(1/\delta)}{a^2 n}\right\}\{1+o_\gamma(1)\} \ ,$$

and with probability larger than $1-4\delta$ over $\mathbf{h}$ and $\mathbf{X}$,

$$L(\mathbf{h};\mathbf{X}) \leqslant \mathbb{E}\|\omega\|^2\left\{\frac{A}{a}\frac{\Gamma(s+1)}{(2\gamma)^s}+\sqrt{\frac{A^2\Gamma(s+1)\log(1/\delta)}{2^{1+2s}a^3\gamma^s n}}+\frac{2^{(1-s)/2}\kappa A\log(1/\delta)}{a^2 n}\right\}\min\left(\frac{1}{\delta}, 1+\xi_{\gamma,n}\right)^{\{1+o_\gamma(1)\}} \ ,$$

where $\xi_{\gamma,n}$ is defined by

$$\xi_{\gamma,n} = \frac{2^{s/2+3}\mathbb{E}_\omega\|\omega\|\sqrt{p\log\left(\frac{2+4p}{\delta}\right)\left\{\frac{\mathbb{E}A(X)\Gamma(s+1)}{\gamma^s}+\frac{3\sqrt{\log(3/\delta)}}{\sqrt{n}}+\frac{3\cdot 2^{\frac{s+3}{2}}\kappa\sqrt{\log(3/\delta)\log(1/\delta)\gamma^s}}{a\Gamma(s+1)n^{3/2}}\right\}}}{\mathbb{E}\|\omega\|^2\left\{\frac{A}{a}\frac{\Gamma(s+1)}{(2\gamma)^s}+\sqrt{\frac{A^2\Gamma(s+1)\log(1/\delta)}{2^{1+2s}a^3\gamma^s n}}+\frac{2^{(1-s)/2}\kappa A\log(1/\delta)}{a^2 n}\right\}(1-\delta/2)}\{1+o_{\gamma,n}(1)\},$$

and converges to 0 when $\gamma, n \to +\infty$ and $\gamma = o(n^{3/(2s)})$.

## 4.B Additional lemmas

### 4.B.1 Covariance eigenvalues in an RBF RKHS

**Lemma 4.B.1.** *Let $(\lambda_r)_r$ be the eigenvalues of the covariance operator $\Sigma_\gamma$. Let* $\mathrm{supp}(P) = \{x \in \mathcal{X} \mid \forall \epsilon > 0,\ \mathbb{P}(d^2(x,X) < \epsilon) > 0\}$. *Assume that the distribution of $X$ admits no point mass and that there exists a function $A : \mathrm{supp}(X) \to \mathbb{R}_+^*$ and $s > 0$ such that*

$$\forall x \in \mathrm{supp}(X)\ ,\ \mathbb{P}\Big(d^2(x,X) \leqslant t\Big) \sim A(x)t^s\ ,\quad \textit{when } t \to 0\ .$$

*Then for any integer $r$,*

$$\lambda_r \sim \frac{[\mathbb{E}_X A(X)]\Gamma(s+1)}{\gamma^s}\ ,\quad \textit{when } \gamma \to +\infty\ ,$$

*where $\Gamma(s) = \int_0^{+\infty} x^{s-1}e^{-x}\mathrm{d}x$ denotes the Gamma function.*

*Proof.* Let $r > 0$ be any integer. The $r$-th eigenvalue $\lambda_r$ of $\Sigma_\gamma$ satisfies the following equation

$$\lambda_r = \mathbb{E}_{X,X'} e^{-\gamma d^2(X,X')}\psi_r(X)\psi_r(X')\ ,$$

where $X'$ is an independent copy of $X$ and $\psi_r$ is the corresponding eigenvector of the covariance function $C(x,x') = \mathbb{E}k_\gamma(X,x)k_\gamma(X,x')$ in $L_2(P)$.

Let $M > 0$ be some arbitrary quantity. By conditioning,

$$\lambda_r = p\xi_- + \xi_+\ , \tag{4.B.27}$$

where

$$p = \mathbb{P}(\mathcal{A}_-)\ ,\ \mathcal{A}_- = \left\{d^2(X,X') \leqslant \frac{M}{\gamma}\right\}\ ,\ \mathcal{A}_+ = \left\{d^2(X,X') > \frac{M}{\gamma}\right\}\ ,$$

$$\xi_- = \mathbb{E}\left(e^{-\gamma d^2(X,X')}\psi_r(X)\psi_r(X') \,\middle|\, \mathcal{A}_-\right)\ ,\ \xi_+ = \mathbb{E}\left(e^{-\gamma d^2(X,X')}\psi_r(X)\psi_r(X')\mathbb{1}_{\mathcal{A}_+}\right)\ .$$

It is straightforward that

$$p = \mathbb{E}_X \mathbb{P}_{X'}\left(d^2(X,X') \leqslant \frac{M}{\gamma}\right) \sim \frac{[\mathbb{E}_X A(X)]M^s}{\gamma^s}\ . \tag{4.B.28}$$

Under the event $\mathcal{A}_-$, $d^2(X,X')$ converges to 0 if $M$ satisfies $M = o_\gamma(\gamma)$. Therefore, the continuity

of the inner product $\langle .,. \rangle_\gamma$ and the kernel $k_\gamma$ implies $\psi_r(X') = \langle \psi_r, k_\gamma(X') \rangle = \langle \psi_r, k_\gamma(X') \rangle + o_\gamma(1) = \psi_r(X) + o_\gamma(1)$ so that

$$\begin{aligned}\xi_- &= \mathbb{E}\left( e^{-\gamma d^2(X,X')} \psi_r(X)\psi_r(X') \,\middle|\, \mathcal{A}_- \right) = \mathbb{E}\left( e^{-\gamma d^2(X,X')} [\psi_r^2(X) + \epsilon] \,\middle|\, \mathcal{A}_- \right) \\ &= \mathbb{E}_X \psi_r^2(X) \mathbb{E}_{X'}\left( e^{-\gamma d^2(X,X')} \,\middle|\, \mathcal{A}_- \right)[1 + o_\gamma(1)] \ , \end{aligned} \tag{4.B.29}$$

where the '$o(\gamma)$' term put outside of the expectation comes from the boundedness of $\epsilon$ since $\epsilon = \psi_r(X)(\psi_r(X') - \psi_r(X)) \leqslant \|\psi_r\| \|k_X\| \|k_\gamma(X',.) - k_\gamma(X,.)\| \leqslant 2\|\psi_r\|^2 < +\infty$ .

Let $g(t;x)$ the function that satisfies $g(t;x) \to 0$ when $t \to 0$ for all $x$ and $\mathbb{P}(d^2(X,X') \leqslant t) = A(x)(1 + g(t;x))t^s$. Fubini's Theorem entails

$$\begin{aligned}\mathbb{E}_{X'}\left( e^{-\gamma d^2(x,X')} \,\middle|\, \mathcal{A}_- \right) &= \mathbb{E}_{X'}\left( [-e^{-y}]_{y=\gamma\|x-X'\|^2}^{y=+\infty} \,\middle|\, \mathcal{A}_- \right) \\ &= \mathbb{E}_{X'}\left( \int_{\gamma\|x-X'\|^2}^{+\infty} e^{-y} \mathrm{d}y \,\middle|\, \mathcal{A}_- \right) \\ &= \int_0^{+\infty} e^{-y} \mathbb{E}_{X'}\left( \mathbb{1}_{\{\gamma d^2(x,X') \leqslant y\}} \,\middle|\, \mathcal{A}_- \right) \mathrm{d}y \\ &= \int_0^{+\infty} e^{-y} \mathbb{P}\left( d^2(x,X') \leqslant y/\gamma \,\middle|\, \mathcal{A}_- \right) \mathrm{d}y \\ &= \int_0^{M} e^{-y} \frac{\mathbb{P}\left( d^2(x,X') \leqslant y/\gamma \right)}{\mathbb{P}\left( d^2(x,X') \leqslant M/\gamma \right)} \mathrm{d}y + \int_M^{+\infty} e^{-y} \mathrm{d}y \\ &= \int_0^{M} e^{-y} \frac{[1 + g(y/\gamma;x)]y^s/\gamma^s}{[1 + g(M/\gamma;x)]M^s/\gamma^s} \mathrm{d}y + e^{-M} \\ &= \frac{1 + o_\gamma(1)}{M^s} \gamma(s+1, M) + e^{-M} \\ &\underset{\gamma \to +\infty}{\sim} \frac{\Gamma(s+1)}{M^s} \ , \end{aligned}$$

and therefore

$$\xi_- \sim \mathbb{E}_X \psi_r^2(X) \mathbb{E}_{X'}\left( e^{-\gamma d^2(x,X')} \psi_r^2(x) \,\middle|\, \mathcal{A}_- \right) \sim \mathbb{E}\psi_r^2(X) \frac{\Gamma(s+1)}{M^s} = \frac{\Gamma(s+1)}{M^s} \ , \tag{4.B.30}$$

if $M$ satisfies $M \to +\infty$ and $M = o_\gamma(\gamma)$. Here $\gamma(.,.)$ denotes the lower incomplete Gamma function $\gamma(s,x) = \int_0^x e^{-y} y^{s-1} \mathrm{d}y$.Combining (4.B.28) and (4.B.30) yields

$$p\xi_- \sim \frac{\mathbb{E}A(X)\Gamma(s+1)}{\gamma^s} \ .$$

It remains to show that $\xi_+$ is negligible in relation to $\gamma^{-s}$.

We use the convexity of $\exp(-x)$ that provides the inequalities $e^{-x} \leqslant e^{-M} - (e^{-M} - e^{-(M+\bar{M})})(x -$

$M)/\bar{M}$ when $M \leqslant x \leqslant M+\bar{M}$ for some $\bar{M} \geqslant M$, and $e^{-x} \geqslant e^{-M} - e^{-M}(x-M)$ for $x \leqslant M+1$.

$$\begin{aligned}
\xi_+ &= \mathbb{E}e^{-\gamma d^2(X,X')}\psi_r^2(X)\mathbb{1}_{\mathcal{A}_+} - \frac{1}{2}\mathbb{E}e^{-\gamma d^2(X,X')}[\psi_r(X)-\psi_r(X')]^2\mathbb{1}_{\mathcal{A}_+} \\
&\leqslant e^{-M}\mathbb{E}\left(1 - \frac{1-e^{-\bar{M}}}{\bar{M}}[\gamma d^2(X,X') - M]\right)\psi^2(X)\mathbb{1}_{\{M/\gamma \leqslant d^2(X,X') \leqslant (M+\bar{M})/\gamma\}} \\
&\quad + \exp(-(M+\bar{M})) \\
&\quad - \frac{1}{2}\mathbb{E}\Big\{[e^{-M} - e^{-M}(\gamma d^2(X,X') - M)][\psi(X)-\psi(X')]^2\mathbb{1}_{\mathcal{B}}\Big\} \\
&\leqslant e^{-M}\left[\mathbb{P}(\bar{\mathcal{B}})\mathbb{E}\left(\psi_r^2(X)\,\middle|\,\bar{\mathcal{B}}\right) + e^{-\bar{M}} + \frac{1}{2}\mathbb{E}\Big([\psi_r(X)-\psi_r(X')]^2\mathbb{1}_{\mathcal{B}}\Big)\right] \\
&\leqslant e^{-M}\left[\mathbb{P}(\bar{\mathcal{B}})\mathbb{E}\left(\psi_r^2(X)\,\middle|\,\bar{\mathcal{B}}\right) + e^{-\bar{M}} + \frac{1}{2}\mathbb{E}\Big([\psi_r(X)-\psi_r(X')]^2\Big)\right] \ , \qquad (4.B.31)
\end{aligned}$$

where $\bar{\mathcal{B}} = \{M/\gamma \leqslant d^2(X,X') \leqslant (M+\bar{M})/\gamma\}$ and $\mathcal{B} = \Big\{M/\gamma \leqslant d^2(X,X') \leqslant (M+1)/\gamma\Big\}$.

It is shown that $\xi_+ = o(e^{-M})$ by checking that the rhs of (4.B.31) tends to 0. First,

$$\mathbb{E}\left(\psi_r^2(X)\,\middle|\,\bar{\mathcal{B}}\right) = \mathbb{E}_X\left(\psi_r^2(X)\mathbb{E}_{X'|\bar{\mathcal{B}}}(1|X)\right) = \mathbb{E}_X\psi_r^2(X) = 1 \ ,$$

and choosing $\bar{M} \to +\infty$ such that $\bar{M} = o_\gamma(\gamma)$ implies $\mathbb{P}(\bar{\mathcal{B}}) \to 0$ and $\exp(-\bar{M}) \to 0$. Finally,

$$\begin{aligned}
\mathbb{E}\Big([\psi_r(X)-\psi_r(X')]^2\Big) &= 2 - 2\mathbb{E}\left[\psi_r(X)\psi_r(X')\,\middle|\,\mathcal{B}\right] \\
&\leqslant 2[\mathbb{E}\psi_r^2(X) + \mathbb{E}\psi_r^2(X')] = 4 \ .
\end{aligned}$$

Hence $\xi_+ = \mathcal{O}_\gamma(e^{-M})$ and setting $M = q\log(\gamma)$ for some $q > s$ leads to $\lambda_r \sim \mathbb{E}A(X)\Gamma(s+1)\gamma^{-s}$.

□

## 4.B.2 Asymptotic orthonormality of Gaussian processes

**Lemma 4.B.2.** *Consider $p$ independent zero-mean Gaussian processes $h_1, \dots, h_p$ with covariance $\Sigma_\gamma = \mathbb{E}_{X\sim P}k_\gamma(X,\cdot)^{\otimes 2}$ for some probability measure $P$. Then these $p$ variables are asymptotically orthonormal when $\gamma \to +\infty$, that is*

$$\forall 1 \leqslant i \leqslant p, \ \ \|h_i\|_\gamma^2 \underset{\gamma\to+\infty}{\longrightarrow} 1 \quad a.s. \ ,$$

*and*

$$\forall 1 \leqslant i,j \leqslant p, \ \ i \neq j, \ \ \langle h_i, h_j\rangle_\gamma \underset{\gamma\to+\infty}{\longrightarrow} 0 \quad a.s. \ .$$

*Proof.* Consider the eigendecomposition $\Sigma_\gamma = \sum_{r\geqslant 1}\lambda_r\Psi_r^{\otimes 2}$ where $(\Psi_r)_r$ form an orthonormal

basis of $H(k_\gamma)$. On the first hand,

$$\mathbb{E}_h \|h\|_\gamma^2 = \mathbb{E}_h \mathrm{Tr}(h^{\otimes 2}) = \mathrm{Tr}(\Sigma_\gamma) = \mathbb{E}_X \mathrm{Tr}(k_X^{\otimes 2}) = \mathbb{E} k_\gamma(X,X) = 1 \ .$$

On the other hand, since $\langle h, \Psi_r \rangle_\gamma$ are independent zero-mean Gaussians of respective variances $\lambda_r$,

$$\mathrm{Var}(\|h\|_\gamma^2) = \mathrm{Var}\left( \sum_{r \geqslant 1} \langle h, \Psi_r \rangle_\gamma^2 \right) = \sum_{r \geqslant 1} \mathrm{Var}\left( \langle h, \Psi_r \rangle_\gamma^2 \right) = 2\mathrm{Tr}(\Sigma_\gamma^2) \ .$$

According to Lemma A.2.2, $\mathrm{Tr}(\Sigma_\gamma^2)$ converges to 0 as $\gamma \to +\infty$. Therefore $\mathrm{Var}(\|h\|_\gamma^2)$ converges to 0 and $\|h\|_\gamma^2$ converges almost surely to its mean 1.

If $i \neq j$, then $\mathbb{E}\langle h_i, h_j \rangle_\gamma = 0$ since $h_i$ and $h_j$ are zero-mean and independent. Also

$$\mathrm{Var}(\langle h_i, h_j \rangle_\gamma) = \mathbb{E}\langle h_i, h_j \rangle_\gamma^2 = \mathbb{E}_{h_i} \langle \Sigma_\gamma h_i, h_i \rangle_\gamma = \mathrm{Tr}(\Sigma_\gamma^2) \to 0 \ ,$$

hence $\langle h_i, h_j \rangle_\gamma \to 0$ almost surely. □

Chapter 5

# A New Method for Online Outlier Detection

In Chapter 4, we have described the probabilistic behaviour of random projections of an embedded variable in the RKHS of a Gaussian RBF kernel. As a reminder, we established that most of those projections converge weakly to a so-called *scale-mixture of isotropic Gaussians* when the hyperparameter of the kernel tends to infinity. The goal of the present chapter is to illustrate the usefulness of the aforementioned result in a practical problem, that is the problem of *outlier detection*. To this aim we describe a new method for outlier detection that is based on random projections in an RKHS.

## 5.1 The problem of outlier detection

Outlier detection (OD) is often mentioned in the literature under several alternative names such as *anomaly detection* or *novelty detection*, yet all refer to the same goal: to identify observations in a given dataset that depart from a "normal" statistical behaviour governing the majority of the dataset. Such "abnormal" observations are called *outliers*, while the non-outliers are called *inliers* in the following.

The latter definition of the problem may seem vague and actually there is no clear consensus on a proper definition of an outlier in the related literature. Indeed, the literature about OD is abundant and consists in many different kinds of approaches as pointed out in the extensive review of [Pim+14]: probabilistic methods, distance-based methods or information-theoretic methods, among others. Each of these different classes of methods stems from a different characterization of outliers. For instance, a probabilistic approach assumes that the observations have been drawn from some probability distribution and outliers are points that lie in low-density regions, whereas a distance-based method labels as outliers observations that are abnormally

distant from most of the dataset and a domain-based approach assumes that the inliers lie in a subset of the space where data take values. All in all, the types of methods mentioned above can be deemed "domain-based" since they all eventually lead to define an acceptation region of the data space that contain most of the inliers and almost no outliers: in a probabilistic point of view, such an acceptation region corresponds to the region where the density function of the "normal" distribution is above a certain threshold; in a distance-based approach, this region consists in points whose distance to the rest of the dataset is smaller than some threshold.

Defining such an acceptation region is subject to two kinds of errors: labeling a inlier as an outlier (false alarm error) and calling an outlier a non-outlier (missed detection error). It is desired to control the probabilities of committing these types of errors. A convenient framework to achieve this goal is that of probabilistic methods for OD, that is the inliers are assumed to have been drawn from a known distribution so that the probability of false alarm error becomes available. However, the shortcoming of such methods is that the working model may not fit the actual data at all. Therefore our concern is to access the error probabilities while avoiding any restrictive assumption about the data distribution.

### 5.1.1 Formalized problem

From now on, the problem is formalized as follows. Consider a stream of $\mathcal{X}$-valued data

$$X_1, X_2, \ldots, X_n, X,$$

for some set $\mathcal{X}$. We assume throughout this chapter that $\mathcal{X}$ is endowed with a metric $d(\cdot,\cdot)$.

The $n$ previous points $X_1, \ldots, X_n$ are assumed to be independently drawn from a common distribution $P$ and represent the inliers. The goal is to test whether the new point $X$ was generated by $P$ or by an alternative distribution $P_1$. Note that this formulation is similar to homogeneity testing, where one tests the null-hypothesis that two samples $X_1, \ldots, X_n$ and $X'_1, \ldots, X'_m$ come from the same underlying distribution. The difference lies in the fact that one of the two samples consists of only one observation — that is reduced to $X$ in our case. Since it is impossible to estimate the underlying distribution of $X$ on the basis of one observation, some assumptions about the possible alternative distribution $P_1$ are necessary to make the problem tractable. Thus outlier detection consists in determining a region of $\mathcal{X}$ that contains most of the inliers and we are naturally led to define an outlier as an observation that lies outside of the support of the inlier distribution $P$ — which is denoted $\mathrm{supp}(P)$ and defined formally as $\mathrm{supp}(P) = \{x \in \mathcal{X} \mid \forall \epsilon > 0,\ \mathbb{P}_{X\sim P}(d(X,x) < \epsilon) > 0\}$. This assumption allows to get consistency results (*i.e.* almost surely detect an outlier as $n$ grows to infinity) as we will see later in Section 5.3.2.

Therefore the outlier problem can be cast as the problem of testing the null-hypothesis $\mathbf{H}_0$

against the alternative hypothesis $\mathbf{H}_1$ which are both defined as

$$\begin{cases} \mathbf{H}_0 : X \sim P \,, \\ \mathbf{H}_1 : X \sim P_1 \quad \text{where } P_1 \in \{Q : \nu(\text{supp}(P) \cap \text{supp}(Q)) = 0\} \,, \end{cases}$$

where $\nu$ is a positive measure on $\mathcal{X}$ such that $P$ and $P_1$ are absolutely continuous with respect to $\nu$.

Note that we have presented the outlier problem in an *online* framework, that is single observations are assumed to arrive one after the other and are tested sequentially. This is opposed to *batch* OD, where a full sample is directly available in its entirety and a small sized subsample of outliers must be identified — see for instance the method proposed by [Rot06] mentioned in Section 3.4.2. Surprisingly, OD is often primarily formulated in the batch version and an online version is designed later — for instance OD with one-class SVM was first conceived in a batch fashion in [Sch+01] then in an online manner as in [GD03]. However, batch OD is typically more difficult than online OD since in the latter case a subsample of inliers is readily given and can serve as a reference, whereas in the former case the whole sample — outliers included — must be used to estimate useful quantities which may lead to erroneous estimations (especially if poorly robust estimators are used). Therefore we chose to formalize the OD problem in an online way as a more natural starting point.

### 5.1.2 Existing methods

As the extensive review by [Pim+14] shows, numerous methods for OD have been proposed in the literature. We focus on two fairly recent ways of performing OD: *one-class Support Vector Machine* (oc-SVM) and *kernel Principal Component Analysis* (kernel PCA). SVM and kernel PCA have been already introduced in Section 2.3.1 and Section 2.3.2 respectively. These two techniques are kernel methods and as such, they are more general than OD methods performed solely in the input space $\mathcal{X}$ — in particular, the set $\mathcal{X}$ may take an arbitrary form.

#### One-class SVM

From a probabilistic point of view, the OD problem is linked to the distribution of the inliers — which may be represented by a density function $f$ if the input space $\mathcal{X}$ is $\mathbb{R}^d$. When it comes to define an acceptation region $\mathscr{A}$, a good candidate is a subset of $\mathcal{X}$ on which the density $f$ take large values, say

$$\mathscr{A} = \left\{ x \in \mathcal{X} \;\middle|\; f(x) > \tau \right\} , \tag{5.1.1}$$

where $\tau > 0$ is some prescribed threshold. Defining $\mathscr{A}$ this way requires to have access to the density function $f$ which is usually unknown in practice. Therefore an estimator for $f$ — based on a sample of inliers $X_1, \ldots, X_n$ — is required. However estimating a density function may be a

difficult task, especially when $\mathcal{X}$ is a high-dimensional space where the curse of dimensionality arises and has to be circumvented by means of Gaussian assumptions [Mog02] or sparsity assumptions [LLW07; Vad+11] . That being said, estimating $f$ is just an intermediate problem thus we should not be concerned about the accuracy of such estimation. More precisely, we are solely interested in the location of small density regions of $\mathcal{X}$, hence even an estimate of $f$ which poorly approximates $f$ on high density regions may eventually allow to correctly detect outliers. All in all, it may be more optimal to bypass the estimation of $f$ and to consider an acceptation region $\mathscr{A}$ where the density $f$ is replaced by some proxy function $g$ that could achieve the goal of detecting outliers as well.

Following this rationale, let us rewrite (5.1.1) by replacing $f$ by such a function $g$. Assuming that $g$ belongs to some RKHS $H(k)$ with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the reproducing property yields $g(x) = \langle g, k(x,\cdot)\rangle_{H(k)}$ for every $x \in \mathcal{X}$. Therefore a point $x \in \mathcal{X}$ belongs or not to $\mathscr{A}$ according to the value of

$$\operatorname{sgn}\Big(\langle g, k(x,\cdot)\rangle_{H(k)} - \tau\Big) .$$

In other words, inliers and outliers are separated in the RKHS by the hyperplane

$$\mathscr{H} = \Big\{ h \in H(k) \;\Big|\; \langle g, h\rangle_{H(k)} - \tau = 0 \Big\} ,$$

which looks like the decision boundary of an SVM (Section 2.3.1). However the difference with standard SVM is that the training set essentially consists of inliers instead of two identified classes with comparable sizes. This leads to a special kind of SVM called *one-class SVM* (oc-SVM) that was introduced by [Sch+01].

Instead of separating two given classes of points, oc-SVM separates the whole given sample from the origin in the RKHS, which is done through the following optimization scheme

$$\begin{array}{lll} \text{minimize} & \frac{1}{2}\|g\|^2_{H(k)} + \frac{1}{\nu n}\sum_{i=1}^n \xi_i - \tau & \text{over } g \in \mathcal{H},\ \tau \in \mathbb{R},\ \xi_1,\dots,\xi_n \in \mathbb{R}_+ \\ \text{such that} & \langle g, k(X_i,\cdot)\rangle_{H(k)} \geqslant \tau - \xi_i & \text{for all } 1 \leqslant i \leqslant n , \end{array} \tag{5.1.2}$$

where $\nu \in (0,1]$ is some parameter whose influence is discussed below.

By maximizing $\tau$, the solution of (5.1.2) maximizes the distance between the embedded data and $\mathscr{H}$, yet allows a few points to be misclassified via the slack variables $\xi_1,\dots,\xi_n$ for a better generalization error as in the standard SVM. Despite the different format, the parameter $\nu$ plays a similar role as $C$ in the standard SVM that is controlling how many observations are allowed to be misclassified. For instance Proposition 4 in [Sch+01] states that $\nu$ is an upper bound on the proportion of declared outliers within the training set. However, this does not imply that the probability of committing a false alarm error when testing new points is also bounded by $\nu$. More recently in [VV06], theoretical error bounds are provided for oc-SVM, however the bounded error is the classification error which encompasses false alarm errors and the error of

missing an outlier altogether. Similarly [SHS09] derives upper bounds for the classification error for density level detection (which is a problem related to outlier detection) through SVM. All in all this lack of control on error probabilities constitutes one weak point of oc-SVM.

**Kernel PCA**

[Hof07] proposed to use kernel PCA to detect outliers in order to improve on the performance of one-class SVM.

In Section 2.3.2, we have introduced kernel PCA as a non-linear extension of principal component analysis. As a reminder, kernel PCA consists in applying standard PCA in the feature space instead of the input space. More precisely, kernel PCA seeks out principal components $\varphi_1, \dots, \varphi_p \in H(k)$ which form an orthonormal family of vectors in $H(k)$ and are such that the subspace $V$ spanned by $\varphi_1, \dots, \varphi_p$ captures most of the variance of the embedded data.

The main idea of [Hof07] is to use the *reconstruction error* to detect outliers. The reconstruction error $\mathrm{RE}(x)$ for a point $x \in \mathcal{X}$ is defined by

$$\mathrm{RE}(x) = \left\| \tilde{k}(x, \cdot) - \Pi_V \tilde{k}(x, \cdot) \right\|^2_{H(k)} ,$$

where $\Pi_V$ denotes the projection operator onto the subspace $V$ and $\tilde{k}(x, x') = \langle k(x, \cdot) - \mu, k(x', \cdot) - \mu \rangle_{H(k)}$ is the re-centered kernel with $\mu = (1/n) \sum_{i=1}^n k(X_i, \cdot)$.

A test point $X$ is declared an outlier if the corresponding reconstruction error $\mathrm{RE}(X)$ is larger than some critical value $\tau > 0$. In other words the acceptation region in the RKHS is the $\tau$-neighborhood of the subspace $V$. [Hof07] argues that when $k$ is a translation-invariant kernel that is $k(x, x') = \kappa(x - x')$ for some $\kappa : \mathcal{X} \to \mathbb{R}$, this acceptation region encloses the embedded inliers more tightly than the acceptation region yielded by oc-SVM. This is due to the fact that embedded points all lie on the same sphere $\mathcal{S}_0$ of $H(k)$ when $k$ is a translation-invariant kernel. In this case, it is known that the intersection of $\mathcal{S}_0$ with the separating hypersphere of oc-SVM coincides with the intersection of $\mathcal{S}_0$ and the decision boundary $\mathcal{S}$ provided by Support Vector Domain Description (SVDD) [TD99] — which is also a sphere. Therefore oc-SVM and SVDD classify points in the same way. Since oc-SVM amounts to define the decision boundary as a sphere, it does not take into account possible heterogeneous variances of the embedded data along different directions in the RKHS, unlike kernel PCA. For this reason, kernel PCA yields a smallest acceptation region that encloses inliers in the RKHS hence kernel PCA is less prone to accepting actual outliers as inliers than oc-SVM.

Note that parametric OD methods that involve density estimation are contained in kernel PCA, since OD with Parzen window density estimation can be seen as a special case of OD with kernel PCA (as long as a positive definite kernel is involved). A Parzen window density estimator $\hat{f}$ is of the form $\hat{f}(x) = (1/nh) \sum_{i=1}^n K\left((x - x_i)/h\right)$ where $K(\cdot)$ is some real positive function satisfying $\int K(t) \mathrm{d}t = 1$ and $h > 0$ is the bandwidth parameter. It turns out that OD with $\hat{f}$ coincides with kernel PCA when choosing a number $p$ of principal component equal to $p = 0$. In this case, the

projector $\Pi_V$ is equal to 0 and the reconstruction error $\mathrm{RE}(x)$ writes

$$\begin{aligned}\mathrm{RE}(x) &= \left\| k(x,\cdot) - \frac{1}{n}\sum_{i=1}^{n} k(x_i,\cdot) \right\|^2_{H(k)} \\ &= k(x,x) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(x_i,x_j) - \frac{2}{n}k(x_i,x)\,.\end{aligned}$$

Since $k$ is chosen as a translation-invariant kernel, $k(x,x) = \kappa(0)$ is constant, and $(1/n^2)\sum_{i,j=1}^{n} k(x_i,x_j)$ does not depend on $x$ as well. Therefore, an outlier $x$ corresponds to a large value of $\mathrm{RE}(x)$ which in turn corresponds to a small value of $(1/n)\sum_{i=1}^{n} k(x_i,x)$. The latter quantity matches the Parzen window density estimator with $K(z) = h\kappa(hz)$. On the other hand, outlier detection with kernel PCA can be seen as a special case of [DRT14].

The main drawback of kernel PCA is its expensive computational cost. Indeed, finding the principal components $\varphi_1,\dots,\varphi_p$ involves the eigendecomposition of the Gram matrix $\mathbf{K} = \left[k(x_i,x_j)\right]_{1\leqslant i,j\leqslant n}$ which is done in $\mathcal{O}(n^3)$ time. Only considering the $p$ first principal components actually requires an execution time of order $\mathcal{O}(n^2 p)$. However the optimal value of $p$ is not known, hence the OD algorithm must be reiterated for several possible values of $p$ (up to the case $p = n$) in order to observe which choice of $p$ yields the best performance — for instance this is done to compute the criterion proposed by [XWX14] to select the optimal $p$.

## 5.2 New approach for outlier detection

### 5.2.1 Principle

We propose a new kernel-based method for outlier detection that circumvents the respective shortcomings of the two methods presented in Section 5.1.2. This new approach mostly relies on the results in Chapter 4 about the distribution of low-dimensional projections in Gaussian RBF kernel spaces, which will allow for a control of the probability of false alarm error.

Using the notation in Section 5.1.1, we assume a sample $X_1,\dots,X_n,X$ is available where $X_1,\dots,X_n \sim P$ are *i.i.d.* $\mathcal{X}$-valued variables representing the inliers and $X$ is a test point. Besides, we consider the Gaussian RBF kernel $k_\gamma$ defined as

$$k_\gamma(x,x') \overset{\Delta}{=} \exp\left(-\gamma d^2(x,x')\right) \quad \text{for every } x,x' \in \mathcal{X}\,,$$

and parameterized by $\gamma > 0$.

Our method is twofold. In the first step, a subspace $V_n$ of $H(k_\gamma)$ is randomly selected and defined as

$$V_n = \mathrm{Span}(h_1,\dots,h_p)\,,$$

where $\mathbf{h} = (h_1, \ldots, h_p)$ are $p$ *i.i.d.* zero-mean Gaussian processes with covariance function $\Sigma_{\gamma,n}(\cdot,\cdot)$:

$$\Sigma_{\gamma,n}(x,x') \triangleq \frac{1}{n}\sum_{i=1}^{n} k_\gamma(X_i,x)k_\gamma(X_i,x') \quad \text{for every } x,x' \in \mathcal{X} ,$$

that is the empirical estimator of $\Sigma_\gamma(x,x') = \mathbb{E}_{X_1\sim P}\Big[k_\gamma(X_1,x)k_\gamma(X_1,x')\Big]$.

In the second step, the embedded test point $k_\gamma(X,\cdot) \in H(k_\gamma)$ is projected onto $V_n$. The projection of $k_\gamma(X,\cdot)$ onto $V_n$ is represented via the $p$-variate vector $p_{V_n}(X)$ whose entries are the coordinates of the projection with respect to the basis $\mathbf{h}$. Since $\mathbf{h}$ forms asymptotically an orthonormal basis of $V_n$ when $\gamma \to +\infty$ (see Lemma 4.3.1), $p_{V_n}(X)$ can be simply defined as

$$\begin{aligned} p_{V_n}(X) &\triangleq \Big(\langle k_\gamma(X,\cdot),h_1\rangle_{H(k_\gamma)} \;\; \ldots \;\; \langle k_\gamma(X,\cdot),h_p\rangle_{H(k_\gamma)}\Big)^\top \\ &= \Big(h_1(X) \;\; \ldots \;\; h_p(X)\Big)^\top . \end{aligned}$$

Finally we are in a position to introduce the test statistic

$$S_n(X) \triangleq \gamma^s \left\| p_{V_n}(X) \right\|^2 , \tag{5.2.3}$$

where $\|\cdot\|$ is the Euclidean norm of $\mathbb{R}^p$ and $s > 0$. When $\mathcal{X} = \mathbb{R}^d$, we set $s = d/2$.

Whether $X$ is an outlier or an inlier, $S_n(X)$ exhibits two distinct behaviours.

If $\mathbf{H}_0$ is true ($X$ is an inlier), Proposition 4.3.5 states that when $n \to +\infty$ and $\gamma$ grows slowly to infinity compared to $n$ , the distribution $\gamma^{s/2} p_{V_n}(X)$ is close to that of a random variable $\mathfrak{s}G$ where $G$ is a $\mathcal{N}(0, I_p)$ Gaussian vector and $\mathfrak{s}^2$ is a copy of $\gamma^s \Sigma_\gamma(X_1,X_1)$ independent of $G$ and where $X_1 \sim P$, which defines a scale-mixture of (isotropic) Gaussians (SMG). Therefore the Continuous Mapping Theorem and the continuity of $\|\cdot\|^2$ entail that $S_n(X)$ is well-approximated by a variable $\mathfrak{s}^2 Q^2$ where $Q^2 \sim \chi^2(p)$ is independent of $\mathfrak{s}$. Note that by Lemma A.2.1, $\mathfrak{s}^2 = \gamma^s \Sigma_\gamma(X_1,X_1) \underset{\gamma\to+\infty}{\asymp} 1$ so that the distribution of $S_n(X)$ is not asymptotically reduced to a Dirac distribution at 0 — hence the $\gamma^s$ term in (5.2.3).

If $\mathbf{H}_1$ is true ($X$ is an outlier), $S_n(X)$ converges almost surely to 0. To see this, note that since $h_1, \ldots, h_p$ have $\Sigma_{\gamma,n}(\cdot,\cdot)$ as a covariance function, then

$$V_n \subseteq V = \mathrm{Span}\Big\{ k_\gamma(x,\cdot) \;\Big|\; x \in \mathrm{supp}(P) \Big\} \quad P-\text{almost surely} .$$

On the other hand, since an outlier $X$ lies outside the support of $P$ by definition, then for any $x \in \mathrm{supp}(P)$

$$\langle k_\gamma(x,\cdot), k_\gamma(X,\cdot)\rangle_{H(k_\gamma)} = \exp\Big(-\gamma d^2(x,X)\Big) = \mathcal{O}(e^{-\gamma d^2(X)}) \underset{\gamma\to+\infty}{\longrightarrow} 0 ,$$

where $d(X) = \inf\Big\{ d(X,x) \;\Big|\; x \in \mathrm{supp}(P) \Big\}$ denotes the distance between $X$ and the support of $P$. It

follows that $k_\gamma(X,\cdot)$ is asymptotically orthogonal to the subspace $V$ — hence also to the subspace $V_n$ almost surely — when $\gamma \to +\infty$ in such a way that

$$S_n(X) = \gamma^s \|p_{V_n}(X)\|^2 = \mathcal{O}_{P_1}(\gamma^s e^{-2\gamma d^2(X)}) \underset{\gamma\to+\infty}{\longrightarrow} 0 \,.$$

Now we are in a position to decide whether $X$ is an outlier or not. Let us introduce the $p$-value

$$\mathrm{PV}(X) = \mathbb{P}_{\mathfrak{s}^2,Q^2}\left(S_n(X) > \mathfrak{s}^2 Q^2\right) , \tag{5.2.4}$$

where the probability is evaluated solely over $\mathfrak{s}^2$ and $Q^2$. Given a confidence level $\alpha \in (0,1)$, the null-hypothesis $\mathbf{H}_0$ is rejected if and only if $\mathrm{PV}(X) < \alpha$.

Note that we did not consider renormalized projections $\tilde{p}_{V_n}(X) = \Sigma_{\gamma,n}^{-1/2}(X,X)p_{V_n}(X)$ whereas we saw in Chapter 4 that they converge weakly to a $\mathcal{N}(0,I_p)$ instead of a SMG, which would have yielded a simpler null-distribution for our statistic. The problem is that when $X$ is an outlier, $\tilde{p}_{V_n}(X)$ is no more guaranteed to converge almost surely to 0 unlike $p_{V_n}(X)$. This is due to the fact that the renormalizer $\Sigma_{\gamma,n}^{1/2}(X,X)$ tends faster to 0 when $X$ is an outlier than otherwise. Indeed when $X$ is an outlier, $\Sigma_{\gamma,n}(X,X) = (1/n)\sum_{i=1}^n k_\gamma^2(X,X_i) \leqslant \exp\left(-2\gamma d^2(X)\right)$ and since $S_n(X) = \mathcal{O}(e^{-\gamma d^2(X)})$ as said above, so that there is no guarantee that $\|\tilde{p}_{V_n}(X)\|^2 = S_n(X)/\Sigma_{\gamma,n}(X,X)$ tends to 0 a.s. under the alternative.

### 5.2.2 Practical implementation

This section presents in detail how our OD method is set up in practice.

For every $j = 1,\dots,p$, $h_j$ can be written as $h_j = (1/\sqrt{n})\sum_{i=1}^n u_i^{(j)} k_\gamma(X_i,\cdot)$ where the variables $\{u_i^{(j)}\}_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}$ are *i.i.d.* $\mathcal{N}(0,1)$ Gaussians. This leads to the following matrix formulation of $p_{V_n}(X)$

$$p_{V_n}(X) = n^{-1/2}\mathbf{U}\mathbf{k}_\mathbf{x} \,,$$

where $\mathbf{U} = [u_i^{(j)}]_{j,i} \in \mathcal{M}_{p,n}(\mathbb{R})$ and $\mathbf{k}_\mathbf{x} = (k_\gamma(X_1,X)\dots k_\gamma(X_n,X))^\top \in \mathbb{R}^n$.

This way the test statistic $S_n(X)$ writes

$$S_n(X) = \gamma^s \|p_{V_n}(X)\|^2 = n^{-1}\gamma^s \mathbf{k}_\mathbf{x}^\top \mathbf{U}^\top \mathbf{U}\mathbf{k}_\mathbf{x} \,.$$

The next step consists in determining the $p$-value $\mathrm{PV}(X)$ of the test as defined in (5.2.4). Introducing $F_{\chi^2(p)}(t) = \mathbb{P}(Q^2 > t)$ the cumulative distribution function (cdf) of the $\chi^2(p)$ distribution, $\mathrm{PV}(X)$ alternatively reads

$$\mathrm{PV}(X) = \mathbb{E}_{\mathfrak{s}^2}\left[F_{\chi^2(p)}\left(\frac{S_n(X)}{\mathfrak{s}^2}\right)\right] , \tag{5.2.5}$$

where the mean is taken only over $\mathfrak{s}^2$. However $\mathfrak{s}^2$ is a copy of $\gamma^s\Sigma_\gamma(X_1,X_1)$ ($X_1 \sim P$) while $\Sigma_\gamma$ and the distribution $P$ of $X$, are unknown, therefore $\mathrm{PV}(X)$ needs to be estimated. To this aim, we can consider for each $i = 1,\dots,n$ the quantities $\hat{\mathfrak{s}}_i = \gamma^s\Sigma_{\gamma,-i}(X_i,X_i)$ where $\Sigma_{\gamma,-i}(X_i,X_i) = 1/(n-1)\sum_{j\neq i} k_\gamma^2(X_j,X_i)$ is an estimate of $\Sigma_\gamma(X_i,X_i)$ (conditionally to $X_i$).

This leads to the estimated $p$-value $\widehat{\mathrm{PV}}(X)$

$$\widehat{\mathrm{PV}}(X) = \frac{1}{n}\sum_{i=1}^{n} F_{\chi^2(p)}\left(\frac{S_n(X)}{\hat{\mathfrak{s}}_i^2}\right).$$

However, the estimation of the $p$-value may be improved upon in terms of execution time. Actually, computing the quantities $\hat{\mathfrak{s}}_i$ requires to access the whole Gram matrix $\mathbf{K} = [k_\gamma(X_1,X_j)]_{1\leqslant i,j\leqslant n}$ which involves the computation of its $n(n-1)/2$ entries, then for each $i = 1,\dots,n$ calculating $\Sigma_{\gamma,-i}(X_i,X_i)$ takes in the order of $n$ operations. All in all, the computational cost is of order $\mathcal{O}(n^2)$. This quadratic time can be reduced to linear time through a low-rank approximation of $\mathbf{K}$ by the methods presented in Section 2.4. Such a method yields an approximation of $\mathbf{K}$ as

$$\mathbf{K} \simeq \mathbf{U}\mathbf{U}^\top \qquad \text{with } \mathbf{U} \in \mathcal{M}_{n,r}(\mathbb{R}), \tag{5.2.6}$$

where $r$ is much smaller than $n$. Using this matrix factorization, each $\Sigma_{\gamma,-i}(X_i,X_i)$ can be approached as follows

$$\begin{aligned}\Sigma_{\gamma,-i}(X_i,X_i) &= \frac{1}{n-1}e_i^\top\mathbf{K}^2e_i - \frac{1}{n-1}\\ &\simeq \frac{1}{n-1}e_i^\top\mathbf{U}\mathbf{W}\mathbf{U}^\top e_i - \frac{1}{n-1},\end{aligned} \tag{5.2.7}$$

where $\mathbf{W} = \mathbf{U}^\top\mathbf{U} \in \mathcal{M}_r(\mathbb{R})$ and $(e_1,\dots,e_n)$ is the canonical base of $\mathbb{R}^n$. $\mathbf{W}$ can be calculated beforehand in $\mathcal{O}(r^2n)$ time. Then the computation of $\Sigma_{\gamma,-i}(X_i,X_i)$ through (5.2.7) takes $\mathcal{O}(r^2)$ operations, hence a total computation time of order $\mathcal{O}(r^2n)$ linear with respect to $n$. The influence of the choice of $r$ is discussed later in Section 5.4.1.

#### Low-rank approximation of the Gram matrix

Two different ways of approximating the Gram matrix as in (5.2.6) can be used. The first one involves the Nyström method (Section 2.4.1). As a reminder, Nyström method consists in randomly picking a subset $\mathcal{I}$ of cardinality $r$ uniformly over $\{I \subseteq \{1,\dots,n\} \mid \mathrm{card}(\mathcal{I}) = r\}$ and approaching $\mathbf{K}$ by

$$\mathbf{K} \simeq \mathbf{K}_{\cdot,\mathcal{I}}\mathbf{K}_{\mathcal{I},\mathcal{I}}^{-1}\mathbf{K}_{\mathcal{I},\cdot},$$

where $\mathbf{K}_{\cdot,\mathcal{I}} = [\mathbf{K}_{ij}]_{1\leqslant i\leqslant n, j\in\mathcal{I}}$, $\mathbf{K}_{\mathcal{I},\mathcal{I}} = [\mathbf{K}_{ij}]_{i,j\in\mathcal{I}}$ and $\mathbf{K}_{\mathcal{I},\cdot} = \mathbf{K}_{\cdot,\mathcal{I}}^\top$. In this case $\mathbf{U} = \mathbf{K}_{\cdot,\mathcal{I}}\mathbf{K}_{\mathcal{I},\mathcal{I}}^{-1/2}$. Let $[\hat{\mathfrak{s}}_i^{nyst}]^2 = \gamma^s\Sigma_{\gamma,-i}^{nyst}(X_i,X_i)$ where $\Sigma_{\gamma,-i}^{nyst}(X_i,X_i)$ denotes the approximation of $\Sigma_{\gamma,-i}(X_i,X_i)$ based on

the Nyström method. We define the corresponding estimated $p$-value as

$$\widehat{PV}_{nyst}(X) = \frac{1}{n}\sum_{i=1}^{n} F_{\chi^2(p)}\left(\frac{S_n(X)}{[\hat{\mathfrak{s}}_i^{nyst}]^2}\right).$$

The second method to approximate $\mathbf{K}$ uses Random Kitchen Sinks (Section 2.4.2). To be able to apply this method, we need to assume that $\mathcal{X} = \mathbb{R}^d$ endowed with the usual Euclidean norm $\|\cdot\|$. Under this assumption, $k_\gamma(x,x') = \exp\left(-\gamma\|x-x'\|^2\right)$ corresponds to the Fourier transform of the $\mathcal{N}(0, 2\gamma I_d)$ Gaussian measure and can be approached as in (5.2.6) with

$$\mathbf{U} = r^{-1}\begin{pmatrix} \cos(w_1^\top X_1) & \sin(w_1^\top X_1) & \dots & \cos(w_r^\top X_1) & \sin(w_r^\top X_1) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \cos(w_1^\top X_n) & \sin(w_1^\top X_n) & \dots & \cos(w_r^\top X_n) & \sin(w_r^\top X_n) \end{pmatrix},$$

where $w_1,\dots,w_r \sim \mathcal{N}(0, 2\gamma I_d)$ *i.i.d.* Let $[\hat{\mathfrak{s}}_i^{RKS}]^2 = \gamma^s \Sigma^{RKS}_{\gamma,-i}(X_i,X_i)$ where each $\Sigma^{RKS}_{\gamma,-i}(X_i,X_i)$ denotes the approximation of $\Sigma_{\gamma,-i}(X_i,X_i)$ based on RKS. We define the corresponding estimated $p$-value as

$$\widehat{PV}_{RKS}(X) = \frac{1}{n}\sum_{i=1}^{n} F_{\chi^2(p)}\left(\frac{S_n(X)}{[\hat{\mathfrak{s}}_i^{RKS}]^2}\right).$$

## 5.3 Theoretical analysis

When testing a point $X$, there are two ways of committing an error: either by labeling $X$ as an outlier while it is not (false alarm error) or by accepting $X$ as an inlier while it is an outlier (missed detection error). In the following, we call *false alarm rate* the probability of a false alarm error, and *missed detection rate* the probability of a missed detection rate. In the hypothesis testing terminology, the false alarm rate corresponds to a Type-I error and missed detection rate to a Type-II error. The aim of this section is to study theoretically how those two kinds of error are controlled for our OD method. In order not to overload theoretical developments, our analysis focuses on the version of our OD procedure that relies on the exact $p$-value $\mathrm{PV}(X)$.

### 5.3.1 False alarm rate

When $X \sim P$, Theorem 4.3.5 in Chapter 4 provides an upper bound of the general form

$$\Delta^2\left(\gamma^{s/2} p_{V_n}(X), \mathfrak{s}G\right) \leqslant C_1\gamma^{-s/2} + C_2\gamma^{s/2}n^{-1/2}, \tag{5.3.8}$$

with high probability over $h_1,\dots,h_p$ and $X_1,\dots,X_n$ and where $\gamma$ denotes the parameter of the kernel $k_\gamma$. Here $\Delta(\cdot,\cdot)$ is some distance between distributions on $\mathbb{R}^p$ and $s > 0$ is some quantity

related to the distribution $P$ — for instance $s = D/2$ when $\mathcal{X} = \mathbb{R}^D$ and $d(\cdot,\cdot)$ is the Euclidean distance.

The bound in (5.3.8) shows that the weak convergence of the null-distribution of $S_n(X)$ to $\mathfrak{s}^2 Q^2$ holds when $\gamma$ and $n$ both grow to infinity and $\gamma = o(n^{1/s})$. More precisely, the right hand side of (5.3.8) is minimized by setting every additive terms of the bound at the same rate $n^{-1/4}$, that is by setting $\gamma = Cn^{1/(2s)}$ for some constant $C > 0$. However this does not tell much about the rate of convergence of the actual false alarm rate $\mathbb{P}(PV(X) < \alpha \mid \mathbf{H}_0)$ to the prescribed level of confidence $\alpha$. Theorem 5.3.1 thereafter fills this gap and its proof can be found in Section 5.A.1.

**Theorem 5.3.1.** *Let $X \sim P$ and assume there exists a bounded, continuous function $A : \mathcal{X} \to \mathbb{R}_+$ and $s > 0$ such that*

$$\forall x \in \operatorname{supp}(P)\,, \quad \mathbb{P}_{X' \sim P}\left(d^2(x, X') \leqslant t\right) \sim A(x) t^s\,, \quad \text{when } t \to 0\,,$$

*where $\operatorname{supp}(P) = \{x \in \mathcal{X} \;\big|\; \forall \epsilon > 0,\ \mathbb{P}_{X' \sim P}(d^2(x, X') < \epsilon) > 0\}$ defines the support of $P$ and $A(x) = 0$ for every $x \notin \operatorname{supp}(P)$. Also assume that $A(\cdot)$ is lower bounded by some $a > 0$ on its support.*

*Then for any $\alpha, \delta \in (0,1)$, the actual false alarm rate is bounded as follows*

$$\mathbb{P}\left(PV(X) < \alpha \;\middle|\; \mathbf{H}_0\right) \leqslant \alpha + C_p \tau_\alpha^{-2} \mathbb{E}_X[A(X)] \sqrt{\frac{2\log(1/\delta)\Gamma(s+1)\gamma^s}{an}} \left[1 + \kappa \sqrt{\frac{\log(1/\delta)(2\gamma)^s}{a\Gamma(s+1)n}}\right] \{1 + o_\gamma(1)\} + 2C_p \tau_\alpha^{-2} \gamma^s \delta\,, \tag{5.3.9}$$

*where $\tau_\alpha$ is the $\alpha$-quantile of the null distribution $\mathfrak{s}G$ with $G \sim \mathcal{N}(0, I_p)$, $\kappa$ is a numerical constant and $C_p$ only depends on $p$.*

One could use the upper bound (5.3.9) as a proxy to control the actual false alarm rate by choosing a (near-)optimal $\gamma$ in a closed form. The problem is that such a closed form expression would involve some terms in the upper bound which depend on quantities related to the inlier distribution $P$ — as an example when $\mathcal{X} = \mathbb{R}^D$ and $d(\cdot,\cdot)$ is the Euclidean distance, $A(\cdot)$ is related to the density function $f$ of $P$. But as we mentioned earlier, OD methods are made to avoid the complete estimation of the density function of $P$. In Section 5.5, we address the problem of selecting $\gamma$ by presenting a procedure based on grid search and a one-sample test statistic.

### 5.3.2 Missed detection rate

We call *missed detection rate* the probability of committing a missed detection error, which is denoted by $\beta_n$

$$\beta_n = \mathbb{P}_X\left(PV(X) > \alpha \;\middle|\; \mathbf{H}_1\right) = \mathbb{P}_X\left(S_n(X) > \tau_\alpha \;\middle|\; \mathbf{H}_1\right),$$

where $\mathbb{P}_X$ is taken with respect to $X$ and $\tau_\alpha$ is the $\alpha$-quantile of $\mathfrak{s}^2Q^2$. In the hypothesis testing terminology, out OD procedure is said to be consistent if $\beta_n \to 0$ when $n$ grows to infinity. Theorem 5.3.2 thereafter provides an upper bound for $\beta_n$ that entails the consistency of our method. The proof of Theorem 5.3.2 is provided in Section 5.A.2.

**Theorem 5.3.2** (Missed detection rate)**.** *Let $X \sim P_1 \neq P$ and assume the same assumptions concerning the distribution $P$ as in Theorem 5.3.1.*

*Let $F$ denote the cdf of $d(X) = \inf\{\|X - x\| : x \in \mathrm{supp}(P)\}$ with $X \sim P_1$.*

*Then for any $\delta \in \left(\frac{np}{(np)^2+1}, np\right)$, there exists an event with probability (over $\mathbf{h}$) larger than $1 - (2 - np/\delta)^{-np/2} e^{-(np-\delta)/2}$ under which*

$$\beta_n \leqslant F\left(\frac{s\log(2\gamma) + \log(n) + \log\left(\frac{1}{a\Gamma(s+1)}\right) + \frac{2\log(1/\alpha)}{p} + 2\sqrt{\frac{2\log(1/\alpha)}{p}} + \log(\delta) + o_{\gamma,n}(1)}{2\gamma}\right), \quad (5.3.10)$$

*where $\alpha \in (0,1)$ is the prescribed confidence level and $o_{\gamma,n}(1) \underset{\gamma,n\to+\infty}{\longrightarrow} 0$.*

Note that Theorem 5.3.2 only assumes that the tested point $X$ was generated by an alternative distribution $P_1$ different than $P$ but not necessarily with a support distinct from the support of $P$. However the upper bound in (5.3.10) decreases to 0 only if the alternative $\mathbf{H}_1$ holds, that is $\mathrm{supp}(P_1) \cap \mathrm{supp}(P)$ is negligible. In this case the distance $d(X)$ between $X$ and $\mathrm{supp}(P)$ is $P_1$-almost surely non-null hence $F(0) = 0$. On the other hand, when the sample size $n$ gets larger, the hyperparameter $\gamma$ is "allowed" to increase as well while still controlling the false alarm rate at level $\alpha$ according to the analysis done in Section 5.3.1. For instance, we saw that setting $\gamma \propto n^{1/(2s)}$ leads to an asymptotic control of the false alarm rate. With this choice of $\gamma$ and when $n$ grows to infinity, the upper bound in (5.3.10) entails that $\beta_n$ decreases to 0 at the rate $\mathcal{O}(F(\log(n)n^{-1/(2s)}))$ which tends to 0 since $F(0) = 0$. This way Theorem 5.3.2 guarantees that our OD procedure is consistent against $\mathbf{H}_1$.

Let us discuss the influence of some key quantities on the upper bound (5.3.10). First of all, since $F$ is a cdf hence an increasing function, the upper bound increases when the lower bound $a$ of $A(\cdot)$ gets smaller. To understand how a small $a$ worsens the performance of our OD method, it may be helpful to resort to the multivariate case $\mathcal{X} = \mathbb{R}^d$ with $d(\cdot,\cdot)$ as the Euclidean distance in which case $A(x) \propto f(x)$ and $f(\cdot)$ is the density of $P$ (Proposition 4.3.2 in Chapter 4). In this case, $a$ must be interpreted as an indicator of the values of $f$ on small density regions. Therefore depending on the form of $f$, a small $a$ may be influential in two different ways. The first case is when the distribution $P$ is highly unbalanced, that is there is a large discrepancy between highest and smallest values of $f$. Thus a small $a$ means that $P$ is mostly concentrated on a smaller subset of its support. In this case our OD algorithm is likely to mistake this subset of $\mathrm{supp}(P)$ as the entire support. In order to keep the false alarm rate at the prescribed level $\alpha$, the OD procedure is compelled to compensate by erroneously "completing" this highly concentrated part of $\mathrm{supp}(P)$ with a "dumb" subset of $\mathcal{X}$ outside of $\mathrm{supp}(P)$, hence a larger missed detection

rate. The second case is when $P$ is concentrated homogeneously across its support — ideally when $P$ is a uniform distribution. In this situation, a small $a$ entails that $\mathrm{supp}(P)$ is a large subset of $\mathbb{R}^d$ hence the inliers sample $X_1, \ldots, X_n$ tends to be sparsely spread across $\mathrm{supp}(P)$ so that it is more difficult for our OD algorithm to determine the boundary of $\mathrm{supp}(P)$, which leads to poor performance.

The shape of the cdf $F$ in the vicinity of 0 also influences the bound of $\beta_n$ in the following sense. If $F(t)$ decreases to 0 when $t \to 0$ at a large speed, then the upper bound decreases to 0 at a faster rate as well hence better performances. On the other hand, a fast decrease of $F(t)$ near $t = 0$ means that outliers tend to lie far from the boundary of $\mathrm{supp}(P)$. Therefore a small missed detection rate is expected, which is confirmed by the bound.

As for the confidence level $\alpha$, reducing $\alpha$ makes the upper bound larger. In hypothesis testing terms, a smaller Type-I error $\alpha$ is supposed to make the Type-II error $\beta_n$ larger. Thus the bound reflects well the expected influence of $\alpha$ on $\beta_n$.

## 5.4 Numerical results

### 5.4.1 False alarm and missed detection rates

In this section, we empirically observe the false alarm and missed detection rates on simulated data, in particular we discuss on the influence of the parameters $\gamma$ and $p$ on those two types of errors.

We apply our OD procedure on the synthetic dataset "spiral" displayed in Figure 5.1. The blue dots represent the inliers, that are distributed on a spiral-shaped subset of $\mathbb{R}^2$. $n = 500$ inliers are generated to form the training set. The test set is made of 200 other inliers and 200 outliers, the latter corresponding to the red dots in Figure 5.1. Each observation in the test set is tested on the basis of the $n = 500$ inliers of the training set. The results on the tested inliers (resp. outliers) are averaged out to get false alarm rates (resp. missed detection rate).

Figure 5.2 shows the false alarm rates obtained when applying our OD method on the 200 test inliers, for several values of $\gamma$ ranging from 0.1 to 3000 and several values of $p$ ranging from 2 to 500. The prescribed level of confidence was set at $\alpha = 0.05$, so that false alarm rates close to 0.05 are expected. For $\gamma$ up to 75 and $p$ smaller than 10, the actual false alarm rates are satisfyingly close to 0.05. However when $\gamma$ takes values beyond $\gamma = 100$, the false alarm rate is no longer controlled, being as high as 0.805 when $\gamma = 3000$ and $p = 500$. This is consistent with Theorem 4.3.5 which implies that $\gamma$ must not be too large compared to $n$ in the empirical setting. On the other hand, the bound of Theorem 4.3.5 also suggested that a larger $p$ had a bad impact on the distributional approximation of $S_n(X)$ and in practice, the control of the type-I error seems to be significantly worsened by larger values of $p$, especially when $\gamma \geqslant 25$.

Figure 5.3 displays the corresponding missed detection rates. As foretold by Theorem 5.3.2, the missed detection rate tends to 0 whenever $\gamma \to +\infty$ and decreases when $p$ increases. Note

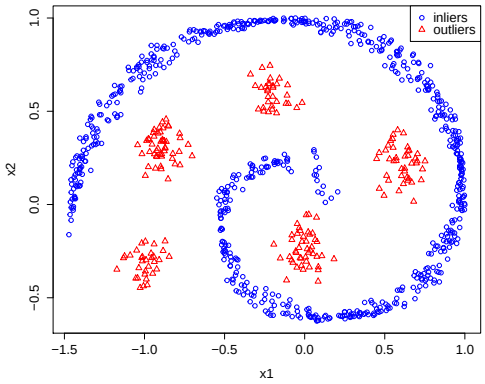


Figure 5.1 – "Spiral" synthetic dataset

| | $\gamma$ | 0.1 | 1 | 5 | 25 | 50 | 75 | 100 | 500 | 1000 | 3000 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $p =$ | 2 | 0.054 | 0.060 | 0.052 | 0.052 | 0.058 | 0.060 | 0.062 | 0.124 | 0.219 | 0.452 |
| | 10 | 0.061 | 0.046 | 0.058 | 0.061 | 0.062 | 0.075 | 0.086 | 0.260 | 0.400 | 0.667 |
| | 25 | 0.043 | 0.050 | 0.067 | 0.068 | 0.079 | 0.080 | 0.096 | 0.296 | 0.450 | 0.730 |
| | 50 | 0.045 | 0.042 | 0.059 | 0.071 | 0.083 | 0.092 | 0.106 | 0.311 | 0.469 | 0.759 |
| | 100 | 0.027 | 0.043 | 0.067 | 0.071 | 0.084 | 0.096 | 0.106 | 0.317 | 0.477 | 0.779 |
| | 500 | 0.054 | 0.034 | 0.068 | 0.075 | 0.090 | 0.101 | 0.114 | 0.320 | 0.488 | 0.805 |

Figure 5.2 – False alarm rates with the "spiral" dataset

| | $\gamma$ | 0.1 | 1 | 5 | 25 | 50 | 75 | 100 | 500 | 1000 | 3000 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $p =$ | 2 | 0.947 | 0.950 | 0.807 | 0.074 | 0.018 | 0.012 | 0.009 | 0.001 | 0 | 0 |
| | 10 | 0.950 | 0.947 | 0.446 | 0.012 | 0.005 | 0.001 | 0 | 0 | 0 | 0 |
| | 25 | 0.972 | 0.920 | 0.245 | 0.007 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 50 | 0.966 | 0.906 | 0.142 | 0.004 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 100 | 0.989 | 0.866 | 0.09 | 0.003 | 0 | 0 | 0 | 0 | 0 | 0 |
| | 500 | 0.988 | 0.798 | 0.054 | 0.001 | 0 | 0 | 0 | 0 | 0 | 0 |

Figure 5.3 – Missed detection rates with the "spiral" dataset

however that the observation of small type-II errors is amplified by loosely controlled type-I errors for large $\gamma$ or $p$, since a larger Type-I error entails a smaller Type-II error.

The same experiments were conducted when using Nyström and Random Kitchen Sinks approximations as in Section 5.2.2. For Nyström method, false alarm and missed detection rates were measured for the same range of values of $\gamma$ and $p$ as above, in a first time for a small rank $r = 50$ and in the second time for a larger rank $r = 400$. The same was done for RKS with $r = 50$ and $r = 300$. The results are shown in Figures 5.4 to 5.7. When $r$ is large enough ($r = 400$ for Nyström and $r = 300$ for RKS), performances comparable to those without Gram matrix approximation are observed. For Nyström with $r = 400$, choosing $25 \leqslant \gamma \leqslant 50$ and $p = 2$ yields false alarm rates of about 0.6 and a missed detection rate below 0.01. Similarly for RKS with $r = 300$, setting $\gamma = 50$ and $p = 2$ yields a false alarm rate of 0.06 and a missed

detection rate of 0.009. As long as $r$ is large enough, the parameters $\gamma$ and $p$ show similar effects as previously, that is increasing $\gamma$ and/or $p$ decreases the missed detection rate and increases the false error rate. In particular the effect of a larger $p$ on the false error rate seems to be more visible when Nyström/RKS approximation are used. This may be due to the fact that such an approximation implicitly consists in projecting embedded data onto a low-dimensional subspace in the RKHS (for instance, the subspace spanned by a small subsample of the embedded dataset for Nyström method). Therefore before projecting onto $V_n$, the projected points does not lie in a high-dimensional RKHS but in a low-dimensional subspace of this RKHS, and this low dimensionality may prevent the weak convergence to a SMG from occurring hence the less controlled false alarm rates when $p$ is not negligible compared to $r$.

All in all when it comes to calibrating $\gamma$ and $p$, only false alarm rate control matters, since $\gamma$ and $p$ can be set as large as wanted when only considering missed detection error. In other words, optimizing the parameters is only constrained by false alarm rate control. This is advantageous since false alarm rate is related to the inlier distribution which is accessible as many inliers are available unlike outliers. A tuning procedure is proposed in details later in Section 5.5.

| | $\gamma$ | 5 | 25 | 50 | 100 | 1000 |
|---|---|---|---|---|---|---|
| $p =$ | 2 | 0.044 | 0.032 | 0.012 | 0.001 | 0 |
| | 10 | 0.060 | 0.010 | 0.001 | 0 | 0 |
| | 25 | 0.058 | 0.008 | 0 | 0 | 0 |
| | 50 | 0.065 | 0.007 | 0 | 0 | 0 |
| | 100 | 0.062 | 0.009 | 0 | 0 | 0 |
| | 500 | 0.056 | 0.009 | 0 | 0 | 0 |

| | $\gamma$ | 5 | 25 | 50 | 100 | 1000 |
|---|---|---|---|---|---|---|
| $p =$ | 2 | 0.819 | 0.124 | 0.065 | 0.058 | 0.119 |
| | 10 | 0.465 | 0.048 | 0.047 | 0.055 | 0.071 |
| | 25 | 0.273 | 0.060 | 0.044 | 0.068 | 0.070 |
| | 50 | 0.153 | 0.046 | 0.063 | 0.068 | 0.077 |
| | 100 | 0.115 | 0.035 | 0.044 | 0.047 | 0.085 |
| | 500 | 0.090 | 0.062 | 0.058 | 0.063 | 0.091 |

Figure 5.4 – False alarm rates (left) and missed detection rates (right) with the "spiral" dataset and the Nyström approximation with $r = 50$

| | $\gamma$ | 5 | 25 | 50 | 100 | 1000 |
|---|---|---|---|---|---|---|
| $p =$ | 2 | 0.053 | 0.060 | 0.061 | 0.098 | 0.137 |
| | 10 | 0.065 | 0.074 | 0.084 | 0.192 | 0.230 |
| | 25 | 0.078 | 0.086 | 0.098 | 0.225 | 0.256 |
| | 50 | 0.068 | 0.087 | 0.102 | 0.238 | 0.260 |
| | 100 | 0.074 | 0.089 | 0.108 | 0.251 | 0.259 |
| | 500 | 0.072 | 0.096 | 0.111 | 0.255 | 0.269 |

| | $\gamma$ | 5 | 25 | 50 | 100 | 1000 |
|---|---|---|---|---|---|---|
| $p =$ | 2 | 0.059 | 0.009 | 0.003 | 0 | 0 |
| | 10 | 0.006 | 0 | 0 | 0 | 0 |
| | 25 | 0.001 | 0 | 0 | 0 | 0 |
| | 50 | 0 | 0 | 0 | 0 | 0 |
| | 100 | 0 | 0 | 0 | 0 | 0 |
| | 500 | 0 | 0 | 0 | 0 | 0 |

Figure 5.5 – False alarm rates (left) and missed detection rates (right) with the "spiral" dataset and the Nyström approximation with $r = 400$

| | $\gamma$ | 5 | 25 | 50 | 100 | 1000 |
|---|---|---|---|---|---|---|
| $p =$ | 2 | 0.062 | 0.072 | 0.080 | 0.100 | 0.406 |
| | 10 | 0.082 | 0.143 | 0.206 | 0.323 | 0.888 |
| | 25 | 0.117 | 0.209 | 0.294 | 0.484 | 0.971 |
| | 50 | 0.110 | 0.255 | 0.392 | 0.566 | 0.991 |
| | 100 | 0.132 | 0.313 | 0.442 | 0.623 | 0.997 |
| | 500 | 0.151 | 0.379 | 0.503 | 0.679 | 1 |

| | $\gamma$ | 5 | 25 | 50 | 100 | 1000 |
|---|---|---|---|---|---|---|
| $p =$ | 2 | 0.801 | 0.044 | 0.007 | 0.001 | 0 |
| | 10 | 0.395 | 0.002 | 0 | 0 | 0 |
| | 25 | 0.165 | 0 | 0 | 0 | 0 |
| | 50 | 0.097 | 0 | 0 | 0 | 0 |
| | 100 | 0.063 | 0 | 0 | 0 | 0 |
| | 500 | 0.043 | 0 | 0 | 0 | 0 |

Figure 5.6 – False alarm rates (left) and missed detection rates (right) with the "spiral" dataset and Random Kitchen Sinks approximation with $r = 50$

| | $\gamma$ | 5 | 25 | 50 | 100 | 1000 |
|---|---|---|---|---|---|---|
| $p =$ | 2 | 0.057 | 0.053 | 0.060 | 0.073 | 0.250 |
| | 10 | 0.060 | 0.080 | 0.094 | 0.127 | 0.524 |
| | 25 | 0.071 | 0.090 | 0.110 | 0.163 | 0.637 |
| | 50 | 0.072 | 0.093 | 0.136 | 0.181 | 0.682 |
| | 100 | 0.073 | 0.098 | 0.135 | 0.189 | 0.705 |
| | 500 | 0.075 | 0.108 | 0.147 | 0.201 | 0.726 |

| | $\gamma$ | 5 | 25 | 50 | 100 | 1000 |
|---|---|---|---|---|---|---|
| $p =$ | 2 | 0.820 | 0.054 | 0.009 | 0.001 | 0 |
| | 10 | 0.446 | 0.005 | 0 | 0 | 0 |
| | 25 | 0.222 | 0.001 | 0 | 0 | 0 |
| | 50 | 0.133 | 0 | 0 | 0 | 0 |
| | 100 | 0.087 | 0 | 0 | 0 | 0 |
| | 500 | 0.058 | 0 | 0 | 0 | 0 |

Figure 5.7 – False alarm rates (left) and missed detection rates (right) with the "spiral" dataset and Random Kitchen Sinks approximation with $r = 300$

### 5.4.2 Comparison with other methods

We compare our OD method with the two existing OD methods presented in Section 5.1.2, that are oc-SVM and kernel PCA. These two latter methods use the same Gaussian RBF kernel

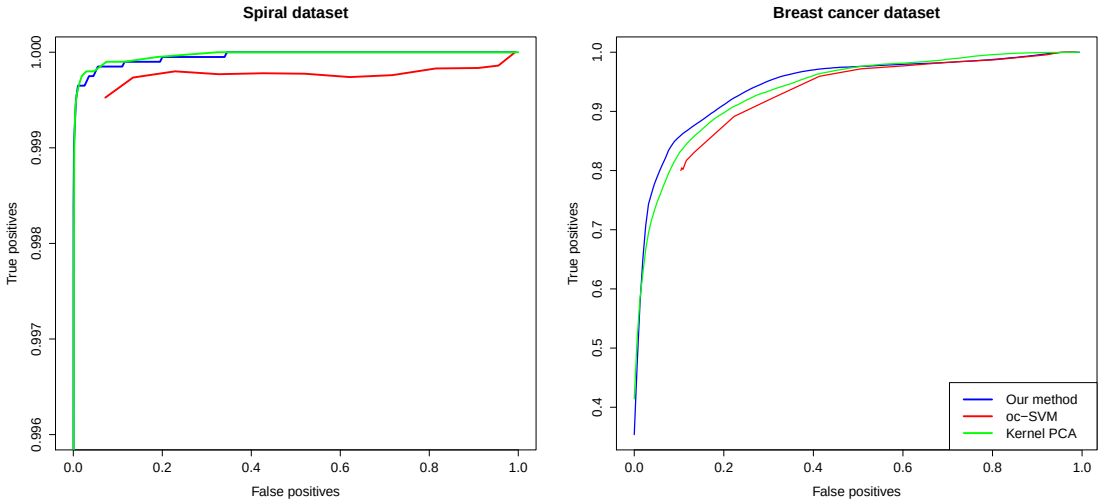


Figure 5.8 – ROC curves for our OD method (blue), oc-SVM (red) and kernel PCA (green), tested on synthetic "Spiral" dataset (left plot) and real-life "Cancer" dataset (right plot).

with parameter $\gamma > 0$ as ours. This comparison is done according to two criteria: detection performance and execution time.

### Detection performance

To assess the detection performances of the three competing methods, we applied them to two different datasets:

- **Synthetic dataset:** it is the same "spiral" dataset as the one used in Section 5.4.1. The training set contains $n = 500$ inliers, and the tested set contains 200 inliers and 200 outliers.
- **Real-life dataset:** it is the "cancer" dataset from the UCI machine learning repository [1]. It consists of 569 instances split into two groups, one with 357 instances labeled "benign" — these will be the inliers — and 112 instances labeled "malignant" — the outliers. Each instance is described by 10 covariates, and each covariate has been renormalized to have unit variance. The training set contains $n = 200$ instances, which yields a tested set with 157 inliers and 112 outliers.

For each method and dataset, a ROC curve is produced. For our method and kernel PCA, it consists in considering a varying threshold $\tau > 0$ and rejects each tested observation $X$ if and only if $S_n(X) < \tau$ or $\mathrm{RE}(X) > \tau$. For the oc-SVM, the parameter $\nu$ is set at several values spanning the interval $(0, 1]$. Varying $\tau$ or $\nu$ yields couples of false positives/true positives pairs that constitutes the ROC curve.

[1] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

Besides, each method is applied several times for several different values of its parameter, and the chosen parameters are those that maximize the AUC (area under curve). Namely, our method is parameterized by the pair $(\gamma, p)$ that has optimal values $(100, 50)$ for the "spiral" dataset and $(0.05, 100)$ for the "cancer" dataset. Oc-SVM is only parameterized by $\gamma$ that is optimal at $\gamma = 3$ for "spiral" and 0.05 for "cancer". Finally kernel PCA is parameterized by $(\gamma, p)$ where $p$ is the number of kept principal component, with optimal values $(50, 50)$ for "spiral" and $(0.8, 5)$ for "cancer".

Obtained ROC curves are displayed in Figure 5.8. One important observation to make is that the ROC curve of oc-SVM does not go below a certain level of false positive rate. For instance, it stays above approximately 0.07 for "spiral" and 0.10 for "cancer". Remember that $\nu$ controls the proportion of declared outliers in the training set. However when $\nu$ is set close to 0, the actual proportion of declared outliers among the tested actual outliers tends to increase back to large values instead of tending to 0. This shows empirically the lack of control on the false alarm error when using oc-SVM.

As for performance, it can be readily observed that oc-SVM yields a smaller AUC than kernel PCA — as expected — and our method on both datasets. Depending on the dataset, kernel PCA achieves a slightly larger AUC ("spiral" dataset) or our method performs better ("cancer" dataset). In either case, the gap between the two methods is relatively small.

In order to decide between our method and kernel PCA, let us take a look at computation costs.

**Execution time**

In Figure 5.9, the execution time (in seconds) of the three competing methods are represented with respect to $n$. In addition, the version of our procedure using Nyström approximation and Random Kitchen Sinks are also represented. The parameter $r$ — the rank of the low-rank approximation of the Gram matrix in Nyström method and RKS — is set at $r = 50$. As expected, kernel PCA is the most time-consuming method, with a running time up to 5 times larger than our procedure when $n = 3000$. Nyström method and RKS improves on the computational cost of our method, reducing the execution times up to about 2 times. Finally, oc-SVM is the fastest of all compared methods, achieving a 2 seconds running time when $n = 3000$.

As discussed in Section 5.1.2 when introducing kernel PCA, the high computational cost of kernel PCA is due to the eigen-decomposition of the Gram matrix which is of order $\mathcal{O}(n^3)$. On the other hand, we saw in Section 5.2.2 that the most time-consuming step in our procedure is the estimation of the $p$-value $\widehat{\mathrm{PV}}(X)$ because of the computation of the quantities $\Sigma_{\gamma,-i}(X_i, X_i)$ for $i = 1, \ldots, n$, which all in all costs $\mathcal{O}(n^2)$ in time but still outperforms kernel PCA computationally speaking. Nyström method or RKS reduces this computational cost down to $\mathcal{O}(nr^2)$ where $r$ is the rank of the approximate Gram matrix. Note that Nyström method or RKS can also be applied to kernel PCA, however using a low-rank approximate Gram matrix restricts the number of possible kept principal components.

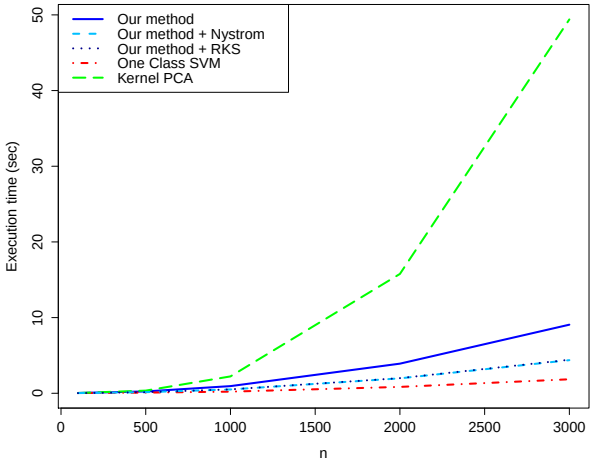


Figure 5.9 – Execution times (in seconds) with respect to $n$ of our OD method (blue, solid), our method combined with Nyström approximation (cyan, dashed), our methods combined with Random Kitchen Sinks (dark blue, dotted), oc-SVM (red, dot-dash) and kernel PCA (green, long-dashed).

## 5.5 Parameter selection

This section presents a procedure to select the parameters $\gamma$ and $p$ of our OD method in practice. In Section 5.4.1, it was observed that optimal parameters are mostly linked to false alarm rate control. Thus the parameters can be chosen solely on the basis of the inliers sample $X_1,\dots,X_n \sim P$.

Let $\mathcal{I} = \{1,\dots,n_0\}$ be indexes of a subsample of $X_1,\dots,X_n$ with $n_0 < n$. For every $i \in \mathcal{I}$, let $s_i = S_{n-n_0}(X_i)$ be the OD statistic applied to $X_i$ when using the subsample $\{X_j\}_{j\notin\mathcal{I}}$ as the inliers sample. Our selection procedure relies on a Kolmogorov-Smirnov-like statistic $\mathrm{KS}(\gamma,p)$ that compares the empirical cumulative distribution function (cdf) of $s_{(1)},\dots,s_{(n_0)}$ to the cdf of $\hat{\mathfrak{s}}^2 Q^2$ where $\hat{\mathfrak{s}}^2$ follows the empirical distribution of $\{\hat{\mathfrak{s}}_i\}_{1\leqslant i\leqslant n}$ as defined in Section 5.2.2 and is independent of $Q^2 \sim \chi^2(p)$. Namely, $\mathrm{KS}(\gamma,p)$ is defined as

$$\begin{aligned}\mathrm{KS}(\gamma,p) &= \sup_{t\geqslant 0}\left|\frac{1}{n_0}\sum_{i=1}^{n_0}\mathbb{1}_{\{s_{(i)}\leqslant t\}} - \frac{1}{n}\sum_{j=1}^{n}F_{\chi^2(p)}\left(\frac{t}{\hat{\mathfrak{s}}_j}\right)\right| \\ &= \max_{i=0}^{n_0}\left|\frac{i}{n_0} - \frac{1}{n}\sum_{j=1}^{n}F_{\chi^2(p)}\left(\frac{s_{(i+1)}}{\hat{\mathfrak{s}}_j}\right)\right| \vee \left|\frac{i}{n_0} - \frac{1}{n}\sum_{j=1}^{n}F_{\chi^2(p)}\left(\frac{s_{(i)}}{\hat{\mathfrak{s}}_j}\right)\right|,\end{aligned}$$

where $F_{\chi^2(p)}$ denotes the cdf of the $\chi^2(p)$ distribution and $s_{(0)} \leqslant s_{(1)} \leqslant \dots \leqslant s_{(n_0+1)}$ are the ordered $s_1,\dots,s_n$ with $s_{(0)} = 0$ and $s_{(n_0+1)} = +\infty$.

Given a range of parameter values $\gamma_1 < \gamma_2 < \dots < \gamma_m$ and $p_1 < p_2 < \dots < p_q$, we compute every quantity $\mathrm{KS}(\gamma_j,p_l)$ for every $1 \leqslant j \leqslant m$ and $1 \leqslant l \leqslant q$. For each fixed $p_l$, the plot $\{\mathrm{KS}(\gamma_j,p_l)\}_{1\leqslant j\leqslant m}$

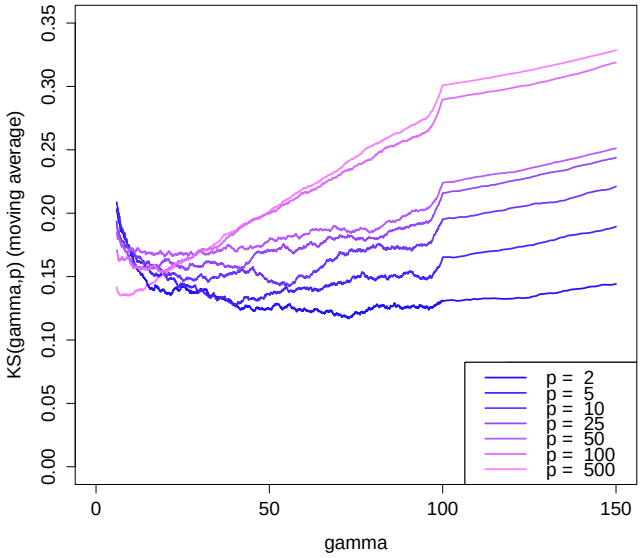


Figure 5.10 – Values of $\widetilde{\mathrm{KS}}(\gamma, p)$ for several values of $\gamma$ and $p$ in the case of the "spiral" dataset

is smoothed out by computing the moving averages $\widetilde{\mathrm{KS}}(\gamma_j, p_l) = (1/2T)\sum_{k=-T}^{T} \mathrm{KS}(\gamma_j + k, p_l)$ for every $1 + T \leqslant j \leqslant m - T$ and some $T > 0$.

We computed the $\mathrm{KS}(\gamma_j, p_l)$ for the "spiral" dataset of section 5.4.1 with $n_0 = n/2 = 250$. The obtained smoothed out curves $\{\widetilde{\mathrm{KS}}(\gamma_j, p_l)\}_{j,l}$ are displayed in Figure 5.10. In our experiment $\gamma$ takes values ranging from 1 to 300 (2000 different values within $[1;100]$, 100 within $[100;200]$ and 50 within $[200;300]$), $p$ is ranging from 2 to 500 and we set $T = 200$.

For each fixed $p$, the curve $\{\widetilde{\mathrm{KS}}(\gamma_j, p)\}_j$ describes a convex function, hence admitting a global minimum at some $\gamma = \gamma^*(p)$. Hence it would be natural to choose $\gamma^*(p)$ as the optimal $\gamma$ for a fixed $p$. However the statistics $s_1, \ldots, s_{n_0}$ were calculated on the basis of an inlier sample of smaller size $n - n_0 < n$. To get a more accurate optimal $\gamma$ for an inlier sample of full size $n$, we propose to multiply $\gamma^*(p)$ by a correction term $(n/(n-n_0))^{1/(2s)}$ where $s$ is the same quantity occurring in (5.3.8). Indeed in Section 5.3.1 it was suggested that the optimal $\gamma$ had the form $\gamma^* = Cn^{1/(2s)}$ for some constant $C > 0$. Finally the optimal pair $(\gamma^*, p^*)$ is chosen as

$$p^* \in \operatorname*{argmin}_{1 \leqslant l \leqslant q} \{\widetilde{\mathrm{KS}}(\gamma^*(p_l), p_l)\} \quad \text{and} \quad \gamma^* = \left(\frac{n}{n-n_0}\right)^{1/(2s)} \gamma^*(p^*)\,.$$

Note that when $\mathcal{X} = \mathbb{R}^D$ endowed with the Euclidean metric, $s$ is readily given as $s = D/2$ according to Proposition 4.3.2 in Chapter 4.

In our simulations, we found $p^* = 2$ and $\gamma^* = (500/250)^{1/2}68 \simeq 96.17$ as optimal values. According to the results in Figure 5.3, this corresponds to a false alarm rate around 0.062 and a

missed detection rate around 0.009.

# 5.A Technical details

## 5.A.1 Proof of Theorem 5.3.1

Let $\alpha \in (0,1)$ be the prescribed level of confidence and $\tau_\alpha$ the $\alpha$-quantile of the distribution of $\mathfrak{s}^2 Q^2$ where $Q^2 \sim \chi^2(p)$.

The goal is to find an upper bound for the gap between $\alpha$ and the actual false alarm rate $\mathbb{P}(S_n(X) < \tau_\alpha)$. Introducing the notation $\mathcal{B}_\alpha$ that denotes the ball of $\mathbb{R}^p$ centered around 0 and of radius $\tau^{1/2}$ and $G \sim \mathcal{N}(0, I_p)$,

$$\begin{aligned}\mathbb{P}(S_n(X) < \tau_\alpha) &= \alpha + \mathbb{P}(S_n(X) < \tau_\alpha) - \alpha \\ &= \alpha + \mathbb{P}(\gamma^{s/2} p_{V_n}(X) \in \mathcal{B}_\alpha) - \mathbb{P}(\mathfrak{s}G \in \mathcal{B}_\alpha) \\ &= \alpha + \mathbb{E}_{X,\mathbf{h},\mathfrak{s},G}\left\{ \mathbb{1}_{\mathcal{B}_\alpha}(\gamma^{s/2} p_{V_n}(X)) - \mathbb{1}_{\mathcal{B}_\alpha}(\mathfrak{s}G) \right\}.\end{aligned}$$

Standard calculations entail that the Fourier transform of $\mathbb{1}_{\mathcal{B}_\alpha}(\cdot)$ is $(2\pi)^{p/2} \int \mathbb{1}_{\mathcal{B}_\alpha}(x) e^{-i\xi^\top x} = \tau_\alpha^{p/2} \|\xi\|^{-p/2} J_{p/2}(\tau_\alpha \|\xi\|) \stackrel{\Delta}{=} g(\xi)$ for every $\xi \in \mathbb{R}^p$, where $J_{p/2}(\cdot)$ denotes the Bessel function of the first kind $J_{p/2}(t) = \frac{(t/2)^{p/2}}{\sqrt{\pi} i \Gamma((p+1)/2)} \int_0^\pi \sin^p(y) e^{-it\cos(y)} \mathrm{d}y$. Since $g(\xi)$ is an integrable function of $\xi$, $\mathbb{1}_{\mathcal{B}_\alpha}(\cdot)$ is the inverse Fourier transform of $g$ that is $\mathbb{1}_{\mathcal{B}_\alpha}(x) = (\tau_\alpha/2\pi)^{p/2} \int \|\xi\|^{-p/2} J_{p/2}(\tau_\alpha \|\xi\|) e^{ix^\top \xi} \mathrm{d}\xi$.

It follows

$$\begin{aligned}\mathbb{P}(S_n(X) < \tau_\alpha) &= \alpha + \mathbb{E}_{X,X_1,\dots,X_n,\mathbf{h},\mathfrak{s},G}\left\{ (\tau_\alpha/2\pi)^{p/2} \int \|\xi\|^{-p/2} J_{p/2}(\tau_\alpha\|\xi\|) \left[ e^{i\gamma^{s/2}\xi^\top p_{V_n}(X)} - e^{i\mathfrak{s}\xi^\top G} \right] \mathrm{d}\xi \right\} \\ &\underset{\text{Fubini}}{=} \alpha + (\tau_\alpha/2\pi)^{p/2} \int \|\xi\|^{-p/2} J_{p/2}(\tau_\alpha\|\xi\|) \left[ \mathbb{E}_{X,X_1,\dots,X_n,\mathbf{h}} e^{i\gamma^{s/2}\xi^\top p_{V_n}(X)} - \mathbb{E}_{\mathfrak{s},G} e^{i\mathfrak{s}\xi^\top G} \right] \mathrm{d}\xi \qquad (5.A.11) \\ &= \alpha + (\tau_\alpha/2\pi)^{p/2} \int \|\xi\|^{-p/2} J_{p/2}(\tau_\alpha\|\xi\|) \left[ \mathbb{E}_{X,X_1,\dots,X_n} e^{-\frac{\gamma^s}{2}\Sigma_{\gamma,n}(X,X)\|\xi\|^2} - \mathbb{E}_X e^{-\frac{\gamma^s}{2}\Sigma_\gamma(X,X)\|\xi\|^2} \right] \mathrm{d}\xi \\ &\leqslant \alpha + \frac{1}{2}(\tau_\alpha/2\pi)^{p/2} \left( \int \|\xi\|^{-p/2+2} |J_{p/2}(\tau_\alpha\|\xi\|)| \mathrm{d}\xi \right) \mathbb{E}_{X,X_1,\dots,X_n}\left[ \gamma^s |\Sigma_{\gamma,n}(X,X) - \Sigma_\gamma(X,X)| \right],\end{aligned}$$

since $x \mapsto e^{-x/2}$ is a 1/2-Lipschitz function, which yields more simply

$$\mathbb{P}(S_n(X) < \tau_\alpha) \leqslant \alpha + C_p \tau_\alpha^{-2} \mathbb{E}_{X,X_1,\dots,X_n}\left[ \gamma^s |\Sigma_{\gamma,n}(X,X) - \Sigma_\gamma(X,X)| \right], \qquad (5.A.12)$$

where $C_p = (1/2)(2\pi)^{-p/2} \int \|\xi\|^{-p/2+2} |J_{p/2}(\|\xi\|)| \mathrm{d}\xi$ after a change of variable was made.

Fixing $X$, Lemma A.2.4 provides the following inequality that holds on the event $\Omega_X$ with

probability larger than $1-\delta$ over the sample $X_1,\dots,X_n$

$$\gamma^s|\Sigma_{\gamma,n}(X,X)-\Sigma_\gamma(X,X)|\leqslant\gamma^s\Sigma_\gamma(X,X)\sqrt{\frac{2\log(1/\delta)\gamma^s}{a\Gamma(s+1)n}}\left[1+\kappa\sqrt{\frac{\log(1/\delta)(2\gamma)^s}{a\Gamma(s+1)n}}[1+o_\gamma(1)]\right],$$

which combined with Lemma A.2.1 yields

$$\gamma^s|\Sigma_{\gamma,n}(X,X)-\Sigma_\gamma(X,X)|\leqslant A(X)\sqrt{\frac{2\log(1/\delta)\Gamma(s+1)\gamma^s}{an}}\left[1+\kappa\sqrt{\frac{\log(1/\delta)(2\gamma)^s}{a\Gamma(s+1)n}}\right]\{1+o_\gamma(1)\}. \tag{5.A.13}$$

Plugging (5.A.13) into (5.A.12) and conditioning with respect to $\Omega_X$ leads to

$$\begin{aligned}\mathbb{P}(S_n(X)<\tau_\alpha)&\leqslant\alpha+C_p\tau_\alpha^{-2}\mathbb{E}_X\left\{\mathbb{E}_{X_1,\dots,X_n}\left[\gamma^s|\Sigma_{\gamma,n}(X,X)-\Sigma_\gamma(X,X)|\,\middle|\,\Omega_X\right]+2\gamma^s\delta\right\}\\&\leqslant\alpha+C_p\tau_\alpha^{-2}\mathbb{E}_X[A(X)]\sqrt{\frac{2\log(1/\delta)\Gamma(s+1)\gamma^s}{an}}\left[1+\kappa\sqrt{\frac{\log(1/\delta)(2\gamma)^s}{a\Gamma(s+1)n}}\right]\{1+o_\gamma(1)\}\\&\quad+2C_p\tau_\alpha^{-2}\gamma^s\delta,\end{aligned} \tag{5.A.14}$$

where we used the inequality $|\Sigma_{\gamma,n}(X,X)-\Sigma_\gamma(X,X)|\leqslant 2$. This leads to the result of the lemma.

Note that we could have tried to plug directly the upper bound of Theorem 4.3.5 into (5.A.11) in the following way

$$\begin{aligned}\mathbb{P}(S_n(X)<\tau_\alpha)&=\alpha+(\tau_\alpha/2\pi)^{p/2}\int\|\xi\|^{-p/2}J_{p/2}(\tau_\alpha\|\xi\|)\left[\mathbb{E}_{X,\mathbf{h}}e^{i\gamma^{s/2}\xi^\top p_{V_n}(X)}-\mathbb{E}_{\mathfrak{s},G}e^{i\mathfrak{s}\xi^\top G}\right]\mathrm{d}\xi\\&\leqslant\alpha+(\tau_\alpha/2\pi)^{p/2}\int\|\xi\|^{-p/2}|J_{p/2}(\tau_\alpha\|\xi\|)|\left|\phi_{\gamma^s p_{V_n}(X)}(\xi)-\phi_{\mathfrak{s}G}(\xi)\right|\mathrm{d}\xi\\&\leqslant\alpha+(\tau_\alpha/2\pi)^{p/2}\left(\int\|\xi\|^{-p}|J_{p/2}(\tau_\alpha\|\xi\|)|^2\left|\phi_{\gamma^s p_{V_n}(X)}(\xi)-\phi_{\mathfrak{s}G}(\xi)\right|^2\mathrm{d}\xi\right)^{1/2}\\&=\alpha+C\Delta(\gamma^s p_{V_n}(X),\mathfrak{s}G),\end{aligned}$$

where $\phi_U$ denotes the characteristic function for any random variable $U$ and $C$ is some constant. Theorem 4.3.5 asserts an inequality of the form $\Delta^2(\gamma^s p_{V_n}(X),\mathfrak{s}G)\leqslant C_1\gamma^{-s/2}+C_2\gamma^{s/2}n^{-1/2}$ for some constants $C_1$ and $C_2$. However setting $\gamma\propto n^{1/(2s)}$ to minimize this upper bound would entail a convergence of the actual false alarm rate to $\alpha$ of order $\mathcal{O}(n^{-1/8})$ which is very slow. On the other hand, (5.A.14) yields for $\gamma\propto n^{1/(2s)}$ a convergence of order $\mathcal{O}(\sqrt{\log(n)}n^{-1/4})$.

## 5.A.2 Proof of Theorem 5.3.2

We must derive an upper bound for the probability $\mathbb{P}\left(PV(X)>\alpha\,\middle|\,\mathbf{H}_1\right)$ which can be also expressed as $\mathbb{P}\left(S_n(X)>\tau_\alpha\,\middle|\,\mathbf{H}_1\right)$ where $\tau_\alpha$ satisfies $\mathbb{P}(\mathfrak{s}^2Q^2<\tau_\alpha)=\alpha$.

For every $j = 1,\dots,p$, let $h_j = (1/\sqrt{n})\sum_{i=1}^n u_k^{(j)} k_\gamma(X_i,\cdot)$ where $u_1^{(j)},\dots,u_n^{(j)} \sim \mathcal{N}(0,1)$ independently so that $h_1,\dots,h_p$ are independent zero-mean Gaussian processes with covariance function $\Sigma_{\gamma,n}(\cdot,\cdot) = (1/n)\sum_{i=1}^n k_\gamma(X_i,\cdot)k_\gamma(X_i,\cdot)$. Let $d(x) = \inf\{\|x-y\| : y \in \mathrm{supp}(P)\}$ denotes the distance between $x$ and the support of $P$.

Then,

$$\begin{aligned} S_n(X) = \gamma^s \|p_{V_n}(X)\|^2 = \gamma^s \sum_{j=1}^p \langle h_j, k_\gamma(X,\cdot)\rangle^2_{H(k_\gamma)} &= (\gamma^s/n)\sum_{j=1}^p \sum_{k,l=1}^n u_k^{(j)} u_l^{(j)} e^{-\gamma(d^2(X,X_i)+d^2(X,X_j))} \\ &\leqslant (\gamma^s/n)\sum_{j=1}^p \sum_{k,l=1}^n |u_k^{(j)}||u_l^{(j)}| e^{-2\gamma d^2(X)} \\ &\leqslant \gamma^s \sum_{j=1}^p \sum_{l=1}^n [u_l^{(j)}]^2 e^{-2\gamma d^2(X)} \\ &\leqslant \gamma^s np[1+\epsilon_n] e^{-2\gamma d^2(X)} , \end{aligned}$$

where $\epsilon_n = o_P(1)$ as $n \to +\infty$ because of the Law of Large Numbers.

Let us bound $1+\epsilon_n$ with high probability. For any $\delta > 0$ and $\lambda \in (0, np/2)$, exponential Markov's inequality yields

$$\mathbb{P}(1+\epsilon > \delta) = \mathbb{P}_{V^2\sim\chi^2(np)}\left(V^2 > np\delta\right) \leqslant \mathbb{E}_{V^2} e^{\lambda V^2} e^{-\lambda np\delta} = (1-2\lambda np)^{-np/2} e^{-\lambda np\delta} ,$$

which is minimized at $\lambda = 1/(2\delta) - 1/(2np)$. Therefore for any $\delta \in \left(\frac{np}{(np)^2+1}, np\right)$, the event $\mathcal{B}_\delta$ defined by

$$\mathcal{B}_\delta = \{1+\epsilon \leqslant \delta\} ,$$

holds with probability larger than $1-(2-np/\delta)^{-np/2}e^{-(np-\delta)/2}$.

From now on we assume that the event $\mathcal{B}_\delta$ holds true. Therefore

$$\mathbb{P}(S_n(X) > \tau_\alpha) \leqslant \mathbb{P}\left(d^2(X) < \frac{\log(\gamma^s np\delta) + \log(1/\tau_\alpha)}{2\gamma}\right) .$$

By Lemma A.2.1 and writing $a = \inf\{A(x) : x \in \mathrm{supp}(P)\} > 0$ and $X_1 \sim P$,

$$\begin{aligned} \mathfrak{s}^2 = \gamma^s \Sigma_\gamma(X_1,X_1) = \frac{A(X_1)\Gamma(s+1)}{2^s}[1 + A^{-1}(X_1)o_\gamma(1)] \\ \geqslant \frac{a\Gamma(s+1)}{2^s}[1+o_\gamma(1)] , \end{aligned}$$

which leads to

$$\tau_\alpha \geqslant \frac{a\Gamma(s+1)F^{-1}_{\chi^2(p)}(\alpha)}{2^s}[1+o_\gamma(1)]\,,$$

where $F^{-1}_{\chi^2(p)}$ denotes the quantile function of the $\chi^2(p)$ distribution.

Therefore introducing $F$ as the c.d.f. of $d^2(X)$

$$\mathbb{P}(S_n(X) > \tau_\alpha) \leqslant \mathbb{E}_{\epsilon_n} F\left(\frac{\log(np\delta) + s\log(2\gamma) + \log\left(\frac{1}{a\Gamma(s+1)}\right) + \log\left(\frac{1}{F^{-1}_{\chi^2(p)}(\alpha)}\right) + o_\gamma(1)}{2\gamma}\right).$$

By the exponential Markov's inequality for any $\lambda > 0$

$$\alpha = F_{\chi^2(p)}(F^{-1}_{\chi^2(p)}(\alpha)) \leqslant e^{\lambda F^{-1}_{\chi^2(p)}(\alpha)}\mathbb{E}_{V^2\sim\chi^2(p)}e^{-\lambda V^2} = e^{\lambda F^{-1}_{\chi^2(p)}(\alpha)}(1+2\lambda)^{-p/2}\,,$$

hence setting $\lambda$ such that $(1/2)\log(1+2\lambda) = \log(1/\alpha)/p + \epsilon$ for some $\epsilon > 0$

$$\begin{aligned}
\log\left(\frac{p}{F^{-1}_{\chi^2(p)}(\alpha)}\right) &\leqslant \log\left(\frac{p\lambda}{(p/2)\log(1+2\lambda) - \log(1/\alpha)}\right) = \log\left(\frac{e^{(2/p)\log(1/\alpha)+2\epsilon} - 1}{2\epsilon}\right) \\
&\leqslant (2/p)\log(1/\alpha) + 2\epsilon + \log\left(1 + \frac{\log(1/\alpha)}{p\epsilon}\right) \\
&\leqslant \frac{(2+1/\epsilon)\log(1/\alpha)}{p} + 2\epsilon \\
&= \frac{2\log(1/\alpha)}{p} + 2\sqrt{\frac{2\log(1/\alpha)}{p}}\,,
\end{aligned}$$

where we used the inequalities $e^x - 1 \leqslant xe^x$ and $\log(1+x) \leqslant x$ and set $\epsilon = \sqrt{\log(1/\alpha)/(2p)}$ to minimize the bound.

This leads to

$$\mathbb{P}(S_n(X) > \tau_\alpha) \leqslant F\left(\frac{\log(n) + s\log(2\gamma) + \log\left(\frac{1}{a\Gamma(s+1)}\right) + \frac{2\log(1/\alpha)}{p} + 2\sqrt{\frac{2\log(1/\alpha)}{p}} + \log(\delta) + o_\gamma(1)}{2\gamma}\right),$$

hence proving Theorem 5.3.2.

Chapter 6

# A One-Sample Test for Normality in Hilbert Spaces

This chapter presents a test for normality in general Hilbert spaces (including RKHS) based on the maximum mean discrepancy (MMD). It can be read independently of Chapter 4 and Chapter 5.

## 6.1 Introduction

Non-vectorial data such as DNA sequences or pictures often require a positive definite kernel $k$ so that further analysis is then carried out in the associated reproducing kernel Hilbert space (RKHS) $H(k)$ where data are embedded into through the feature map $x \mapsto k(x,.)$. For many applications, kernelized data — or even more general high-dimensional data — are assumed to fit a specific type of distribution, often a Gaussian distribution. For instance supervised and unsupervised classification are performed in [BFG15] by modeling each class as a Gaussian process. In [Rot06], outliers are detected by modelling embedded data as a Gaussian random variable and by removing points lying in the tails of that Gaussian distribution. This key assumption is also made in [SKK13] where a mean equality test is used in high-dimensional setting. Moreover, Principal Component Analysis (PCA) and its kernelized version Kernel PCA [SSM97] are known to be optimal for Gaussian data as these methods rely on second-order statistics (covariance). Besides, a Gaussian assumption allows to use Expectation-Minimization (EM) techniques to speed up PCA [Row98].

Depending on the (finite or infinite dimensional) structure of the RKHS, Cramer-von-Mises-type normality tests can be applied, such as Mardia's skewness test [Mar70], the Henze-Zirkler test [HZ90] and the Energy-distance test [SR05]. However these tests become less powerful as dimension increases (see Table 3 in [SR05]). An alternative approach consists in randomly

projecting high-dimensional objects on one-dimensional directions and then applying univariate test on a few randomly chosen marginals [Cue+06]. This projection pursuit method has the advantage of being suited to high-dimensional settings. On the other hand, such approaches also suffer a lack of power because of the limited number of considered directions (see Section 4.2 in [Cue+06]).

[Gre+07a] introduced the Maximum Mean Discrepancy (MMD) which quantifies the gap between two distributions through distances between two corresponding elements in an RKHS. The MMD approach has been used for two-sample testing [Gre+07a] and for independence testing (Hilbert Space Independence Criterion, [Gre+07b]). However to the best of our knowledge, MMD has not been applied in a one-sample goodness-of-fit testing framework.

Our main contribution presented in this chapter is to devise a one-sample statistical test of normality for data taking values in an RKHS, by means of the MMD principle. This implies that we will consider a second kernel $\bar{k}$ defined on $H(k)$ and handle elements in the second RKHS $H(\bar{k})$ to build our MMD-based test. However it must be remarked that our test also suits the case where $H(k)$ is replaced by a more general Hilbert space. Therefore to avoid confusion due to this two levels of RKHS, we present our test as dealing with data in a general Hilbert space.

Our test features two possible applications: testing the normality of the data but also testing parameters (mean and covariance) if data are assumed Gaussian. The latter application encompasses many current methods that assume normality to make inferences on parameters, for instance to test the nullity of the mean [SKK13] or to assess the sparse structure of the covariance [SW03; BT11].

Once the test statistic is defined, a critical value is needed to decide whether to accept or reject the Gaussian hypothesis. In goodness-of-fit testing, this critical value is typically estimated by parametric bootstrap. Unfortunately, parametric bootstrap requires parameters to be computed several times, hence heavy computational costs (*i.e.* diagonalization of covariance matrices). Our test bypasses the recomputation of parameters by implementing a faster version of parametric bootstrap. Following the idea of [KY12], this fast bootstrap method "linearizes" the test statistic through a Fréchet derivative approximation and thus can estimate the critical value by a *weighted* bootstrap (in the sense of [Bur00]) which is computationally more efficient. Furthermore our version of this bootstrap method allows parameters estimators that are not explicitly "linear" (*i.e.* that consist of a sum of independent terms) and that take values in possible infinite-dimensional Hilbert spaces.

Finally, we illustrate our test and present a sequential procedure that assesses the rank of a covariance operator. The problem of covariance rank estimation is adressed in several domains: functional regression [CJ10; BMR13], classification [Zwa05] and dimension reduction methods such as PCA, Kernel PCA and Non-Gaussian Component Analysis [Bla+06; Die+10; DJS13] where the dimension of the kept subspace is a crucial problem.

Here is the outline of the chapter. Section 6.2 sets our framework and Section 6.3 recalls the notion of MMD and how it is used for our one-sample test. The new normality test is described

in Section 6.4, while both its theoretical and empirical performances are detailed in Section 6.5 in terms of control of Type-I and Type-II errors. A sequential procedure to select covariance rank is presented in Section 6.6.

## 6.2 Framework

Let $(\mathcal{H}, \mathcal{A})$ be a measurable space, and $Y_1, \dots, Y_n \in \mathcal{H}$ denote a sample of *independent and identically distributed* (*i.i.d.*) random variables drawn from an unknown distribution $P \in \mathcal{P}$, where $\mathcal{P}$ is a set of distributions defined on $\mathcal{A}$.

In our framework, $\mathcal{H}$ is a separable Hilbert space endowed with a dot product $\langle \, . \, , \, . \, \rangle_{\mathcal{H}}$ and the associated norm $\|.\|_{\mathcal{H}}$ (defined by $\|h\|_{\mathcal{H}} = \langle h, h \rangle_{\mathcal{H}}^{1/2}$ for any $h \in \mathcal{H}$). Our goal is to test whether $Y_i$ is a *Gaussian random element (r.e.)* of $\mathcal{H}$. A Gaussian r.e. in a general Hilbert space is defined similarly as a Gaussian r.e. in an RKHS (see Definition 3.1.4).

**Definition 6.2.1.** *(Gaussian random element in a Hilbert space)*
*Let $(\Omega, \mathcal{F}, \mathbb{P})$ a measure space, $(\mathcal{H}, \mathcal{F}')$ a measurable space where $\mathcal{H}$ is a Hilbert space, and $Y : \Omega \to \mathcal{H}$ a measurable map.*
*$Y$ is a Gaussian r.e. of $\mathcal{H}$ if $\langle Y, h \rangle_{\mathcal{H}}$ is a univariate Gaussian random variable for any $h \in \mathcal{H}$.*
*Assuming that $\mathbb{E}_Y \|Y\|_{\mathcal{H}} < +\infty$, there exists $m \in \mathcal{H}$ such that:*

$$\forall h \in \mathcal{H}, \;\; \langle m, h \rangle_{\mathcal{H}} = \mathbb{E}_Y \langle Y, h \rangle_{\mathcal{H}} \;\; ,$$

*and a (finite trace) operator $\Sigma : \mathcal{H} \to \mathcal{H}$ satisfying:*

$$\forall h, h' \in \mathcal{H}, \;\; \langle \Sigma h, h' \rangle_{\mathcal{H}} = \mathrm{cov}(\langle Y, h \rangle_{\mathcal{H}}, \langle Y, h' \rangle_{\mathcal{H}}) \;\; .$$

*$m$ and $\Sigma$ are respectively the mean and the covariance operator of $Y$. The distribution of $Y$ is denoted $\mathcal{N}(m, \Sigma)$.*

More precisely, the tested hypothesis is that $Y_i$ follows a Gaussian distribution $\mathcal{N}(m_0, \Sigma_0)$, where $(m_0, \Sigma_0) \in \Theta_0$ and $\Theta_0$ is a subset of the parameter space $\Theta$. [1] Following [LR05], let us define the null hypothesis $\mathbf{H}_0 : \; P \in \mathcal{P}_0$, and the alternative hypothesis $\mathbf{H}_1 : \; P \notin \mathcal{P} \setminus \mathcal{P}_0$ where the subset of null-hypotheses $\mathcal{P}_0 \subseteq \mathcal{P}$ is

$$\mathcal{P}_0 = \{ \mathcal{N}(m_0, \Sigma_0) \mid (m_0, \Sigma_0) \in \Theta_0 \} \;\; .$$

The purpose of a statistical test $T(Y_1, \dots, Y_n)$ of $\mathbf{H}_0$ against $\mathbf{H}_1$ is to distinguish between the null ($\mathbf{H}_0$) and the alternative ($\mathbf{H}_1$) hypotheses. It requires two elements: a statistic $n\hat{\Delta}^2$ (which

[1] The parameter space $\Theta$ is endowed with the dot product $\langle (m, \Sigma), (m', \Sigma') \rangle_\Theta = \langle m, m' \rangle_{\mathcal{H}} + \langle \Sigma, \Sigma' \rangle_{HS(\mathcal{H})}$, where $HS(\mathcal{H})$ is the space of Hilbert-Schmidt (finite trace) operators $\mathcal{H} \to \mathcal{H}$ and $\langle \Sigma, \Sigma' \rangle_{HS(\mathcal{H})} = \sum_{i \geq 1} \langle \Sigma e_i, \Sigma' e_i \rangle_{\mathcal{H}}$ for any complete orthonormal basis $(e_i)_{i \geq 1}$ of $\mathcal{H}$. Therefore, for any $\theta \in \Theta$, the tensor product $\theta^{\otimes 2}$ is defined as the operator $\Theta \to \Theta, \theta' \mapsto \langle \theta, \theta' \rangle_\Theta \theta$. For any $\theta \in \Theta$ and $\bar{h} \in H(\bar{k})$, the tensor product $\bar{h} \otimes \theta$ is the operator $\Theta \to H(\bar{k}), \theta' \mapsto \langle \theta, \theta' \rangle_\Theta \bar{h}$.

we define in Section 6.4.1) that measures the gap between the empirical distribution of the data and the considered family of normal distributions $\mathcal{P}_0$, and a rejection region $\mathcal{R}_\alpha$ (at a level of confidence $0 < \alpha < 1$). $\mathbf{H}_0$ is accepted if and only if $n\hat{\Delta}^2 \notin \mathcal{R}_\alpha$. The rejection region is determined by the distribution of $n\hat{\Delta}^2$ under the null-hypothesis such that the probability of wrongly rejecting $\mathbf{H}_0$ (Type-I error) is controlled by $\alpha$.

## 6.3 Characteristic kernels on a Hilbert space

Within our framework the goal is to compare $P$ the true distribution of the data with a Gaussian distribution $P_0 = \mathcal{N}(m_0, \Sigma_0)$ for some $(m_0, \Sigma_0) \in \Theta_0$. To achieve this goal, we rely on the notion of Maximum Mean Discrepancy (MMD) introduced previously in Section 2.3.3. Given a characteristic kernel $\bar{k}$ on $\mathcal{H}$, we consider the following Hilbert space embedding of distributions of $\mathcal{H}$

$$\bar{\mu} : \mathcal{M}^+_{1,\bar{k}}(\mathcal{H}) \to H(\bar{k}), \quad P \mapsto \bar{\mu}_P = \int \bar{k}(h, \cdot) P(\mathrm{d}h) \ ,$$

where $\mathcal{M}^+_{1,\bar{k}}(\mathcal{H}) \subseteq \mathcal{M}^+_1(\mathcal{H})$ where $\mathcal{M}^+_1(\mathcal{H})$ denotes the set of probability measures on $\mathcal{H}$ and $\mathcal{M}^+_{1,\bar{k}}(\mathcal{H})$ is the subset of probability measures $P$ on $\mathcal{H}$ for which $\bar{\mu}_P$ is well defined.

Hence the quantity of interest is

$$\Delta^2 = \left\| \bar{\mu}_P - \bar{\mu}_{P_0} \right\|^2_{H(\bar{k})} \ . \tag{6.3.1}$$

For the sake of simplicity, we use the notation

$$\bar{\mu}_{\mathcal{N}(m,\Sigma)} = N[m, \Sigma]$$

to denote the Hilbert space embedding of a Gaussian distribution.

It remains to choose a kernel $\bar{k}$ defined on the Hilbert space $\mathcal{H}$ that is characteristic. Several criteria for a kernel to be characteristic have been investigated [Fuk+09; Sri+10; CS10; SFL11]. However most of these criteria are relevant for input spaces such as compact topological spaces, $\mathbb{R}^d$ with a finite $d \in \mathbb{N}^*$ or more generally locally compact Haussdorf spaces. The only exception is the case of integrally strictly positive definite kernels [2](see Theorem 7 in [Sri+10]) defined on a general topological space, but proving that a kernel is integrally strictly positive definite may be a difficult task.

In the following, we introduce two examples of kernels defined on the Hilbert space $\mathcal{H}$ and show that they are characteristic. Firstly, we consider the *exponential kernel*

$$\forall x, y \in \mathcal{H}, \quad \bar{k}(x, y) = \exp(\langle x, y \rangle_{\mathcal{H}}) \ .$$

[2] A kernel $\bar{k} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is said integrally strictly definite positive iff $\int \bar{k}(x, y)\mu(\mathrm{d}x)\mu(\mathrm{d}y) > 0$ for every finite non-zero signed Borel measure $\mu$ on $\mathcal{H}$.

Secondly, we cover the case of the Gaussian RBF kernel

$$\forall x, y \in \mathcal{H}, \quad \bar{k}(x,y) = \exp\left(-\gamma \|x-y\|_{\mathcal{H}}^2\right) ,$$

parameterized by $\gamma$.

**Proposition 6.3.1** (Exponential kernel)**.** *The exponential kernel $\bar{k}(x,y) = \exp(\langle x, y\rangle_{\mathcal{H}})$ defined on a separable Hilbert space $\mathcal{H}$ is characteristic (for the class of probability measures $P$ for which $\bar{\mu}_P$ is well defined).*

*Proposition 6.3.1.* Let $C > 1$. Cauchy-Schwarz's inequality entails

$$\begin{aligned} C\|\bar{\mu}_P - \bar{\mu}_Q\|_{H(\bar{k})}^2 &= \sup_{g \in H(\bar{k}),\ \|g\|_{H(\bar{k})} \leqslant 1} C\left|\langle g, \bar{\mu}_P - \bar{\mu}_Q\rangle_{H(\bar{k})}\right| \\ &= \sup_{g \in H(\bar{k}),\ \|g\|_{H(\bar{k})} \leqslant C} \left|\mathbb{E}_{X\sim P} g(X) - \mathbb{E}_{Y\sim Q} g(Y)\right| \\ &\geqslant \sup_{\|w\|_{\mathcal{H}}=1,\ 0 \leqslant t \leqslant \sqrt{2\log(C)}} \left|\mathbb{E}_{X\sim P} e^{t\langle X, w\rangle_{\mathcal{H}}} - \mathbb{E}_{Y\sim Q} e^{t\langle Y, w\rangle_{\mathcal{H}}}\right| , \end{aligned}$$

where we restricted ourselves to functions $g$ of the form $g = e^{t\langle w, \cdot\rangle_{\mathcal{H}}}$ to get the last inequality.

Therefore if $\Delta(P,Q) = 0$ then $\left|\mathbb{E}_{X\sim P} e^{t\langle X, w\rangle_{\mathcal{H}}} - \mathbb{E}_{Y\sim Q} e^{t\langle Y, w\rangle_{\mathcal{H}}}\right| = 0$ for every $\|w\|_{\mathcal{H}} = 1$ and $0 \leqslant t \leqslant \sqrt{2\log(C)}$. It follows that every marginals $\langle X, w\rangle_{\mathcal{H}}$ and $\langle Y, w\rangle_{\mathcal{H}}$ follow the same univariate distribution. Finally since $\mathcal{H}$ is a separable Hilbert space, the Cramer-Wald theorem ([Cue+06], Proposition 2.1) entails that $P = Q$. □

**Proposition 6.3.2** (Gaussian kernel)**.** *Let $\bar{k}(x,y) = \exp(-\gamma\|x-y\|_{\mathcal{H}}^2)$ be a Gaussian kernel with $\gamma > 0$ and assume that $\mathcal{H}$ is a separable Hilbert space. Then $\bar{k}$ is a characteristic kernel.*

*Proof of Proposition 6.3.2.* Since $\mathcal{H}$ is assumed separable, there exists an orthonormal Hilbertian basis $(e_i)_{i\geqslant 1}$ of $\mathcal{H}$. Consider some operator $T : \mathcal{H} \to \mathcal{H}$ defined by $T = \sum_{i\geqslant 1} \varphi_i e_i^{\otimes 2}$ where $\sum_{i\geqslant 1} \varphi_i < +\infty$ and $\varphi_i > 0$ for every $i \geqslant 1$. Then there exists a zero-mean Gaussian process $G$ with covariance operator $2\gamma T$ that is well defined in $\mathcal{H}$ (since $2\gamma T$ has a finite trace). The first step of the proof consists in proving that the "proxy" kernel $\bar{k}_T(x,y) = \exp(-\gamma\|T^{1/2}(x-y)\|_{\mathcal{H}}^2)$ defined on $\mathcal{H}$ is characteristic — here $T^{1/2} = \sum_{i\geqslant 1} \varphi_i^{1/2} e_i^{\otimes 2}$. Following Theorem 7 from [Sri+10], it suffices to prove that $\bar{k}_T$ is integrally strictly positive definite, that is $\int\int \bar{k}_T(x,y)\nu(\mathrm{d}x)\nu(\mathrm{d}y) > 0$ for every non-zero signed Borel measure $\nu$ on $\mathcal{H}$. Let $\nu$ be such a measure,

$$\begin{aligned} \int\int \bar{k}_T(x,y)\nu(\mathrm{d}x)\nu(\mathrm{d}y) &= \int\int \mathbb{E}_G\left\{e^{i\langle x-y, G\rangle_{\mathcal{H}}}\right\}\nu(\mathrm{d}x)\nu(\mathrm{d}y) \\ &= \mathbb{E}_G|\varphi_\nu(G)|^2 \geqslant 0 \end{aligned} \tag{6.3.2}$$

where $\varphi_\nu(G) = \int e^{i\langle x, G\rangle_{\mathcal{H}}}\nu(\mathrm{d}x)$ denotes the Fourier transform of the measure $\nu$.

**Algorithm 1** Kernel Normality Test procedure

**Input:** $Y_1, \ldots, Y_n \in \mathcal{H}$, $\bar{k} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ (kernel) and $0 < \alpha < 1$ (test level).

1. Compute $K = \left[\langle Y_i, Y_j \rangle\right]_{i,j}$ (Gram matrix).
2. Compute $n\hat{\Delta}^2$ (test statistic) from (6.4.3) that depends on $K$ and $\bar{k}$ (Section 6.4.1)
3. (a) Draw $B$ (approximate) independent copies of $n\hat{\Delta}^2$ under $\mathbf{H}_0$ by fast parametric bootstrap (Section 6.4.2).
   (b) Compute $\hat{q}_{\alpha,n}$ ($1-\alpha$ quantile of $n\hat{\Delta}^2$ under $\mathbf{H}_0$) from these replications.

**Output:** Reject $\mathbf{H}_0$ if $n\hat{\Delta}^2 > \hat{q}_{\alpha,n}$, and accept otherwise.

If (6.3.2) is null, then $\varphi_G$ is null almost everywhere (with respect to the distribution of $G$). Since the eigenvalues $\ell_i$ of $T$ are non-null, the Gaussian distribution of $G$ is supported on the entire $\mathcal{H}$, therefore $\nu$ is a null measure which leads to a contradiction. Thus $\bar{k}_T$ is integrally strictly positive definite, hence characteristic.

In the second step, we relate the embedding $\bar{\mu}_T$ corresponding to $\bar{k}_T$ to the embedding $\mu$ corresponding to the Gaussian kernel $\bar{k}(x,y) = \exp(-\gamma\|x-y\|^2_{\mathcal{H}})$. Note that for every probability measure $P$ on $\mathcal{H}$,

$$\bar{\mu}_T[P](\cdot) = \mathbb{E}_{X\sim P}\bar{k}_T(X,\cdot) = \mathbb{E}_{X\sim P}\bar{k}(T^{1/2}X, T^{1/2}(\cdot)) = \bar{\mu}[T^{1/2}P](T^{1/2}(\cdot)) \ ,$$

that is $\bar{\mu}_T = A \ o \ \bar{\mu} \ o \ B$ where $A : f \in H(\bar{k}) \mapsto foT^{1/2}$ and $B : P \mapsto T^{1/2}P$ where $T^{1/2}P$ is the probability measure of a random variable $T^{1/2}X$ with $X \sim P$. Since $\bar{k}_T$ is characteristic, $\bar{\mu}_T$ is an injective map, which entails that $\bar{\mu}$ is also injective and that $\bar{k}$ is characteristic. □

## 6.4 Kernel normality test

This section introduces our one-sample test for normality based on the quantity (6.3.1). As said in Section 6.2, we test the null-hypothesis $\mathbf{H}_0 : P \in \{\mathcal{N}(m_0, \Sigma_0) \mid (m_0, \Sigma_0) \in \Theta_0\}$ where $\Theta_0$ is a subset of the parameter space. Therefore our procedure may be used as test for normality or a test on parameter if data are assumed Gaussian. The test procedure is summed up in Algorithm 1.

### 6.4.1 Test statistic

As in [Gre+07a], $\Delta^2$ can be estimated by replacing $\bar{\mu}_P$ with the sample mean

$$\hat{\bar{\mu}}_P = \bar{\mu}_{\hat{P}} = (1/n)\sum_{i=1}^{n} \bar{k}(Y_i, .) \ ,$$

where $\hat{P} = (1/n)\sum_{i=1}^{n} \delta_{Y_i}$ is the empirical distribution. The null-distribution embedding $N[m_0, \Sigma_0]$ is estimated by $N[\tilde{m}, \tilde{\Sigma}]$ where $\tilde{m}$ and $\tilde{\Sigma}$ are appropriate and consistent (under $\mathbf{H}_0$) estimators of $m_0$ and $\Sigma_0$. This yields the estimator

$$\hat{\Delta}^2 = \|\hat{\bar{\mu}}_P - N[\tilde{m}, \tilde{\Sigma}]\|^2_{H(\bar{k})} \ ,$$

which can be written by expanding the square of the norm and using the reproducing property of $H(\bar{k})$ as follows

$$\hat{\Delta}^2 = \frac{1}{n^2}\sum_{i,j=1}^{n} \bar{k}(Y_i, Y_j) - \frac{2}{n}\sum_{i=1}^{n} N[\tilde{m}, \tilde{\Sigma}](Y_i) + \|N[\tilde{m}, \tilde{\Sigma}]\|^2_{H(\bar{k})} \ . \tag{6.4.3}$$

Proposition 6.4.1 ensures the consistency of the statistic (6.4.3).

**Proposition 6.4.1.** *Assume that $P$ is Gaussian $\mathcal{N}(m_0, \Sigma_0)$ where $(m_0, \Sigma_0) \in \Theta_0$ and $(\tilde{m}, \tilde{\Sigma})$ are consistent estimators of $(m_0, \Sigma_0)$. Also assume that $\mathbb{E}_P \bar{k}(Y, Y) < +\infty$ and $N[m, \Sigma]$ is a continuous function of $(m, \Sigma)$ on $\Theta_0$. Then $\hat{\Delta}^2$ is a consistent estimator of $\Delta^2$.*

*Proof.* First, note that $\bar{\mu}_P$ exists since $\mathbb{E}\bar{k}(Y, Y) < +\infty$ implies $\mathbb{E}\bar{k}^{1/2}(Y, Y) < +\infty$. By the Law of Large Numbers in Hilbert Spaces [HP76], $\hat{\bar{\mu}}_P \underset{n\to\infty}{\longrightarrow} \bar{\mu}_P$ $P$-almost surely since $\mathbb{E}\|\bar{k}(Y, \cdot) - \bar{\mu}_P\|^2_{H(\bar{k})} = \mathbb{E}\bar{k}(Y, Y) - \mathbb{E}\bar{k}(Y, Y') \leq \mathbb{E}\bar{k}(Y, Y) + \mathbb{E}^2\bar{k}(Y, Y') < +\infty$. The continuity of $N[m, \Sigma]$ (with respect to $(m, \Sigma)$) and the consistency of $(\tilde{m}, \tilde{\Sigma})$ yield $N[\tilde{m}, \tilde{\Sigma}] \overset{P-a.s.}{\underset{n\to\infty}{\longrightarrow}} N[m_0, \Sigma_0]$ $P$-a.s.. Finally, the continuity of $\|\cdot\|^2_{\mathcal{H}}$ leads to $\hat{\Delta}^2 \overset{P-a.s.}{\underset{n\to\infty}{\longrightarrow}} \Delta^2$. □

The expressions for $N[\tilde{m}, \tilde{\Sigma}](Y_i)$ and $\|N[\tilde{m}, \tilde{\Sigma}]\|^2_{H(\bar{k})}$ in (6.4.3) depend on the choice of $\bar{k}$. Those are given by Propositions 6.4.2 and 6.4.3 when $\bar{k}$ is Gaussian and exponential. Note that in these cases, the continuity assumption of $N[m, \Sigma]$ required by Proposition 6.4.1 is satisfied.

Before stating Propositions 6.4.2 and 6.4.3, the following notation is introduced. For a symmetric operator $L : \mathcal{H} \to \mathcal{H}$ with eigenexpansion $L = \sum_{r\geq 1} \lambda_r \Psi^{\otimes 2}$, its determinant is denoted $|L| = \prod_{r\geq 1} \lambda_r$. For any $q \in \mathbb{R}$, the operator $L^q$ is defined as $L^q = \sum_{r\geq 1} \lambda_r^q \mathbb{1}_{\{\lambda_r > 0\}} \Psi_r^{\otimes 2}$.

**Proposition 6.4.2.** *(Gaussian kernel case) Let $\bar{k}(.,.) = \exp(-\sigma\|. - .\|^2_{\mathcal{H}})$ where $\sigma > 0$. Then,*

$$N[\tilde{m}, \tilde{\Sigma}](.) = |I + 2\sigma\tilde{\Sigma}|^{-1/2} \exp\left(-\sigma\|(I + 2\sigma\tilde{\Sigma})^{-1/2}(. - \tilde{m})\|^2_{\mathcal{H}}\right) \ ,$$
$$\|N[\tilde{m}, \tilde{\Sigma}]\|^2_{H(\bar{k})} = |I + 4\sigma\tilde{\Sigma}|^{-1/2} \ .$$

**Proposition 6.4.3.** *(Exponential kernel case) Let $\bar{k}(.,.) = \exp(\langle ., . \rangle_{\mathcal{H}})$. Assume that the largest eigenvalue of $\tilde{\Sigma}$ is smaller than 1. Then,*

$$N[\tilde{m}, \tilde{\Sigma}](.) = \exp\left(\langle \tilde{m}, . \rangle_{\mathcal{H}} + \frac{1}{2}\langle \tilde{\Sigma}., . \rangle_{\mathcal{H}}\right) \ ,$$
$$\|N[\tilde{m}, \tilde{\Sigma}]\|^2 = |I - \tilde{\Sigma}^2|^{-1/2} \exp\left(\|(I - \tilde{\Sigma}^2)^{-1/2}\tilde{m}\|^2_{\mathcal{H}}\right) \ .$$

The proofs of Propositions 6.4.2 and 6.4.3 are provided in Appendix 6.B.1.

For most estimators $(\tilde{m}, \tilde{\Sigma})$, the quantities provided in Propositions 6.4.3 and 6.4.2 are computable via the Gram matrix $K = \left[\langle Y_i, Y_j \rangle_{\mathcal{H}}\right]_{1 \leq i,j \leq n}$. For instance, assume that $(\tilde{m}, \tilde{\Sigma})$ are the classical estimators $(\hat{m}, \hat{\Sigma})$ where $\hat{m} = (1/n)\sum_{i=1}^n Y_i$ and $\hat{\Sigma} = (1/n)\sum_{i=1}^n (Y_i - \hat{m})^{\otimes 2}$. Let $I_n$ and $J_n$ be respectively the $n \times n$ identity matrix and the $n \times n$ matrix whose all entries equal 1, $H = I_n - (1/n)J_n$ and $K_c = HKH$ be the centered Gram matrix. Then for any $\square \in \mathbb{R}$,

$$\left|I + \square\hat{\Sigma}\right| = \det\left(I_n + \frac{\square}{n} K_c\right) \ ,$$

where det(.) denotes the determinant of a matrix and

$$\left\|(I + \square\hat{\Sigma})^{-1/2} Y_i\right\|_{\mathcal{H}}^2 = \left[(I_n + \frac{\square}{n} K_c)^{-1}\right]_{i,i} \ ,$$

where $[.]_{ii}$ denotes the entry in the $i$-th row and the $i$-th column of a matrix.

### 6.4.2 Estimation of the critical value

Designing a test with confidence level $0 < \alpha < 1$ requires to compute the $1 - \alpha$ quantile of the $n\hat{\Delta}^2$ distribution under $\mathbf{H}_0$ denoted by $q_{\alpha,n}$. Thus $q_{\alpha,n}$ serves as a critical value to decide whether the test statistic $\hat{\Delta}^2$ is significantly close to 0 or not, so that the probability of wrongly rejecting $\mathbf{H}_0$ (Type-I error) is at most $\alpha$.

#### Classical parametric bootstrap

In the case of a goodness-of-fit test, a usual way of estimating $q_{\alpha,n}$ is to perform a parametric bootstrap. Parametric bootstrap consists in generating $B$ samples of $n$ *i.i.d.* random variables $Y_1^{(b)}, \ldots, Y_n^{(b)} \sim \mathcal{N}(\tilde{m}, \tilde{\Sigma})$ $(b = 1, \ldots, B)$. Each of these $B$ samples is used to compute a bootstrap replication

$$[n\hat{\Delta}^2]^b = n\|\hat{\mu}_P^b - N[\tilde{m}^b, \tilde{\Sigma}^b]\|_{H(\tilde{k})}^2 \ , \tag{6.4.4}$$

where $\hat{\mu}_P^b$, $\tilde{m}^b$ and $\tilde{\Sigma}^b$ are the estimators of $\mu_P$, $m$ and $\Sigma$ based on $Y_1^b, \ldots, Y_n^b$.

It is known that parametric bootstrap is asymptotically valid [SMQ93]. Namely, under $\mathbf{H}_0$,

$$\forall b = 1, \ldots, B, \quad \left(n\hat{\Delta}^2, \ [n\hat{\Delta}^2]^b\right) \xrightarrow[n \to +\infty]{\mathcal{L}} (U, U') \ ,$$

where $U$ and $U'$ are *i.i.d.* random variables. In a nutshell, (6.4.4) is approximately an independent copy of the test statistic $n\hat{\Delta}^2$ (under $\mathbf{H}_0$). Therefore $B$ replications $[n\hat{\Delta}^2]^b$ can be used to estimate the $1 - \alpha$ quantile $q_{\alpha,n}$ of $n\hat{\Delta}^2$ under the null-hypothesis.

However, this approach suffers heavy computational costs. In particular, each bootstrap replication involves the estimators $(\tilde{m}^b, \tilde{\Sigma}^b)$. In our case, this leads to compute the eigendecomposition

of the $B$ Gram matrices $K^b = [\langle Y_i^b, Y_j^b \rangle]_{i,j}$ of size $n \times n$ hence a complexity of order $\mathcal{O}(Bn^3)$.

**Fast parametric bootstrap**

This computational limitation is alleviated by means of another strategy described in [KY12]. Let us consider in a first time the case when the estimators of $m$ and $\Sigma$ are the classical empirical mean and covariance $\hat{m} = (1/n)\sum_{i=1}^n Y_i$ and $\hat{\Sigma} = (1/n)\sum_{i=1}^n (Y_i - \hat{m})^{\otimes 2}$. Introducing the Fréchet derivative [FSG08] $D_{(m,\Sigma)}N$ at $(m,\Sigma)$ of the function

$$N : \Theta \to H(\bar{k}), \;\; (m,\Sigma) \mapsto N[m,\Sigma] \;\; ,$$

our bootstrap method relies on the following approximation

$$\begin{aligned}
\sqrt{n}\Big(\hat{\bar{\mu}}_P - N[\hat{m},\hat{\Sigma}]\Big) \simeq \;& \sqrt{n}\left(\hat{\bar{\mu}}_P - \underbrace{N[m_0,\Sigma_0]}_{=\bar{\mu}_P \text{ under } \mathcal{H}_0} - D_{(m_0,\Sigma_0)}N[\hat{m}-m_0, \hat{\Sigma}-\Sigma_0]\right) \\
\simeq \;& \frac{1}{\sqrt{n}}\sum_{i=1}^n [\bar{k}(Y_i,.) - \bar{\mu}_P] \\
& - D_{(m_0,\Sigma_0)}N[Y_i - m_0, (Y_i-m_0)^{\otimes 2} - \Sigma_0] \;\; . \qquad (6.4.5)
\end{aligned}$$

Since (6.4.5) consists of a sum of centered independent terms (under $\mathbf{H}_0$), it is possible to generate approximate independent copies of this sum via *weighted* bootstrap [Bur00]. Given $Z_1^b, \ldots, Z_n^b$ *i.i.d.* real random variables of mean zero and unit variance and $\bar{Z}^b$ their empirical mean, a bootstrap replication of (6.4.5) is given by

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (Z_i^b - \bar{Z}^b)\Big\{\bar{k}(Y_i,.) - D_{(m_0,\Sigma_0)}N[Y_i, (Y_i-m_0)^{\otimes 2}]\Big\} \;\; . \qquad (6.4.6)$$

Taking the square of the norm of (6.4.6) in $H(\bar{k})$ and replacing the unknown true parameters $m_0$ and $\Sigma_0$ by their estimators $\hat{m}$ and $\hat{\Sigma}$ yields the bootstrap replication $[n\hat{\Delta}^2]_{fast}^b$ of $n\hat{\Delta}^2$

$$[n\hat{\Delta}^2]_{fast}^b \stackrel{\Delta}{=} \left\| \sqrt{n}\Big(\hat{\bar{\mu}}_P^b - D_{(\hat{m},\hat{\Sigma})}N[\hat{m}^b, \hat{\Sigma}^b]\Big)\right\|_{H(\bar{k})}^2 \;\; , \qquad (6.4.7)$$

where

$$\begin{aligned}
\hat{\bar{\mu}}_P^b &= (1/n)\sum_{i=1}^n (Z_i^b - \bar{Z}^b)\bar{k}(Y_i,.) \;\; , \\
\hat{m}^b &= (1/n)\sum_{i=1}^n (Z_i^b - \bar{Z}^b) Y_i \;\; ,
\end{aligned}$$

$$\hat{\Sigma}^b = (1/n)\sum_{i=1}^{n}(Z_i^b - \bar{Z}^b)(Y_i - \hat{m}^b)^{\otimes 2}.$$

Therefore this approach avoids the re-computation of parameters for each bootstrap replication, hence a computational cost of order $\mathcal{O}(Bn^2)$ instead of $\mathcal{O}(Bn^3)$. This is illustrated empirically in the right half of Figure 6.1.

**Fast parametric bootstrap for general parameter estimators**

The bootstrap method proposed by [KY12] used in Section 6.4.2 requires that the estimators $(\tilde{m}, \tilde{\Sigma})$ can be written as a sum of independent terms with an additive term which converges to 0 in probability. Formally, $(\tilde{m}, \tilde{\Sigma}) = (m_0, \Sigma_0) + (1/n)\sum_{i=1}^{n}\psi(Y_i) + \epsilon$ where $\mathbb{E}\psi(Y) = 0$, $\mathrm{Var}(\psi(Y)) < +\infty$ and $\epsilon \xrightarrow[n\to+\infty]{\mathbb{P}} 0$. However there are some estimators which cannot be written in this form straightforwardly. This is the case for instance if we test whether data follow a Gaussian with covariance of fixed rank $r$ (as in Section 6.6). In this example, the associated estimators are $\tilde{m} = \hat{m} = (1/n)\sum_{i=1}^{n} Y_i$ (empirical mean) and $\tilde{\Sigma} = \hat{\Sigma}_r = \sum_{s=1}^{r}\hat{\lambda}_s\hat{\Psi}_s^{\otimes 2}$ where $(\hat{\lambda}_s)_s$ and $(\hat{\Psi}_s)_s$ are the eigenvalues and eigenvectors of the empirical covariance operator $\hat{\Sigma} = (1/n)\sum_{i=1}^{n}(Y_i - \hat{\mu})^{\otimes 2}$.

We extend (6.4.7) to the general case when $\Theta_0 \neq \Theta$ and the estimators $(\tilde{m}, \tilde{\Sigma})$ are not the classical $(\hat{m}, \hat{\Sigma})$. We assume that the estimators $(\tilde{m}, \tilde{\Sigma})$ are functions of the empirical estimators $\hat{m}$ and $\hat{\Sigma}$, namely there exists a differentiable mapping $\mathcal{T}$ such that

$$(\tilde{m}, \tilde{\Sigma}) = \mathcal{T}(\hat{m}, \hat{\Sigma}), \text{ where } \mathcal{T}(\Theta) \subseteq \Theta_0 \text{ and } \mathcal{T}|_{\Theta_0} = \mathrm{Id}_{\Theta_0}.$$

Under this definition, $(\tilde{m}, \tilde{\Sigma})$ are consistent estimators of $(m, \Sigma)$ when $(m, \Sigma) \in \Theta_0$. This kind of estimators are met for various choices of the null-hypothesis:

- **Unknown mean and covariance:** $(\tilde{m}, \tilde{\Sigma}) = (\hat{m}, \hat{\Sigma})$ and $\mathcal{T}$ is the identity map $\mathrm{Id}_\Theta$,
- **Known mean and covariance:** $(\tilde{m}, \tilde{\Sigma}) = (m_0, \Sigma_0)$ and $\mathcal{T}$ is the constant map $\mathcal{T}(m, \Sigma) = (m_0, \Sigma_0)$,
- **Known mean and unknown covariance:** $(\tilde{m}, \tilde{\Sigma}) = (m_0, \hat{\Sigma})$ and $\mathcal{T}(m, \Sigma) = (m_0, \Sigma)$,
- **Unknown mean and covariance of known rank $r$:** $(\tilde{m}, \tilde{\Sigma}) = (\hat{m}, \hat{\Sigma}_r)$ and $\mathcal{T}(m, \Sigma) = (m, \Sigma_r)$ where $\Sigma_r$ is the rank $r$ truncation of $\Sigma$.

By introducing $\mathcal{T}$, we get a similar approximation as in (6.4.5) by replacing the mapping $N : \Theta_0 \to H(\bar{k})$ with $N o \mathcal{T} : \Theta_0 \to H(\bar{k})$. This leads to the bootstrap replication

$$[n\hat{\Delta}^2]_{fast}^b := \left\| \sqrt{n}\left(\hat{\mu}_P^b - D_{(\hat{m},\hat{\Sigma})}(N o \mathcal{T})[\hat{m}^b, \hat{\Sigma}^b]\right)\right\|_{H(\bar{k})}^2 . \tag{6.4.8}$$

The validity of this bootstrap method is justified in Section 6.4.2.

Finally we define an estimator $\hat{q}_{\alpha,n}$ of $q_{\alpha,n}$ from the generated $B$ bootstrap replications $[n\hat{\Delta}^2]^1_{fast} < \ldots < [n\hat{\Delta}^2]^B_{fast}$ (assuming they are sorted)

$$\hat{q}_{\alpha,n} = [n\hat{\Delta}^2]^{(\lfloor (1-\alpha)B \rfloor)} \ ,$$

where $\lfloor . \rfloor$ stands for the integer part. The rejection region is defined by

$$\mathcal{R}_\alpha = \{ n\hat{\Delta}^2 > \hat{q}_{\alpha,n} \} \ .$$

**Validity of the fast parametric bootstrap**

Proposition 6.4.4 hereafter shows the validity of the fast parametric bootstrap as presented in Section 6.4.2. The proof of Proposition 6.4.4 is provided in Section 6.B.2.

**Proposition 6.4.4.** *Assume* $\mathbb{E}_P \bar{k}^{1/2}(Y,Y)$, $\mathrm{Tr}(\Sigma)$ *and* $\mathbb{E}_P \|Y - m_0\|^4$ *are finite. Also assume that* $\mathcal{T}$ *is continuously differentiable on* $\Theta_0$.

*If* $\mathbf{H}_0$ *is true, then for each* $b = 1, \ldots, B$,

*(i)* $\sqrt{n}\left(\hat{\hat{\mu}}_P - N[\tilde{m}, \tilde{\Sigma}]\right) \xrightarrow[n\to+\infty]{\mathcal{L}} G_P - D_{(m_0,\Sigma_0)}(No\mathcal{T})[U_P]$

*(ii)* $\sqrt{n}\left(\hat{\hat{\mu}}^b_P - D_{(\hat{m},\hat{\Sigma})}(No\mathcal{T})[\hat{m}^b, \hat{\Sigma}^b]\right) \xrightarrow[n\to+\infty]{\mathcal{L}} G'_P - D_{(m_0,\Sigma_0)}(No\mathcal{T})[U'_P]$

*where* $(G_P, U_P)$ *and* $(G'_P, U'_P)$ *are i.i.d. random variables in* $H(\bar{k}) \times \Theta$.
*If otherwise* $\mathbf{H}_0$ *is false,* $(ii)$ *still holds true (as long as* $m_0$ *and* $\Sigma_0$ *are well defined with respect to* $P$*). Furthermore,* $G_P$ *and* $U_P$ *are zero-mean Gaussian r.v. with covariances*

$$\begin{aligned}
\mathrm{Var}(G_P) &= \mathbb{E}_{Y\sim P}(\bar{k}(Y,.) - \bar{\mu}_P)^{\otimes 2} \\
\mathrm{Var}(U_P) &= \mathbb{E}_{Y\sim P}\left[ Y - m_0, (Y-m_0)^{\otimes 2} - \Sigma \right]^{\otimes 2} \\
\mathrm{cov}(G_P, U_P) &= \mathbb{E}_{Y\sim P}(\bar{k}(Y,.) - \bar{\mu}_P) \otimes \left[ Y - m_0, (Y-m_0)^{\otimes 2} - \Sigma_0 \right] \ .
\end{aligned}$$

By the Continuous Mapping Theorem and the continuity of $\|.\|^2_{H(\bar{k})}$, Proposition 6.4.4 guarantees that the estimated quantile converges almost surely to the true one as $n, B \to +\infty$, so that the type-I error equals $\alpha$ asymptotically.

Note that in [KY12] the parameter subspace $\Theta_0$ must be a subset of $\mathbb{R}^p$ for some integer $p \geq 1$. Proposition 6.4.4 allows $\Theta_0$ to be a subset of a possibly infinite-dimensional Hilbert space ($m$ belongs to $\mathcal{H}$ and $\Sigma$ belongs to the space of finite trace operators $\mathcal{H} \to \mathcal{H}$).

Figure 6.1 (left plot) compares empirically the bootstrap distribution of $[n\hat{\Delta}^2]^b_{fast}$ and the distribution of $n\hat{\Delta}^2$. When $n = 1000$, the two corresponding densities are superimposed and a two-sample Kolmogorov-Smirnov test returns a p-value of 0.978 which confirms the strong similarity between the two distributions. Therefore the fast bootstrap method seems to provide a very good approximation of the distribution of $n\hat{\Delta}^2$ even for a moderate sample size $n$.

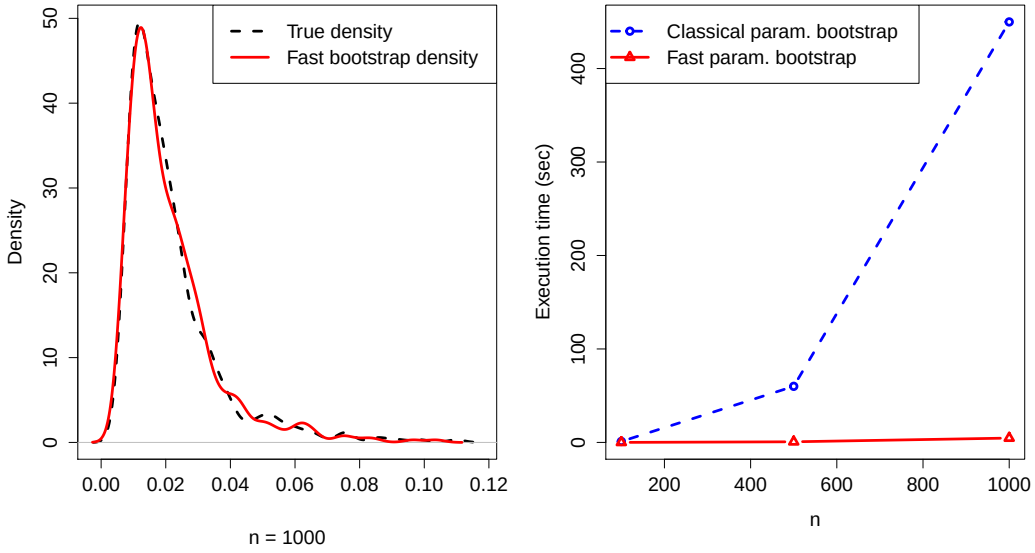


Figure 6.1 – **Left:** Comparison of the distributions of $n\hat{\Delta}^2$ (test statistic) and $[n\hat{\Delta}^2]^b_{fast}$ (fast bootstrap replication) when $n = 1000$. A Kolmogorov-Smirnov two-sample test applied to our simulations returns a p-value of 0.978 which confirms the apparent similarity between the two distributions. **Right:** Comparison of the execution time (in seconds) of both classical and fast bootstrap methods.

## 6.5 Test performances

### 6.5.1 An upper bound for the Type-II error

Let us assume the null-hypothesis is false, that is $P \neq \mathcal{N}(m_0, \Sigma_0)$ or $(m_0, \Sigma_0) \notin \Theta_0$. Theorem 6.5.1 gives the magnitude of the Type-II error, that is the probability of wrongly accepting $\mathbf{H}_0$. The proof can be found in Appendix 6.B.3.

Before stating Theorem 6.5.1, let us introduce or recall useful notation :

- $\Delta = \left\| \bar{\mu}_P - (NoT)[m_0, \Sigma_0] \right\|_{H(\bar{k})}$,
- $q_{\alpha,n} = \mathbb{E}\hat{q}_{\alpha,n}$ ,
- $m_P^2 = \mathbb{E}_P \| D_{(m_0,\Sigma_0)}(NoT)[\Psi(Y)] - \bar{k}(Y,.) + \bar{\mu}_P \|^2_{H(\bar{k})}$,

where $\Psi(Y) = (Y - m_0, [Y - m_0]^{\otimes 2} - \Sigma_0)$ and $D_{(m_0,\Sigma_0)}(NoT)$ denotes the Fréchet derivative of $NoT$ at $(m_0, \Sigma_0)$. According to Proposition 6.4.4 and the continuous mapping theorem, $\hat{q}_{\alpha,n}$ corresponds to an order statistic of a random variable which converges weakly to $\left\| G'_P - D_{(m_0,\Sigma_0)}(NoT)[U'_P] \right\|^2$

(as defined in Proposition 6.4.4). Therefore, its mean $q_{\alpha,n}$ tends to a finite quantity as $n \to +\infty$. $L$ and $m_P^2$ do not depend on $n$ as well.

**Theorem 6.5.1.** *(Type II error) Assume* $\sup_{x,y\in\mathcal{H}_0} |\bar{k}(x,y)| = M < +\infty$ *where* $Y \in \mathcal{H}_0 \subseteq \mathcal{H}$ *P-almost surely and* $\hat{q}_{\alpha,n}$ *is independent of* $n\hat{\Delta}^2$.
*Then, for any* $n > q_{\alpha,n}\Delta^{-2}$

$$\mathbb{P}\left(n\hat{\Delta}^2 \leq \hat{q}_{\alpha,n}\right) \leq \exp\left(-\frac{n\left(\Delta - \frac{q_{\alpha,n}}{n\Delta}\right)^2}{2m_P^2 + Cm_P M^{1/2}(\Delta^2 - q_{\alpha,n}/n)}\right) f(\alpha, B, M, \Delta) \ , \tag{6.5.9}$$

*where*

$$f(\alpha, B, M, \Delta) = (1 + o_n(1))\left(1 + \frac{C_{P^b}}{C'\Delta^2 M^{1/2} m_P \sqrt{\alpha B}} + \frac{o_B(B^{-1/2})}{C''\Delta^4 M m_P^2}\right) \ ,$$

*and* $C, C', C''$ *are absolute constants and* $C_{P^b}$ *only depends on the distribution of* $[n\hat{\Delta}^2]^b_{fast}$.

The first implication of Proposition (6.5.1) is that our test is consistent, that is

$$\mathbb{P}(n\hat{\Delta}^2 \leq \hat{q}_{\alpha,n} \,|\, \mathbf{H}_0 \text{ false}) \underset{n\to+\infty}{\longrightarrow} 0 \ .$$

Furthermore, the upper bound in (6.5.9) reflects the expected behaviour of the Type-II error with respect to meaningful quantities. When $\Delta$ decreases, the bound increases (alternative more difficult to detect). When $\alpha$ (Type-I error) decreases, $q_{\alpha,n}$ gets larger and $n$ has to be larger to get the bound. The variance term $m_P^2$ encompasses the difficulty of estimating $\bar{\mu}_P$ and of estimating the parameters as well. In the special case when $m$ and $\Sigma$ are known, $\mathcal{T} = Id$ and the chain rule yields $D_{(m_0,\Sigma_0)}(N o \mathcal{T}) = (D_{\mathcal{T}(m_0,\Sigma_0)}N) o (D_{(m_0,\Sigma_0)}\mathcal{T}) = 0$ so that $m_P^2 = \mathbb{E}\|\bar{\phi}(Y) - \bar{\mu}_P\|^2$ reduces to the variance of $\hat{\bar{\mu}}_P$. As expected, a large $m_P^2$ makes the bound larger. Note that the estimation of the critical value which is related to the term $f(\alpha, B, M, \Delta)$ in (6.5.9) does not alter the asymptotic rate of convergence of the bound.

Remark that assuming that $\hat{q}_{\alpha,n}$ is independent of $n\hat{\Delta}^2$ is reasonable for a large $n$, since $n\hat{\Delta}^2$ and $\hat{q}_{\alpha,n}$ are asymptotically independent according to Proposition 6.4.4.

### 6.5.2 Empirical study of type-I/II errors

Empirical performances of our test are inferred on the basis of synthetic data. For the sake of brevity, our test is referred to as KNT (Kernel Normality Test) in the following.

One main concern of goodness-of-fit tests is their drastic loss of power as dimensionality increases. Empirical evidences (see Table 3 in [SR05]) prove ongoing multivariate normality tests suffer such deficiencies. The purpose of the present section is to check if KNT displays a good behavior in high or infinite dimension.

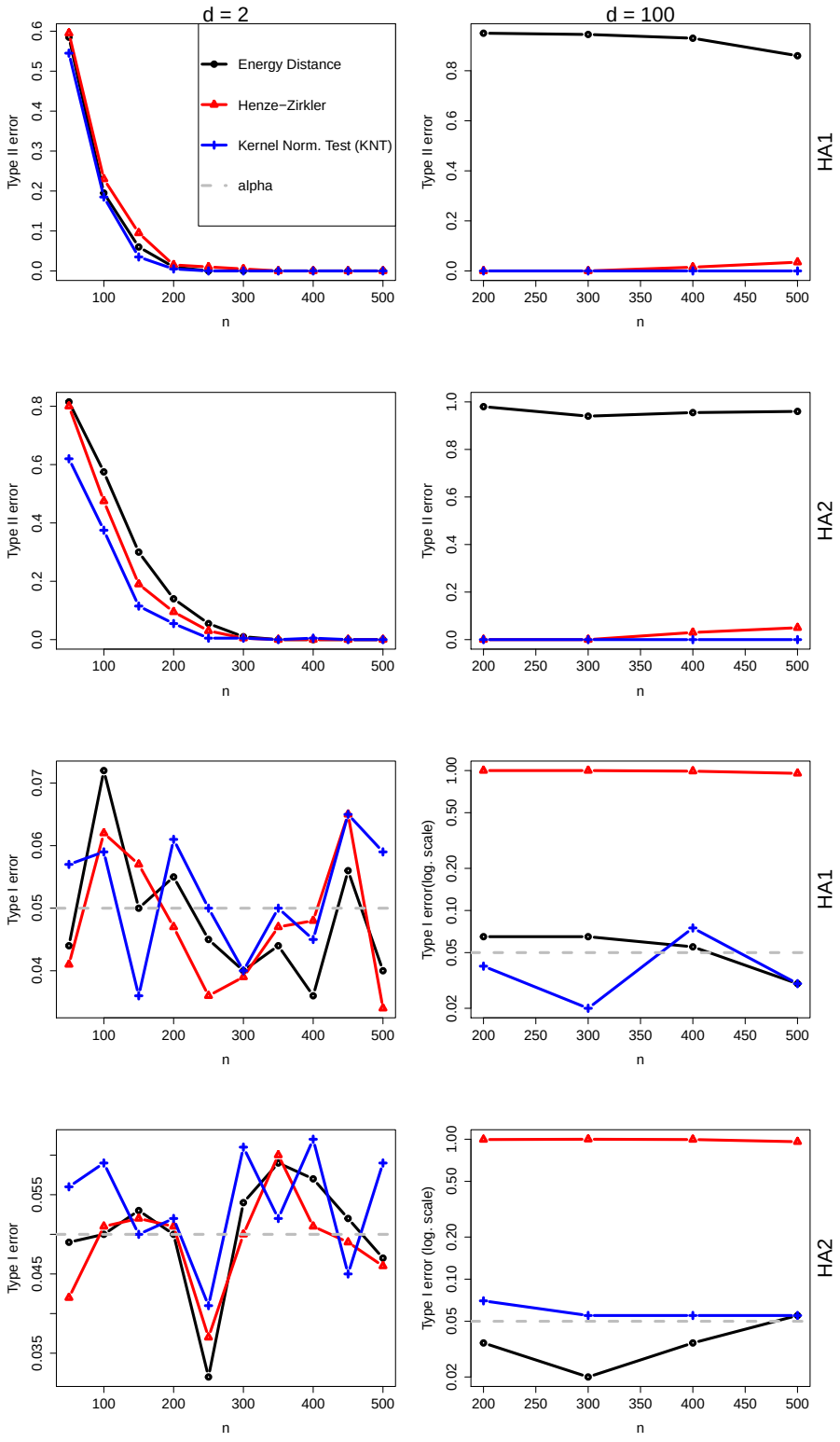


Figure 6.2 – Type-I and type-II errors of KNT (+ blue), Energy Distance (∘ black), and Henze-Zirkler (△ red). Two alternative distributions are considered: HA1 (rows 1 and 3) and HA2 (rows 2 and 4). Two settings are considered: $d = 2$ (left) and $d = 100$ (right).

Besides, note that in the following section, $\bar{k}$ is defined as a Gaussian kernel (see Proposition 6.4.2) with arbitrarily fixed parameter ($\sigma = 1$).

**Finite-dimensional case (Synthetic data)**

*Reference tests.* The power of our test is compared with that of two multivariate normality tests: the Henze-Zirkler test (HZ) [HZ90] and the energy distance (ED) test [SR05]. The main idea of these tests is briefly recalled in Appendix 6.A.1 and 6.A.2.

*Null and alternative distributions.* Two alternatives are considered: a mixture of two Gaussians with different means ($\mu_1 = 0$ and $\mu_2 = 1.5\ (1, 1/2, \ldots, 1/d)$) and same covariance $\Sigma = 0.5\ \mathrm{diag}(1, 1/4, \ldots, 1/d^2)$, whose mixture proportions equals either $(0.5, 0.5)$ (alternative HA1) or $(0.8, 0.2)$ (alternative HA2). Furthermore, two different cases for $d$ are considered: $d = 2$ (small dimension) and $d = 100$ (large dimension).

*Simulation design.* 200 simulations are performed for each test, each alternative and each $n$ (ranging from 100 to 500). $B$ is set at $B = 250$ for KNT. The test level is set at $\alpha = 0.05$ for all tests.

*Results.* In the small dimension case (Figure 6.2, left column), the actual Type-I error of all tests remain more or less around $\alpha$ ($\pm 0.02$). Their Type-II errors are superimposed and quickly decrease down to 0 when $n \geq 200$. On the other hand, experimental results reveal different behaviors as $d$ increases (Figure 6.2, right column). Whereas ED test lose power, KNT and HZ still exhibits small Type-II error values. Besides, ED and KNT Type-I errors remain around the prescribed level $\alpha$ while that of HZ is close to 1, which shows that its small Type-II error is artificial. This seems to confirm that HZ and ED tests are not suited to high-dimensional settings unlike KNT.

**Infinite-dimensional case (real data)**

*Dataset and chosen kernel.* Let us consider the USPS dataset which consists of handwritten digits represented by a vectorized $8 \times 8$ greyscale matrix ($\mathcal{X} = \mathbb{R}^{64}$). A Gaussian kernel $k_G(\cdot, \cdot) = \exp(-\tau^2 \| \cdot - \cdot \|^2)$ is used with $\tau^2 = 10^{-4}$. Comparing sub-datasets "Usps236" (keeping the three classes "2", "3" and "6", 541 observations) and "Usps358" (classes "3", "5" and "8", 539 observations), the 3D-visualization (Figure 6.3, top panels) suggests three well-separated Gaussian components for “Usps236” (left panel), and more overlapping classes for “Usps358” (right panel).

*References tests.* KNT is compared with Random Projection (RP) test, specially designed for infinite-dimensional settings. RP is presented in Appendix 6.A.3. Several numbers of projections $p$ are considered for the RP test : $p = 1, 5$ and 15.

*Simulation design.* We set $\alpha = 0.05$ and 200 repetitions have been done for each sample size.

*Results.* (Figure 6.3, bottom plots) RP is by far less powerful KNT in both cases, no matter how many random projections $p$ are considered. Indeed, KNT exhibits a Type-II error near 0 when $n$ is barely equal to 100, whereas RP still has a relatively large Type-II error when $n = 400$. On the other hand, RP becomes more powerful as $p$ gets larger as expected. A large

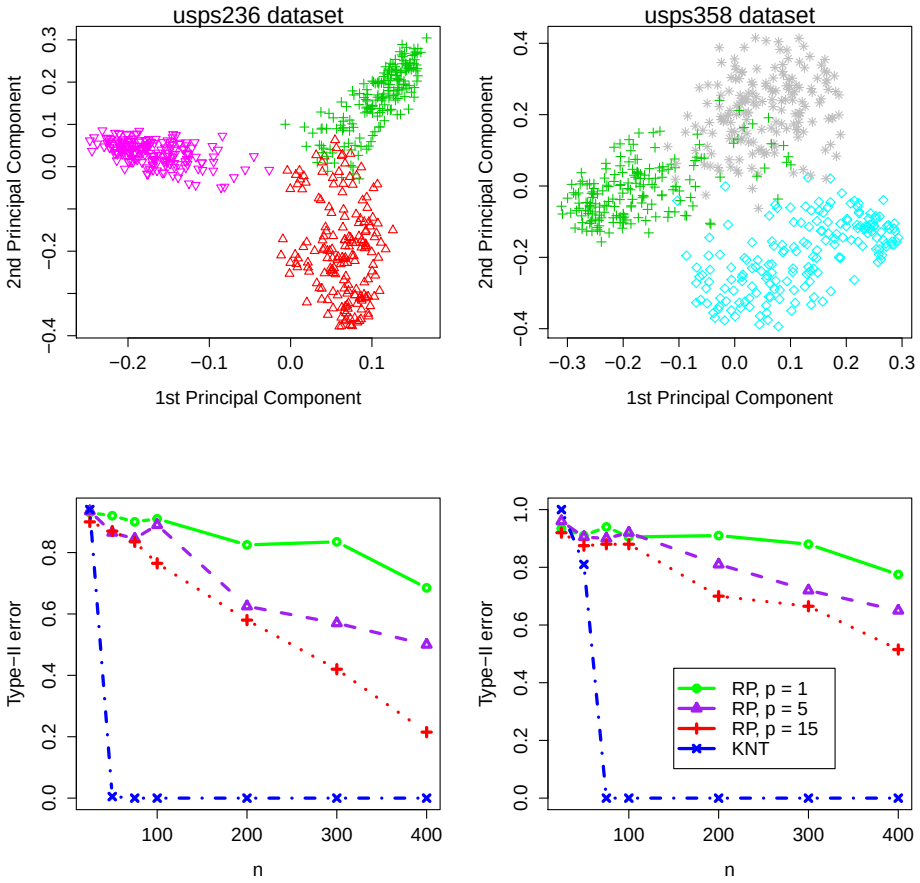


Figure 6.3 – 3D-Visualization (Kernel PCA) of the "Usps236" (top row, left) and "Usps358" (top row, right) datasets; comparison of Type-II error (bottom row, left: "Usps236", right: "Usps358") for: KNT (× blue) and Random Projection with $p = 1$ (• green), $p = 5$ (△ purple) and $p = 15$ (+ red) random projections.

enough number of random projections may allow RP to catch up KNT in terms of power. But RP has a computational advantage over KNT only when $p = 1$ where the RP test statistic is distribution-free. This is no longer the case when $p \geq 2$ and the critical value for the RP test is only available through Monte-Carlo methods.

## 6.6 Application to covariance rank selection

### 6.6.1 Covariance rank selection through sequential testing

Under the Gaussian assumption, the null hypothesis becomes

$$\mathbf{H}_0 : (m_0, \Sigma_0) \in \Theta_0 \ ,$$

and our test reduces to a test on parameters.

We focus on the estimation of the rank of the covariance operator $\Sigma$. Namely, we consider a collection of models $(\mathcal{M}_r)_{1 \leq r \leq r_{max}}$ such that, for each $r = 1, \dots, r_{max}$,

$$\mathcal{M}_r = \{P = \mathcal{N}(m, \Sigma_r) \mid m \in H(k) \text{ and } \mathrm{rk}(\Sigma_r) = r\} \ .$$

Each of these models correspond respectively to the following null hypotheses

$$H_{0,r} : \mathrm{rank}(\Sigma) = r, \quad r = 1, \dots, r_{max} \ ,$$

and the corresponding tests can be used to select the most reliable model.

This can be seen as a case of multiple hypothesis testing, that could be solved through methods such as Benjamini-Hochberg procedure [BH95]. However such procedures usually impose to perform all of the tests and to sort all of the corresponding p-values. In order to eventually perform only a few tests, the null-hypotheses $H_{0,r}$ are tested in a sequential procedure which is summarized in Algorithm 2. This sequential procedure yields an estimator $\hat{r}$ defined as

$$\hat{r} \stackrel{\Delta}{=} \min_{\tilde{r}} \{H_{0,r} \text{ rejected for } r = 1, \dots, \tilde{r} - 1 \text{ and } H_{0,\tilde{r}} \text{ accepted}\} \ .$$

or $\hat{r} \stackrel{\Delta}{=} r_{max}$ if all of the hypotheses are rejected.

Sequential testing to estimate the rank of a covariance matrix (or more generally a noisy matrix) is mentioned in [Rat03] and [RS00]. Both of these papers focus on the probability to select a wrong rank, that is $\mathbb{P}(\hat{r} \neq r^*)$ where $r^*$ denotes the true rank. The goal is to choose a level of confidence $\alpha$ such that this probability of error converges almost surely to 0 when $n \to +\infty$.

There are two ways of guessing a wrong rank : either by overestimation or by underestimation. Getting $\hat{r}$ greater than $r^*$ implies that the null-hypothesis $H_{0,r^*}$ was tested and wrongly rejected, hence a probability of overestimating $r^*$ at most equal to $\alpha$. Underestimating means that at least

**Algorithm 2** Sequential selection of covariance rank

**Input:** Gram matrix $K = [\bar{k}(Y_i, Y_j)]_{i,j}$, confidence level $0 < \alpha < 1$

1. Set $r = 1$ and test $H_{0,r}$
2. If $H_{0,r}$ is rejected and $r < r_{max}$, set $r = r + 1$ and return to 1.
3. Otherwise, set the estimator of the rank $\hat{r} = r$.

**Output:** estimated rank $\hat{r}$

one of the false null-hypothesis $H_{0,1}, \ldots, H_{0,r^*-1}$ was wrongly accepted (Type-II error). Let $\beta_r(\alpha)$ denote the Type-II error of testing $H_{0,r}$ with confidence level $\alpha$ for each $r < r^*$. Thus by a union bound argument,

$$\mathbb{P}(\hat{r} \neq r^*) \leq \sum_{r=1}^{r^*-1} \beta_r(\alpha) + \alpha \ . \tag{6.6.10}$$

The bound in (6.6.10) decreases to 0 only if $\alpha$ converges to 0 but at a slow rate. Indeed, the Type-II errors $\beta_r(\alpha)$ grow with decreasing $\alpha$ but converge to zero when $n \to +\infty$. For instance in the case of the sequential tests mentioned in [Rat03] and [RS00], the correct rate of decrease for $\alpha$ must satisfy $(1/n)\log(1/\alpha) = o_n(1)$.

### 6.6.2 Empirical performances

In this section, the sequential procedure to select covariance rank (as presented in Section 6.6.1) is tested empirically on synthetic data.

**Dataset** A sample of $n$ zero-mean Gaussian with covariance $\Sigma_{r^*}$ are generated, where $n$ ranges from 100 to 5000. $\Sigma_{r^*}$ is of rank $r^* = 10$ and its eigenvalues decrease either polynomially ($\lambda_r = r^{-1}$ for all $r \leq r^*$) or exponentially ($\lambda_r = \exp(-0.2r)$ for all $r \leq r^*$).

**Benchmark** To illustrate the level of difficulty, we compare our procedure with an oracle procedure which uses the knowledge of the true rank. Namely, the oracle procedure follows our sequential procedure at a level $\alpha_{oracle}$ defined as follows

$$\alpha_{oracle} = \max_{1 \leq r \leq r^*-1} \mathbb{P}_Z(n\hat{\Delta}_r^2 \leq Z_r) \ ,$$

where $n\hat{\Delta}_r^2$ is the observed statistic for the $r$-th test and $Z_r$ follows the distribution of this statistic under $H_{0,r}$. Hence $\alpha_{oracle}$ is chosen such that the true rank $r^*$ is selected whenever it is possible.

**Simulation design** To get a consistent estimation of $r^*$, the confidence level $\alpha$ must decrease with $n$ and is set at $\alpha = \alpha_n = \exp(-0.125n^{0.45})$. Each time, 200 simulations are performed.

**Results** The top panels of Figure 6.4 display the proportion of cases when the target rank is found, either for our sequential procedure or the oracle one. When the eigenvalues decay polynomially, the oracle shows that the target rank cannot be almost surely guessed until

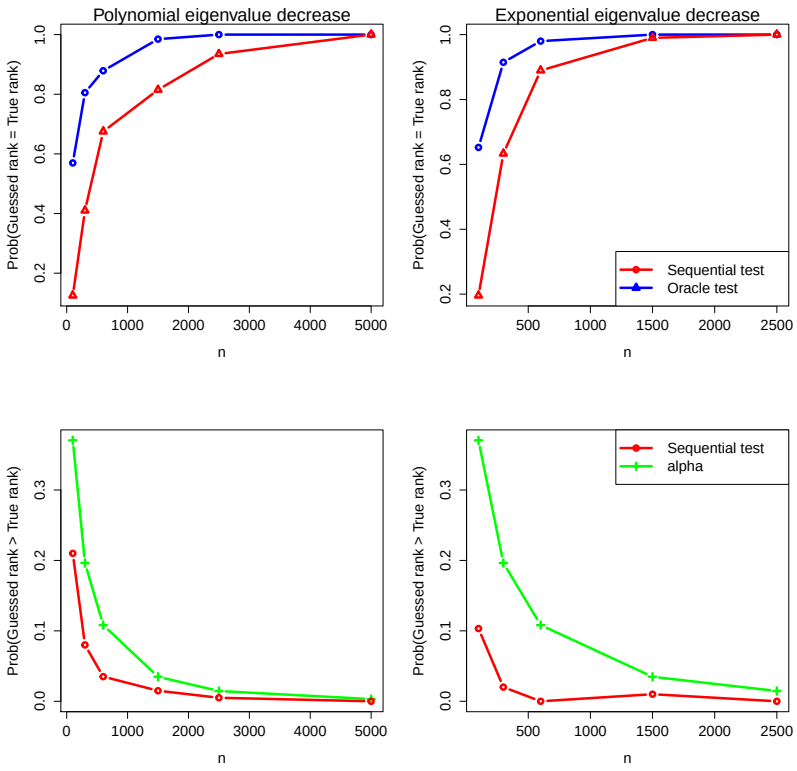


Figure 6.4 – **Top half:** Probabilities of finding the right rank with respect to $n$ for our sequential test (• red) and the oracle procedure (△ blue); **bottom half:** probabilities of overestimating the true rank with the sequential procedure compared with fixed alpha (+ green). In each case, two decreasing rate for covariance eigenvalues are considered : polynomial (left column) and exponential (right column).

$n = 1500$. When $n \leq 1500$, our procedure finds the true rank with probability at least 0.8 and quickly catches up to the oracle as $n$ grows. In the exponential decay case, a similar observation is made. This case seems to be easier, as our procedure performs almost as well as the oracle when $n \geq 600$. In all cases, the consistency of our procedure is confirmed by the simulations.

The bottom panels of Figure 6.4 compare $\alpha$ with the probability of overestimating $r^*$ (denoted by $p_+$). As noticed in Section 6.6.1, the former is an upper bound of the latter. But we must check empirically whether the gap between those two quantities is not too large, otherwise the sequential procedure would be too conservative and lead to excessive underestimation of $r^*$. In the polynomial decay case, the difference between $\alpha$ and $p_+$ is small, even when $n = 100$. The gap is larger in the exponential case but gets narrower when $n \geqslant 1500$.

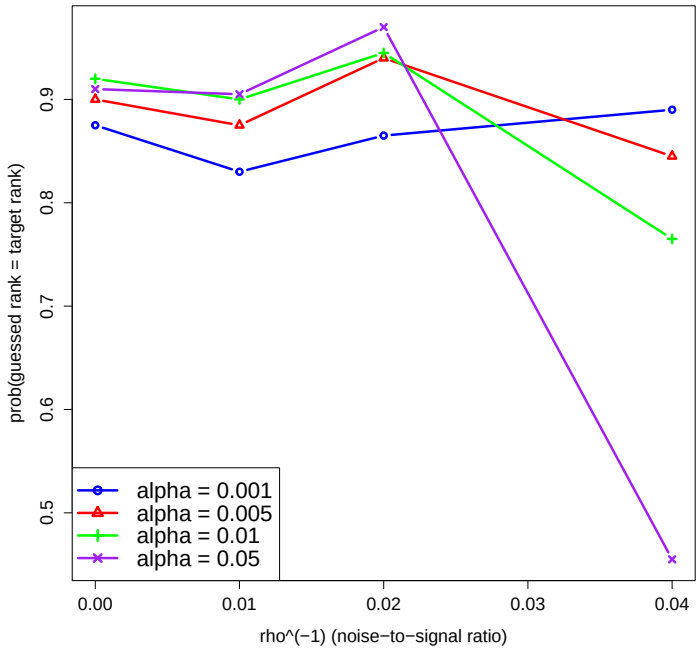


Figure 6.5 – Illustration of the robustness of our sequential procedure under a noisy model.

### 6.6.3 Robustness analysis

In practice, none of the models $\mathcal{M}_r$ is true. An additive full-rank noise term is often considered in the literature [CTT14; JH12]. Namely, we set in our case

$$Y = Z + \epsilon \tag{6.6.11}$$

where $Z \sim \mathcal{N}(m, \Sigma_{r^*})$ with $\mathrm{rk}(\Sigma_{r^*}) = r^*$ and $\epsilon$ is the error term independent of $Z$. Note that the Gaussian assumption concerns the main signal $Z$ and not the error term whereas usual models assume the converse [CTT14; JH12].

Figure 6.5 illustrates the performance of our sequential procedure under the noisy model (6.6.11). We set $\mathcal{H} = \mathbb{R}^{100}$, $n = 600$, $r^* = 3$ and $\Sigma_{r^*} = \Sigma_3 = \mathrm{diag}(\lambda_1, \ldots, \lambda_3, 0, \ldots, 0)$ where $\lambda_r = \exp(-0.2r)$ for $r \leq 3$. The noise term is $\epsilon = (\lambda_3 \rho^{-1} \eta_i)_{1 \leq i \leq 100}$ where $\eta_1, \ldots, \eta_{100}$ are *i.i.d.* Student random variables with 10 degrees of freedom and $\rho > 0$ is the *signal-to-noise ratio*.

As expected, the probability of guessing the target rank $r^*$ decreases down to 0 as the signal-to-noise ratio $\rho$ diminishes. However, choosing a smaller level of confidence $\alpha$ allows to improve the probability of right guesses for a fixed $\rho$. without sacrificing much for smaller signal-to-noise ratios. This is due to the fact that each null-hypothesis $H_{0,r}$ is false, hence the need for a smaller

$\alpha$ (smaller Type-I error) which yields greater Type-II errors and avoids the rejection of all of the null-hypotheses.

## 6.A Goodness-of-fit tests

### 6.A.1 Henze-Zirkler test

The Henze-Zirkler test [HZ90] relies on the following statistic

$$HZ = \int_{\mathbb{R}^d} \left|\hat{\Psi}(t) - \Psi(t)\right|^2 \omega(t) dt \ , \tag{6.A.12}$$

where $\Psi(t)$ denotes the characteristic function of $\mathcal{N}(0, I)$, $\hat{\Psi}(t) = n^{-1} \sum_{j=1}^{n} e^{i\langle t, Y_j\rangle}$ is the empirical characteristic function of the sample $Y_1, \dots, Y_n$, and $\omega(t) = (2\pi\beta)^{-d/2} \exp(-\|t\|^2/(2\beta))$ with $\beta = 2^{-1/2}[(2d+1)n)/4]^{1/(d+4)}$. The $\mathbf{H}_0$-hypothesis is rejected for large values of $HZ$. Note that the sample $Y_1, \dots, Y_n$ must be whitened (centered and renormalized) beforehand.

### 6.A.2 Energy distance test

The energy distance test [SR05] is based on

$$\mathcal{E}(P, P_0) = 2\mathbb{E}\|Y - Z\| - \mathbb{E}\|Y - Y'\| - \mathbb{E}\|Z - Z'\| \tag{6.A.13}$$

which is called the *energy distance*, where $Y, Y' \sim P$ and $Z, Z' \sim P_0$. Note that $\mathcal{E}(P, P_0) = 0$ if and only if $P = P_0$. The test statistic is given by

$$\begin{aligned} \hat{\mathcal{E}} = & \frac{2}{n} \sum_{i=1}^{n} \mathbb{E}_Z \|Y_i - Z\| - \mathbb{E}_{Z,Z'} \|Z - Z'\| \\ & - \frac{1}{n^2} \sum_{i,j=1}^{n} \|Y_i - Y_j\| \ , \end{aligned} \tag{6.A.14}$$

where $Z, Z' \overset{i.i.d.}{\sim} \mathcal{N}(0, I)$ (null-distribution). HZ and ED tests set the $\mathbf{H}_0$-distribution at $P_0 = \mathcal{N}(\hat{\mu}, \hat{\Sigma})$ where $\hat{\mu}$ and $\hat{\Sigma}$ are respectively the standard empirical mean and covariance. As for the Henze-Zirkler test, data must be centered and renormalized.

### 6.A.3 Projection-based statistical tests

In the high-dimensional setting, several approaches share a common idea consisting in projecting on one-dimensional spaces. This idea relies on the Cramer-Wold theorem extended to infinite dimensional Hilbert space.

**Proposition 6.A.1.** *(Prop. 2.1 from [Cue+06]) Let $\mathcal{H}$ be a separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$, and $Y, Z \in \mathcal{H}$ denote two random variables with respective Borel probability measures $P_Y$ and $P_Z$. If for every $h \in \mathcal{H}$, $\langle Y, h\rangle = \langle Z, h\rangle$ weakly then $P_Y = P_Z$.*

[Cue+06] suggest to randomly choose directions $h$ from a Gaussian measure and perform a Kolmogorov-Smirnov test on $\langle Y_1, h\rangle, \dots, \langle Y_n, h\rangle$ for each $h$, leading to the test statistic

$$D_n(h) = \sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F_0(x)| \tag{6.A.15}$$

where $\hat{F}_n(x) = (1/n)\sum_{i=1}^{n} \mathbb{1}_{\{\langle Y_i,h\rangle \le x\}}$ is the empirical cumulative distribution function (cdf) of $\langle Y_1, h\rangle, \dots, \langle Y_n, h\rangle$ and $F_0(x) = \mathbb{P}(\langle Y, h\rangle \le x)$ denotes the cdf of $\langle Z, h\rangle$.

Since [Cue+06] proved too few directions lead to a less powerful test, this can be repeated for several randomly chosen directions $h$, keeping then the largest value for $D_n(h)$. However the test statistic is no longer distribution-free (unlike the univariate Kolmogorov-Smirnov one) when the number of directions is larger than 2.

## 6.B Proofs

Throughout this section, $\langle .,. \rangle$ (resp. $\|.\|$) denotes either $\langle .,. \rangle_{\mathcal{H}}$ or $\langle \ . \ ,. \rangle_{H(\bar{k})}$ (resp. $\|.\|_{\mathcal{H}}$ or $\|.\|_{H(\bar{k})}$) depending on the context.

### 6.B.1 Proof of Propositions 6.4.2 and 6.4.3

Consider the eigenexpansion of $\tilde{\Sigma} = \sum_{i\ge1} \lambda_i \Psi_i^{\otimes 2}$ where $\lambda_1, \lambda_2, \dots$ is a decreasing sequence of positive reals and where $\{\Psi_i\}_{i\ge1}$ form a complete orthonormal basis of $\mathcal{H}$.

Let $Z \sim \mathcal{N}(\tilde{m}, \tilde{\Sigma})$. The orthogonal projections $\langle Z, \Psi_i\rangle$ are $\mathcal{N}(0, \lambda_i)$ and for $i \ne j$, $\langle Z, \Psi_i\rangle$ and $\langle Z, \Psi_j\rangle$ are independent. Let $Z'$ be an independent copy of $Z$.

• **Gaussian kernel case :** $\bar{k}(.,.) = \exp(-\sigma\|\cdot - \cdot\|_{\mathcal{H}}^2)$

Let us first expand the following quantity

$$\begin{aligned}
N[\tilde{m}, \tilde{\Sigma}](y) &= \mathbb{E}_Z \exp\left(-\sigma\|Z - y\|^2\right) \\
&= \mathbb{E}_Z \exp\left(-\sigma \sum_{r\ge1} \langle Z - y, \Psi_r\rangle^2\right) \\
&= \prod_{r=1}^{+\infty} \mathbb{E}_Z \exp\left(-\sigma \sum_{r\ge1} \langle Z - y, \Psi_r\rangle^2\right) \\
&= \prod_{r=1}^{+\infty} \mathbb{E}_Z \exp\left(-\sigma \langle Z - y, \Psi_r\rangle^2\right) \\
&= \prod_{r=1}^{+\infty} (1 + 2\sigma\lambda_r)^{-1/2} \exp\left(-\sigma \frac{\langle \tilde{m} - y, \Psi_r\rangle^2}{1 + 2\sigma\lambda_r}\right) \\
&= \left|I + 2\sigma\tilde{\Sigma}\right|^{-1/2} \exp(-\sigma\langle (I + 2\sigma\tilde{\Sigma})(y - \tilde{m}), y - \tilde{m}\rangle)
\end{aligned} \tag{6.B.16}$$

We can switch the mean and the limit in (6.B.16) by using the Dominated Convergence theorem since for every $N \geq 1$

$$\left| \prod_{r=1}^{N} \exp\left(-\sigma \langle Z - y, \Psi_r \rangle^2\right) \right| \leq 1 < +\infty \ .$$

The second quantity $\|N[\tilde{m}, \tilde{\Sigma}]\|^2$ is computed likewise

$$\begin{aligned}
\|N[\tilde{m}, \tilde{\Sigma}]\|^2 &= \mathbb{E}_{Z,Z'} \exp\left(-\sigma \|Z - Z'\|^2\right) \\
&= \mathbb{E}_{Z,Z'} \exp\left(-\sigma \sum_{r \geq 1} \langle Z - Z', \Psi_r \rangle^2\right) \\
&= \prod_{r=1}^{+\infty} \mathbb{E}_{Z,Z'} \exp\left(-\sigma \sum_{r \geq 1} \langle Z - Z', \Psi_r \rangle^2\right) \\
&= \prod_{r=1}^{+\infty} \mathbb{E}_{Z} \mathbb{E}_{Z'} \left(\exp\left(-\sigma \langle Z - Z', \Psi_r \rangle\right) \mid Z\right) \\
&= \prod_{r=1}^{+\infty} \mathbb{E}_{Z} (1 + 2\sigma\lambda_r)^{-1/2} \exp\left(-\sigma \frac{\langle \tilde{m} - Z, \Psi_r \rangle^2}{1 + 2\sigma\lambda_r}\right) \\
&= \prod_{r=1}^{+\infty} (1 + 2\sigma\lambda_r)^{-1/2} \mathbb{E}_{U \sim \mathcal{N}(0, \lambda_r)} \exp\left(-\frac{\sigma U^2}{1 + 2\sigma\lambda_r}\right) \\
&= \prod_{r=1}^{+\infty} (1 + 2\sigma\lambda_r)^{-1/2} \left(1 + \frac{2\sigma\lambda_r}{1 + 2\sigma\lambda_r}\right)^{-1/2} \\
&= \prod_{r=1}^{+\infty} (1 + 4\sigma\lambda_r)^{-1/2} \\
&= \left|I + 4\sigma\tilde{\Sigma}\right|^{-1/2}
\end{aligned}$$

• **Exponential kernel case :** $\bar{k}(.,.) = \exp(\langle \cdot, \cdot \rangle_{\mathcal{H}})$

Let us first expand the following quantity

$$\begin{aligned}
N[\tilde{m}, \tilde{\Sigma}](y) &= \mathbb{E}_Z \exp\left(\langle Z, y \rangle\right) \\
&= \mathbb{E}_{U \sim \mathcal{N}(\langle \tilde{m}, y \rangle, \langle \tilde{\Sigma} y, y \rangle)} \exp\left(U\right) \\
&= \exp(\langle \tilde{m}, y \rangle + (1/2) \langle \tilde{\Sigma} y, y \rangle).
\end{aligned}$$

Expanding $\|N[\tilde{m}, \tilde{\Sigma}]\|^2$,

$$\begin{aligned}
\|N[\tilde{m}, \tilde{\Sigma}]\|^2 &= \mathbb{E}_{Z,Z'} \exp\left(\langle Z, Z' \rangle\right) \\
&= \mathbb{E}_{Z,Z'} \exp\left(\sum_{i \geq 1} \langle Z, \Psi_i \rangle \langle Z', \Psi_i \rangle\right)
\end{aligned}$$

$$= \prod_{i=1}^{\infty} \mathbb{E}_{Z,Z'} \exp\left(\langle Z, \Psi_i\rangle\langle Z', \Psi_i\rangle\right) \ . \tag{6.B.17}$$

We can switch the mean and the limit by using the Dominated Convergence theorem since

$$\left|\prod_{i=1}^{N} \exp\left(\langle Z, \Psi_i\rangle\langle Z', \Psi_i\rangle\right)\right| \xrightarrow[N\to\infty]{a.s.} \exp\left(\langle Z, Z'\rangle\right) \leq \exp\left(\frac{\|Z\|^2 + \|Z'\|^2}{2}\right) \ ,$$

and

$$\mathbb{E}_{Z,Z'} \exp\left(\frac{\|Z\|^2 + \|Z'\|^2}{2}\right) = \left[\mathbb{E}_Z \exp\left(\frac{\|Z\|^2}{2}\right)\right]^2 = \mathbb{E}\bar{k}^{1/2}(Z,Z) < +\infty \ .$$

The integrability of $\bar{k}^{1/2}(Z,Z)$ is necessary to ensure the existence of the embedding $N[\tilde{m}, \tilde{\Sigma}]$ related to the distribution of $Z$. As we will see thereafter, it is guaranteed by the condition $\hat{\lambda}_1 < 1$.

For each $i$,

$$\begin{aligned}
\mathbb{E}_{Z,Z'} \exp\left(\langle Z, \Psi_i\rangle\langle Z', \Psi_i\rangle\right) &= \mathbb{E}_Z \mathbb{E}_{Z'} \left(\exp\left(\langle Z, \Psi_i\rangle\langle Z', \Psi_i\rangle\right) \mid Z\right) \\
&= \mathbb{E}_Z \exp\left(\frac{\lambda_i}{2}\langle Z, \Psi_i\rangle^2\right) \\
&= (1-\lambda_i^2)^{-1/2}
\end{aligned} \tag{6.B.18}$$

Plugging (6.B.18) into Equation (6.B.17),

$$\|N[\tilde{m}, \tilde{\Sigma}]\|^2 = \prod_{i=1}^{\infty} (1-\gamma_i^2)^{-1/2} = \left|I - \Sigma^2\right|^{-1/2} \ . \tag{6.B.19}$$

(6.B.19) is well defined only if $\left|I - \Sigma^2\right| > 0$. As we have assumed that $\lambda_i < 1$ for all $i$, it is positive. The non-nullity also holds since

$$\begin{aligned}
\prod_{i=1}^{\infty} [1-\lambda_i^2] = \exp\left(-\sum_{i=1}^{\infty} \log\left(\frac{1}{1-\lambda_i^2}\right)\right) &\geq \exp\left(-\sum_{i=1}^{\infty}\left(\frac{1}{1-\lambda_i^2} - 1\right)\right) \\
&= \exp\left(-\sum_{i=1}^{\infty} \frac{\lambda_i^2}{1-\lambda_i^2}\right) \\
&\geq \exp\left(\frac{-\mathrm{Tr}(\Sigma^2)}{1-\lambda_1^2}\right) \ .
\end{aligned} \tag{6.B.20}$$

where we used the inequality: $\log(x) \leq x - 1$. Since $\Sigma$ is a finite trace operator, its eigenvalues converge towards 0 and $\lambda_i^2 \leq \lambda_i < 1$ for $i$ large enough. Thus, $\mathrm{Tr}(\Sigma^2)$ is finite and it follows from (6.B.20) that $\left|I - \Sigma^2\right| > 0$.

### 6.B.2 Proof of Theorem 6.4.4

The proof of Proposition 6.4.4 follows the same idea as the original paper [KY12] and broadens its field of applicability. Namely, the parameter space does not need to be a subset of $\mathbb{R}^d$ anymore. The main ingredient for our version of the proof is to use the CLT in Banach spaces [HP76] instead of the multiplier CLT for empirical processes ([Kos07], Theorem 10.1).

We introduce the following notation :

- $\theta_0 = (m_0, \Sigma_0)$ denotes the true parameters
- $\theta_n = (\hat{m}, \hat{\Sigma})$ denotes the empirical parameters
- For any $\theta = (m, \Sigma)$ and $y \in \mathcal{H}$, $\Psi_\theta(y) = (y - m, (y-m)^{\otimes 2} - \Sigma)$.
- $\theta^b_{0,n} = (1/n)\sum_{i=1}^n (Z_i^b - \bar{Z}^b)\Psi_{\theta_0}(Y_i)$ and $\theta^b_n = (1/n)\sum_{i=1}^n (Z_i^b - \bar{Z}^b)\Psi_{\theta_n}(Y_i)$

As a first step, we prove that

$$\left(\sqrt{n}(\hat{\bar{\mu}}_P - \bar{\mu}_P), n^{-1/2}\sum_{i=1}^n \Psi_{\theta_0}(Y_i), \sqrt{n}\bar{\mu}^b_{\hat{P}}, \sqrt{n}\theta^b_{0,n}\right) \xrightarrow[n\to+\infty]{\mathcal{L}} (G_P, U_P, G'_P, U'_p) \ , \tag{6.B.21}$$

where $(G_P, U_P)$ and $(G'_P, U'_P)$ are *i.i.d.* (jointly) Gaussian random variables in $H(\bar{k}) \times \Theta$. To do this, we need to write the left hand side of (6.B.21) as a sum of *i.i.d.* terms to apply the CLT for Hilbert spaces [HP76]. As a preliminary step, we have to rewrite $\sqrt{n}\bar{\mu}^b_{\hat{P}}$ and $\sqrt{n}\theta^b_{0,n}$ as follows

$$\begin{aligned}
\sqrt{n}\bar{\mu}^b_{\hat{P}} &= n^{-1/2}\sum_{i=1}^n Z_i^b(\bar{k}(Y_i,\cdot) - \bar{\mu}) + \sqrt{n}\bar{Z}^b(\bar{\mu} - \hat{\bar{\mu}}) = n^{-1/2}\sum_{i=1}^n Z_i^b(\bar{k}(Y_i,\cdot) - \bar{\mu}) + o_P(1) \\
\sqrt{n}\theta^b_{0,n} &= n^{-1/2}\sum_{i=1}^n Z_i^b \Psi_{\theta_0}(Y_i) - \sqrt{n}\bar{Z}^b(m - \hat{m}, \hat{\Sigma} - \Sigma) = \sqrt{n}\bar{Z}^b + o_P(1) \ ,
\end{aligned}$$

where $\sqrt{n}\bar{Z}^b \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$ (from the classical CLT) and $(\hat{m}, \hat{\Sigma}) \to (m_0, \Sigma_0)$ almost surely from the Law of Large Numbers (LLN) in Banach spaces (Theorem 2.1, [HP76]). More precisely, in our case, the LLN holds if $\mathbb{E}\|Y - m_0\|^2$ and $\mathbb{E}\|(Y-m)^{\otimes 2} - \Sigma_0\|^2$ are finite, which is the case by assumption since $\mathbb{E}\|Y - m_0\|^2 = \mathrm{Tr}(\Sigma_0) < +\infty$ and $\mathbb{E}\|(Y-m_0)^{\otimes 2} - \Sigma_0\|^2 = \mathbb{E}\|Y - m_0\|^4 + \mathrm{Tr}(\Sigma_0 - \Sigma_0^2) < +\infty$. Likewise $\hat{\bar{\mu}} \to \bar{\mu}$ a.s. since $\mathbb{E}\|\bar{k}(Y,\cdot) - \bar{\mu}\|^2 = \mathbb{E}\bar{k}(Y,Y) - \|\bar{\mu}\|^2 < +\infty$ by assumption.

Therefore the lhs in (6.B.21) can be written as

$$\left(\sqrt{n}(\hat{\bar{\mu}}_P - \bar{\mu}_P), n^{-1/2}\sum_{i=1}^n \Psi_{\theta_0}(Y_i), \sqrt{n}\bar{\mu}^b_{\hat{P}}, \sqrt{n}\theta^b_{0,n}\right) = n^{-1/2}\sum_{i=1}^n W_i + o_P(1) \ , \tag{6.B.22}$$

where $W_i = \left(\bar{k}(Y_i,\cdot) - \bar{\mu}_P, \Psi_{\theta_0}(Y_i), Z_i^b(\bar{k}(Y_i,\cdot) - \bar{\mu}), Z_i^b\Psi_{\theta_0}(Y_i)\right)$ for $i = 1, \dots, n$ are *i.i.d.* terms.

Now we are in a position to apply the CLT in Hilbert spaces ([HP76],Theorem 3.5) to $n^{-1/2}\sum_{i=1}^n W_i$ so that (6.B.22) entails (6.B.21). In particular the CLT holds if $W_i$ are of mean 0

and finite variance. Clearly, $\mathbb{E}W_i = 0$. As for the finite variance condition, let us first introduce the covariance operators $C_1, C_1^b : H(\bar{k}) \to H(\bar{k})$, $C_2, C_2^b : \Theta \to \Theta$ as follows

$$\begin{aligned}
C_1 &= \mathbb{E}_Y(\bar{k}(Y,.) - \bar{\mu}_P)^{\otimes 2} \\
C_2 &= \mathbb{E}_Y \Psi_{\theta_0}(Y)^{\otimes 2} \\
C_1^b &= \mathbb{E}_{Z^b,Y}[Z^b]^2(\bar{k}(Y,\cdot) - \bar{\mu})^{\otimes 2} = C_1 \\
C_2^b &= \mathbb{E}_{Z^b,Y}[Z^b]^2 \Psi_{\theta_0}^{\otimes 2}(Y) = C_2 \ .
\end{aligned}$$

$W_i$ is of finite variance if for every $(g,u,g',u') \in H(\bar{k}) \times \Theta \times H(\bar{k}) \times \Theta$ , $\mathbb{E}\langle (g,u,g',u'), W\rangle^2 < +\infty$. Remark that $\mathbb{E}\langle (g,u,g',u'), W\rangle^2 \leqslant 4\Big(2\mathrm{Tr}(C_1)(\|g\|^2 + \|g'\|^2) + 2\mathrm{Tr}(C_2)(\|u\|^2 + \|u'\|^2)\Big)$ so it suffices to show that $\mathrm{Tr}(C_1)$ and $\mathrm{Tr}(C_2)$ are finite. This is the case since $\mathrm{Tr}(C_1) = \mathbb{E}_P \bar{k}(Y,Y) - \|\bar{\mu}_P\|^2 < +\infty$ and $\mathrm{Tr}(C_2) = \mathbb{E}_P\|Y - \mu_0\|^4 + \mathrm{Tr}(\Sigma_0 - \Sigma_0^2) < +\infty$ by assumption. Therefore, the CLT in Hilbert spaces yields

$$\begin{aligned}
\sqrt{n}(\hat{\bar{\mu}}_P - \bar{\mu}_P) &\xrightarrow[n\to+\infty]{\mathcal{L}} G_P \sim \mathcal{GP}(0, C_1) \\
n^{-1/2}\sum_{i=1}^n \Psi_{\theta_0}(Y_i) &\xrightarrow[n\to+\infty]{\mathcal{L}} U_P \sim \mathcal{GP}(0, C_2) \\
\sqrt{n}\hat{\bar{\mu}}_P^b &\xrightarrow[n\to+\infty]{\mathcal{L}} G_P' \sim \mathcal{GP}(0, C_1) \\
\sqrt{n}\theta_{0,n}^b &\xrightarrow[n\to+\infty]{\mathcal{L}} U_P' \sim \mathcal{GP}(0, C_2) \ .
\end{aligned}$$

Furthermore since $(G_P, U_P)$ and $(G_P', U_P')$ are jointly Gaussian and

$$\mathrm{cov}\Big(\sqrt{n}\hat{\bar{\mu}}_P^b, \sqrt{n}(\hat{\bar{\mu}}_P - \bar{\mu}_P)\Big) = 0, \qquad \mathrm{cov}\Big(\sqrt{n}\theta_{0,n}^b, n^{-1/2}\textstyle\sum_{i=1}^n \Psi_{\theta_0}(Y_i)\Big) = 0$$

$$\mathrm{cov}\Big(\sqrt{n}\hat{\bar{\mu}}_P^b, n^{-1/2}\textstyle\sum_{i=1}^n \Psi_{\theta_0}(Y_i)\Big) = 0, \quad \mathrm{cov}\Big(\sqrt{n}\theta_{0,n}^b, \sqrt{n}(\hat{\bar{\mu}}_P - \bar{\mu}_P)\Big) = 0 \ ,$$

the limit Gaussian processes $(G_P, U_P)$ and $(G_P', U_P')$ are independent. Besides, $(G_P, U_P)$ and $(G_P', U_P')$ share the same cross-covariance operator $C_{1,2} : \Theta \to H(\bar{k})$ since

$$\begin{aligned}
C_{1,2} &= \mathrm{cov}\left(\sqrt{n}(\hat{\bar{\mu}}_P - \bar{\mu}_P), n^{-1/2}\sum_{i=1}^n \Psi_{\theta_0}(Y_i)\right) = \mathbb{E}_Y(\bar{k}(Y,.) - \bar{\mu}_P) \otimes \Psi_{\theta_0}(Y) \\
C_{1,2}^b &= \mathrm{cov}\Big(Z^b(\bar{k}(Y,\cdot) - \bar{\mu}), Z^b\Psi_{\theta_0}(Y)\Big) = \mathbb{E}[Z^b]^2 C_{1,2} = C_{1,2} \ ,
\end{aligned}$$

therefore $(G_P, U_P)$ and $(G_P', U_P')$ are *i.i.d.* .

Since $D_{\theta_0}(No\mathcal{T})$ is assumed continuous w.r.t. $\theta_0$, we get by the continuous mapping theorem

that

$$\left(\sqrt{n}(\hat{\bar{\mu}}_P - \bar{\mu}_P) - D_{\theta_0}(NoT)[n^{-1/2}\sum_{i=1}^{n}\Psi_{\theta_0}(Y_i)], \sqrt{n}\bar{\mu}^b_{\hat{P}} - D_{\theta_0}(NoT)[\sqrt{n}\theta^b_{0,n}]]\right) ,$$

converges weakly to

$$\left(G_P - D_{\theta_0}(NoT)[U_P], G_P^{'} - D_{\theta_0}(NoT)[U_P^{'}]\right) .$$

To get the final conclusion of Proposition 6.4.4, we have to prove two things.
First, under the assumption that $P = \mathcal{N}(\mathcal{T}(\theta_0))$ ($\mathbf{H}_0$ is true), considering the Fréchet derivative $D_{\theta_0}(NoT)$ of $NoT$ at $\theta_0$ yields the following Taylor approximation

$$\begin{aligned}
\sqrt{n}(\hat{\bar{\mu}}_P - NoT[\theta_n]) &= \sqrt{n}(\hat{\bar{\mu}}_P - \bar{\mu}_P) - \sqrt{n}(NoT[\theta_n] - NoT[\theta_0]) \\
&= \sqrt{n}(\hat{\bar{\mu}}_P - \bar{\mu}_P) - D_{\theta_0}(NoT)[\sqrt{n}(\theta_n - \theta_0)] + o_P(\sqrt{n}\|\theta_n - \theta_0\|) \\
&= \sqrt{n}(\hat{\bar{\mu}}_P - \bar{\mu}_P) - D_{\theta_0}(NoT)[n^{-1/2}\sum_{i=1}^{n}\Psi_{\theta_0}(Y_i)] \\
&\qquad + o_P(\sqrt{n}\|\theta_n - \theta_0\|) \\
&\xrightarrow[n\to+\infty]{\mathcal{L}} G_P - D_{\theta_0}(NoT)[U_P] ,
\end{aligned}$$

because $\sqrt{n}(\theta_n - \theta_0)$ converges weakly to a zero-mean Gaussian and by using the continuous mapping theorem with the continuity of $\theta_0 \mapsto D_{\theta_0}N$ and of $\|.\|_\Theta$.
Secondly, whether $\mathbf{H}_0$ is true or not,

$$\begin{aligned}
\sqrt{n}\hat{\bar{\mu}}^b_P - D_{\theta_n}(NoT)[\sqrt{n}\theta^b_n] = \sqrt{n}\hat{\bar{\mu}}^b_P - D_{\theta_0}(NoT)[\sqrt{n}\theta^b_{0,n}] + \underbrace{D_{\theta_n}(NoT)[\sqrt{n}(\theta^b_{0,n} - \theta^b_n)]}_{:=(a)} \\
+ \underbrace{D_{\theta_0}(NoT)[\sqrt{n}\theta^b_{0,n}] - D_{\theta_n}(NoT)[\sqrt{n}\theta^b_{0,n}]}_{:=(b)} .
\end{aligned}$$

we must check that both $(a)$ and $(b)$ converge $P$-almost surely to 0.
Since $\theta_n(\omega) \to \theta_0$ $P$-almost surely and $D_\theta(NoT) = (D_{T(\theta)}N)\ o\ (D_\theta\mathcal{T})$ is continuous w.r.t. $\theta$, then

$$(a) = \underbrace{\left(n^{-1/2}\sum_{i=1}^{n}(Z_i - \bar{Z})\right)}_{\longrightarrow\mathcal{N}(0,1)} D_{\theta_n}(NoT)[\underbrace{(m_0 - m_n, \Sigma_0 - \Sigma_n)}_{\to 0 \text{ a.s.}}] \xrightarrow[n\to+\infty]{P-a.s.} 0 ,$$

and

$$(b) = D_{\theta_0}(NoT)[\sqrt{n}\theta^b_{0,n}] - D_{\theta_n}(NoT)[\sqrt{n}\theta^b_{0,n}] \longrightarrow 0 \ \ P\text{-almost surely} ,$$

so that it follows

$$\sqrt{n}\bar{\mu}^b_{\hat{P}} - D_{\theta_n}(NoT)[\sqrt{n}\theta^b_n] \longrightarrow G'_P - D_{\theta_0}(NoT)[U'_P] \ ,$$

hence the conclusion of Proposition 6.4.4.

### 6.B.3 Proof of Theorem 6.5.1

The goal is to get an upper bound for the Type-II error

$$\mathbb{P}(n\hat{\Delta}^2 \leq \hat{q} \mid \mathbf{H}_1) \ . \tag{6.B.23}$$

In the following, the feature map from $H(k)$ to $H(\bar{k})$ will be denoted as

$$\bar{\phi} : H(k) \to H(\bar{k}), \ \ y \mapsto \bar{k}(y,.) \ .$$

Besides, we use the shortened notation $q := \mathbb{E}\hat{q}_{\alpha,n,B}$ for the sake of simplicity (see Section 6.5.1 for definitions).

1. **Reduce $n\hat{\Delta}^2$ to a sum of independent terms**

   The first step consists in getting a tight upper bound for (6.B.23) which involve a sum of independent terms. This will allow the use of a Bennett concentration inequality in the next step.
   Introducing the Fréchet derivative $D_{\theta_0}(NoT)$ of $NoT$ at $\theta_0$, $n\hat{\Delta}^2$ is expanded as follows

$$\begin{aligned}
n\hat{\Delta}^2 =& n\left\|\hat{\bar{\mu}}_P - (NoT)(\theta_n)\right\|^2 \\
=& n\left\|\hat{\bar{\mu}}_P - \bar{\mu}_P + (NoT)(\theta_0) - (NoT)(\theta_n) + \bar{\mu}_P - (NoT)(\theta_0)\right\|^2 \\
=& n\left\|\hat{\bar{\mu}}_P - \bar{\mu}_P + D_{\theta_0}(NoT)(\theta_0 - \theta_n) + o(\|\theta_0 - \theta_n\|) + \bar{\mu}_P - (NoT)(\theta_0)\right\|^2 \\
=& \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \bar{\phi}(Y_i) - \bar{\mu}_P - D_{\theta_0}(NoT)(\Psi(Y_i)) \right\} + \sqrt{n}\left(\bar{\mu}_P - (NoT)(\theta_0)\right) + o_n\left(\|\sqrt{n}(\theta_0 - \theta_n)\|\right) \right\|^2 \\
:=& n\hat{\Delta}_0^2 + n\Delta^2 + 2nS_n + \epsilon \ .
\end{aligned} \tag{6.B.24}$$

   where

$$\hat{\Delta}_0^2 = \frac{1}{n^2} \sum_{i,j=1}^{n} \langle \Xi(Y_i), \Xi(Y_j) \rangle \ ,$$

$$S_n = \frac{1}{n} \sum_{i=1}^{n} \langle \bar{\mu}_P - NoT(\theta_0), \Xi(Y_i) \rangle \ ,$$

$$\Xi(Y_i) = \bar{\phi}(Y_i) - \bar{\mu}_P - D_{\theta_0}(No\mathcal{T})(\Psi(Y_i)) \ ,$$

and $\epsilon = o_P(1)$ almost surely.

$n\hat{\Delta}_0^2$ is a degenerate U-statistics so it converges weakly to a sum of weighted chi-squares ([Ser80], page 194). $\sqrt{n}S_n$ converges weakly to a zero-mean Gaussian by the classic CLT as long as $\mathbb{E}\langle \bar{\mu}_P - No\mathcal{T}(\theta_0), \Xi(Y_i)\rangle^2$ is finite (which is true since $\bar{k}$ is bounded). It follows that $\hat{\Delta}_0^2$ becomes negligible with respect to $S_n$ when $n$ is large. Therefore, we consider a surrogate for the Type-II error (6.B.23) by removing $\hat{\Delta}_0^2$ with a negligible loss of accuracy.

Plugging (6.B.24) into (6.B.23)

$$\begin{aligned}\mathbb{P}(n\hat{\Delta}^2 \leq \hat{q}) =& \mathbb{P}(n\hat{\Delta}_0^2 + n\Delta^2 + 2nS_n \leq \hat{q} - \epsilon) \\ \leqslant& \mathbb{P}(n\hat{\Delta}_0^2 + n\Delta^2 + 2nS_n \leq \hat{q}) \ . \end{aligned} \tag{6.B.25}$$

Finally, using $\hat{\Delta}_0^2 \geq 0$ and conditioning on $\hat{q}$ yield the upper bound

$$\mathbb{P}(n\hat{\Delta}^2 \leq \hat{q} \mid \hat{q}) \ \leq \ \mathbb{P}\left(\sum_{i=1}^n f(Y_i) \geq n\hat{s}\right) \ , \tag{6.B.26}$$

where

$$f(Y_i) := \ \langle \bar{\mu}_P - No\mathcal{T}(\theta_0), \Xi(Y_i)\rangle \ , \quad \hat{s} := \ \Delta^2 - \frac{\hat{q}}{n} \ . \tag{6.B.27}$$

2. **Apply a concentration inequality**

We now want to find an upper bound for (6.B.26) through a concentration inequality, namely Lemma A.1.2 with $\xi_i = f(Y_i)$, $t = n\hat{s}$, $\nu^2 = \mathrm{Var}(f(Y_i))$ and $f(Y_i) \leq c = \bar{M}$ ($P$-almost surely).

Lemma A.1.2 combined with Lemma 6.B.3 and 6.B.2 yields the upper bound

$$\begin{aligned}\mathbb{P}(\sum_{i=1}^n f(Y_i) \geq n\hat{s} \mid \hat{q}) \leq& \exp\left(-\frac{n\hat{s}^2}{2\nu^2 + (2/3)\overline{M}\nu\hat{s}}\right)\mathbb{1}_{\hat{s}\geq 0} + \mathbb{1}_{\hat{s}<0} \\ \leq& \exp\left(-\frac{n\hat{s}^2}{2\vartheta^2 + (2/3)\overline{M}\vartheta\hat{s}}\right)\mathbb{1}_{\hat{s}\geq 0} + \mathbb{1}_{\hat{s}<0} \\ :=& \exp(g(\hat{s}))\mathbb{1}_{\hat{s}\geq 0} + \mathbb{1}_{\hat{s}<0} := h(\hat{s}) \ , \end{aligned} \tag{6.B.28}$$

where

$$\overline{M} := \left(4 + \sqrt{2}\right)\Delta M^{1/2} \ , \qquad \nu^2 \leq \vartheta^2 := \ \Delta^2 m_P^2 \ ,$$

and $m_P^2 = \mathbb{E}\|\Xi(Y)\|^2$.

3. **"Replace" the estimator $\hat{q}_{\alpha,n}$ with the true quantile $q_{\alpha,n}$ in the bound**

It remains to take the expectation with respect to $\hat{q}$. In order to make it easy, $\hat{q}$ is pull out of the exponential term of the bound. This is done through a Taylor-Lagrange expansion (Lemma 6.B.4).

Lemma 6.B.4 rewrites the bound in (6.B.28) as

$$\exp\left(-\frac{ns^2}{2\vartheta^2+(2/3)\overline{M}\vartheta s}\right)\left\{1+\frac{3n}{2\overline{M}\vartheta}\exp\left(\frac{3|\tilde{q}-q|}{2\overline{M}\vartheta}\right)\mathbb{1}_{\tilde{s}\geq 0}|\hat{s}-s|\right\} \ , \tag{6.B.29}$$

where

$$s=\Delta^2-\frac{q}{n} \ , \quad \tilde{s}=\Delta^2-\frac{\tilde{q}}{n} \ , \quad \tilde{q}\in(q\wedge\hat{q},q\vee\hat{q}) \ ,$$

and $s\geq 0$ because of the assumption $n>qL^{-2}$.

The mean (with respect to $\hat{q}$) of the right-side multiplicative term of (6.B.29) is bounded by

$$1+\frac{3n}{2\overline{M}\vartheta}\left\{\mathbb{E}_{\hat{q}}\left(\exp\left(\frac{3|\tilde{q}-q|}{\overline{M}\vartheta}\right)\mathbb{1}_{\tilde{s}\geq 0}\right)\right\}^{1/2}\sqrt{\mathbb{E}_{\hat{q}}(\hat{s}-s)^2} \ ,$$

because of the Cauchy-Schwarz inequality.
On one hand, $\mathbb{E}(\hat{q}-q)^2 \underset{B\to+\infty}{\to} 0$ (Lemma 6.B.1) implies $\mathbb{E}(\tilde{q}-q)^2 \underset{B,n\to+\infty}{\to} 0$ and thus $\tilde{q} \underset{B,n\to+\infty}{\to} q_\infty$ weakly where $q_\infty=\lim_{n\to+\infty} q_{\alpha,n}$ (that is almost surely for $q_\infty$ is a constant). Hence

$$\begin{aligned}\mathbb{E}_{\hat{q}}\left(\exp\left(\frac{3|\tilde{q}-q|}{\overline{M}\vartheta}\right)\mathbb{1}_{\tilde{s}\geq 0}\right) &= \mathbb{E}_{\hat{q}}\left(\left[1+\frac{o_B(|\hat{q}-q|)}{\overline{M}\vartheta}\right]\mathbb{1}_{\tilde{s}\geq 0}\right)\\ &\leq 1+\frac{\mathbb{E}_{\hat{q}}(o_B(|\hat{q}-q|)\mathbb{1}_{\tilde{s}\geq 0})}{\overline{M}\vartheta} \ .\end{aligned}$$

Since the variable $|\tilde{q}-q|\mathbb{1}_{\tilde{s}\geq 0}$ is bounded by the constant $|nL^2-q|\vee|q|$ for every $B$, it follows from the Dominated Convergence Theorem

$$1+\frac{\mathbb{E}_{\hat{q}}(o_B(|\hat{q}-q|)\mathbb{1}_{\tilde{s}\geq 0})}{\overline{M}\vartheta} = 1+\frac{o_B(1)}{\overline{M}\vartheta} \tag{6.B.30}$$

On the other hand, Lemma 6.B.1 provides

$$\mathbb{E}(\hat{s}-s)^2=\frac{\mathbb{E}(\hat{q}-q)^2}{n^2}\leq\frac{C_{1,P}+\alpha C_{2,P}/B}{n^2\alpha B}\leq\frac{C_{P^b}}{n^2\alpha B} \ . \tag{6.B.31}$$

so that an upper bound for the Type-II error is given by

$$\left(\frac{3}{2}+o_n(1)\right)\exp\left(-\frac{ns^2}{2\vartheta^2+(2/3)\overline{M}\vartheta s}\right)\left\{1+\frac{3C_{P^b}}{2\overline{M}\vartheta\sqrt{\alpha B}}+\frac{o_B(B^{-1/2})}{\overline{M}^2\vartheta^2}\right\} \ . \tag{6.B.32}$$

which one rewrites as

$$\exp\left(-\frac{n\left(\Delta-\frac{q}{n\Delta}\right)^2}{2m_P^2+Cm_PM^{1/2}(\Delta^2-q/n)}\right)f_1(B,M,\Delta)\ ,$$

where

$$f_1(B,M,\Delta)=(3/2+o_n(1))\left(1+\frac{C_{P^b}}{C'\Delta^2M^{1/2}m_P\sqrt{\alpha B}}+\frac{o_B(B^{-1/2})}{C''\Delta^4Mm_P^2}\right)\ ,$$

and $C=(2/3)(4+\sqrt{2})$, $C'=2(4+\sqrt{2})/3$ and $C''=(4+\sqrt{2})^2$.

Theorem 6.5.1 is proved.

### 6.B.4 Auxiliary results

**Lemma 6.B.1.** *(Theorem 2.9. in [BT12]) Assume $\alpha<1/2$. Then,*

$$\mathrm{Var}(\hat{q}_{\alpha,n})\leq\frac{C_{P^b}}{\alpha B}\ , \tag{6.B.33}$$

*where $C_{P^b}$ only depends on the bootstrap distribution (of $[n\hat{\Delta}^2]^b_{fast}$).*

**Lemma 6.B.2.** *Let $f$ be defined as in* (6.B.27)*. If $Y\in\mathcal{H}_0\subseteq\mathcal{H}$ $P$-almost surely and $\sup_{x,y\in\mathcal{H}_0}|\bar{k}(x,y)|=M$, then for any $y\in\mathcal{H}_0$*

$$|f(y)|\leq\overline{M}:=(4+\sqrt{2})\Delta\sqrt{M}\ . \tag{6.B.34}$$

*Proof.*

$$\begin{aligned}
|f(y)|=&\ \left|\langle\bar{\mu}_P-NoT(\theta_0),D_{\theta_0}(NoT)(\Psi(y))-\bar{\phi}(y)+\bar{\mu}_P\rangle\right|\\
\leq&\Delta\left\|D_{\theta_0}(NoT)(\Psi(y))-\bar{\phi}(y)+\bar{\mu}_P\right\|\\
\leq&\ \Delta\left(\left[\lim_{t\to0}\mathbb{E}_{Z,Z'\sim NoT(\theta_0+t\Psi(y))}\bar{k}(Z,Z)+\bar{k}(Z',Z')-2\bar{k}(Z,Z')\right]^{1/2}\right.\\
&\left.+\sqrt{\bar{k}(y,y)}+\sqrt{\mathbb{E}\bar{k}(Y,Y')}\right)\\
\leq&\ \Delta\left(\sqrt{M}+\sqrt{M}+\sqrt{2M}+\sqrt{M}+\sqrt{M}\right)\\
\leq&\ \Delta(4+\sqrt{2})\sqrt{M}:=\overline{M}\ .
\end{aligned}$$

□

**Lemma 6.B.3.**

$$v^2 \leq \vartheta^2 := \Delta^2 m_P^{(2)} \ . \tag{6.B.35}$$

*Proof.*

$$\begin{aligned} v^2 := \mathrm{Var}(f(Y)) = \ & \mathbb{E} f^2(Y_i) \\ =& \mathbb{E}\langle \bar{\mu}_P - NoT(\theta_0), D_{\theta_0}(NoT)(\Psi(Y)) - \bar{\phi}(Y) + \bar{\mu}_P \rangle^2 \\ \leq \ & \Delta^2 \underbrace{\mathbb{E}\left\| D_{\theta_0}(NoT)(\Psi(Y)) - \bar{\phi}(Y) + \bar{\mu}_P \right\|^2}_{:= m_P^2} \ . \end{aligned}$$

□

**Lemma 6.B.4.** *Let $h$ be defined as in* (6.B.28)*. Then,*

$$h(\hat{s}) \leq \exp\left( -\frac{ns^2}{2\vartheta^2 + (2/3)\overline{M}\vartheta s} \right) \left\{ 1 + \frac{3n}{2\overline{M}\vartheta} \exp\left( \frac{3|\tilde{q} - q|}{2\overline{M}\vartheta} \right) \mathbb{1}_{\tilde{s} \geq 0} |\hat{s} - s| \right\} \ , \tag{6.B.36}$$

*where*

$$s = \Delta^2 - \frac{q}{n} \ , \quad \tilde{s} = \Delta^2 - \frac{\tilde{q}}{n} \ , \quad \tilde{q} \in (q \wedge \hat{q}, q \vee \hat{q}) \ .$$

*Proof.* A Taylor-Lagrange expansion of order 1 can be derived for $h(\hat{s})$ since the derivative of $h$

$$h'(x) = -\frac{(2/3)n\overline{M}\vartheta x^2 + 4n\vartheta^2 x}{\left(2\vartheta^2 + (2/3)\overline{M}\vartheta x\right)^2} \exp\left( -\frac{nx^2}{2\vartheta^2 + (2/3)\overline{M}\vartheta x} \right) \mathbb{1}_{x \geq 0} \ ,$$

is well defined for every $x \in \mathbb{R}$ (in particular, the left-side and right-ride derivatives at $x = 0$ coincide).
Therefore $h(\hat{s})$ equals

$$\begin{aligned} & h(s) + h'(\tilde{s})(\hat{s} - s) \\ & \quad = \exp\left( -\frac{ns^2}{2\vartheta^2 + (2/3)\overline{M}\vartheta s} \right) \left[ 1 + \exp\left( g(s) - g(\tilde{s}) \right) g'(\tilde{s}) \mathbb{1}_{\tilde{s} \geq 0} (\hat{s} - s) \right] \ , \end{aligned} \tag{6.B.37}$$

where

$$s = \Delta^2 - \frac{q}{n} \ , \quad \tilde{s} = \Delta^2 - \frac{\tilde{q}}{n} \ , \quad \tilde{q} \in (q \wedge \hat{q}, q \vee \hat{q}) \ ,$$

and $s \geq 0$ because of the assumption $n > q\Delta^{-2}$.

For every $x, y > 0$, $|g'(x)| \leq 3n/(2\overline{M}\vartheta)$ and then $|g(x) - g(y)| \leq 3n|x-y|/(2\overline{M}\vartheta)$. It follows

$$|g'(\tilde{s})| \leq \frac{3n}{2\overline{M}\vartheta} \ ,$$

and

$$\exp\left(g(s) - g(\tilde{s})\right) \leq \exp\left(\frac{3n}{2\overline{M}\vartheta}|s-\tilde{s}|\right) = \exp\left(\frac{3|\tilde{q}-q|}{2\overline{M}\vartheta}\right) \ .$$

Lemma 6.B.4 is proved. □

Chapter 7

# General Conclusion and Perspectives

## 7.1 Nearly Gaussian marginals in an RKHS

In Chapter 4, we have evidenced that most $p$-dimensional projections of an embedded distribution in the RKHS of a Gaussian RBF kernel are close to some scale-mixture of Gaussians and derived a practical application to outlier detection from this observation in Chapter 5. Besides we also showed that with a proper renormalization of such a kernel, those projections become close to a $\mathcal{N}(0, I_p)$ Gaussian distribution instead. However a straightforward application of this result is yet to be devised. In the following, we consider a few practical issues for which said result could be useful:

- **Dimension reduction.** In the multivariate setting, Non-Gaussian Component Analysis or NGCA [Bla+06; Die+10; DJS13] is a method of dimension reduction that relies on the assumption that a $d$-variate random vector $X$ (with large $d$) can be decomposed as $X = X_0 + X_\perp$ where $X_0$ lies on a $p$-dimensional subspace $V$ of $\mathbb{R}^d$ and the noisy term $X_\perp$ is Gaussian. Hence the goal of NGCA is to reconstruct $V$. However as mentioned in Chapter 4, existing results in $\mathbb{R}^d$ state that most marginals of such $X$ when $d$ is large are close to a scale-mixture of Gaussians instead of a Gaussian, which is not compatible with the assumptions of NGCA. That being said, embedding $X$ into the RKHS of a renormalized Gaussian RBF kernel would yield nearly Gaussian marginals in most directions in the kernel space. Therefore it would be of interest to devise a kernel extension of NGCA to find the few remaining directions with non-Gaussian marginals. Such an extension is not straightforward since multivariate NGCA relies on a rewriting of the assumed decomposition of $X$ in terms of its density, but in an infinte-dimensional RKHS there exists no canonical Haar measure (*i.e.* a shift-invariant measure) in the RKHS to define a density as in $\mathbb{R}^d$. NGCA could be adapted to the kernel case by considering for instance moment-generating functions instead of densities.

- **Sampling from a known distribution.** Given a density $f$ of a distribution in $\mathbb{R}^d$, sampling a sequence $x_1, \dots, x_n \in \mathbb{R}^d$ that emulates an *i.i.d.* sample from $f$ is usually done through Monte Carlo Markov Chain methods [LLC10]. This class of methods generates a Markov chain $x_1, x_2, \dots$ where the distribution of the $x_i$ converges weakly to an equilibrium distribution of density $f$. Albeit popular, these methods suffer some issues, typically a slow convergence of the Markov chain to the desired distribution, either because of a strong dependence between the successive samples or because of small acceptation rates. In such cases, a long Markov chain must be generated in practice, which leads to an intensive computational burden. Our result could be the basis of an alternative take on this sampling problem without using Markov chains. Indeed we could sample an *i.i.d.* sample from a $\mathcal{N}(0, I_p)$ Gaussian directly in a subspace of the RKHS and revert to the input space to find the corresponding sample that would look like an *i.i.d.* sample from $f$ [1]. This "revert" step would be the challenging part of such a sampling procedure since a given point in an RKHS does not admit an exact pre-image in the input space. However there exist several approaches to solve this problem [SSM97; BWS04].

## 7.2 Outlier detection

In Chapter 5, we introduced a new kernel-based method based on random projections in an RKHS (induced by a Gaussian RBF kernel) to perform outlier detection in an online setting. Our proposed method has the advantage of bypassing shortcomings of one-class SVM and kernel PCA — that are the lack of false alarm rate control for one-class SVM and a high computational cost for kernel PCA. Furthermore, we have provided theoretical guarantees about performances of our algorithm — in particular for missed detection rate — whilst such guarantees are lacking for kernel PCA to the best of our knowledge.

The next step would be to transpose our OD method to the batch framework, where inliers and outliers are all provided at once in one sample. The problem to solve is that the random generation of the projection subspace $V_n$ in the RKHS depends on the empirical covariance function $\Sigma_{\gamma,n}(\cdot,\cdot)$ which estimates the covariance of the embedded inliers. However in a batch setting, this covariance would be estimated on the basis of inliers *and* outliers all together. This problem could be tackled by resorting to robust covariance estimators. In this case one should check that asymptotic results provided in Chapter 4 still holds with such estimators.

## 7.3 Normality test in Hilbert spaces

In Chapter 6, we introduced a new normality test suited to high-dimensional Hilbert spaces, that was based on an MMD statistic constructed from a characteristic kernel $\bar{k}$. Investigating the

[1] The density $f$ would appear in the computations through the equivalent term given in Lemma A.2.1 combined with Proposition 4.3.2

influence of the kernel $\bar{k}$ on the performance of the test would be of interest. In the case where $\bar{k}$ is a Gaussian RBF kernel for instance, a method to optimize the Type-II error with respect to the hyperparameter of $\bar{k}$ would be welcomed. It turns out to be a challenging task. In the literature about hypothesis testing with the MMD, one of the few papers addressing this issue is [Gre+12b] which optimizes a homogeneity test over convex combinations of a basis of kernels, but at the cost of modifying the MMD statistic to resort to a less powerful test.

We applied successfully a fast parametric bootstrap procedure that reduces the computational costs of our test from cubic to quadratic time with respect to the sample size $n$. However a quadratic execution time may still be too high when $n$ becomes very large. For two-sample MMD tests, a linear time statistic can be devised to speed up the test (see Section 6 in [Gre+12a]). However there are two issues: firstly the corresponding test is typically less powerful, and secondly it cannot be straightforwardly applied to the one-sample setting because of the estimation of the parameters (mean and covariance).

Finally, the choice of the level $\alpha$ for the sequential procedure (covariance rank selection) is another subject for future research. Indeed, an asymptotic regime for $\alpha$ has been exhibited to get consistency, but setting the value of $\alpha$ when $n$ is fixed remains an open question.

# Bibliography


[Aro50] N. Aronszajn. "Theory of Reproducing Kernels". In: *Transactions of the American Mathematical Society* 68 (May 1950), pp. 337–404.

[Bac13] Francis R. Bach. "Sharp analysis of low-rank kernel matrix approximations." In: *COLT*. Vol. 30. 2013, pp. 185–209.

[Bas08] J. Basak. "A least square kernel machine with box constraints". In: *19th International Conference on Pattern Recognition, 2008. ICPR 2008*. Dec. 2008, pp. 1–4. DOI: 10.1109/ICPR.2008.4761717.

[BFG15] C. Bouveyron, M. Fauvel, and S. Girard. "Kernel discriminant analysis and clustering with parsimonious Gaussian process models". In: *Statistics & Computing* 25 (2015), pp. 1143–1162.

[BGV92] B. E. Boser, I. M. Guyon, and V. N. Vapnik. "A Training Algorithm of Optimal Margin Classifiers". In: *COLT '92 Proceedings of the fifth annual workshop on Computational learning theory* (1992), pp. 144–152.

[BH95] Y. Benjamini and Y. Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the royal statistical society. Series B (Methodological)* (1995), pp. 289–300.

[BJ05] F. R. Bach and M. I. Jordan. "Predictive low-rank decomposition for kernel methods". In: *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 33–40.

[BK03] S. G. Bobkov and A. Koldobsky. "On the central limit property of convex bodies". In: *Geometric aspects of functional analysis*. Springer, 2003, pp. 44–52.

[Bla+06] G. Blanchard et al. "Non-Gaussian Component analysis: a Semi-parametric Framework for Linear Dimension Reduction". In: *Advances in Neural Information Processing Systems* (2006).

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, USA, 2013.

[BMR13] E. Brunel, A. Mas, and A. Roche. "Non-asymptotic Adaptive Prediction in Functional Linear Models". In: *arXiv preprint arXiv:1301.3017* (2013).

[Bor+06] K. M. Borgwardt et al. "Integrating structured biological data by kernel maximum mean discrepancy". In: *Bioinformatics* 22.14 (2006), e49–e57.

[BT04] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US, 2004.

[BT11] J. Bien and R. J. Tibshirani. "Sparse estimation of a covariance matrix". In: *Biometrika* 98.4 (2011), pp. 807–820.

[BT12] S. Boucheron and M. Thomas. "Concentration Inequalities for Order Statistics". In: *Electronic Communications in Probability* 17 (2012).

[Bur00] M. D. Burke. "Multivariate tests-of-fit and uniform confidence bands using a weighted bootstrap". In: *Statistics & Probability Letters* 46.1 (Jan. 2000), pp. 13–20. ISSN: 0167-7152. DOI: 10.1016/S0167-7152(99)00082-6.

[BWS04] Gokhan H. Bakir, Jason Weston, and Bernhard Schölkopf. "Learning to find pre-images". In: *Advances in neural information processing systems* 16.7 (2004), pp. 449–456.

[C+10] T. Cai, C.-H. Zhang, H. Zhou, et al. "Optimal rates of convergence for covariance matrix estimation". In: *The Annals of Statistics* 38.4 (2010), pp. 2118–2144.

[CD01] M. Collins and N. Duffy. "Convolution Kernels for Natural Language". In: *Advances in Neural Information Processing Systems* (2001).

[CJ10] H. Cardot and J. Johannes. "Thresholding projection estimators in functional linear models". In: *Journal of Multivariate Analysis* 101.2 (2010), pp. 395–408.

[CS10] A. Christmann and I. Steinwart. "Universal kernels on non-standard input spaces". In: *Advances in neural information processing systems*. 2010, pp. 406–414.

[CTT14] Y. Choi, J. Taylor, and R. Tibshirani. "Selecting the number of principal components: estimation of the true rank of a noisy matrix". In: *arXiv preprint arXiv:1410.8260* (2014).

[Cue+06] J.A. Cuesta-Albertos et al. "The Random Projection Method in Goodness of Fit for Functional Data". In: *Computational Statistics and Data Analysis* 51 (June 2006), pp. 4864–4877.

[CV95] C. Cortes and V. Vapnik. "Support-vector networks". In: *Machine learning* 20.3 (1995), pp. 273–297.

[DF84] P. Diaconis and D. Freedman. "Asymptotics of graphical projection pursuit". In: *The annals of statistics* (1984), pp. 793–815.

[DHV06] S. Dasgupta, D. Hsu, and N. Verma. "A concentration theorem for projections". In: *Conference on Uncertainty in Artificial Intelligence (UAI)*. 2006.

[Die+10] E. Diederichs et al. "Sparse non-Gaussian component analysis". In: *Information Theory, IEEE Transactions on* 56.6 (2010), pp. 3033–3047.

[DJS13] E. Diederichs, A. Juditsky A.and Nemirovski, and V. Spokoiny. "Sparse non Gaussian component analysis by semidefinite programming". In: *Machine learning* 91.2 (2013), pp. 211–238.

[DM05] P. Drineas and M. W. Mahoney. "On the Nyström method for approximating a Gram matrix for improved kernel-based learning". In: *Journal of Machine Learning Research* 6.Dec (2005), pp. 2153–2175.

[Dri73] M. F. Driscoll. "The Reproducing Kernel Hilbert Space Structure of the Sample Paths of a Gaussian Process". In: *Probability Theory and Related Fields* 26 (1973), pp. 309–316.

[DRT14] E. De Vito, L. Rosasco, and A. Toigo. "Learning Sets with Separating Kernels". In: (2014).

[Fis36] R. A. Fisher. "The use of multiple measurements in taxonomic problems". In: *Annals of eugenics* 7.2 (1936), pp. 179–188.

[FSG08] B. A. Frigyik, S. Srivastava, and M. R. Gupta. "An introduction to functional derivatives". In: *Dept. Electr. Eng., Univ. Washington, Seattle, WA, Tech. Rep* 1 (2008).

[Fuk+09] K. Fukumizu et al. "Characteristic Kernels on Groups and Semigroups". In: *Advances in Neural Information Processing Systems* 21 (2009).

[GD03] A. Gretton and F. Desobry. "On-line one-class support vector machines. an application to signal segmentation". In: *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. Vol. 2. IEEE, 2003, pp. II–709.

[Gre+07a] A. Gretton et al. "A Kernel Method for the Two-Sample-Problem". In: *Advances in Neural Information Processing Systems*. Ed. by B. Schoelkopf, J. Platt, and T. Hoffman. Vol. 19. MIT Press, Cambridge. 2007, pp. 513–520.

[Gre+07b] A. Gretton et al. "A Kernel Statistical Test of Independence". In: *Advances in Neural Information Processing Systems* 21 (2007).

[Gre+09] A. Gretton et al. "A Fast, Consistent Kernel Two-Sample Test". In: *Advances in Neural Information Processing Systems* 22 (2009).

[Gre+12a] A. Gretton et al. "A Kernel Two-Sample Test". In: *Journal of Machine Learning Research* (Mar. 2012).

[Gre+12b] A. Gretton et al. "Optimal kernel choice for large-scale two-sample tests". In: *Advances in Neural Information Processing Systems*. 2012, pp. 1205–1213.

[Gre15] A. Gretton. "A simpler condition for consistency of a kernel independence test". In: *arXiv:1501.06103 [stat]* (Jan. 2015). arXiv: 1501.06103.

[HJ80] F. Hernandez and R. A. Johnson. "The large-sample behavior of transformations to normality". In: *Journal of the American Statistical Association* 75.372 (1980), pp. 855–861.

[Hof07] H. Hoffmann. "Kernel PCA for novelty detection". In: *Pattern Recognition* 40.3 (2007), pp. 863–874.

[HP76] J. Hoffmann-Jorgensen and G. Pisier. "The Law of Large Numbers and the Central Limit Theorem in Banach Spaces". In: *The Annals of Probability* 4 (1976), pp. 587–599.

[HZ90] N. Henze and B. Zirkler. "A Class of Invariant and Consistent Tests for Multivariate Normality". In: *Comm. Statist. Theory Methods* 19 (1990), pp. 3595–3617.

[JDH99] T. Jaakkola, M. Diekhans, and D. Haussler. "Using the Fisher kernel method to detect remote protein homologies". In: *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology* (1999), pp. 149–158.

[JH12] J. Josse and F. Husson. "Selecting the number of components in principal component analysis using cross-validation approximations". In: *Computational Statistics & Data Analysis* 56.6 (2012), pp. 1869–1879.

[Joa02] T. Joachims. *Learning to Classify Text Using Support Vector Machines*. Springer US, 2002.

[JR92] M. C. Jones and J. A. Rice. "Displaying the Important Features of Large Collections of Similar Curves". In: *The American Statistician* 46 (1992), pp. 140–145.

[Kos07] Michael R. Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media, 2007.

[KW71] G. Kimbeldorf and G. Wahba. "Some Results on Tchebycheffian Spline Functions". In: *Journal of Mathematical Analysis and Applications* 33 (1971). DOI: 10.1016/0022-247X(71)90184-3.

[KY12] I. Kojadinovic and J. Yan. "Goodness-of-fit Testing Based On A Weighted Bootstrap: A Fast Large-Sample Alternative to the Parametric Bootstrap". In: *Canadian Journal of Statistics* 40 (2012), pp. 480–500.

[LB01] M. N. Lukic and J. H. Beder. "Stochastic Processes wih Sample Paths in Reproducing Kenel Hilbert Spaces". In: *Transactions of the American Mathematical Society* 353 (May 2001).

[LF00] P. L. Lai and C. Fyfe. "Kernel and nonlinear canonical correlation analysis". In: *International Journal of Neural Systems* 10.05 (2000), pp. 365–377.

[LLC10] F. Liang, C. Liu, and R. Carroll. *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*. Wiley, 2010.

[LLW07] H. Liu, J. D. Lafferty, and L. Wasserman. "Saprse Nonparametric Density estimation in High Dimensions Using the Rodeo". In: *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS-07)* (2007), pp. 283–290.

[LR05] E.L. Lehmann and J. P. Romano. *Testing Statistical hypotheses*. Ed. by George Casella Stephen Fienberg Ingram Olkin. Springer, 2005.

[LS01] Wenbo V. Li and Q.-M. Shao. "Gaussian processes: inequalities, small ball probabilities and applications". In: *Handbook of Statistics* 19 (2001), pp. 533–597.

[LSS13] Q. Le, T. Sarlós, and A. Smola. "Fastfood-approximating kernel expansions in loglinear time". In: *Proceedings of the international conference on machine learning*. 2013.

[LV12] E. Levina and R. Vershynin. "Partial estimation of covariance matrices". In: *Probability Theory and Related Fields* 153.3-4 (2012), pp. 405–419.

[LW04] O. Ledoit and M. Wolf. "A well-conditioned estimator for large-dimensional covariance matrices". In: *Journal of multivariate analysis* 88.2 (2004), pp. 365–411.

[Mam92] E. Mammen. *When Does Bootstrap Work?* Springer-Verlag New York, 1992.

[Mar70] K.V. Mardia. "Measures of Multivariate Skewness and Kurtosis with Applications". In: *Biometrika* 57 (1970), pp. 519–530.

[Mer09] J. Mercer. "Functions of Positive and Negative Type, and their Connections with the Theory of Integral Equations". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 209 (1909), pp. 415–446. DOI: 10.1098/rsta.1909.0016.

[Mik+99] S. Mika et al. "Fisher Discriminant Analysis with Kernels". In: *Neural Networks for Signal Processing* 9 (1999), pp. 41–48.

[MM08] P. D. McNicholas and T. B. Murphy. "Parsimonious Gaussian mixture models". In: *Statistics and Computing* 18.3 (2008), pp. 285–296.

[Mog02] B. Moghaddam. "Principal Manifolds and Probabilistic Subspaces for Visual Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24 (2002), pp. 780–788.

[Pim+14] M. A. F. Pimentel et al. "A review of novelty detection". In: *Signal Processing* 99 (2014), pp. 215–249.

[Rat03] Z. Ratsimalahelo. "Strongly consistent determination of the rank of matrix". In: *Econometrics* (2003).

[Rot06] V. Roth. "Kernel Fisher Discriminants for Outlier Detection". In: *Neural Computation* 18.4 (2006), pp. 942–960. DOI: 10.1162/neco.2006.18.4.942.

[Row98] S. Roweis. "EM Algorithms for PCA and SPCA". In: *in Advances in Neural Information Processing Systems*. MIT Press, 1998, pp. 626–632.

[RR07] A. Rahimi and B. Recht. "Random features for large-scale kernel machines". In: *Advances in neural information processing systems*. 2007, pp. 1177–1184.

[RS00] J.-M. Robin and R. J. Smith. "Tests of Rank". In: *Econometric Theory* 16 (2000), pp. 151–175.

[RS80] M. Reed and B. Simon. *Methods of Modern Mathematical Physics I: Functional Analysis*. Academic Press, Inc., 1980.

[RW06] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[SC08] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer-Verlag New York, 2008.

[Sch+01] B. Schölkopf et al. "Estimating the support of a high-dimensional distribution". In: *Neural computation* 13.7 (2001), pp. 1443–1471.

[Sej+13] D. Sejdinovic et al. "Equivalence of Distance-Based and RKHS-based Statistics in Hypothesis Testing". In: *The Annals Of Statistics* 41.5 (2013), pp. 2263–2291.

[Ser80] R. J. Serfling. *Approximation Theorems for Mathematical Statistics*. John Wiley & Sons, 1980.

[SFL11] B. K. Sriperumbudur, K. Fukumizu, and G. R. G. Lanckriet. "Universality, characteristic kernels and RKHS embedding of measures". In: *The Journal of Machine Learning Research* 12 (2011), pp. 2389–2410.

[SGV98] Craig Saunders, Alexander Gammerman, and Volodya Vovk. "Ridge regression learning algorithm in dual variables". In: *(ICML-1998) Proceedings of the 15th International Conference on Machine Learning*. Morgan Kaufmann, 1998, pp. 515–521.

[SHS01] Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. "A generalized representer theorem". In: *Computational learning theory*. Springer, 2001, pp. 416–426.

[SHS09] I. Steinwart, D. Hush, and C. Scovel. "Density Level Detection is Classification". In: *Advances in Neural Information Processing Systems* (2009).

[SKK13] M.S. Srivastava, S. Katayama, and Y. Kano. "A Two-Sample Test in High Dimensional Data". In: *Journal of Multivariate Analysis* (2013), pp. 349–358.

[SMQ93] W. Stute, W. G. Manteiga, and M. P. Quindimil. "Bootstrap based goodness-of-fit-tests". In: *Metrika* 40.1 (1993), pp. 243–256.

[SR05] G.J. Szekely and R.L. Rizzo. "A New Test For Multivariate Normality". In: *Journal of Multivariate Analysis* 93 (2005), pp. 58–80.

[Sri+10] B. K. Sriperumbudur et al. "Hilbert Space Embeddings and Metrics on Probability Measures". In: *Journal of Machine Learning Research* (2010), pp. 1517–1561.

[SS00] A. J. Smola and B. Schölkopf. "Sparse greedy matrix approximation for machine learning". In: *ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning* (2000), pp. 911–918.

[SS12] I. Steinwart and C. Scovel. "Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHS". In: *Constructive Approximation* 35.3 (2012), pp. 363–417.

[SSM97] B. Schölkopf, A. Smola, and K.-R. Müller. "Nonlinear Component Analysis as a Kernel Eigenvalue Problem". In: *Neural Computation* 10 (1997), pp. 1299–1319.

[SW03] T. Svantesson and J. W. Wallace. "Tests for assessing multivariate normality and the covariance structure of MIMO data". In: *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. Vol. 4. IEEE, 2003, pp. IV–656.

[Tan+11] A. Tanaka et al. "Theoretical Analyses on a Class of Nested RKHS's". In: *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2072–2075.

[TD99] D. M. J. Tax and R. P. W Duin. "Support vector domain description". In: *Pattern recognition letters* 20.11 (1999), pp. 1191–1199.

[V+97] V. Vapnik, S. E. Golowich, A. Smola, et al. "Support vector method for function approximation, regression estimation, and signal processing". In: *Advances in neural information processing systems* (1997), pp. 281–287.

[Vad+11] K. S. Vadivel et al. "Generalized Subspace Based High Dimensional Density Estimation". In: *International Conference on Image Processing* (2011).

[Vap00] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.

[VV06] R. Vert and J.-P. Vert. "Consistency and Convergence Rates of One-Class SVMs and Related Algorithms". In: *Journal of Machine Learning Research* 7 (2006), pp. 817–854.

[Wei97] H. von Weizsäcker. "Sudakov's typical marginals, random linear functionals and a conditional central limit theorem". In: *Probability theory and related fields* 107.3 (1997), pp. 313–324.

[WS01] C. Williams and M. Seeger. "Using the Nyström method to speed up kernel machines". In: *Proceedings of the 14th annual conference on neural information processing systems*. 2001, pp. 682–688.

[XWX14] Y. Xiao, H. Wang, and W. Xu. "Model selection of Gaussian kernel PCA for novelty detection". en. In: *Chemometrics and Intelligent Laboratory Systems* 136 (Aug. 2014), pp. 164–172. ISSN: 01697439. DOI: 10.1016/j.chemolab.2014.05.015.

[YJ00] I.-K. Yeo and R. A. Johnson. "A new family of power transformations to improve normality or symmetry". In: *Biometrika* 87.4 (2000), pp. 954–959.

[Zwa05] L. Zwald. "Performances d'Algorithmes Statistiques d'Apprentissage: "Kernel Projection Machine" et Analyse en Composantes Principales à Noyaux". PhD thesis. 2005.

Appendix A

# Technical lemmas

## A.1 Some useful concentration inequalities

**Lemma A.1.1** (McDiarmid's inequality)**.** *Let $X_1,\dots,X_p$ be* i.i.d. *random variables taking values in some set $\mathcal{X}$. Let $L : \mathcal{X}^p \to \mathbb{R}$ be some real-valued function and assume there exists there exists $c_1,\dots,c_p > 0$ such that*

$$\forall j = 1,\dots,p, \quad \sup_{x_1,\dots,x_j,x_j',\dots,x_p\in\mathcal{X}} |L(x_1,\dots,x_j,\dots) - L(x_1,\dots,x_j',\dots)| \leqslant c_j \ .$$

*Then for any $t > 0$*

$$\mathbb{P}\Big(L(X_1,\dots,X_p) - \mathbb{E}L(X_1,\dots,X_p) > t\Big) \leqslant \exp\left(-\frac{2t^2}{\sum_{j=1}^p c_i^2}\right) \ .$$

**Lemma A.1.2.** *(Bennett's inequality, Theorem 2.9 in [BLM13]) Let $\xi_1,\dots,\xi_n$ i.i.d. variables bounded by $c$ and with variance bounded by $\nu$.*
*Then, for any $\epsilon > 0$*

$$\mathbb{P}\left(\sum_{i=1}^n \xi_i - n\mathbb{E}\xi_1 \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2n\nu + 2c\epsilon/3}\right) \ . \tag{A.1.1}$$

*In other words, with probability larger than $1-\delta$,*

$$\sum_{i=1}^n \xi_n \leqslant \mathbb{E}\xi_1 + \sqrt{\frac{2\nu\log(1/\delta)}{n}}\left[1 + \kappa c\sqrt{\frac{\log(1/\delta)}{n\nu}}\right] \ ,$$

*where $\kappa$ is an absolute constant.*

## A.2 Auxilary lemmas used in Chapter 4 and Chapter 5

**Lemma A.2.1.** *Let $X \sim P$ and assume there exists a function $A : \mathcal{X} \to \mathbb{R}$ and $s > 0$ such that $\mathbb{P}(d^2(X,x) \leqslant t) = A(x)t^s[1 + o(1)]$ as $t \to 0$ for any $x \in \mathrm{supp}(P)$ and $A(x) = 0$ for any $x \notin \mathrm{supp}(P)$. Also assume that $\mathrm{supp}(P)$ is a compact subset of $\mathcal{X}$ and $(t,x) \mapsto \mathbb{P}(d^2(x,X) \leqslant t)$ is continuous. Then for any $x \in \mathrm{supp}(P)$,*

$$\Sigma_\gamma(x,x) = \frac{A(x)\Gamma(s+1)}{(2\gamma)^s}\left\{1 + A^{-1}(x)\, o_\gamma(1)\right\} \quad a.s. \quad ,$$

*when $\gamma \to +\infty$, where the '$o_\gamma(1)$' term is uniformly bounded with respect to $x$.*

*Proof.* By definition

$$\Sigma_\gamma(x,x) = \mathbb{E}_X k_\gamma^2(X,x) = \mathbb{E}e^{-2\gamma d^2(x,X)} \quad .$$

Fubini's Theorem entails

$$\begin{aligned}
\mathbb{E}_X e^{-2\gamma d^2(x,X)} &= \mathbb{E}_X\left[-e^{-y}\right]_{y=2\gamma d^2(x,X)}^{y=+\infty} \\
&= \mathbb{E}_X \int_{2\gamma d^2(x,X)}^{+\infty} e^{-y}\mathrm{d}y \\
&= \int_0^{+\infty} e^{-y}\mathbb{E}_X \mathbb{1}_{\{2\gamma d^2(x,X) \leqslant y\}}\mathrm{d}y \\
&= \int_0^{+\infty} e^{-y}\mathbb{P}\left(d^2(x,X) \leqslant \frac{y}{2\gamma}\right)\mathrm{d}y
\end{aligned}$$

Introduce the function $g(t;x)$ satisfying $\mathbb{P}(d^2(x,X) \leqslant t) = A(x)(1 + g(t;x))t^s$ and $g(t;x) \to 0$ when $t \to 0$.

$$\begin{aligned}
\mathbb{E}_X e^{-2\gamma d^2(x,X)} &= \frac{A(x)}{(2\gamma)^s}\int_0^{+\infty} e^{-y}y^s\mathrm{d}y + \frac{A(x)}{(2\gamma)^s}\int_0^{+\infty} e^{-y}y^s g(y/2\gamma;x)\mathrm{d}y \\
&= \frac{A(x)}{(2\gamma)^s}\Gamma(s+1) + \frac{A(x)}{(2\gamma)^s}\int_0^{f(\gamma)} e^{-y}y^s g(y/2\gamma;x)\mathrm{d}y \\
&\qquad + \frac{A(x)}{(2\gamma)^s}\int_{f(\gamma)}^{+\infty} e^{-y}y^s g(y/2\gamma;x)\mathrm{d}y \quad , \qquad \text{(A.2.2)}
\end{aligned}$$

where $f(\gamma)$ is some arbitrary function such that $f(\gamma) = o(\gamma)$ when $\gamma \to +\infty$.

On the first hand, $g : \mathbb{R}_+ \times \mathrm{supp}(P) \to \mathbb{R}$ is continuous since $(t,x) \mapsto \mathbb{P}_X(d^2(X,x) \leqslant t)$ and $x \mapsto A(x)$ are continuous. Therefore since $\mathrm{supp}(P)$ is assumed compact, $[0, f(\gamma)/(2\gamma)] \times \mathrm{supp}(P)$ is a compact for the product topology of $\mathbb{R}_+ \times \mathcal{X}$ and by the continuity of $g$ hence of $|g|$, the image $|g|([0, f(\gamma)/(2\gamma)] \times \mathrm{supp}(P))$ is a compact of $\mathbb{R}_+$, namely $|g|([0, f(\gamma)/(2\gamma)] \times \mathrm{supp}(P)) = [0, C_\gamma]$ with $C_\gamma \geqslant 0$. $C_\gamma$ is decreasing with respect to $\gamma$ and decreases to 0 when $\gamma \to +\infty$ since $g$ is continuous and $g(0;x) = 0$ for every $x \in \mathrm{supp}(P)$. It follows

$$\int_0^{f(\gamma)} e^{-y}y^s g(y/2\gamma;x)\mathrm{d}y \leqslant C_\gamma \int_0^{f(\gamma)} e^{-y}y^s\mathrm{d}y \leqslant C_\gamma \int_0^{+\infty} e^{-y}y^s\mathrm{d}y \leqslant C_\gamma\Gamma(s+1) = o_\gamma(1) \quad ,$$

where the '$o_\gamma(1)$' term is uniformly bounded with respect to $x$.

On the other hand, we show that the last additive term in (A.2.2) is negligible before $\gamma^{-s}$ for a proper choice of $f(\gamma)$.

$$\begin{aligned}\int_{f(\gamma)}^{+\infty} e^{-y} y^s g(y/2\gamma; x)\mathrm{d}y &\leqslant \int_{f(\gamma)}^{+\infty} e^{-y} y^s \left(\frac{(2\gamma)^s}{A(x) f^s(\gamma)} - 1\right)\mathrm{d}y \\ &\leqslant \frac{(2\gamma)^s}{A(x) f^s(\gamma)} \Gamma(s+1, f(\gamma)) \ , \end{aligned} \tag{A.2.3}$$

where $\Gamma(.,.)$ denotes the upper incomplete Gamma function.

Equivalent terms for $\Gamma(.,t)$ when $t \to +\infty$ are well-known and yield

$$\Gamma(s+1, f(\gamma)) \sim f^s(\gamma) e^{-f(\gamma)} \ ,$$

hence an equivalent for (A.2.3) is $A(x)^{-1}(2\gamma)^s \exp(-f(\gamma))$ which converges to 0 if $f(\gamma) = q\log(\gamma)$ with $q > s$ for instance. (Note that such a choice for $f(\gamma)$ still fulfills the condition $f(\gamma) = o(\gamma)$) □

**Lemma A.2.2.** *With the same assumption as Lemma A.2.1,*

$$\mathrm{Tr}(\Sigma_\gamma^2) \sim \frac{\mathbb{E}A(X)\Gamma(s+1)}{(2\gamma)^s} \ ,$$

*as* $\gamma \to +\infty$.

*Proof.* $\mathrm{Tr}(\Sigma_\gamma^2)$ is expressed as follows

$$\mathrm{Tr}(\Sigma_\gamma^2) = \mathbb{E}_X \langle \Sigma_\gamma k_X, k_X \rangle_\gamma \overset{\Delta}{=} \mathbb{E}\delta_X^2 \ .$$

By Lemma A.2.1 and Slutsky's lemma, $(2\gamma)^{2s}\delta_X^2$ converges weakly to $A(X)\Gamma(s+1)$ when $\gamma \to +\infty$. Therefore, since $A$ is assumed continuous and bounded, the weak convergence implies $\mathbb{E}(2\gamma)^{2s}\delta_X^2 \to \mathbb{E}A(X)\Gamma(s+1)$ hence the result of the lemma. □

**Lemma A.2.3.** *Write* $\delta_{XX'} = \Sigma_\gamma(X, X')$ *where* $X, X'$ *are two* i.i.d. *random variables. When* $\gamma \to +\infty$,

$$\mathbb{E}\delta_{XX'} \sim \frac{\mathbb{E}_X A^2(X)\Gamma^2(s+1)}{\gamma^{2s}} \ .$$

*Proof.* By definition of $\delta_{XX'}$,

$$\begin{aligned}\mathbb{E}\delta_{XX'} &= \mathbb{E}\langle \Sigma_\gamma k_X, k'_X \rangle_\gamma \\ &= \mathbb{E}_{X,X',X''} \langle k_{X''}^{\otimes 2} k_X, k_{X'} \rangle_\gamma \\ &= \mathbb{E}_{X,X',X''} k_\gamma(X', X) k_\gamma(X'', X) \\ &= \mathbb{E} e^{-\gamma\{d^2(X,X') + d^2(X,X'')\}} \\ &= \mathbb{E}_X \left[ \left( \mathbb{E}_{X'} e^{-\gamma d^2(X,X')} \right)^2 \right] \\ &= \mathbb{E}_X \left[ \Sigma_{\gamma/2}(X, X) \right]^2 \ . \end{aligned}$$

Lemma A.2.1 entails

$$\begin{aligned}\mathbb{E}\delta_{XX'} &= \mathbb{E}_X\left\{\frac{A^2(X)\Gamma^2(s+1)}{\gamma^{2s}}[1+A^{-1}(X)o_\gamma(1)]^2\right\}\\ &= \frac{\mathbb{E}A^2(X)\Gamma^2(s+1)}{\gamma^{2s}}[1+o_\gamma(1)]\ ,\end{aligned}$$

since $\mathbb{E}A(X) < +\infty$ because of the boundedness of $A(.)$. □

**Lemma A.2.4.** *Assume* $a = \inf\{A(x) : x \in \mathrm{supp}(P)\} > 0$ *and* $A = \sup\{A(x) : x \in \mathrm{supp}(P)\} < +\infty$. *Then*

$$\left|\frac{\Sigma_{\gamma,n}(x,x)}{\Sigma_\gamma(x,x)} - 1\right| \leqslant \sqrt{\frac{2\log(1/\delta)\gamma^s}{a\Gamma(s+1)n}}\left[1+\kappa\sqrt{\frac{\log(1/\delta)(2\gamma)^s}{a\Gamma(s+1)n}}\{1+o_\gamma(1)\}\right]\ ,$$

*with probability at least* $1-\delta$. *Here* $\kappa$ *is an absolute constant.*

*Proof.* Since $\Sigma_{\gamma,n}(x,x) = (1/n)\sum_{i=1}^n k_\gamma^2(X_i,x)$, Lemma A.1.2 can be applied with $c = 1$ and $\mathrm{Var}_X(k_\gamma^2(X,x)) \leqslant \mathbb{E}k_\gamma^4(X,x) = \Sigma_{2\gamma}(x,x) = \nu$ which entails with probability at least $1-\delta$

$$\left|\Sigma_{\gamma,n}(x,x) - \Sigma_\gamma(x,x)\right| \leqslant \sqrt{\frac{2\nu\log(1/\delta)}{n}}\left[1+\kappa\sqrt{\frac{\log(1/\delta)}{n\nu}}\right]\ ,$$

where $K$ is an absolute constant.

Therefore with high probability

$$\begin{aligned}\left|\frac{\Sigma_{\gamma,n}(x,x)}{\Sigma_\gamma(x,x)} - 1\right| &\leqslant \sqrt{\frac{2\log(1/\delta)}{n}}\frac{\sqrt{\Sigma_{2\gamma}(x,x)}}{\Sigma_\gamma(x,x)}\left[1+\kappa\sqrt{\frac{\log(1/\delta)}{n\nu}}\right]\\ &= \sqrt{\frac{2\log(1/\delta)\gamma^s}{a\Gamma(s+1)n}}\left[1+\kappa\sqrt{\frac{\log(1/\delta)(2\gamma)^s}{a\Gamma(s+1)n}}\{1+o_\gamma(1)\}\right]\ ,\end{aligned}$$

since by Lemma A.2.1 and introducing $a = \inf\{f(x) : x \in \mathrm{supp}(f)\} > 0$,

$$\frac{\sqrt{\Sigma_{2\gamma}(x,x)}}{\Sigma_\gamma(x,x)} = \frac{(4\gamma)^{-s/2}}{\sqrt{A(x)\Gamma(s+1)}(2\gamma)^{-s}}[1+o_\gamma(1)] \leqslant \frac{\gamma^{s/2}}{\sqrt{a\Gamma(s+1)}}[1+o_\gamma(1)]$$

and

$$\nu = \Sigma_{2\gamma}(x,x) \geqslant a\Gamma(s+1)(2\gamma)^{-s}[1+o_\gamma(1)]\ .$$

□

**Gaussian Models and Kernel Methods**

**Abstract**

Kernel methods have been extensively used to transform initial datasets by mapping them into a so-called kernel space or RKHS, before applying some statistical procedure onto transformed data. In particular, this kind of approach has been explored in the literature to try and make some prescribed probabilistic model more accurate in the RKHS, for instance Gaussian mixtures for classification or mere Gaussians for outlier detection. Therefore this thesis studies the relevancy of such models in kernel spaces. In a first time, we focus on a family of parameterized kernels - Gaussian RBF kernels - and study theoretically the distribution of an embedded random variable in a corresponding RKHS. We managed to prove that most marginals of such a distribution converge weakly to a so-called "scale-mixture" of Gaussians - basically a Gaussian with a random variance - when the parameter of the kernel tends to infinity. This result is used in practice to device a new method for outlier detection. In a second time, we present a one-sample test for normality in an RKHS based on the Maximum Mean Discrepancy. In particular, our test uses a fast parametric bootstrap procedure which circumvents the need for re-estimating Gaussian parameters for each bootstrap replication.

**Keywords:** kernel methods, rkhs, normality test, outlier detection

---

**Modèles Gaussiens et Méthodes à Noyaux**

**Résumé**

Les méthodes à noyaux ont été beaucoup utilisées pour transformer un jeu de données initial en les envoyant dans un espace dit « à noyau » ou RKHS, pour ensuite appliquer une procédure statistique sur les données transformées. En particulier, cette approche a été envisagée dans la littérature pour tenter de rendre un modèle probabiliste donné plus juste dans l'espace à noyaux, qu'il s'agisse de mélanges de gaussiennes pour faire de la classification ou d'une simple gaussienne pour de la détection d'anomalie. Ainsi, cette thèse s'intéresse à la pertinence de tels modèles probabilistes dans ces espaces à noyaux. Dans un premier temps, nous nous concentrons sur une famille de noyaux paramétrée - la famille des noyaux radiaux gaussiens - et étudions d'un point de vue théorique la distribution d'une variable aléatoire projetée vers un RKHS correspondant. Nous établissons que la plupart des marginales d'une telle distribution est asymptotiquement proche d'un « scale-mixture » de gaussiennes - autrement dit une gaussienne avec une variance aléatoire - lorsque le paramètre du noyau tend vers l'infini. Une nouvelle méthode de détection d'anomalie utilisant ce résultat théorique est introduite. Dans un second temps, nous introduisons un test d'adéquation basé sur la Maximum Mean Discrepancy pour tester des modèles gaussiens dans un RKHS. En particulier, notre test utilise une procédure de bootstrap paramétrique rapide qui permet d'éviter de ré-estimer les paramètres de la distribution gaussienne à chaque réplication bootstrap.

**Mots clés :** méthodes à noyaux, rkhs, test de normalité, détection d'anomalie

---

**Laboratoire de Mathématiques Paul Painlevé**
UMR 8524 CNRS - Université Lille 1 – 59655, Villeneuve d'Ascq Cedex – France