Kernel-based testing and their application to single-cell data

Anthony Ozier-Lafontaine

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

CENTRALE NANTES

MÉMOIRE DE DOCTORAT DE

L'ÉCOLE CENTRALE DE NANTES

ÉCOLE DOCTORALE N° 641
Mathématiques et Sciences et Technologies du numérique, de l'Information et de la Communication
Spécialité : *Mathématiques et leurs Interactions*

Par

Anthony OZIER-LAFONTAINE

Kernel-based testing and their application to single-cell data

Projet de recherche doctoral présenté et soutenu à l'École Centrale de Nantes, le 24 novembre 2023
Unité de recherche : UMR 6629, Laboratoire de Mathématiques Jean Leray

Rapporteurs avant soutenance :

Nathalie VIALANEIX, Directrice de recherche, Centre INRAE Occitanie-Toulouse
Zoltán SZABÓ Full professor, The London School of Economics and Political Science, University of London, UK

Composition du Jury :

Président :	Stéphane MINVIELLE	Directeur de recherche CNRS, Nantes Université
Examinateurs :	Nathalie VIALANEIX	Directrice de recherche, Centre INRAE Occitanie-Toulouse
	Zoltan SZABÓ	Full professor, The London School of Economics and Political Science, University of London, UK
	Céline VALLOT	Directrice de recherche CNRS, Institut Curie, Paris
	Jean-Philippe VERT	Dr, Chief R&D Officer HDR, Owkin, Paris
Directeur de recherches doctorales :	Bertrand MICHEL	Professeur des universités, École Centrale de Nantes
Co-dir. de recherches doctorales :	Franck PICARD	Directeur de recherche CNRS, Ecole normale supérieure de Lyon

Remerciements

Bertrand et Franck, je mesure ma chance de vous avoir rencontré et de vous avoir eu comme directeurs de thèse. Votre bienveillance sincère a rendu ces années épanouissantes scientifiquement et humainement. Vous avez accepté et encouragé ma manière d'être et de faire sans jamais la remettre en question, je me suis senti à la fois libre et soutenu. Bertrand, ta passion pour les statistiques est contagieuse, ton écoute et ton humanité sont inspirantes. Franck, ton hospitalité m'a fait adorer mes nombreux séjours Lyonnais, ton soutien sans faille et tes encouragements ont été des moteurs, je garde précieusement tes nombreux conseils qui me serviront de boussole dans les années à venir. Je vous remercie profondément pour tout, et je souhaite que notre amitié perdure au delà de cette thèse.

La recherche est tout sauf une activité solitaire, j'ai une pensée pour chaque personne rencontrée. Lise et Cathy, merc d'avoir suivi le déroulement de ma thèse, et de m'avoir parfois guidé, depuis votre rôle de comité de suivi. Stéphane et Florence, vous m'avez initié à la recherche en biologie en m'accueillant en stage dans votre équipe au Centre de Recherche en Cancérologie et Immunologie Intégrée Nantes Angers. Jean-Baptiste, Benjamin, Jennifer et Victor, vos conseils sur la ligne de départ m'ont servi tout au long de ma thèse. Magalie, Nathalie et Elise, merci de m'avoir montré votre virtuosité à manipuler les cellules, et à créer une ambiance d'équipe aussi bienveillante. Jonathan, je suis content de t'avoir rencontré et de t'avoir vu évoluer en parallèle de moi. J'ai aussi passé beaucoup de temps à Lyon. Merci Olivier et Philippe de m'avoir si bien accueilli dans vos équipes respectives, au Laboratoire de Biologie et Modélisation de la Cellule et au Centre de Recherche en Cancérologie de Lyon. François, c'est grâce à toi que mes codes tournent aussi vite. Claire et Antoine, j'ai adoré découvrir les méthodes à noyaux avec vous. Antoine, c'était un plaisir de te retrouver à chaque nouveau Workshop. Camille, ton enthousiasme à faire de la recherche et à t'intéresser à tout sont géniaux, merci de m'avoir fait visiter les laboratoires de biologie. Ellias, Hugues, Maxime, Thao, merci pour tous ces bons moments. Ghislain, merci d'avoir été de si bon conseil, et de m'avoir permis de faire exister le package *ktest*. Perrine, merci d'être venue donner un second souffle à nos projets, et d'avoir été un soutien aussi solide et encourageant dans mes derniers mois de

thèse, j'ai aimé travailler avec toi. Enfin, tous ces séjours n'auraient jamais été possibles sans votre accueil, Thierry, Solène, Maxence, merci d'avoir toujours eu une chambre ou un canapé pour moi.

C'est au Laboratoire de Mathématiques Jean Leray que j'ai passé la majeure partie de mon temps. Je garderai un souvenir très doux de l'ambiance de ce laboratoire. Aymeric, merci de m'avoir initié au CCIPL et au métier d'ingénieur de recherche, et aussi d'avoir persévéré avec tes invitations au bar. Françoise, Anthony, Anne, Frédéric, merci pour vos encouragements tout au long de ces quatre années. Claire, nos TP du lundi soir vont me manquer. Adrian et Antoine, pour m'avoir partagé vos doutes sur l'avenir, et aidé à clarifier ma pensée si souvent. Alexandre, pour notre tour de l'île à vélo. Arthur, pour m'avoir accompagné dans mes premiers pas à la radio. Saad, pour nous avoir fait visiter Milan. Silvère, pour m'avoir confié Rima. Lucas, pour avoir autant d'enthousiasme à chacune de nos conversations. Zihad, je suis content d'avoir croisé ta route. Damien, Gurvan, Thomas, Klervi, Flavien, pour toutes ces pauses et ces jeux. Lucile, pour ton ouverture d'esprit. Gaspard, Jean, Ludovic, et toutes les personnes qui donnent vie aux couloirs de ce bâtiment. J'ai réalisé à quel point rien ne pourrait bien se passer dans la recherche sans toutes les personnes support qui s'occupent de tout. Je remercie énormément Bénédicte, pour avoir toujours accepté de prendre mes billets de trains dans l'urgence de ma procrastination et pour m'avoir si gentiment relancé chaque fois que j'oubliais de te demander mes remboursements. Je remercie aussi Eric, Saïd, Stéphanie, Caroline et Béatrice pour tout faire fonctionner aussi bien.

Et surtout, quel bonheur d'avoir pu partager mon bureau avec des personnes aussi exceptionnelles. J'ai savouré ces rencontres auxquelles je ne m'attendais pas. Germain, pour nos débats sur la nature du réel. Samuel, pour ces longues soirées à refaire le monde, j'espère vraiment que tu y trouveras ta place. Charbella, pour ta générosité et ta capacité exemplaire à tout réussir. Elric, pour ta justesse fascinante, ta bonté et ton enthousiasme. Destin, pour avoir été mon binôme de statistiques, de programmation, et de tout le reste, c'est aussi grâce à toi que le résultat est tel qu'il est.

C'est une chance d'avoir cet espace de remerciements pour témoigner mon affection aux personnes qui rendent mon existence si agréable. Je n'ai pas les mots pour dire à quel point les personnes qui suivent, vous êtes des repères pour moi.

Je crois que j'aurais explosé, si je n'avais pas trouvé des activité épanouissantes en dehors du travail. Benjamin, Ken, j'ai adoré devenir Chevallié, puis Androspectre avec vous, Kharô et Seb. Vous quatre méritez amplement votre place dans ces remerciements, car l'idée de réaliser nos projets ambitieux m'a souvent motivé à être efficace la journée, et tout ce qu'on a accompli, ainsi que les personne que vous êtes, m'ont tellement transformé et enrichi. Bref, vous avez été ma bouée, et j'ai hâte de voir ce qu'on créera ensemble dans le futur. Arthur, merci d'avoir toujours été un ami et un voisin si fiable et compréhensif. Antoine, je suis heureux que ma thèse est coïncidé avec tes années d'internat, j'espère qu'on aura d'autres occasions d'habiter dans la même ville. Thierry, merci pour ta présence inconditionnelle et ta capacité à partager tes passions. Solène, merci d'avoir toujours été fan dans tout ce que j'entreprenais. Anne, je suis tellement content d'avoir partagé mes études avec toi, depuis la première année à Centrale, jusqu'à la fin de la thèse. Simon, merci d'être la personne aussi sensible, ambitieuse et créative que tu es, j'aime ta définition de l'amitié, et j'adore faire partie de tes amis. Baptiste, Alexandre, Baptiste, merci de continuer à être là après tout ce temps. Megan, Benjamin, Jonathan, Estelle, Jahed, Laurène, je me considère comme extrêmement chanceux de vous avoir dans ma vie, il y a peu d'endroits où je me sens autant à ma place qu'avec vous, et c'est tellement ressourçant. Notre amitié dure depuis si longtemps, vous êtes devenus ma famille, et j'aime les belles personnes que vous êtes.

Je remercie évidemment tous les membres de ma famille, vous avez toujours eu un mot d'encouragement, une présence bienveillante. Papa, Maman, c'était chaotique, mais vous m'avez permis d'être la personne que je suis devenu, chacun à votre manière, je suis heureux de vous avoir. Mon plus grand privilège est d'être entouré de grands-parents aussi inspirants. Réjane, merci d'être si aimante, résiliente et pragmatique. Bernard, merci d'être un exemple de sagesse et de poésie. Luc, merci d'être l'artiste le plus talentueux que je connaisse, ta vie est une oeuvre. Arlette, c'est toi qui nous a transmis, à Maxence, Maëva et moi, cette curiosité sans fin, cette confiance optimiste qui rend toute ambition réalisable, et surtout, le courage d'accepter et d'aimer toutes les différence, en soi et autour de soi, merci pour tout, ma thèse est pour la grande scientifique que tu es. Maxence, il n'y a rien de plus rassurant dans la vie que de savoir qu'on est pas seul pour affronter le chaos. Tu le sais, tout a commencé avec toi, j'ai appris à être et à penser à tes côtés. On s'est mutuellement sauvé la vie, en créant ensemble notre univers si unique et pertinent. Notre relation est ce que j'ai de plus précieux.

Rahma, tu as vécu le quotidien de ces quatre ans avec moi. On a souvent cru que ça ne finirai jamais. J'ai adoré ces années parce que j'étais avec toi tout le temps. Tu as stressé pour moi à chacune de mes présentations, tu m'as soutenu, encouragé, rassuré, conseillé, au point que c'est aussi un peu ta thèse. Je suis un peu fier de terminer ce projet, mais pas autant que d'avoir construit une relation aussi parfaite avec toi. Je ne souhaite rien d'autre que de continuer à être heureux ensemble, je t'aime tellement.

Table of Contents

Résumé Substantiel en Français

Les travaux décrits dans cette thèse se situent à l'interface entre les statistiques, l'informatique et la biologie. Ils sont motivés par l'ambition double de contribuer à la recherche en statistique tout en répondant à un besoin méthodologique pour l'analyse de données expérimentales. Une partie importante de ces travaux se sont déroulés au contact de collaborateurs biologistes, dans une démarche de compréhension de leurs problématiques et de leurs données. C'est pendant ces échanges que nous avons identifié la nécessité de développer les outils de test et d'exploration de données issues de mesures de séquençage en cellule unique qui font l'objet de ce manuscrit.

Le développement de technologies de séquençage en cellule unique constitue une révolution pour la recherche en biologie moléculaire, car il permet de mesurer l'activité spécifique de chaque cellule d'une population. On s'intéresse en particulier à la technologie de séquençage des ARN transcrits en cellule unique (scRNA-Seq) et marginalement au séquençage par immunoprécipitation de la Chromatine en cellule unique (scChIP-Seq) qui nous informent sur deux aspects différents de l'activité cellulaire. Le scRNA-Seq permet de mesurer précisément le transcriptome de chaque cellule d'une population de cellules, et le scChIP-Seq permet de mesurer les modifications de la chromatine, ce qui permet notamment d'étudier la régulation de l'expression des gènes.

En contrepartie de cette précision sans précédent, les données de séquençage en cellule unique et en séquençage d'ARN, sont des données de comptage massives et en grande dimension, qui contiennent typiquement plusieurs milliers d'observations (les cellules) et plusieurs milliers de variables (les gènes). De plus, elles sont sur-dispersées, sporadiques (elles contiennent beaucoup de zéros), et contiennent de nombreux biais liés aux nombreuses étapes nécessaires à une mesure. Des outils d'analyse et algorithmiques adaptés sont nécessaires afin d'étudier les mécanismes cellulaires liés à un phénotype à partir des données de scRNA-Seq. Pour développer ce genre d'outils, il est nécessaire de comprendre les interrogations scientifiques et les problématiques biologiques qui sous tendent l'analyse de ces données, et de traduire la méthodologie de la recherche en biologie en questions

statistiques pertinentes. Cela n'est possible qu'à travers des collaborations étroites entre biologistes et statisticiens, à l'instar de ce travail de thèse.

Nous avons donc identifié en collaboration avec des collègues spécialistes de l'analyse de données transcriptomique, l'absence d'un outil de comparaison globale de transcriptome, et avons choisi de formuler cette question comme un test statistique de comparaison de distributions. Pour y répondre, nous nous sommes focalisés sur les méthodes de test de comparaison de deux échantillons basés sur les méthodes à noyaux [63]. Ces tests à noyaux, dont le représentant le plus connu est le test Maximum Mean Discrepancy (MMD) ainsi que le test basé sur l'Analyse Discriminante de Fisher à noyaux (KFDA), existent depuis les années 2000 [52, 64], leurs performances sont attestées et il s'agit d'un sujet de recherche en statistique actif sur lequel il existe de nombreuses questions ouvertes. Aussi, aucune implémentation pratique de ces méthode n'est publiée, donc l'outil que nous avons proposé rend possible l'accès aux tests à noyau à des non spécialistes. Motivés par cet objectif pragmatique, nos travaux peuvent être décomposés en quatre contributions principales.

La première contribution est le développement d'un package nommé *ktest* dans lequel sont efficacement implémentés ces test à noyaux, ce qui permet de faire l'analyse complète d'une comparaison de transcriptomes en seulement quelques lignes de code et quelques minutes de calculs, avec un large panel d'outils de visualisation et de diagnostic permettant d'interpréter les résultats. L'implémentation pose aussi des questions pratiques en termes de structuration des données et d'efficacité computationnelle. Le package *ktest* est donc implémenté dans les langages *R* et *Python*, et compatible avec les frameworks d'analyse de données en cellule unique existants, tels que *Seurat*, *Single-cell experiment* et *Scanpy*. Enfin, les méthodes d'approximation de Nyström y sont implémentées de manière optionnelle pour les échantillons qui contiendraient plusieurs milliers de cellules.

La seconde contribution consiste en la mise en pratique opérationnelle des tests à noyaux. Nous avons éprouvé leurs performances sur des données simulées et sur des données de transcriptomique en cellule unique. Sur les données simulées, nous avons démontré la compétitivité des tests à noyaux par rapport aux méthodes de test habituellement utilisées pour ce type de données. Sur les données de séquençage, nous avons à la fois pu analyser les différences entre les résultats de notre méthode avec d'autres, et publier des

résultats originaux obtenus à partir de notre approche. Nous avons grandement approfondi l'interprétation géométrique des tests à noyaux, ce qui nous a permis de proposer des outils de diagnostic et de visualisation, qui permettent une interprétabilité complète des résultats. Enfin, nous avons identifié des heuristiques pratiques pour le choix des hyperparamètres sous-jacents de façon adaptée au problème étudié.

Nous proposons aussi trois contributions théoriques. Tout d'abord, nous appliquons les méthodes d'approximation de Nyström aux tests à noyaux, ce qui n'avait pas clairement été décrit dans une publication jusqu'à présent. De plus, nous introduisons un cadre de modélisation général pour les tests à noyaux. Ce cadre de modélisation nous permet de proposer un test à noyaux qui généralise les tests de comparaison de deux distributions à des tests pour les designs quelconques, adaptés à des données issues de protocoles expérimentaux plus complexes. Ce test est inspiré du test de la trace de Hotelling-Lawley pour les modèles linéaires multivariés. Ce lien avec les modèles linéaires nous permet de proposer plusieurs outils de diagnostic qui enrichissent l'interprétation des resultats des tests à noyaux. Grâce à des outils issus de la théorie de la perturbation des opérateurs, nous obtenons la distribution nulle asymptotique de la statistique proposée. Notre troisième contribution théorique est l'application des fonctions d'influence issues du domaine des statistiques robustes à nos statistiques de test.

INTRODUCTION

This work is motivated by the ambition to both contribute to research in statistics and deliver a practical framework that supports the analysis of high-throughput single-cell sequencing data. Thus, it lies at the interface between statistics, informatics, and biology. We identified the need to develop an interpretable tool dedicated to the comparison of single-cell datasets thanks to several fruitful collaborations. These collaborations were not all successful but they all contributed to shape a tool highly compatible with our collaborator's methodology and issues.

In this Introduction, we first present the main principles of the single-cell technology, the challenges related to single-cell data analysis and motivate the need for an interpretable method that compares single-cell datasets globally. Then we introduce the solution we propose to fill this gap as four main contributions. We conclude by presenting the organization of the manuscript.

Single-Cell Sequencing

Cell Activity

The cell is considered as the basis of living organisms. In Eukaryotes, the genetic material is contained in the nucleus in the form of chromosomes, and the cytoplasm contains other organelles of the cell.

Cell activity is mainly determined by the genes encoded in the nuclear DNA that are expressed during the cell life. Inside the cell nucleus, coding genes are transcribed from DNA to messenger RNA (mRNA) molecules. The mRNAs migrate from the nucleus to the cytoplasm where their information is translated into proteins until the mRNA degradation. Regulation mechanisms govern this activity. These steps are subject to random variations and that supports the stochastic nature of cell fate. The field of molecular biology focuses on the mechanisms underlying gene activity. The development of single-cell sequencing technologies initiated a revolution in the field, as it allowed to

observe cell populations at the unprecedented single-cell resolution and study the cell-to-cell variability, when previous techniques referred as bulk sequencing only accessed to an averaged information.

Single-Cell Sequencing

Theoretical and technical advances in droplet-based microfluidics in the 2000s allowed the development of instruments able to encapsulate cells in droplets isolated from each other. Coupled with techniques from molecular biology, it became possible to uniquely identify each cell of a sample from a sequencing measurement. These high-throughput droplet-based techniques allowed to access to the information related to the activity of hundreds to thousand of cells in a few hours only [45].

Many techniques are based on the single-cell approach. For instance, the single-cell version of Assay for Transposase Accessible Chromatin using sequencing (scATAC-Seq) measures the gene accessibility of each cell [23]. The single-cell Chromatin immunoprecipitation sequencing (scChIP-Seq) measures protein-DNA interactions in the nucleus, it can be used to study gene regulation [120, 57]. In addition to an application on scChIP-Seq data, the methods we developed are mainly motivated by the analysis of single-cell RNA sequencing (scRNA-Seq) measures, that capture a representative subset of the mRNAs that are located in the cytoplasm of a cell [68, 89, 152]. Recently, spatial scRNA-Seq was developed to append the 3D position of each sequenced cell [136].

Single-Cell RNA Sequencing Data Analysis

A scRNA-Seq dataset contains the number of detected mRNAs associated to each transcript in each cell. It is encoded as a table where the rows are cells and the columns are transcripts. From a statistical point of view, cells are called observations and transcripts are called variables or features. Generally, scRNA-Seq datasets contain hundreds to thousands of observations and up to tens of thousands of variables. A dataset that contains a lot of observations is considered as large. That raises practical issues in terms of data storage and computational cost. A dataset that contains a lot of variables is said to be high-dimensional. There is a theoretical difficulty related to the statistical analysis of high-dimensional datasets, refered as the curse of dimensionality [49].

A scRNA-Seq dataset is the output of a succession of many steps among which we have the sample preparation, the droplet-based single-cell isolation, the PCR signal amplification, the mRNA library preparation, the sequencing, the gene identification based on a reference genome and the quality controls and data filtering. Each of these steps induces technical biaises [150, 95], that are difficult to distinguish from the biological variability related to the uniqueness of each cell and the intrinsic stochasticity of transcription. Moreover, as a cell does not express every gene and the expressed genes are not expressed continuously, scRNA-Seq datasets contain a lot of zeros. We say they are sparse [137].

The analysis of the distributions observed in scRNA-Seq dataset is challenging. It requires efficient algorithms able to run on large datasets and accessible to non-specialists. A large panel of such methods have been developed [87]. For instance, it is possible to determine cell types [6], differentiation trajectories [80], cell-cell communication [37] or cell cycles [140] based on a scRNA-Seq dataset. A lot of work has been done to obtain synthetic visual representations of these high-dimensional datasets, non-linear dimentionality reduction methods such as t-SNE [88] or UMAP [96] are the most popular. We are especially interested in the comparison of scRNA-Seq datasets.

Motivation for a Global Comparison Method

Most experiments in single-cell transcriptomics consist in comparing cell activity in several conditions of interest in order to determine what affects cell activity and how. Conditions comparison has been the basis of transcriptomics methodology for a long time and the historical approach on bulkRNA-Seq data is called Differential Expresssion Analysis (DEA) [119]. DEA on bulkRNA-Seq data compares the averaged expressions gene by gene in order to eventually detect a significant difference for some genes, refered as differentially expressed genes. DEA has been adapted to scRNA-Seq data and is now used in a large majority of the scRNA-Seq publications [134]. Many new methods were developed to take advantage of the distributional information available for each gene in scRNA-Seq datasets to increase the performance of DEA [32]. The application of DEA to scRNA-Seq data raises issues on the probabilistic distribution underlying scRNA-Seq data [78].

The DEA paradigm is particularly suited to detect variables of interest taking part in a response phenotype under specific conditions. The detected differentially expressed genes

can then be confirmed by cutting them off or inhibiting them in further experiments. We noticed that the DEA paradigm is also employed in order to identify observations of interest, such as a sub-population with a specific phenotype hidden in only one of the two compared populations that contributed to the difference. However, DEA is in general not able to fulfill this task because it is restricted to compare marginal distributions. For instance, the sub-population can be characterized by a slight difference in the expression of many variables, that do not induce significant gene-wise differences between the two compared datasets. Moreover, if the sub-population is too small compared to the global population, even a high difference in the expression of some genes could remain undetected and considered as random noise. In both situations, DEA would fail to detect the difference induced by the sub-population that would remain invisible from the data analysis.

If such a homogeneous sub-population exists in a dataset, a multivariate comparison between the two samples would have more chance to detect the existing difference than the univariate DEA approach. Indeed, several slight marginal differences may lead to a large joint difference. While such a global comparison tool would be particularly suited to many experiments in molecular biology, we identified that it did not exist back in 2019. Since, some methods have been published but they lack of a fundamental feature: interpretability to identify the sub-population of interest defined as the cells supporting the difference [16, 101]. Thus, we developed this interpretable global comparison approach that we called Differential Transcriptome Analysis (DTA), complementary to DEA. Our DTA approach relies on kernel testing. It is promising as the interpretation of the test results allowed us twice to discover a sub-population of interest conjectured by our collaborators but not detected with existing techniques such as DEA or clustering algorithms.

Our Contributions

Mathematical Background : Kernel Methods and Kernel Tests

All the contributions developed in this manuscript belong to the field of kernel methods. Kernel methods are popular, competitive and interpretable methods developed in the Machine Learning community [125, 129]. The principle of kernel methods is to embed observations in a high-dimensional Reproducing Kernel Hilbert Space (RKHS) called

feature space, and to apply linear statistical methods on the embeddings. When the embedding function called the feature map is non-linear, the resulting kernel method is non-linear. While embedding data in high-dimensional spaces could increase the computational cost of kernel methods, the kernel trick makes the computational cost polynomial in the number of observations and thus generally affordable. Even when this polynomial cost is too expensive, as it can be the case for the large datasets encountered in scRNA-Seq data analysis, matrix approximation methods such as the Nyström approximation can be coupled to the kernel trick to reduce the computational cost [147].

The principle of embedding observations can be generalized to embedding probability distributions [100]. This technique allows to define a metric between probability distributions and it is at the basis of several pair-wise comparison methods, refered as kernel two-sample tests [63]. The most famous kernel two-sample test is the MMD test, based on the MMD statistic that is defined as the distance between the embeddings of the two compared probability distributions in the feature space [53]. Several variants exist for this test, in particular, the kernel Fisher Discriminant Analysis test is very close to the MMD test as the KFDA statistic is a normalized version of the MMD statistic [64]. The MMD and KFDA tests do not require any distributional assumption and can theoretically detect any existing difference between two distributions. The first Chapter of this manuscript contains a more detailed presentation of kernel methods and kernel testing.

A Package for the Implementation and Interpretation of Kernel Testing

One of our main contribution is to propose a user-oriented implementation of these kernel tests in a package called *ktest* to perform interpretable sample comparisons in a few lines of code. Indeed, despite the interest of kernel tests, to the best of our knowledge, there is no such implementation of these tests. *ktest* is implemented in both *Python* and *R*, it is designed to be compatible with existing single-cell analysis frameworks such as Seurat [24], SingleCellExperiment [4] and Scanpy [148], and we project to integrate *ktest* to the *scverse* consortium in the future [143]. *ktest* is not limited to single-cell data and could be applied on any research data to compare conditions. To deal with the large datasets encountered in scRNA-Seq data analysis, Nyström approximations are implemented as optional features, most importantly, all the methods described in this manuscript are

implemented in the *ktest* package.

Operational Aspects of Kernel Testing

Our second main contribution is the operational application of kernel testing in real conditions. For DEA, we compared the performances of kernel two-sample tests to the large set of existing DEA methods on simulated data and through the systematic analysis of the detected differentially expressed genes on published data. The protocol on simulated data was shared with us by the authors of [44], and it was inspired from the four categories of alternatives for single-cell RNA sequencing data described in [78]. The KFDA test shows competitive performances compared to other existing DEA methods and is even the most powerful to detect the most complex alternative, which the MMD test fails to detect. The systematic analysis of the results on published data deepens and discusses the analyses made in [134]. These practical applications raised issues on the kernel function choice and on hyperparameter tuning for the KFDA test. As hyperparameter tuning is related to statistical issues that are beyond the scope of this manuscript, we discuss how to choose a heuristic and some perspectives in the Conclusion chapter.

In complement to kernel-testing, we developed a large range of data exploration tools and diagnostic tools. Our main data exploration tool is called the discriminant axis and relies on the geometrical concepts underlying the KFDA framework. The discriminant analysis is basically a non-linear dimension reduction that summarizes the main differences existing between two compared samples on a one-dimensional axis. The principal difference of this visualization tool compared to popular scRNA-Seq data visualization tools is that it is based on discrimination instead of being based on description. The discriminant axis allowed us to identify sub-populations of interest undetected from previous analyses on both scRNA-Seq data and scChIP-Seq data. Diagnostic tools aim at monitoring what led to the test outcome. Diagnostic graphs inspired from tests on multivariate linear models allow to assert that the test assumptions are true. Kernelized Cook distances and influence functions are different ways to identify influential observations that weighted on the results. Finally, a diagnostic graph describing how the dimension reduction related to the KFDA test captured the inner variability of both samples can give some intuition on hyperparameter tuning. To summarize, we enriched kernel testing to obtain a comprehensive, efficient and interpretable framework suited to give a deep understanding of the existing differences between several conditions, as the marginal DEA and

the joint DTA approaches allow to detect variable-wise differences and observation-wise differences respectively.

Theoretical Contributions

A first theoretical contribution is to detail the application of the Nyström method to kernel testing. We also propose a model framework for kernel testing called kernel linear model, that allows to define a new kernel test that generalizes kernel two-sample tests to any experimental design. The kernel linear model is inspired from the multivariate linear model, and it allows to enrich kernel tests with a set of diagnostic tools to interpret the test outcome. We derive the asymptotic null distribution of our generalized kernel test with analytic tools from operator perturbation theory inspired from [22]. This contribution is detailed in Chapter 3 and is a work in progress that has not been the object of a publication yet. Another work in progress have been initiated on the detection of the observations that have a strong influence on the test results. We propose to apply on our test statistics a tool that comes from robust statistics called Gateaux derivatives. It has already been applied to the KFDA classifier and the Kernel Principal Component Analysis (KPCA) that is a dimension reduction method, the first steps of this work are presented in Chapter 4. All these contributions are implemented in our package.

Successful Applications to Transcriptomic Research

Our exploration of kernel testing applied to single-cell data was nourished by several collaborations with molecular biologists. Each collaboration became the building block of an aspect of our final methodology. Our first collaboration with the team of Stephane Minvielle at the Centre de Recherche en Cancerologie et Immunologie Intégrée de Nantes-Angers (CRCI^{2}NA) on the Multiple Myeloma was fundamental to identify and formulate the need for a global comparison framework. Our first application of the early implementation of kernel tests on the Pituitary Adenoma with the team of Philippe Bertolino at the Centre de Recherche en Cancerologie de Lyon (CRCL) deeply enhanced our understanding of the method and largely motivated the development of diagnostic and visualization tools. Moreover, their complex experimental design confirmed the need for a generalization of kernel two-sample tests. Finally, these two collaborations were very fruitful as they shaped our approach from a theoretical kernel test to a practical framework able to give answers to concrete research issues.

The experience gathered from these exchanges allowed us to apply an advanced version of our tool on the Reversion dataset from the team of Olivier Gandrion at Ecole Normale Supérieure de Lyon. As we only pretended to confirm an already published similarity between two datasets with our method [153], our visualization tool detected the subpopulation of interest they were looking for. Finally, the second detection of a subpopulation of interest on a scChIP-Seq dataset measured on Breast Cancer cells from the team of Celine Vallot at Institut Curie was somehow easy as we knew exactly how to use our package *ktest*. These collaborations correspond to the most practical and exciting contributions of this work as they had concrete impact on their research and they most fundamentally participated to build and strenghten a bridge between molecular biology and statistics.

Organisation of the Manuscript

This introduction presented our motivations from molecular biology and the context of single-cell data analysis. The following chapter will detail the statistical concepts underlying our framework. The first chapter (Chapter 1) is a mathematical introduction to kernel methods and kernel testing. The second chapter (Chapter 2) is an adapted version of our manuscript [106] supplemented with a presentation of how to apply the Nyström method to kernel testing. The third chapter (Chapter 3 is a presentation of the kernel linear model and the kernel test for general designs with a demonstration of its asymptotic distribution. It also introduce some diagnostic tools derived from the multivariate linear model. The fourth chapter (Chapter 4) relates our initiated work on influence functions for kernel testing. We discuss about the perspectives and hyperparameter tuning in the Conclusion chapter.

Chapter 1

Introduction to Kernel Methods and Kernel Testing

Kernel methods are popular tools in many areas of Applied Mathematics. Their theoretical origin goes back to the seminal work of Mercer [97], and their practical implementation relies on the work of Aronszajn [9]. However, they became popular in Statistics in the late 90's, with the Support Vector Machine algorithm [29].

In statistical learning, the idea of kernel methods is to embed the observations in a high-dimensional linear space to apply linear methods on the embeddings. The embedding function is called the feature map. It is often chosen to be non-linear in order to catch non-linear relationships between the observations that would remain undetected with a linear approach. The high-dimensional space that contains the embeddings is called the feature space. The kernel trick allows to avoid the principal difficulty that underlies this type of approach, that is the explicit computation of the embeddings. Thanks to the kernel trick, we can even work in an infinite dimensional feature space without any computational limitation. From a statistical point of view, the interest of kernel methods also lies in the possibility to use the feature map to embed probability distributions in the feature space. Then we can characterize a probability distribution by its embedding, and use this representant as a proxy to study the distribution. The literature on embedding probability distributions with kernel methods was recently reviewed in [100].

In this chapter, we are particularly interested in applying kernel methods to infer statistical quantities and then perform hypothesis testing. Many linear statistical methods have their kernelized version, as the kernel principal component analysis [124], kernel canonical correlation analysis [11], kernel k-means clustering [10]. It is a simple way to obtain a non-linear method from a linear method.

The first section of this chapter is dedicated to the introduction of the general notions related to kernel methods such as Hilbert spaces and kernel functions (Section 1.1). In the second Section, we describe how to embed a probability distribution in the feature space, with a particular focus on the kernel trick used to diagonalize an empirical operator related to this embedding in practice (Section 1.2). We are especially interested in applying kernel methods for non-parametric hypothesis testing. This range of applications was initiated by the seminal work of Gretton [55] who introduced the now famous Maximum Mean Discrepancy statistic for two-sample testing. The third section of this chapter is dedicated to the presentation of non-parametric hypothesis testing and kernel testing, with a focus on the MMD test that is based on the concept of embedding probability distributions in the feature space (Section 1.3). The fourth Section details the KFDA framework and two regularized KFDA statistics for two-sample testing (Section 1.4).

1.1 Introduction to Kernel Methods

Let $(\mathcal{Y}, \mathfrak{Y})$ a measurable space called the input space. The core idea of kernel methods is to map the elements of $\mathcal{Y}$ into a high-dimensional reproducing kernel Hilbert space $\mathcal{H}$ using a function $\phi(\cdot) : \mathcal{Y} \to \mathcal{H}$. The function $\phi(\cdot)$ is called the feature map, and the space $\mathcal{H}$ is called the feature space, it can be of infinite dimension. The image $\phi(y) \in \mathcal{H}$ of $y \in \mathcal{Y}$ by the feature map $\phi(\cdot)$ is called the embedding of y. Kernel methods are linear methods applied in the feature space $\mathcal{H}$, that correspond to non-linear methods in the input space $\mathcal{Y}$ when the feature map is non-linear. Thus, these methods are able to catch non-linear relationships in $\mathcal{Y}$.

1.1.1 Introduction to Hilbert Spaces

A common interpretation of kernel methods is to consider that we embed the observed data from the input space to a high-dimensional feature space in order to study them. Most manipulations done on the embeddings in the feature space are possible because it is a Hilbert space, and the reproducing property allows the practical computation of the quantities of interest. A Hilbert space is a vector space in which we can easily define distances, orthogonality and convergence, that are at the root of many statistical methods, it is thus particularly suited to the development of statistical methods. In this subsection, we describe some basic notions relative to Hilbert spaces.

Definition 1 (Hilbert Space)**.** *A Hilbert Space $\mathcal{H}$ is a vector space endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ in which any Cauchy sequence has a limit in $\mathcal{H}$.*

Let $\mathcal{H}$ be a Hilbert space. The existence of an inner product in $\mathcal{H}$ allows the definition of the orthogonality of two elements of $\mathcal{H}$, as $h, h' \in \mathcal{H}$ are said to be orthogonal if and only if $\langle h, h' \rangle_{\mathcal{H}} = 0$. In addition, the limit of any convergent sequence of $\mathcal{H}$ belongs to $\mathcal{H}$, this is useful for describing the asymptotic properties of empirical estimators constructed in $\mathcal{H}$. Note that the Hilbert norm of a Hilbert space is defined with respect to the inner product. Such that for h in $\mathcal{H}$, we have $\|h\|_{\mathcal{H}} = \sqrt{\langle h, h \rangle_{\mathcal{H}}}$.

Some kernel tests rely on decompositions on orthonormal bases of the feature space $\mathcal{H}$, that are well defined for separable Hilbert spaces.

Definition 2 (Separable Hilbert Space)**.** *A Hilbert space $\mathcal{H}$ is separable if and only if there exists an orthonormal sequence $(e_s)_{s \geq 1}$ of $\mathcal{H}$ such that for any $h \in \mathcal{H}$, we have:*

$$\|h\|_{\mathcal{H}} = \sqrt{\sum_{s \geq 1} \langle h, e_s \rangle_{\mathcal{H}}^2}. \tag{1.1}$$

When $\mathcal{H}$ is separable, any orthonormal sequence $(e_s)_{s \geq 1}$ that fulfills condition (1.1) is an orthonormal basis of $\mathcal{H}$.

Feature spaces are in general not endowed with a canonical orthonomal basis. In statistical learning, a common approach to work with an orthonormal basis of $\mathcal{H}$ is to use the orthonormal sequence of eigenfunctions of a Hilbert-Schmidt operator. Hilbert-Schmidt operators are a particular class of linear operators from $\mathcal{H}$ to $\mathcal{H}$.

Definition 3 (Hilbert-Schmidt operator)**.** *Let $\mathcal{H}$ a separable Hilbert space and $(e_s)_{s \geq 1}$ an orthonormal basis of $\mathcal{H}$. A linear operator $L : \mathcal{H} \to \mathcal{H}$ is a Hilbert-Schmidt operator if and only if:*

$$\sum_{s \geq 1} \|Le_s\|_{\mathcal{H}}^2 < +\infty.$$

Then, the sum is independent from the orthonormal basis $(e_s)_{s \geq 1}$, and it is called the Hilbert-Schmidt norm:

$$\|L\|_{\mathrm{HS}(\mathcal{H})}^2 = \sum_{s \geq 1} \|Le_s\|_{\mathcal{H}}^2 .$$

Hilbert-Schmidt operators form a separable Hilbert space $\mathrm{HS}(\mathcal{H})$ endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathrm{HS}(\mathcal{H})}$ such that for $L, N \in \mathrm{HS}(\mathcal{H})$, we have:

$$\langle L, N \rangle_{\mathrm{HS}(\mathcal{H})} = \sum_{s \geq 1} \langle Le_s, Ne_s \rangle_{\mathcal{H}} .$$

Hilbert-Schmidt operators are compact and have countable spectra. Every non-zero eigenvalue of a Hilbert-Schmidt operator is associated to an eigenspace of finite dimension. Every Hilbert-Schmidt operator can be decomposed in an orthonormal basis of its eigenfunctions. Let L a positive Hilbert-Schmidt operator and $(f_s)_{s \geq 1}$ in $\mathcal{H}$ be the sequence of its orthonormal eigenfunctions associated to the sequence of non-negative and non-increasing eigenvalues $(\lambda_s)_{s \geq 1}$ in $\mathbb{R}$, we have:

$$L = \sum_{s \geq 1} \lambda_s f_s \otimes f_s,$$

where $\otimes$ is the tensor product, that defines a rank-one Hilbert-Schmidt operator, such that if $f, g \in \mathcal{H}$ are non-zero and $h \in \mathcal{H}$, we have:

$$(f \otimes g)h = \langle g, h \rangle_{\mathcal{H}} f.$$

From the definitions, we can show that:

$$\langle L, f \otimes g \rangle_{\mathrm{HS}(\mathcal{H})} = \langle f, Lg \rangle_{\mathcal{H}} .$$

1.1.2 Kernel Functions and Reproducing Kernel Hilbert Spaces

Both the feature map and the feature space are actually defined with respect to a kernel function, that is basically a measure of the similarity between pairs of observations from the input space. Kernel methods are named with respect to the central role of the kernel function.

Definition 4 (Kernel function)**.** *Let $(\mathcal{Y}, \mathfrak{Y})$ a measurable space. A kernel function $k(\cdot, \cdot)$ over $\mathcal{Y}$ is a function $k(\cdot, \cdot) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$.*

Kernel methods are in fact based on a particular subset of kernel functions called positive definite (p.d.) kernels.

Definition 5 (Positive Definite Kernel)**.** *A kernel $k(\cdot,\cdot)$ is positive definite if and only if for all $n \geq 1$ and for all $\mathbf{Y} = (Y_1, \ldots, Y_n)$ in $\mathcal{Y}^n$, the Gram matrix $\mathbf{K_Y} = \Big(k(Y_i, Y_j)\Big)_{i,j\in\{1,\ldots,n\}} \in \mathcal{M}_n(\mathbb{R})$ is symmetric and positive semi-definite.*

The use of kernel methods in statistics is due to Aronszajn's theorem [9] that proposed a characterization of p.d. kernel with respect to Hilbert spaces.

Theorem 1 (Aronszajn [9])**.** *Let $(\mathcal{Y}, \mathfrak{Y})$ a measurable space. A function $k(\cdot,\cdot) : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ is a positive definite kernel if and only if there exists a Hilbert Space $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ of functions from $\mathcal{Y}$ to $\mathbb{R}$ and a mapping $\phi(\cdot)$ from $\mathcal{Y}$ to $\mathcal{H}$ such that for $y, y' \in \mathcal{Y}$, we have:*

$$k(y, y') = \langle\phi(y), \phi(y')\rangle_{\mathcal{H}}. \tag{1.2}$$

Aronszajn's theorem shows that given a p.d. kernel $k(\cdot,\cdot)$, one can find a function $\phi(\cdot)$ and a Hilbert space $\mathcal{H}$ so that the relation (1.2) is verified. Thus, any statistical method based on inner-product evaluations can be applied on the embeddings in the Hilbert space $\mathcal{H}$ by using kernel evaluations. This is the basis of the kernel trick, as it means that the kernel function can be seen as a shortcut that avoids the explicit determination of the embeddings of the observations to compute their inner product in the feature space. Moreover, when the feature space is infinite dimensional, the explicit determination of the embeddings is impossible. The popularity of kernel methods relies on the fact that the algorithms of most statistical methods can be rewritten with inner-product evaluations only. Then kernelizing a statistical method consists in replacing every inner product of its algorithm by a kernel evaluation. When the input space $\mathcal{Y}$ is not endowed with an inner-product, it suffices to have a p.d. kernel that measures pair-wise similarities between elements of $\mathcal{Y}$ to be able to run a statistical methods on the embeddings, thus kernel methods also allow to do statistics on any kind of data [129].

Aronszajn's theorem does not state the unicity of the Hilbert space and feature map associated to a p.d. kernel, and indeed, a p.d. kernel may be associated to several Hilbert spaces. To avoid any risk of confusion, we need to associate each p.d. kernel to a unique Hilbert space. To do so, we introduce the notion of reproducing kernel Hilbert space (RKHS). The following definition comes from [125].

Definition 6 (Reproducing Kernel Hilbert Space)**.** *Let $(\mathcal{Y}, \mathfrak{Y})$ a non-empty space and $\mathcal{H}$ a Hilbert space of functions from $\mathcal{Y}$ to $\mathbb{R}$. The space $\mathcal{H}$ is a reproducing kernel Hilbert*

space if and only if there exists a function $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ *such that for* $h \in \mathcal{H}$*, we have the reproducing property:*

$$h(y) = \langle h, \phi(y) \rangle_{\mathcal{H}} , \tag{1.3}$$

where $\phi(y) = k(y, \cdot)$*, and* $\mathcal{H}$ *is the completion*[1] *of the image of* $k : y \in \mathcal{Y} \mapsto \phi(y)$*, i.e.*

$$\mathcal{H} = \overline{\text{Span}\left(\phi(y),\ y \in \mathcal{Y}\right)}.$$

It can be easily shown that the kernel associated to a RKHS is unique and it is called the reproducing kernel. Note that it is sufficient to say that for any $y \in \mathcal{Y}$, the linear operator $A_y : h \in \mathcal{H} \mapsto h(y) \in \mathbb{R}$ is bounded[2] to define a RKHS. Then the feature map $\phi(\cdot)$ is defined thanks to the Riesz representation theorem.

Theorem 2 (Riesz representation theorem)**.** *If* $L : \mathcal{H} \to \mathbb{R}$ *is a bounded linear operator, then there exists a unique element* $\ell \in \mathcal{H}$ *such that for all* $h \in \mathcal{H}$*, we have:*

$$Lh = \langle h, \ell \rangle_{\mathcal{H}} .$$

Indeed, for any $y \in \mathcal{Y}$, the unique element associated to A_y is the feature map $\phi(y)$, and we obtain the reproducing property. For $h \in \mathcal{H}$, we have:

$$A_y(h) = \langle \phi(y), h \rangle_{\mathcal{H}} .$$

The element $\phi(y)$ is the unique representer of y in $\mathcal{H}$, this is the reason why it is considered as the embedding of y in $\mathcal{H}$ [100]. Then the kernel function follows from the application of $\phi(y)$ to another element $y' \in \mathcal{Y}$:

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

The following theorem states that p.d. kernels are reproducing kernels.

[1] The completion $\overline{\mathcal{G}}$ of a space $\mathcal{G}$ is the space obtained by adding all the limits of Cauchy sequences to $\mathcal{G}$.

[2] A linear operator L from $\mathcal{H}$ to $\mathcal{H}$ is said to be bounded if there exist a constant $C > 0$ such that forall $h \in \mathcal{H}$, we have $\|Lh\|_{\mathcal{H}} \leq C \|h\|_{\mathcal{H}}$.

Theorem 3 (Positive Definite Kernels are Reproducing Kernels)**.** *Let $(\mathcal{Y}, \mathfrak{Y})$ a measurable space and $k(\cdot,\cdot)$ a kernel on $\mathcal{Y}$. The kernel $k(\cdot,\cdot)$ is a positive definite kernel if and only if it is a reproducing kernel.*

As the reproducing kernel of a RKHS is unique, any p.d. kernel is associated to a unique RKHS and conversely, any RKHS is associated to a unique p.d. kernel. An important result in kernel methods that we do not use here is the Representer theorem.

Theorem 4. *Let $k(\cdot,\cdot)$ be a p.d. kernel on a measurable space $(\mathcal{Y}, \mathfrak{Y})$ and $\mathcal{H}$ the associated RKHS. Let $Y_1, \ldots, Y_n \in \mathcal{Y}$ and $R : \mathbb{R}^{n+1} \to \mathbb{R}$ a function strictly increasing with respect to its $(n+1)^{th}$ variable. Then the minimum of the following empirical risk:*

$$h^\star = \min_{h \in \mathcal{H}} R\Big(h(Y_1), \ldots, h(Y_n), \|h\|_{\mathcal{H}}\Big),$$

belongs to $\text{Span}\Big(\phi(Y_1), \ldots, \phi(Y_n)\Big)$*, the span of the embeddings associtated to $Y_1, \ldots, Y_n$.*

We notify this theorem as it is a central and practical results for kernel methods in general because it allows to solve any optimization problem as a linear combination of the embeddings. However we will not need this theorem in this manuscript as every quantity defined as the solution of an optimization problem can be determined explicitly.

1.2 Linking Kernels with Probability Distributions

Let $(\mathcal{Y}, \mathfrak{Y})$ a measurable space and $k(\cdot,\cdot)$ a p.d. kernel on $\mathcal{Y}$ associated to the separable RKHS $\mathcal{H}$ and the feature map $\phi(\cdot)$. As there exists a unique representer $\phi(y)$ of $y \in \mathcal{Y}$ called the embedding of y, we can define a unique representer of any probability distribution $\mathbb{P}$ defined on $\mathcal{Y}$. This representer is called the kernel mean embedding of $\mathbb{P}$ and is denoted μ. The kernel mean embedding is defined through the Riesz representation theorem and is equal to the expectation of $\phi(Y) \in \mathcal{H}$ where $Y \sim \mathbb{P}$. Following the same idea, we can also define a unique representer of the probability distribution $\mathbb{P}$ in $\text{HS}(\mathcal{H})$, the Hilbert space of Hilbert-Schmidt operators from $\mathcal{H}$ to $\mathcal{H}$. This representer is called the kernel covariance operator and is equal to the covariance operator of $\phi(Y)$ under $\mathbb{P}$ and is denoted Σ. Then, for a statistical sample $Y_1, \ldots, Y_n \in \mathcal{Y}$ from $\mathbb{P}$, we can define the empirical kernel mean embedding $\widehat{\mu}$ and the empirical kernel covariance operator $\widehat{\Sigma}$. These quantities allow us to develop statistical methods in the RKHS.

The first part of this section is dedicated to the embedding of distributions in $\mathcal{H}$ and HS($\mathcal{H}$), and the second part is dedicated to the definition of their empirical estimators.

1.2.1 Kernel Mean Embedding and Kernel Covariance Operator

The existence and unicity of the kernel mean embedding and kernel covariance operator rely on the Riesz representation theorem.

Kernel Mean Embedding

The definition of the kernel mean embedding relies on the definition of an appropriate bounded linear functional to apply the Riesz representation theorem.

Proposition 1. *If $\mathbb{E}_{\mathbb{P}}\sqrt{k(Y,Y)} < +\infty$, then there exists a unique element $\mu \in \mathcal{H}$ such that for all $h \in \mathcal{H}$, we have:*

$$\mathbb{E}_{\mathbb{P}}\Big(h(Y)\Big) = \langle h, \mu \rangle_{\mathcal{H}}\,. \tag{1.4}$$

Proof. Consider the operator:

$$\begin{aligned} \mathcal{H} &\longrightarrow \mathbb{R} \\ \mathbb{E}_{\mathbb{P}} : h &\longmapsto \mathbb{E}_{\mathbb{P}}\Big(h(Y)\Big), \end{aligned}$$

where $\mathbb{E}_{\mathbb{P}}\Big(h(Y)\Big) = \int_{y\in\mathcal{Y}} h(y)d\mathbb{P}(y)$. The function $\mathbb{E}_{\mathbb{P}}$ is linear by the linearity of the expectation. The reproducing property and the Cauchi-Schwarz inequality allow to show that it is bounded, for $h \in \mathcal{H}$, we have:

$$\begin{aligned} |\mathbb{E}_{\mathbb{P}}\Big(h(Y)\Big)| \leq& \mathbb{E}_{\mathbb{P}}\Big(\,|\,\langle h, \phi(Y)\rangle_{\mathcal{H}}\,|\,\Big) \\ \leq& \mathbb{E}_{\mathbb{P}}\Big(\,\|h\|_{\mathcal{H}}\,\|\phi(Y)\|_{\mathcal{H}}\,\Big) \\ \leq& \|h\|_{\mathcal{H}}\,\mathbb{E}_{\mathbb{P}}\sqrt{k(Y,Y)}. \end{aligned}$$

This upper bound is finite by assumption. According to the Riesz representation theorem, there exists a unique $\mu \in \mathcal{H}$ such that for $h \in \mathcal{H}$, we have:

$$\mathbb{E}_{\mathbb{P}}\Big(h(Y)\Big) = \langle h, \mu \rangle_{\mathcal{H}}\,.$$

□

The assumption $\mathbb{E}\sqrt{k(Y,Y)} < +\infty$ is needed to ensure the existence of the kernel mean embedding. It is common to work with a bounded kernel function ($\sup_{y\in\mathcal{Y}} k(y,y) < +\infty$) in order to have $\mathbb{E}_{\mathbb{P}}\Big(k(Y,Y)^{\beta}\Big) < +\infty$ for any $\beta \geq 1$. Equation (1.4) is a reproducing property on the evaluation of the expectation, thus the kernel mean embedding μ of $\mathbb{P}$ can be interpreted as the embedding of $\mathbb{P}$ in $\mathcal{H}$. Note that the kernel mean embedding is also the expectation of the embedded random variable $\phi(Y)$. Indeed, for $h \in \mathcal{H}$, we have:

$$\mathbb{E}_{\mathbb{P}}\Big(h(Y)\Big) = \mathbb{E}_{\mathbb{P}}\Big(\,\langle h, \phi(Y)\rangle_{\mathcal{H}}\,\Big) = \Big\langle h, \mathbb{E}_{\mathbb{P}}\Big(\phi(Y)\Big)\Big\rangle_{\mathcal{H}}.$$

Thus, by unicity of μ, we conclude that $\mu = \mathbb{E}_{\mathbb{P}}\Big(\phi(Y)\Big)$. Note that this quantity is well defined as a Bochner integral.

Covariance Operator

Similarly to the kernel mean embedding, the kernel covariance operator is the covariance operator of the embedded random variable $\phi(Y)$.

Proposition 2 (Kernel Covariance Operator)**.** *If $\mathbb{E}_{\mathbb{P}}\Big(k(Y,Y)\Big) < +\infty$, then there exists a unique Hilbert-Schmidt operator $\Sigma \in \mathrm{HS}(\mathcal{H})$ such that for $g, h \in \mathcal{H}$, we have:*

$$\mathrm{Cov}\,\Big(g(Y), h(Y)\Big) = \langle g, \Sigma h\rangle_{\mathcal{H}}$$

Proof. Let $g, h \in \mathcal{H}$, observe that:

$$\begin{aligned}
\mathrm{Cov}\,\Big(g(Y), h(Y)\Big) =& \mathbb{E}_{\mathbb{P}}\Bigg(\Big(g(Y) - \mathbb{E}_{\mathbb{P}}\Big(g(Y)\Big)\Big)\Big(h(Y) - \mathbb{E}_{\mathbb{P}}\Big(h(Y)\Big)\Big)\Bigg)\\
=& \mathbb{E}_{\mathbb{P}}\Bigg(\Big\langle g, \phi(Y) - \mu\Big\rangle_{\mathcal{H}}\Big\langle h, \phi(Y) - \mu\Big\rangle_{\mathcal{H}}\Bigg)\\
=& \mathbb{E}_{\mathbb{P}}\Bigg(\Big\langle \Big(g \otimes h\Big)(\phi(Y) - \mu), \phi(Y) - \mu\Big\rangle_{\mathcal{H}}\Bigg)
\end{aligned}$$

Then we define $A_{\mathrm{cov}} : g \otimes h \in \mathrm{HS}(\mathcal{H}) \mapsto \mathrm{Cov}\,\Big(g(Y), h(Y)\Big) \in \mathbb{R}$. The operator A_{cov} is

bilinear by bilinearity of the covariance. We also have:

$$\begin{aligned}
|A_{\text{cov}}(g,h)| =& \left|\mathbb{E}_{\mathbb{P}}\Big(\Big(g(Y)-\mathbb{E}_{\mathbb{P}}\big(g(Y)\big)\Big)\Big(h(Y)-\mathbb{E}_{\mathbb{P}}\big(h(Y)\big)\Big)\Big)\right| \\
\leq& \left|\mathbb{E}_{\mathbb{P}}\Big(\big\langle g,\phi(Y)-\mu\big\rangle_{\mathcal{H}}\big\langle h,\phi(Y)-\mu\big\rangle_{\mathcal{H}}\Big)\right| \\
\leq& \|g\|_{\mathcal{H}}\|h\|_{\mathcal{H}}\,\mathbb{E}_{\mathbb{P}}\|\phi(Y)-\mu\|_{\mathcal{H}}^2 \\
\leq& \|g\|_{\mathcal{H}}\|h\|_{\mathcal{H}}\,\mathbb{E}_{\mathbb{P}}\|\phi(Y)\|_{\mathcal{H}}^2 .
\end{aligned}$$

As $\|\phi(Y)\|_{\mathcal{H}}^2 = \langle\phi(Y),\phi(Y)\rangle_{\mathcal{H}} = k(Y,Y)$, the upper-bound is finite by hypothesis and the operator A_{cov} is bounded. We apply the Riesz representation theorem on the bounded linear operator A_{cov}, there exists a unique Hilbert-Schmidt operator Σ such that:

$$\begin{aligned}
\text{Cov}\Big(g(Y),h(Y)\Big) &= \langle\Sigma, g\otimes h\rangle_{\text{HS}(\mathcal{H})} \\
&= \langle g,\Sigma h\rangle_{\mathcal{H}} .
\end{aligned}$$

□

The kernel covariance operator has several properties that will be used for the proofs of this manuscript. It is self-adjoint, thus for any $g,h\in\mathcal{H}$:

$$\langle g,\Sigma h\rangle_{\mathcal{H}} = \langle\Sigma g,h\rangle_{\mathcal{H}} .$$

It is trace-class, as for any orthonormal basis $(e_s)_{s\geq 1}$ of $\mathcal{H}$, $\sum_{s\geq 1}\langle e_s,\Sigma e_s\rangle_{\mathcal{H}} < +\infty$. As a positive self-adjoint Hilbert-Schmidt operator, the kernel covariance operator can be diagonalized, i.e. there exists an orthonormal basis $(f_s^{\mathbb{P}})_{s\geq 1}$ of eigenfunctions of Σ in $\mathcal{H}$ and a sequence of decreasing and non-negative eigenvalues $(\lambda_s^{\mathbb{P}})_{s\geq 1}$ with $\sum_{s\geq 1}\lambda_s^{\mathbb{P}} < +\infty$ such that:

$$\Sigma = \sum_{s\geq 1}\lambda_s^{\mathbb{P}}(f_s^{\mathbb{P}}\otimes f_s^{\mathbb{P}}).$$

The eigenfunctions of Σ associated to the highest eigenvalues can be interpreted as the directions of the feature space that carry the largest part of the variance of the embedded random variable $\phi(Y)$. Thus, the eigen-decomposition of a covariance operator can be used to focus on a low dimensional subspace of the feature space that carry a large proportion of the variability of $\phi(Y)$.

1.2.2 Empirical Aspects and Kernel Trick

Assume that $\mathbb{E}k(Y,Y) < +\infty$. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ in $\mathcal{Y}$ a set of n i.i.d. random variables from $\mathbb{P}$. All the quantities presented above have their empirical counterparts with respect to $\mathbf{Y}$. Here, we present the expressions of the empirical estimators $\widehat{\mu}$ and $\widehat{\Sigma}$ of μ and Σ respectively, with respect to $\mathbf{Y}$. Then, we present how to determine the eigenvalues $(\widehat{\lambda}_s^{\mathbb{P}})_{s\in\{1,\ldots,n\}}$ and associated eigenfunctions $(\widehat{f}_s^{\mathbb{P}})_{s\in\{1,\ldots,n\}}$ of $\widehat{\Sigma}$ with respect to $\mathbf{Y}$, which are the empirical estimators of $(\lambda_s^{\mathbb{P}})_{s\geq 1}$ and $(f_s^{\mathbb{P}})_{s\geq 1}$ respectively.

Empirical Mean Embedding and Covariance Operator

The empirical kernel mean embedding of $\mathbb{P}$ associated to $\mathbf{Y}$ is defined such that:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\phi(Y_i). \tag{1.5}$$

This is an unbiased estimator of the kernel mean embedding μ. The empirical kernel covariance operator of $\mathbb{P}$ associated to $\mathbf{Y}$ is defined such that:

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\Big(\phi(Y_i) - \widehat{\mu}\Big)^{\otimes 2}. \tag{1.6}$$

This is a biased estimator of the kernel covariance operator Σ. An unbiased estimator would be obtained by replacing the factor n^{-1} by a factor $(n-1)^{-1}$.

Diagonalization of the Empirical Kernel Covariance Operator

Assume that the non-zero eigenvalues of $\widehat{\Sigma}$ are distinct for simplicity. To diagonalize $\widehat{\Sigma}$, we define $\Phi(\mathbf{Y}) = (\phi(Y_1), \ldots, \phi(Y_n))' \in \mathcal{H}^n$. In this approach, we generalize matrix algebra and consider elements of $\mathcal{H}$ as vectors even if they are infinite dimensional. This approach is adapted from [62]. By an abuse of notation, we have that:

$$\Phi(\mathbf{Y})'\Phi(\mathbf{Y}) = \begin{pmatrix}\phi(Y_1) & \ldots & \phi(Y_n)\end{pmatrix}\begin{pmatrix}\phi(Y_1)' \\ \vdots \\ \phi(Y_n)'\end{pmatrix} = \sum_{i=1}^{n}\phi(Y_i)\otimes\phi(Y_i),$$

and that:

$$\Phi(\mathbf{Y})\Phi(\mathbf{Y})' = \begin{pmatrix} \phi(Y_1)' \\ \vdots \\ \phi(Y_n)' \end{pmatrix} \begin{pmatrix} \phi(Y_1) & \dots & \phi(Y_n) \end{pmatrix} = \underbrace{\begin{pmatrix} \langle \phi(Y_1), \phi(Y_1) \rangle_{\mathcal{H}} & \dots & \langle \phi(Y_1), \phi(Y_n) \rangle_{\mathcal{H}} \\ \vdots & \ddots & \vdots \\ \langle \phi(Y_n), \phi(Y_1) \rangle_{\mathcal{H}} & \dots & \langle \phi(Y_n), \phi(Y_n) \rangle_{\mathcal{H}} \end{pmatrix}}_{\mathbf{K_Y}},$$

where $\phi(Y)'$ is defined such that $\phi(Y)'\phi(Y) = \langle \phi(Y), \phi(Y) \rangle_{\mathcal{H}}$ and $\phi(Y)\phi(Y)' = \phi(Y) \otimes \phi(Y)$. Moreover, for any matrix $\mathbf{M} = (m_{i,j})_{i,j \in \{1,\dots,n\}} \in \mathcal{M}_n(\mathbb{R})$, then $\mathbf{M}\Phi(\mathbf{Y}) \in \mathcal{H}^n$ and its i^{th} element is such that $\Big(\mathbf{M}\Phi(\mathbf{Y})\Big)_i = \sum_{j=1}^n m_{i,j}\phi(Y_j) \in \mathcal{H}$. The first step of this approach is to write the operator to diagonalize with respect to $\Phi(\mathbf{Y})$. According to Equation (1.6), we directly have:

$$\widehat{\Sigma} = \frac{1}{n}\Big(\mathbf{\Pi}_n\Phi(\mathbf{Y})\Big)'\Big(\mathbf{\Pi}_n\Phi(\mathbf{Y})\Big),$$

where $\mathbf{\Pi}_n = \mathbf{I}_n - n^{-1}\mathbf{J}_n \in \mathcal{M}_n(\mathbb{R})$, with $\mathbf{I}_n$ and $\mathbf{J}_n$ being the identity matrix and the matrix full of ones of $\mathcal{M}_n(\mathbb{R})$. Let $\widehat{f}^{\mathbb{P}}$ be an eigenfunction of $\widehat{\Sigma}$ associated to the eigenvalue $\widehat{\lambda}^{\mathbb{P}}$, we have:

$$\frac{1}{n}\Big(\mathbf{\Pi}_n\Phi(\mathbf{Y})\Big)'\Big(\mathbf{\Pi}_n\Phi(\mathbf{Y})\Big)\widehat{f}^{\mathbb{P}} = \widehat{\lambda}^{\mathbb{P}}\widehat{f}^{\mathbb{P}}. \tag{1.7}$$

Now we force the appearance of $\mathbf{K_Y} = \Phi(\mathbf{Y})\Phi(\mathbf{Y})'$ in Equation (1.7). For instance, we can multiply both sides of the equation by $\mathbf{\Pi}_n\Phi(\mathbf{Y})$ on the left to obtain:

$$\frac{1}{n}\mathbf{\Pi}_n\mathbf{K_Y}\mathbf{\Pi}_n\Big(\mathbf{\Pi}_n\Phi(\mathbf{Y})\widehat{f}^{\mathbb{P}}\Big) = \widehat{\lambda}^{\mathbb{P}}\Big(\mathbf{\Pi}_n\Phi(\mathbf{Y})\widehat{f}^{\mathbb{P}}\Big), \tag{1.8}$$

where:

$$\mathbf{\Pi}_n\Phi(\mathbf{Y})\widehat{f}^{\mathbb{P}} = \begin{pmatrix} \Big(\phi(Y_1) - \widehat{\mu}\Big)' \\ \vdots \\ \Big(\phi(Y_n) - \widehat{\mu}\Big)\prime \end{pmatrix} (\widehat{f}^{\mathbb{P}}) = \begin{pmatrix} \Big\langle \phi(Y_1) - \widehat{\mu}, \widehat{f}^{\mathbb{P}} \Big\rangle_{\mathcal{H}} \\ \vdots \\ \Big\langle \phi(Y_n) - \widehat{\mu}, \widehat{f}^{\mathbb{P}} \Big\rangle_{\mathcal{H}} \end{pmatrix} \in \mathbb{R}^n.$$

The matrix $\mathbf{K}_\Sigma = n^{-1}\mathbf{\Pi}_n\mathbf{K_Y}\mathbf{\Pi}_n \in \mathcal{M}_n(\mathbb{R})$ is the matrix to diagonalize and for each eigenfunction $\widehat{f}^{\mathbb{P}} \in \mathcal{H}$ of $\widehat{\Sigma}$, the vector $\mathbf{\Pi}_n\Phi(\mathbf{Y})\widehat{f}^{\mathbb{P}} \in \mathbb{R}^n$ is an eigenvector of the matrix $\mathbf{K}_\Sigma$ associated to the same eigenvalue $\widehat{\lambda}^{\mathbb{P}}$. Since this is true for any eigenfunction $\widehat{f}^{\mathbb{P}}$ associated to an eigenvalue $\widehat{\lambda}^{\mathbb{P}}$ of $\widehat{\Sigma}$, it shows that the spectrum of $\widehat{\Sigma}$ is included in the

spectrum of $\mathbf{K}_\Sigma$.

The second step consists in showing that the spectrum of $\mathbf{K}_\Sigma$ is included in the spectrum of $\widehat{\Sigma}$ and at the same time obtaining an expression of any eigenfunction $\widehat{f}^{\mathbb{P}}$ of $\widehat{\Sigma}$ with respect to the unit eigenvector u associated to the same eigenvalue. Let $u = (u_1, \ldots, u_n)' \in \mathbb{R}^n$ an eigenvector of $\mathbf{K}_\Sigma$ associated to the eigenvalue $\widehat{\lambda}^{\mathbb{P}}$. We have that:

$$\frac{1}{n}\Big(\mathbf{\Pi}_n\Phi(\mathbf{Y})\Big)\Big(\mathbf{\Pi}_n\Phi(\mathbf{Y})\Big)'u = \widehat{\lambda}^{\mathbb{P}}u.$$

We force the appearance of $\widehat{\Sigma}$ by multiplying both sides by $\Big(\mathbf{\Pi}_n\Phi(\mathbf{Y})\Big)'$ on the left:

$$\widehat{\Sigma}\Big(\mathbf{\Pi}_n\Phi(\mathbf{Y})\Big)'u = \widehat{\lambda}^{\mathbb{P}}\Big(\mathbf{\Pi}_n\Phi(\mathbf{Y})\Big)'u. \tag{1.9}$$

Thus for each eigenvector u of $\mathbf{K}_\Sigma$, the function $\Big(\mathbf{\Pi}_n\Phi(\mathbf{Y})\Big)'u \in \mathcal{H}$ is an eigenfunction of $\widehat{\Sigma}$ associated to the same eigenvalue $\widehat{\lambda}^{\mathbb{P}}$. In addition the spectrum of $\mathbf{K}_\Sigma$ is included in the spectrum of $\widehat{\Sigma}$, thus they both share the same spectrum. We have the following explicit formulations:

$$\mathbf{K}_\Sigma = \left(\frac{1}{n}\Big\langle \phi(Y_i) - \widehat{\mu}, \phi(Y_j) - \widehat{\mu}\Big\rangle_{\mathcal{H}}\right)_{i,j\in\{1,\ldots,n\}} \in \mathcal{M}_n(\mathbb{R}),$$

$$\mathbf{\Pi}_n\Phi(\mathbf{Y})'u = \Big(\phi(Y_1) - \widehat{\mu} \quad \ldots \quad \phi(Y_n) - \widehat{\mu}\Big)\begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = \sum_{i=1}^{n} u_i\Big(\phi(Y_i) - \widehat{\mu}\Big) \in \mathcal{H}.$$

Note that these expressions are written with respect to the embeddings $\phi(Y)$ and can thus be used in further kernel tricks. In practice, we determine the matrix to diagonalize $\mathbf{K}_\Sigma$ to obtain the eigenvalues $\widehat{\lambda}_1^{\mathbb{P}}, \ldots, \widehat{\lambda}_n^{\mathbb{P}} \in \mathbb{R}$ of both $\widehat{\Sigma}$ and $\mathbf{K}_\Sigma$, and the eigenvectors $u_1, \ldots, u_n \in \mathbb{R}^n$ of $\mathbf{K}_\Sigma$. Then we replace every appearance of the eigenfunctions $\widehat{f}_1^{\mathbb{P}}, \ldots, \widehat{f}_n^{\mathbb{P}}$ of $\widehat{\Sigma}$ by their expression with respect to the eigenvectors $u_1, \ldots, u_n$.

Although this matrix formalism is justified because the RKHS $\mathcal{H}$ is separable, some argue that an analytic formalism is more suited and rigorous to study elements of an infinite dimensional Hilbert space [22]. Thus, we confirm the results with the analytic

formalism. Let $\widehat{f}^{\mathbb{P}}$ be an eigenfunction of $\widehat{\Sigma}$ associated to the eigenvalue $\widehat{\lambda}^{\mathbb{P}}$, we have:

$$\widehat{\Sigma}\widehat{f}^{\mathbb{P}} = \frac{1}{n}\sum_{i=1}^{n}\Big\langle \phi(Y_i) - \widehat{\mu}, \widehat{f}^{\mathbb{P}}\Big\rangle_{\mathcal{H}}\Big(\phi(Y_i) - \widehat{\mu}\Big) = \widehat{\lambda}^{\mathbb{P}}\widehat{f}^{\mathbb{P}}$$

We consider $u = \mathbf{\Pi}_n\Phi(\mathbf{Y})\widehat{f}^{\mathbb{P}} = \left(\left\langle \phi(Y_1) - \widehat{\mu}, \widehat{f}^{\mathbb{P}}\right\rangle_{\mathcal{H}}, \ldots, \left\langle \phi(Y_n) - \widehat{\mu}, \widehat{f}^{\mathbb{P}}\right\rangle_{\mathcal{H}}\right)' \in \mathbb{R}^n$. The i^{th} coordinate of $\mathbf{K}_\Sigma u$ is such that:

$$\begin{aligned}
(\mathbf{K}_\Sigma u)_i &= \sum_{j=1}^{n}\frac{1}{n}\Big\langle \phi(Y_i) - \widehat{\mu}, \phi(Y_j) - \widehat{\mu}\Big\rangle_{\mathcal{H}}\Big\langle \phi(Y_j) - \widehat{\mu}, \widehat{f}^{\mathbb{P}}\Big\rangle_{\mathcal{H}} \\
&= \left\langle \phi(Y_i) - \widehat{\mu}, \frac{1}{n}\sum_{j=1}^{n}\Big\langle \phi(Y_j) - \widehat{\mu}, \widehat{f}^{\mathbb{P}}\Big\rangle_{\mathcal{H}}\Big(\phi(Y_j) - \widehat{\mu}\Big)\right\rangle_{\mathcal{H}} \\
&= \Big\langle \phi(Y_i) - \widehat{\mu}, \widehat{\lambda}^{\mathbb{P}}\widehat{f}^{\mathbb{P}}\Big\rangle_{\mathcal{H}} \\
&= \widehat{\lambda}^{\mathbb{P}}u_i
\end{aligned}$$

It confirms Equation (1.8).

Oppositely, let $u = (u_1, \ldots, u_n)$ in $\mathbb{R}^n$ an eigenvector of $\mathbf{K}_\Sigma$ associated to the eigenvalue $\widehat{\lambda}^{\mathbb{P}}$ and $\widehat{f}^{\mathbb{P}} = \mathbf{\Pi}_n\Phi(\mathbf{Y})'u = \sum_{i=1}^{n} u_i\Big(\phi(Y_i) - \widehat{\mu}\Big)$. The eigen-relation $\mathbf{K}_\Sigma u = \widehat{\lambda}^{\mathbb{P}}u$ gives that for $i \in \{1, \ldots, n\}$ we have the relation:

$$\widehat{\lambda}^{\mathbb{P}}u_i = \frac{1}{n}\sum_{j=1}^{n}\Big\langle \phi(Y_i) - \widehat{\mu}, \phi(Y_j) - \widehat{\mu}\Big\rangle_{\mathcal{H}}u_j.$$

We then have:

$$\begin{aligned}
\widehat{\Sigma}\widehat{f}^{\mathbb{P}} =& \frac{1}{n}\sum_{i=1}^{n}\Big(\Big(\phi(Y_i) - \widehat{\mu}\Big) \otimes \Big(\phi(Y_i) - \widehat{\mu}\Big)\Big)\Big(\sum_{j=1}^{n}u_j\Big(\phi(Y_j) - \widehat{\mu}\Big)\Big) \\
=& \sum_{i=1}^{n}\underbrace{\left(\frac{1}{n}\sum_{j=1}^{n}\Big\langle \phi(Y_i) - \widehat{\mu}, \phi(Y_j) - \widehat{\mu}\Big\rangle_{\mathcal{H}}u_j\right)}_{=\widehat{\lambda}^{\mathbb{P}}u_i}\Big(\phi(Y_i) - \widehat{\mu}\Big) \\
=& \widehat{\lambda}^{\mathbb{P}}\widehat{f}^{\mathbb{P}}
\end{aligned}$$

This confirms Equation (1.9). Observe that:

$$\begin{aligned}\|\mathbf{\Pi}_n\Phi(\mathbf{Y})'u\|_{\mathcal{H}}^2 &= \sum_{i,j=1}^{n} u_i u_j \left\langle \phi(Y_i) - \widehat{\mu}, \phi(Y_j) - \widehat{\mu} \right\rangle_{\mathcal{H}} \\ &= n u' \mathbf{K}_\Sigma u \\ &= n\widehat{\lambda}^{\mathbb{P}} \|u\|_2^2 \\ &= n\widehat{\lambda}^{\mathbb{P}}.\end{aligned}$$

Thus the function $(\widehat{\lambda}^{\mathbb{P}} n)^{-\frac{1}{2}} \mathbf{\Pi}_n \Phi(\mathbf{Y})' u \in \mathcal{H}$ is a unit eigenfunction of $\widehat{\Sigma}$ associated to the eigenvalue $\widehat{\lambda}^{\mathbb{P}}$.

The embedding of a probability distribution is at the root of many kernel tests. In particular, the idea behind the Maximum Mean Discrepancy (MMD) statistic is to consider the distance between two kernel mean embedding in Hilbert norm as a meaningful measure of the distance between two probability distribution. The KFDA statistic go even further by using the kernel covariance operator to compute a normalized distance between two kernel mean embeddings as another insightful distance between two probability distributions. This two kernel statistics are used to perform powerful and computationally efficient non-parametric two-sample tests, as presented in the next section.

1.3 Kernel Testing

Some statistical methods use the kernel mean embedding as a way to study the underlying probability distribution. In particular, for two-sample testing that consists in comparing two probability distributions, some kernel tests propose to directly compare the kernel mean embeddings. When the kernel mean embedding is obtained through a particular type of kernel called characteristic kernel, it characterizes the probability distribution. This allows to develop non-asymptotic two-sample tests that do not need any assumption on the probability distribution, which is particularly useful when the analysed data are too complex and difficult to model with simple distributions, as it is the case for single-cell data.

The first part of this section is dedicated to the presentation of non-parametric two-sample tests and it will recall some basic notions about hypothesis testing. In a second

part, we will present kernel testing and in particular the Maximum Mean Discrepancy test, that is the most famous kernel test. The third section focuses on a test derived from the Kernel Fisher Discriminant Analysis (KFDA), the KFDA test, that can be seen as a studentization of the MMD test. This presentation of the KFDA test constitute the building block of the next chapter that is focused on the application of the KFDA test in practice.

In this section, we consider two i.i.d. samples $\mathbf{Y}_1 = (Y_{1,1}, \ldots, Y_{1,n_1})$ and $\mathbf{Y}_2 = (Y_{2,1}, \ldots, Y_{2,n_2})$ in $\mathcal{Y}$, drawn from probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ respectively. We have n_1 observations in $\mathbf{Y}_1$ and n_2 observations in $\mathbf{Y}_2$, with $n = n_1 + n_2$. Let $k(\cdot, \cdot)$ a p.d. kernel and $\mathcal{H}$ the associated RKHS. For simplicity, we assume that the kernel $k(\cdot, \cdot)$ is bounded to ensure the existence of every quantity of interest such as kernel mean embeddings and kernel covariance operators, however, this assumption will be reminded when it is necessary. For $y \in \mathcal{Y}$ we denote $\phi(y) = k(y, \cdot)$ the feature map of y with respect to the kernel $k(\cdot, \cdot)$. Let μ_1 and μ_2 be the kernel mean embedding of $\mathbb{P}_1$ and $\mathbb{P}_2$ respectively, and $\widehat{\mu}_1$ and $\widehat{\mu}_2$ are their empirical counterparts defined with respect to $\Phi(\mathbf{Y}_1) = \Big(\phi(Y_{1,1}), \ldots, \phi(Y_{1,n_1})\Big)$ and $\Phi(\mathbf{Y}_2) = \Big(\phi(Y_{2,1}), \ldots, \phi(Y_{2,n_2})\Big)$. Let Σ_1 and Σ_2 be the kernel covariance operator associated to $\mathbb{P}_1$ and $\mathbb{P}_2$ respectively, and $\widehat{\Sigma}_1$ and $\widehat{\Sigma}_2$ be their respective empirical counterpart.

1.3.1 Hypothesis Testing

The two sample problem consists in assessing if the two observed samples $\mathbf{Y}_1$ and $\mathbf{Y}_2$ in $\mathcal{Y}$ drawn from probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ are different enough to consider that the underlying distributions are different [82]. This problem is generally adressed with the framework of hypothesis testing. The null and alternative hypothesis are such that:

$$H_0 : \mathbb{P}_1 = \mathbb{P}_2 \quad \text{versus} \quad H_1 : \mathbb{P}_1 \neq \mathbb{P}_2.$$

Existing Two-Sample Tests

The approaches tackling this problem can be separated in two types of approaches. When assumption are made on the form of the probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, the approach is parametric.

The most common assumption in parametric two-sample testing is to consider that the two distributions are Gaussian. The historical parametric two-sample test are the Student t-test [135] for univariate observations, and its multivariate generalization, the Hotelling T^2 test [66]. These are parametric two-sample tests that assume that the two distributions are univariate Gaussian and multivariate Gaussian resectively, and that share the same covariance structure. The Hotelling T^2 has recently been generalized to Hilbert spaces [126, 108].

A non-parametric two-sample test is a general testing framework supposed to be able to detect a difference between two distributions. The historical univariate non-parametric test are the Kolmogorov-Smirnov test [132, 93], the Wald-Wolfowitz run test [144], the Mann-Whitney rank sum test [90], also known as Wilkoxon rank sum test, and the Cramer-Von Mises test [5]. These tests are still considered as powerful tests to be used in practice for univariate data. In particular, the Wilkoxon rank sum test is applied in the vast majority of single-cell differential analysis [134], that is the univariate comparison of the gene expression of two samples measured through single-cell RNA sequencing.

The multivariate generalization of these classical tests suppose to define a rank on multidimensional spaces, several such tests have been proposed [146, 21]. Recently, a test based on a multivariate rank defined using optimal transport has been proposed [47]. Graph-based tests constitute a popular type of non-parametric multivariate two-sample tests. The first graph-based test was introduced by [42], and it was the first non-parametric multivariate two-sample test to be computationally efficient. The literature on graph-based tests is reviewed in detail in [16] and [74]. Another review of non-parametric two-sample tests can be found in [1].

Another class of non-parametric multivariate two-sample tests is based on the definition of a distance or a divergence between two probability distributions. Energy based tests [17, 138] and kernel tests [63] belong to this category. It has been shown that the former are a particular case of the later [127].

Performances of a Test

The response of a two-sample test is obtained by comparing the value of a test statistic computed with respect to the data to strictly positive testing threshold and to reject the

null hypothesis when the test statistic exceeds this threshold. The outcome of a test is wrong when the null hypothesis is wrongly rejected, it is called the type I error, and when the null hypothesis is wrongly accepted, it is called the type II error. The type I error rate of a test is equal to $P(\text{reject } H_0|H_0)$ and the type II error rate of a test is equal to $P(\text{accept } H_0|H_1)$. It is not possible to control for both the type I and type II error rate when constructing a testing procedure. Thus, the convention is to control for the type I error rate.

For $\alpha \in [0, 1]$, a test is a level α test if α is an upper bound on the type one error rate, such that:

$$P(\text{reject } H_0|H_0) \leq \alpha.$$

A test is said to be conservative when the type I error rate is lower than the chosen value of α, i.e. $P(\text{reject } H_0|H_0) < \alpha$. Testing procedures are compared for fixed α only. Among several testing procedures, the tests with level α can be compared, then lower is the probability of making type II errors, better is the test procedure. The power of a test is defined as the probability of a test of not making type II errors:

$$\pi = P(\text{reject } H_0|H_1) = 1 - P(\text{accept } H_0|H_1).$$

The level and power of a test may be theoretically studied or numerically estimated.

A test statistic is the evaluation of a random variable, its probability distribution differs whether H_0 is true or not. The $p - value$ of a test statistic is defined as the likelihood of it being drawn from the distribution under the null hypothesis. This value is often used to construct a test procedure, by rejecting the null hypothesis when $p - value < \alpha$ and accepting it otherwise. The second part of this section describes several approaches to construct a testing procedure based on the MMD test statistic.

1.3.2 Maximum Mean Discrepancy Test

The kernel mean embedding is at the root of many kernel tests. The key point is that for a particular set of kernel functions called characteristic kernels, the operator $A : \mathbb{P} \mapsto \mu$ is injective, thus two proability distributions having the same kernel mean embedding

are equal. The first methods that benefited from this property are the Maximum Mean Discrepancy test for two samples testing [55] and the HSIC test for independence testing [55]. Many variants of the MMD test have been proposed to tackle different issues of kernel testing. The relation between the MMD test and other families of test has been highlighted, such as the link with independence tests based on distance correlation [115].

The MMD as a Metric Between Probability Distributions

A standard approach for constructing two-sample test statistics is to define a metric between distributions. Let $\mathcal{F}$ be a space of functions from $\mathcal{Y}$ to $\mathbb{R}$ in which expectations with respect to $\mathbb{P}_1$ and $\mathbb{P}_2$ are well defined. The quantity:

$$\mathcal{M}_{\mathcal{F}}(\mathbb{P}_1, \mathbb{P}_2) = \sup_{f \in \mathcal{F}} |\mathbb{E}_{\mathbb{P}_1} f(Y) - \mathbb{E}_{\mathbb{P}_2} f(Y)|,$$

can be a metric in the space of probability measures over $\mathcal{Y}$ if it fulfills the separability property:

$$\mathbb{P}_1 = \mathbb{P}_2 \Leftrightarrow \mathcal{M}_{\mathcal{F}}(\mathbb{P}_1, \mathbb{P}_2) = 0. \tag{1.10}$$

When $\mathcal{F}$ is too general, it is difficult to compute the metric. The objective if to find a function space $\mathcal{F}$ that both makes $\mathcal{M}_{\mathcal{F}}$ a metric and where algorithms may be developed to compute $\mathcal{M}_{\mathcal{F}}$. Some well-known metric between distributions are associated to specific choices of $\mathcal{F}$.

- $\mathcal{F} = \{\mathbb{1}_{(-\infty,t]} | \quad t \in \mathbb{R}\}$ gives the Kolmogorov distance between $\mathbb{P}_1$ and $\mathbb{P}_2$.
- $\mathcal{F} = \{f | \quad \sup_{y \in \mathcal{Y}} |f(y)| \leq 1\}$ gives the total variation distance between $\mathbb{P}_1$ and $\mathbb{P}_2$.
- If ρ is a metric and $\mathcal{Y}$ is compact, $\mathcal{F} = \{f | \quad \sup_{y,y' \in \mathcal{Y}} \frac{|f(y)-f(y')|}{\rho(y,y')} \leq 1\}$ gives the Wasserstein distance between $\mathbb{P}_1$ and $\mathbb{P}_2$.

The MMD is the metric obtained when the space $\mathcal{F}$ is constrained to be included in a RKHS. When $\mathcal{F}$ is the unit ball, the metric can be determined explicitely with respect to the kernel mean embeddings of the two compared distributions:

$$\mathrm{MMD}(\mathbb{P}_1, \mathbb{P}_2) = \sup_{\substack{h \in \mathcal{H} \\ \|h\|_{\mathcal{H}}=1}} |\mathbb{E}_{\mathbb{P}_1} h(Y) - \mathbb{E}_{\mathbb{P}_2} h(Y)|.$$

By definition of the kernel mean embedding we have:

$$\mathrm{MMD}(\mathbb{P}_1, \mathbb{P}_2) = \sup_{\substack{h \in \mathcal{H} \\ \|h\|_{\mathcal{H}}=1}} |\langle h, \mu_1 - \mu_2 \rangle_{\mathcal{H}}|.$$

Then the suppremum is reached for h colinear to $\mu_1 - \mu_2$, so that:

$$\begin{aligned}\mathrm{MMD}(\mathbb{P}_1, \mathbb{P}_2) &= \left\langle \frac{\mu_1 - \mu_2}{\|\mu_1 - \mu_2\|_{\mathcal{H}}}, \mu_1 - \mu_2 \right\rangle_{\mathcal{H}} \\ &= \|\mu_1 - \mu_2\|_{\mathcal{H}}.\end{aligned}$$

The function reaching this supremum is called the MMD witness function [79].

The MMD defines a distance in the space of probability distributions on $\mathcal{Y}$ if and only if the kernel $k(\cdot, \cdot)$ is a characteristic kernel.

Definition 7 (Characteristic kernel)**.** *The p.d. kernel $k(\cdot, \cdot)$ is characteristic if the operator $A_\mu : \mathbb{P} \mapsto \mu$ is injective.*

A famous example of characteristic kernel is the Gaussian kernel, also known as radial basis function (RBF) kernel. The Gaussian kernel between two observations $y, y' \in \mathcal{Y}$ is defined such that:

$$k_{\sigma^2}(y, y') = e^{-\frac{1}{\sigma^2}\|y - y'\|_{\mathcal{Y}}^2},$$

where $\|\cdot\|_{\mathcal{Y}}$ is a norm over $\mathcal{Y}$ and $\sigma^2 > 0$ is the bandwidth parameter to be tuned. The RKHS associated to the Gaussian kernel is an infinite dimensional separable Hilbert space. Most of our analyses are done with the Gaussian kernel. When a kernel is characteristic, the separability property (1.10) of the MMD is ensured by the injectivity of A_μ. An intuitive reason for the Gaussian kernel to be characteristic is that the Taylor expansion of the exponential function in the Gaussian kernel makes the kernel mean embedding an infinite sum of every moment of the original distribution. In general, the key idea for a kernel to be characteristic is that the associated RKHS is a function space that contains functions complex enough to be able to represent any probability distribution. This property is related to the concept of universal kernels, that are beyond the scope of this manuscript. The relations between characteristic kernels and universal kernels are detailed in [100].

Testing with MMD

The true value of the MMD between $\mathbb{P}_1$ and $\mathbb{P}_2$ is unknown in general. Thus MMD tests are defined with respect to the MMD test statistic that is an empirical estimate of MMD^2:

$$\widehat{\text{MMD}}_b^2 = \|\widehat{\mu}_1 - \widehat{\mu}_2\|_{\mathcal{H}}^2 ,$$

where the subscript b stands for biased MMD. If we replace the empirical mean embeddings by their value according to Equation (1.5) and correct for the biais, we obtain an unbiased estimator of MMD^2, that is the MMD test statistic:

$$\widehat{\text{MMD}}^2 = \frac{1}{n_1(n_1-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n_1} k(Y_{1,i}, Y_{1,j}) - \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} k(Y_{1,i}, Y_{2,j}) + \frac{1}{n_2(n_2-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n_2} k(Y_{2,i}, Y_{2,j})$$

The key idea of MMD two-sample testing is to consider that when the null hypothesis is true, the estimator of the MMD should be close to zero. Given a level α, different procedures exist in order to determine a testing threshold $s_\alpha > 0$ such that we reject the null hypothesis when $\widehat{\text{MMD}}^2 > s_\alpha$.

Such a threshold can be obtained through the quantiles of the asymptotic distribution of the MMD test statistic under the null hypothesis, presented in [54].

Theorem 5. *Under H_0, we have $\mathbb{P}_1 = \mathbb{P}_2 = \mathbb{P}$. If $\sup_{y \in \mathcal{Y}} k(y,y) = M_k < +\infty$ and $\frac{n_1}{n} \underset{n,n_1 \to \infty}{\longrightarrow} \rho_1$ and $\frac{n_2}{n} \underset{n,n_2 \to \infty}{\longrightarrow} \rho_2$, then:*

$$n\widehat{\text{MMD}}^2 \underset{n \to \infty}{\xrightarrow{\mathcal{D}}} \sum_{s \geq 1} \lambda_s^{\mathbb{P}} \left((\frac{a_s}{\sqrt{\rho_1}} - \frac{b_s}{\sqrt{\rho_2}})^2 - \frac{1}{\rho_1 \rho_2} \right)$$

where $(a_s)_{s \geq 1}$ and $(b_s)_{s \geq 2}$ are two sequences of i.i.d. standard Gaussian random variables and $(\lambda_s^{\mathbb{P}})_{s \geq 1}$ is the sequence of eigenvalues of the kernel covariance operator of $\mathbb{P}$.

Some methods exist to approximate the quantiles of this asymptotic distribution in order to obtain a testing threshold. The resulting test is called an asymptotic test. The asymptotic distribution of the MMD test statistic under the null hypothesis is independant from the p.d. kernel. Typically, a non-characteristic kernel could have this asymptotic distribution under an alternative hypothesis. The asymptotic distribution may be approx-

imated with its empirical version where the theoretical eigenvalues $(\lambda_s^{\mathbb{P}})_{s\geq 1}$ are replaced by their empirical counterpart $(\widehat{\lambda}_s^{\mathbb{P}})_{s\geq 1}$, thanks to the following theorem from [51]:

Theorem 6. *Let $(a_s)_{s\geq 1}$ and $(b_s)_{s\geq 1}$ two sequences of i.i.d standard Gaussian random variables. Assume that $\sum_{s\geq 1} \lambda^{\mathbb{P}\,\frac{1}{2}} < +\infty$, then for $\frac{n_1}{n} \underset{n,n_1\to\infty}{\longrightarrow} \rho_1$ and $\frac{n_2}{n} \underset{n,n_2\to\infty}{\longrightarrow} \rho_2$, we have:*

$$\sum_{s\geq 1} \widehat{\lambda}_s^{\mathbb{P}} \left((\frac{a_s}{\sqrt{\rho_1}} - \frac{b_s}{\sqrt{\rho_2}})^2 - \frac{1}{\rho_1\rho_2} \right) \xrightarrow[n\to\infty]{D} \sum_{s\geq 1} \lambda_s^{\mathbb{P}} \left((\frac{a_s}{\sqrt{\rho_1}} - \frac{b_s}{\sqrt{\rho_2}})^2 - \frac{1}{\rho_1\rho_2} \right).$$

And we also have:

$$\sup_x \left| P(n\widehat{\mathrm{MMD}}^2 > x) - P(\sum_{s\geq 1} \widehat{\lambda}_s^{\mathbb{P}} \left((\frac{a_s}{\sqrt{\rho_1}} - \frac{b_s}{\sqrt{\rho_2}})^2 - \frac{1}{\rho_1\rho_2} \right) > x) \right| \underset{n\to\infty}{\longrightarrow} 0$$

Then it suffices to compute the eigenvalues of the empirical covariance operator associated to one of the two samples, or a combination of both, and to simulate Gaussian random variables to be able to simulate random variables following the approximated asymptotic distribution. This allows to empirically determine the testing threshold s_α for a given level α. This procedure has competitive performances, according to the authors. Moreover, it costs $O(n^3)$ to obtain the empirical eigenvalues. Some algorithms are able to compute a nice approximation of the eigenvalues for $O(n^2)$ operations, it is thus both fast and competitive. However, in practice, asymptotic testing with MMD is unpopular.

Another approach to determine a testing threshold s_α is to use a non-asymptotic large deviation bound of the MMD test statistic, as presented in [54].

Theorem 7. *Under H_0, we have $\mathbb{P}_1 = \mathbb{P}_2 = \mathbb{P}$. If $\sup_{y\in\mathcal{Y}} k(y,y) = M_k < +\infty$, then we have with probability greater than $1 - e^{-\frac{\xi^2 n_1 n_2}{2M_k n}}$:*

$$\widehat{\mathrm{MMD}} \leq \sqrt{\frac{M_k n}{n_1 n_2}} + \xi$$

We obtain a testing threshold by choosing ξ such that the result it true with probability $1-\alpha$. This concentration bound refers to the worst possible situation and the resulting test is thus very conservative. It seems to be the case here according to the simulation studies performed in [54]. This testing procedure costs $O(n^2)$ to compute the MMD test

statistic, and is then $O(1)$. However, regarding its low performance, it is also unpopular in practice.

In fact, the computational cost of the MMD test statistic is relatively low, it needs $O(n^2)$ operations to be computed. This advantage makes the permutation testing procedure the most popular. With this approach, the quantiles of the non-asymptotic distribution of the MMD test statistic are estimated through a permutation procedure, and the level α is guaranted from the procedure. The permutation procedure consists in computing $B \gg 1$ statistics on the observed data with randomly permuted labels. A dataset with randomly permuted labels contains two samples following the same distribution that is a mixture of the two original probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$. With this approach the distribution of the MMD test statistic under the null hypothesis is estimated by B observations. The testing threshold is then defined as the $(1-\alpha)$ quantile of this empirical distribution. The permutation procedure is the most popular approach for two-sample testing with MMD.

More generally, the MMD metric between distributions has been used in a wide variety of machine learning problems. For instance, transfer learning is a method that allows to adapt an already trained learning algorithm on an application (target domain) close but different from the task it was trained on (source domain). Transfer learning is promising as it is less computationally expensive than training algorithms from scratch. This task can be formulated as an optimization problem aiming to match the distributions of the source and target domains, thus the quantity to optimize has often be written with respect to the MMD metric [149, 84, 107].

Motivation for a Normalized MMD Test Statistic

The value of the MMD test statistic depends both on the distance between the true kernel mean embeddings μ_1 and μ_2, and on the variability of the embeddings. This is the reason why the asymptotic null distribution of the MMD test statistic depends on the variability of the embeddings, through the eigenvalues of the kernel covariance operator $(\lambda_s^{\mathbb{P}})_{s\geq 1}$. It can be insightful to consider these two informations separately, by comparing the two spectra on one hand, and by comparing a normalized test statistic on the other hand. This issue is well illustrated in [63], where the authors show that the orthonormal basis of eigenfunctions $(f_s^{\mathbb{P}})_{s\geq 1}$ associated to the eigenvalues $(\lambda_s^{\mathbb{P}})_{s\geq 1}$ of a kernel covariance operator Σ associated to a probability distribution $\mathbb{P}$ is well suited to study and normalize

a test statistic. Indeed, for $s \geq 1$, we have

$$\mathrm{Var}\left(\left\langle \widehat{\mu} - \mu, f_s^{\mathbb{P}} \right\rangle_{\mathcal{H}}\right) = \lambda_s^{\mathbb{P}}.$$

Thus the quantity $\widehat{\mu} - \mu$ can be normalized direction by direction on this basis. Moreover, they argue that it also allows to compare projections on different directions of the feature space. Then a studentized MMD test statistic under $H_0 : \mathbb{P}_1 = \mathbb{P}_2 = \mathbb{P}$ would have a form like this:

$$\left\langle (\widehat{\mu}_1 - \widehat{\mu}_2), \widehat{\Sigma}^{-1}(\widehat{\mu}_1 - \widehat{\mu}_2) \right\rangle_{\mathcal{H}} = \left\| \widehat{\Sigma}^{-\frac{1}{2}}(\widehat{\mu}_1 - \widehat{\mu}_2) \right\|_{\mathcal{H}}, \tag{1.11}$$

leaving aside for the moment the fact that $\widehat{\Sigma}$ is generally not invertible. Additionally, we would expect that the asymptotic null distribution of this normalized statistic does not depend on the variability of the embeddings. In fact, such a normalized test statistic may be obtained by deriving a test statistic from the KFDA classifier, namely the KFDA test statistic [64].

1.4 Kernel Fisher Discriminant Analysis for Two-Sample Testing

Now we present the Kernel Fisher Discriminant Analysis test statistic that was introduced in [64]. It is a normalized distance between the empirical kernel mean embeddings associated to $\mathbf{Y}_1$ and $\mathbf{Y}_2$, thus it can be considered as a studentized version of the MMD test statistic. According to the expression of the KFDA test statistic, it is also considered as a kernelized Hotelling T^2 test statistic. Despite these interesting links, the KFDA test statistic is named after the KFDA classifier because of the geometrical ideas underlying its construction. The fact that the KFDA test statistic is derived from a classifier makes it belong to the family of classifier two-sample tests [41, 85, 75], however, these approaches generally use the test error as a test statistic, based on the intuitive idea that a classifier should fail to discriminate a distribution from itself. The KFDA test statistic is not directly linked to the classification error, as it relies in fact on the definition of the Fisher Linear Discriminant. It has been shown that the KFDA test can be considered as an optimal kernel test in some sense [79], in particular, in the minimax sense [60]. A generalization to the k-sample test have also been proposed [15].

The Fisher Discriminant Analysis aims at finding the optimal one-dimensional axis of the space to discriminate between two samples, called the Fisher Linear Discriminant. It is usually used as a classifier as the observations are then projected on this axis and their label is predicted according to the position of their projection. The kernelized version of the FDA was introduced in [99], and it simply consists in applying the FDA to the embeddings in $\mathcal{H}$, and to derive a kernel trick to compute the quantities of interest. The interest is that for non-linear kernels, the Fisher Linear Discriminant in the feature space is a non-linear function of the input space $\mathcal{Y}$.

In this section, we first present the KFDA classifier, then we show how to derive a test in a second part, with the presentation of the two regularized version of the test statistic. The third part is dedicated to the kernel trick for the computation of the test statistic and some numerical analyses.

1.4.1 The Fisher Discriminant Analysis in the Feature Space

The KFDA classifier introduced by [99] was an application of kernel methods to obtain a non-linear classifier. The first non-linear kernel classifier is known as the Support Vector Machine (SVM) [29]. Compared to the SVM, the KFDA classifier has the advantage to take the variability of the embeddings into account to construct the one-dimensional axis that separates the two groups. Now, we show how to define the Fisher Linear Discriminant axis in the feature space $\mathcal{H}$ associated to two probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$. The Fisher Linear Discriminant axis is defined as the optimal axis to discriminate between two samples $\mathbf{Y}_1$ and $\mathbf{Y}_2$ drawn from $\mathbb{P}_1$ and $\mathbb{P}_2$ respectively. This presentation is strongly inspired from the tutorial on the FDA and the KFDA proposed in [46].

Let $h \in \mathcal{H}$ an element of $\mathcal{H}$ that represents a one-dimensional axis in $\mathcal{H}$, in what follows, we refers to h as an axis. For $g \in \mathcal{H}$, the projection g^h of g onto the axis supported by h is such that:

$$g^h = \langle g, h \rangle_{\mathcal{H}} \, h.$$

The intuition is that h is a discriminant axis if the two groups of embeddings $\Phi(\mathbf{Y}_1)$ and $\Phi(\mathbf{Y}_2)$ projected on it are (i) far from each other and (ii) there is no or minimal overlap. The key idea of the FDA is to propose quantitative definitions for these two criteria.

Distance Between the Projected Samples

A quantitative measure of the distance between the projected samples can be defined as the distance between the expected projections of the two samples, that is the distance between the projected kernel mean embeddings. We have:

$$\begin{aligned}\left\|\mathbb{E}_{\mathbb{P}_1}\left(\phi(Y)^h\right) - \mathbb{E}_{\mathbb{P}_2}\left(\phi(Y)^h\right)\right\|_{\mathcal{H}}^2 &= \left\|\mu_1^h - \mu_2^h\right\|_{\mathcal{H}}^2 \\ &= \langle \langle \mu_1 - \mu_2, h\rangle_{\mathcal{H}}\, h, \langle \mu_1 - \mu_2, h\rangle_{\mathcal{H}}\, h\rangle_{\mathcal{H}} \\ &= \langle \mu_1 - \mu_2, h\rangle_{\mathcal{H}}^2 \, \|h\|_{\mathcal{H}}^2 \\ &= \left\langle (\mu_1 - \mu_2)^{\otimes 2} h, h\right\rangle_{\mathcal{H}} \|h\|_{\mathcal{H}}^2 \\ &= \left\langle (\mu_1 - \mu_2)^{\otimes 2}, h \otimes h\right\rangle_{\mathrm{HS}(\mathcal{H})} \|h\|_{\mathcal{H}}^2 .\end{aligned}$$

When $\|h\|_{\mathcal{H}} = 1$, we have:

$$\left\|\mathbb{E}_{\mathbb{P}_1}\left(\phi(Y)^h\right) - \mathbb{E}_{\mathbb{P}_2}\left(\phi(Y)^h\right)\right\|_{\mathcal{H}}^2 = \left\langle (\mu_1 - \mu_2)^{\otimes 2}, h \otimes h\right\rangle_{\mathrm{HS}(\mathcal{H})} .$$

We recognize the form of a reproducing property. The distance between the expected projections of embeddings associated to $\mathbb{P}_1$ and $\mathbb{P}_2$ is represented by the Hilbert-Schmidt operator $(\mu_1 - \mu_2)^{\otimes 2}$. We then define the between-group covariance operator as:

$$\Sigma_B = \frac{n_1 n_2}{n^2} (\mu_1 - \mu_2)^{\otimes 2}.$$

Then for $h \in \mathcal{H}$ such that $\|h\|_{\mathcal{H}} = 1$, the following expression is a measure of the distance between the two projected samples:

$$\langle h, \Sigma_B h\rangle_{\mathcal{H}} = \frac{n_1 n_2}{n^2} \left\|\mathbb{E}_{\mathbb{P}_1}\left(\phi(Y)^h\right) - \mathbb{E}_{\mathbb{P}_2}\left(\phi(Y)^h\right)\right\|_{\mathcal{H}}^2 . \tag{1.12}$$

Note that we also have the following expression:

$$\langle h, \Sigma_B h\rangle_{\mathcal{H}} = \frac{n_1 n_2}{n^2} \langle h, \mu_1 - \mu_2\rangle_{\mathcal{H}}^2 . \tag{1.13}$$

An illustration of what represents $\langle h, \Sigma_B h\rangle_{\mathcal{H}}$ on a simple example is given in Figure 1.1

Overlap Between the Projected Samples

A quantitative measure of the overlap may be indirectly defined with respect to the inertia of the embeddings, that describes the spread of the projected embeddings around their projected mean embeddings. The inertia of group $i \in \{1, 2\}$ is defined as:

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_i}\left(\left\|\Phi(\mathbf{Y})^h - \mu_i^h\right\|_{\mathcal{H}}^2 \right) =& \mathbb{E}_{\mathbb{P}_i}\left(\langle \Phi(\mathbf{Y}) - \mu_i, h\rangle_{\mathcal{H}}^2 \right) \|h\|_{\mathcal{H}}^2 \\ =& \left\langle \mathbb{E}_{\mathbb{P}_i}\left((\Phi(\mathbf{Y}) - \mu_i)^{\otimes 2}\right), h \otimes h \right\rangle_{\mathcal{H}} \|h\|_{\mathcal{H}}^2 \\ =& \langle \Sigma_i, h \otimes h\rangle_{\mathcal{H}} \|h\|_{\mathcal{H}}^2 .\end{aligned}$$

When $\|h\|_{\mathcal{H}} = 1$, we have:

$$\mathbb{E}_{\mathbb{P}_i}\left(\left\|\Phi(\mathbf{Y})^h - \mu_i^h\right\|_{\mathcal{H}}^2 \right) = \langle \Sigma_i, h \otimes h\rangle_{\mathcal{H}} .$$

Thus, the kernel covariance operator of a sample can describe its inertia. Then the global inertia of the two groups can be described by a weighted sum of the two kernel covariance operators, we define the within-group covariance operator as:

$$\Sigma_W = \frac{n_1}{n}\Sigma_1 + \frac{n_2}{n}\Sigma_2.$$

Then the following quantity captures the global spread of the projected embeddings around their respective projected means:

$$\langle h, \Sigma_W h\rangle_{\mathcal{H}} = \frac{n_1}{n}\mathbb{E}_{\mathbb{P}_1}\left(\left\|\Phi(\mathbf{Y})^h - \mu_1^h\right\|_{\mathcal{H}} \right) + \frac{n_2}{n}\mathbb{E}_{\mathbb{P}_2}\left(\left\|\Phi(\mathbf{Y})^h - \mu_2^h\right\|_{\mathcal{H}} \right). \tag{1.14}$$

As a measure of the spread of the projected embeddings, this quantity is informative about the overlap between the two projected samples, because the more spread are the projections, the higher is the probability to observe an overlap. The notion of inertia is illustrated on a single example in Figure 1.1

An Optimization Problem

Based on the two quantitative measures of Equations (1.12) and (1.14), a discriminant axis can be defined as an axis that finds a trade off between maximizing $\langle h, \Sigma_B h\rangle_{\mathcal{H}}$ and minimizing $\langle h, \Sigma_W h\rangle_{\mathcal{H}}$. In particular the Fisher Linear Discriminant axis $h^\star$ of the feature

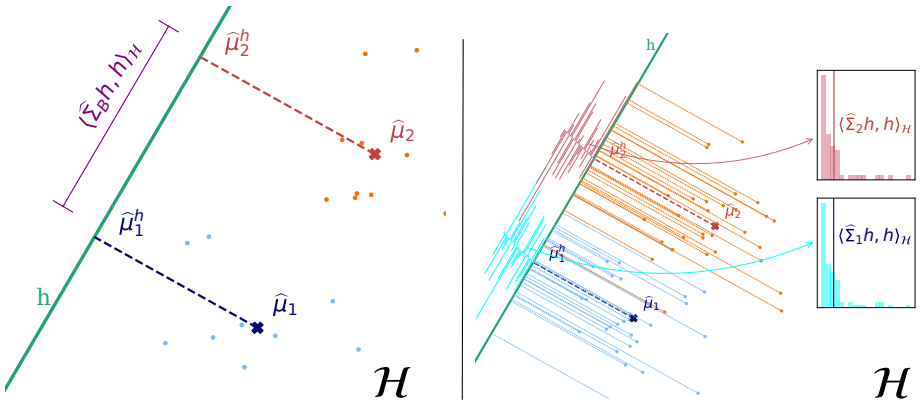

Figure 1.1: Left : Representation of the distance between the projected mean embeddings (left) and the inertia (right) on a toy dataset.

space $\mathcal{H}$ is defined as the axis that maximizes the Kernel Fisher Discriminant Ratio (KFDR):

$$h^\star = \underset{h \in \mathcal{H}}{\operatorname{argmax}} \frac{\langle h, \Sigma_B h \rangle_{\mathcal{H}}}{\langle h, \Sigma_W h \rangle_{\mathcal{H}}}.$$

According to Equation (1.13), we may rewrite this expression as:

$$h^\star = \underset{h \in \mathcal{H}}{\operatorname{argmax}} \frac{n_1 n_2}{n^2} \frac{\langle h, \mu_1 - \mu_2 \rangle_{\mathcal{H}}^2}{\langle h, \Sigma_W h \rangle_{\mathcal{H}}}.$$

Note that the KFDR is always positive as Σ_W is positive definite and the numerator is a squared quantity. If we assume that Σ_W is invertible, $h^\star$ can be determined explicitely by substituting h by $\Sigma_W^{-\frac{1}{2}} g$ with $\|g\|_{\mathcal{H}} = 1$. As Σ_W is a self-adjoint operator, we then have:

$$\begin{aligned} \langle h, \Sigma_W h \rangle_{\mathcal{H}} &= \left\langle \Sigma_W^{\frac{1}{2}} h, \Sigma_W^{\frac{1}{2}} h \right\rangle_{\mathcal{H}} \\ &= \|g\|_{\mathcal{H}}^2 \\ &= 1. \end{aligned}$$

Thus, $h^\star$ is equal to $\Sigma_W^{-\frac{1}{2}} g^\star$ where $g^\star$ is such that:

$$\begin{aligned} g^\star =& \underset{\substack{g \in \mathcal{H} \\ \|g\|_{\mathcal{H}}=1}}{\operatorname{argmax}} \frac{n_1 n_2}{n^2} \left\langle \Sigma_W^{-\frac{1}{2}} g, \mu_1 - \mu_2 \right\rangle_{\mathcal{H}}^2 \\ =& \underset{\substack{g \in \mathcal{H} \\ \|g\|_{\mathcal{H}}=1}}{\operatorname{argmax}} \frac{n_1 n_2}{n^2} \left\langle g, \Sigma_W^{-\frac{1}{2}} (\mu_1 - \mu_2) \right\rangle_{\mathcal{H}}^2 . \end{aligned}$$

The maximum of the inner product is reached for $g^\star = \left\| \Sigma_W^{-\frac{1}{2}} (\mu_1 - \mu_2) \right\|_{\mathcal{H}}^{-1} \Sigma_W^{-\frac{1}{2}} (\mu_1 - \mu_2)$. Thus, the Fisher Linear Discriminant axis is supported by:

$$h^\star = \frac{\Sigma_W^{-1} (\mu_1 - \mu_2)}{\left\| \Sigma_W^{-\frac{1}{2}} (\mu_1 - \mu_2) \right\|_{\mathcal{H}}}. \tag{1.15}$$

Then, a classifier is obtained by using the projections on the Fisher Linear Discriminant axis in the feature space to predict the labels of new observations. For a new observation $Y \in \mathcal{Y}$, we have:

$$\text{predicted label} = \underset{i \in \{1,2\}}{\operatorname{argmin}} \left\| \phi(Y)^{h^\star} - \mu_i^{h^\star} \right\|_{\mathcal{H}} .$$

The practical aspects of the implementation of the KFDA classifier are developed in Chapter 2.

1.4.2 KFDA Two-Sample Tests

The optimization of the KFDR may also be the basis of a two-sample test [64]. To do so, we focus on its maximal value, that is a measure of the possible discrimination between the two probability distributions. The highest is this value, the easier it is to discriminate between the two samples. Similarly to the MMD, when the kernel is characteristic, the maximal value of the KFDR is equal to zero if and only if the two probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ are equal. Thus, a non-parametric two-sample test can be obtained by assessing if the maximal value of the KFDR is null. Thus, the KFDA test statistic is defined as an empirical estimator of the following rescaled version of the maximal value of the KFDR:

$$\mathrm{D}^2 \left(\mathbb{P}_1, \mathbb{P}_2 \right) = n \max_{h \in \mathcal{H}} \frac{\langle h, \Sigma_B h \rangle_{\mathcal{H}}}{\langle h, \Sigma_W h \rangle_{\mathcal{H}}}.$$

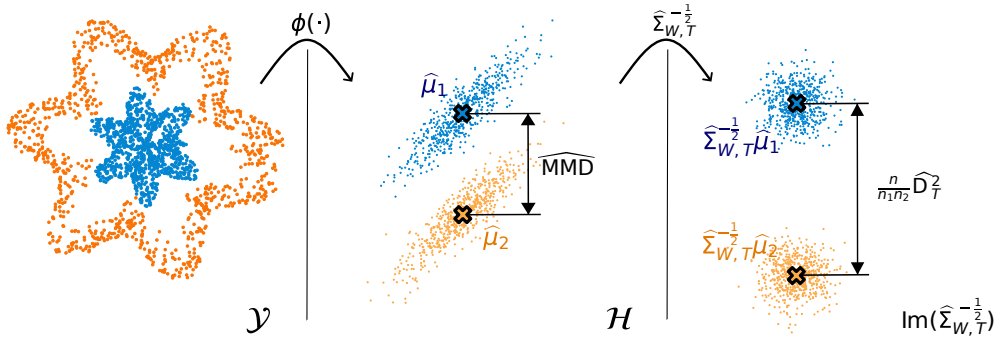

Figure 1.2: Illustration of a metric point of view on the KFDA statistic. From left to right, raw data are embedded in the input space through the feature map $\phi(\cdot)$ where we can compute the MMD, then we transform the embedding with $\widehat{\Sigma}_W^{-\frac{1}{2}}$ to normalize them and then compute the KFDA statistic.

By injecting the Fisher Linear Discriminant axis in the ratio, we obtain the following expression:

$$\mathrm{D}^2\left(\mathbb{P}_1, \mathbb{P}_2\right) = \frac{n_1 n_2}{n} \left\| \Sigma_W^{-\frac{1}{2}} (\mu_1 - \mu_2) \right\|_{\mathcal{H}}^2 .$$

Note that the Fisher Linear Discriminant axis $h^\star$ is called the KFDA witness function in this context. We recognise the form of a normalized MMD of Equation (1.11), indeed, under H_0, the quantity in the norm is the same. We illustrated the succession of transformation applied to the data to obtain the MMD and KFDA statistics from a fictional dataset on Figure 1.2

The quantity D^2 is actually ill-defined. In the previous subsection, we assumed that Σ_W was invertible, which is not the true in general. In the context of kernel testing, two regularizations have been proposed.

Regularization of the Within-Group Covariance

The ridge regularization consists in substituting Σ_W by the ridge within-group covariance operator:

$$\Sigma_{W,\gamma} = \Sigma_W + \gamma I_{\mathcal{H}},$$

where $\gamma > 0$ is an regularization hyperparameter and $I_{\mathcal{H}}$ is the identity operator from $\mathcal{H}$ to $\mathcal{H}$. The ridge within-group covariance operator is invertible.

As Σ_1 and Σ_2 are Hilbert-Schmidt operators, $\Sigma_W \in \mathrm{HS}(\mathcal{H})$. Thus, there exists an orthonormal basis $(f_t)_{t\geq 1}$ of eigenfunctions of Σ_W associated to the decreasing sequence of eigenvalues $(\lambda_t)_{t\geq 1}$. Then the spectral regularization consists in substituting Σ_W by the truncated within-group covariance operator, that is defined with respect to the spectral decomposition of Σ_W:

$$\Sigma_{W,T} = \sum_{t=1}^{T} \lambda_t \Big(f_t \otimes f_t \Big),$$

where T is such that $\lambda_T > 0$. The pseudo-inverse of this operator is then defined as follows:

$$\Sigma_{W,T}^{-1} = \sum_{t=1}^{T} \lambda_t^{-1} \Big(f_t \otimes f_t \Big).$$

Regularized Test Statistics

Now we can define two test statistic associated to these two regularizations. The empirical within-group kernel covariance operator is defined such that:

$$\widehat{\Sigma}_W = \frac{n_1}{n} \widehat{\Sigma}_1 + \frac{n_2}{n} \widehat{\Sigma}_2.$$

Then the empirical ridge within-group kernel covariance operator is equal to $\widehat{\Sigma}_{W,\gamma} = \widehat{\Sigma}_W + \gamma I_{\mathcal{H}}$, and the empirical truncated within-group kernel covariance operator is equal to:

$$\widehat{\Sigma}_{W,T} = \sum_{\substack{t=1 \\ \widehat{\lambda}_t > 0}}^{T} \widehat{\lambda}_t \Big(\widehat{f}_t \otimes \widehat{f}_t \Big),$$

where $T \geq 1$ is such that $\widehat{\lambda}_t > 0$ and $(\widehat{\lambda}_t)_{t\in\{1,\dots,n\}}$ are the decreasing non-negative eigenvalues of $\widehat{\Sigma}_W$ and $(\widehat{f}_t)_{t\in\{1,\dots,n\}}$ are are the associated orthonormal eigenfunctions in $\mathcal{H}$. These eigenquantities can be determined through a kernel trick similar to (1.2.2) that is developed at the end of this section. Then the ridge KFDA test statistic and the truncated

KFDA test statistic are such that:

$$\widehat{\mathrm{D}_\gamma}^2 = \frac{n_1 n_2}{n} \left\| \widehat{\Sigma}_{W,\gamma}^{-\frac{1}{2}} (\widehat{\mu}_1 - \widehat{\mu}_2) \right\|_{\mathcal{H}}^2, \tag{1.16}$$

$$\widehat{\mathrm{D}_T}^2 = \frac{n_1 n_2}{n} \left\| \widehat{\Sigma}_{W,T}^{-\frac{1}{2}} (\widehat{\mu}_1 - \widehat{\mu}_2) \right\|_{\mathcal{H}}^2. \tag{1.17}$$

We have the following theorems about their asymptotic distributions:

Theorem 8 (Asymptotic distribution of the ridge KFDA test statistic [64])**.** *Assume that* $\sum_{s\geq 1} \lambda_s^{\frac{1}{2}} < +\infty$ *and that* $\sup_{y\in\mathcal{Y}} k(y,y) = M_k < +\infty$*. Under* H_0*, we have:*

$$\widehat{\mathrm{D}_\gamma}^2 \xrightarrow[n\to\infty]{\mathcal{D}} \frac{1}{\sqrt{2\sum_{t\geq 1} \frac{\lambda_t^2}{(\lambda_t+\gamma)^2}}} \sum_{t\geq 1} \frac{\lambda_t}{\lambda_t+\gamma}(Z_t^2 - 1)$$

where $(Z_t)_{t\geq 1}$ *is a sequence of i.i.d. standard Gaussian random variables.*

Theorem 9. *Asymptotic distribution of the truncated KFDA test statistic [62] Assume that* $\sum_{s\geq 1} \lambda_s^{\frac{1}{2}} < +\infty$ *and that* $\sup_{y\in\mathcal{Y}} k(y,y) = M_k < +\infty$*. Under* H_0*, we have:*

$$\widehat{\mathrm{D}_T}^2 \xrightarrow[n\to\infty]{\mathcal{D}} \sum_{t=1}^{T} Z_t^2 \sim \chi^2(T)$$

where $(Z_t)_{t\in\{1,\dots,T\}}$ *is a set of i.i.d. standard Gaussian random variables and* $\chi^2(T)$ *denote the chi-squared distribution with* T *degrees of freedom.*

1.4.3 Kernel Trick

The kernel trick for the ridge KFDA test statistic is detailed in [64], contrary to the kernel trick of the truncated KFDA test statistic that is not detailed in [62]. We take the occasion to detail it here and to highlight that their formula of the truncated KFDA test statistic contains an error.

Consider $\Phi(\mathbf{Y}) = \Big(\Phi(\mathbf{Y}_1), \Phi(\mathbf{Y}_2)\Big) = \Big(\phi(Y_{1,1}), \dots, \phi(Y_{1,n_1}), \phi(Y_{2,1}), \dots, \phi(Y_{2,n_2})\Big)$ of $\mathcal{H}^n$. We recall that for $m \geq 1$, $\mathbf{\Pi}_m = \mathbf{I}_m - m^{-1}\mathbf{J}_m \in \mathcal{M}_m(\mathbb{R})$, where $\mathbf{I}_m$ is the identity matrix

and $\mathbf{J}_m$ is the matrix full of ones. Then we define the bi-centering matrix:

$$\mathbf{\Pi}_W = \begin{pmatrix} \mathbf{\Pi}_{n_1} & 0 \\ 0 & \mathbf{\Pi}_{n_2} \end{pmatrix},$$

such that we have:

$$\widehat{\Sigma}_W = \frac{1}{n}(\mathbf{\Pi}_W \Phi(\mathbf{Y}))'(\mathbf{\Pi}_W \Phi(\mathbf{Y})).$$

Reproducing the kernel trick of (1.2.2), we can show that $\widehat{\Sigma}_W$ has the same eigenvalues $(\widehat{\lambda}_t)_{t\geq 1}$ than the matrix:

$$\mathbf{K}_W = \frac{1}{n}(\mathbf{\Pi}_W \Phi(\mathbf{Y}))(\mathbf{\Pi}_W \Phi(\mathbf{Y}))'.$$

Let $\mathbf{U}_W = (u_{W,1}, \ldots, u_{W,n}) \in \mathcal{M}_n(\mathbb{R})$ be the matrix of orthonormal eigenvectors of $\mathbf{K}_W$ such that for $t \in \{1, \ldots, n\}$, the column $u_{W,t} \in \mathbb{R}^n$ is an eigenvector associated to the eigenvalue $\widehat{\lambda}_t$. Then the function $\widehat{f}_t = (n\widehat{\lambda}_t)^{-\frac{1}{2}} \Phi(\mathbf{Y})' \mathbf{\Pi}_W u_{W,t} \in \mathcal{H}$ is a unit eigenfunction of $\widehat{\Sigma}_W$ associated to $\widehat{\lambda}_t$. Denote $\widehat{f}_W = (\widehat{f}_1, \ldots, \widehat{f}_n)'$ the vector of $\mathcal{H}^n$ containing the orthonormal eigenfunctions of $\widehat{\Sigma}_W$ and $\mathbf{\Lambda}_W = \mathrm{Diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_n) \in \mathcal{M}_n(\mathbb{R})$ the diagonal matrix of eigenvalues of $\widehat{\Sigma}_W$ such that:

$$\widehat{f}_W = \frac{1}{\sqrt{n}} \mathbf{\Lambda}_W^{-\frac{1}{2}} \mathbf{U}_W' \mathbf{\Pi}_W \Phi(\mathbf{Y}). \tag{1.18}$$

Then the spectral decomposition of $\widehat{\Sigma}_W$ can be written in matrix form:

$$\widehat{\Sigma}_W = \sum_{t=1}^{n} \widehat{\lambda}_t \Big(\widehat{f}_t \otimes \widehat{f}_t \Big) = \widehat{f}_W' \mathbf{\Lambda}_W \widehat{f}_W.$$

Then for $T \in \{1, \ldots, n\}$, the truncated within-group covariance operator $\widehat{\Sigma}_{W,T}$ is such that $\widehat{\Sigma}_{W,T} = \widehat{f}_{W,T}' \mathbf{\Lambda}_{W,T} \widehat{f}_{W,T}$ and its pseudo-inverse is such that:

$$\widehat{\Sigma}_{W,T}^{-1} = \sum_{t=1}^{T} \widehat{\lambda}_t^{-1} \Big(\widehat{f}_t \otimes \widehat{f}_t \Big) = \widehat{f}_{W,T}' \mathbf{\Lambda}_{W,T}^{-1} \widehat{f}_{W,T}, \tag{1.19}$$

where $\widehat{f}_{W,T} = (\widehat{f}_1, \ldots, \widehat{f}_T)'$ and $\mathbf{\Lambda}_{W,T} = \mathrm{Diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_T)$. Note that we have $\widehat{f}_{W,T} = n^{-\frac{1}{2}} \mathbf{\Lambda}_{W,T}^{-\frac{1}{2}} \mathbf{U}_{W,T}' \mathbf{\Pi}_W \Phi(\mathbf{Y})$ with $\mathbf{U}_{W,T} = (u_{W,1}, \ldots, u_{W,T})$. Now let $\omega = (n_1^{-1} \mathbb{1}_{n_1}', -n_2^{-1} \mathbb{1}_{n_2}')' \in$

$\mathbb{R}^n$, such that:

$$\Phi(\mathbf{Y})'\omega = \widehat{\mu}_1 - \widehat{\mu}_2. \tag{1.20}$$

We obtain a formulation of the kernel trick to compute the truncated KFDA test statistic by using Equations (1.19) and (1.20):

$$\begin{aligned}
\widehat{\mathrm{D}_T}^2 =& \frac{n_1 n_2}{n} \left\| \widehat{\Sigma}_{W,T}^{-\frac{1}{2}} \left(\widehat{\mu}_1 - \widehat{\mu}_2 \right) \right\|_{\mathcal{H}}^2 \\
=& \frac{n_1 n_2}{n} \left\langle \widehat{\mu}_1 - \widehat{\mu}_2, \widehat{\Sigma}_{W,T}^{-1} \left(\widehat{\mu}_1 - \widehat{\mu}_2 \right) \right\rangle_{\mathcal{H}} \\
=& \frac{n_1 n_2}{n} \omega' \Phi(\mathbf{Y}) \widehat{f}'_{W,T} \mathbf{\Lambda}_{W,T}^{-1} \widehat{f}_{W,T} \Phi(\mathbf{Y})' \omega \\
=& \frac{n_1 n_2}{n^2} \omega' \mathbf{K_Y} \mathbf{\Pi}_W \mathbf{U}_{W,T} \mathbf{\Lambda}_{W,T}^{-2} \mathbf{U}'_{W,T} \mathbf{\Pi}_W \mathbf{K_Y} \omega.
\end{aligned}$$

This formula differs from the one proposed in [62] that is:

$$\widehat{\mathrm{D}_T}^2 = \frac{n_1 n_2}{n^2} \omega' \mathbf{K_Y} \mathbf{U}_{W,T} \mathbf{\Lambda}_{W,T}^{-1} \mathbf{U}'_{W,T} \mathbf{K_Y} \omega.$$

This difference is due to the fact that they did not properly determined a set of orthonormal eigenfunctions of $\widehat{\Sigma}_W$ with respect to the orthonormal eigenvectors of $\mathbf{K}_W$ and considered instead that $(\Phi(\mathbf{Y})' u_{W,t})_{t \in \{1,\dots,n\}}$ was an orthonormal set of eigenfunctions of $\widehat{\Sigma}_W$, which is not true. A limitation of this approach is that a new matrix product has to be computed for each value of T. In practice, we prefer the following expression based on the spectral decomposition of $\widehat{\Sigma}_{W,T}$:

$$\begin{aligned}
\widehat{\mathrm{D}_T}^2 =& \frac{n_1 n_2}{n} \left\| \sum_{t=1}^{T} \widehat{\lambda}_t^{-\frac{1}{2}} \left(\widehat{f}_t \otimes \widehat{f}_t \right) \left(\widehat{\mu}_1 - \widehat{\mu}_2 \right) \right\|_{\mathcal{H}}^2 \\
=& \frac{n_1 n_2}{n} \left\| \sum_{t=1}^{T} \widehat{\lambda}_t^{-\frac{1}{2}} \left\langle \widehat{f}_t, \widehat{\mu}_1 - \widehat{\mu}_2 \right\rangle_{\mathcal{H}} \widehat{f}_t \right\|_{\mathcal{H}}^2 \\
=& \frac{n_1 n_2}{n} \sum_{t,t'=1}^{T} \widehat{\lambda}_t^{-\frac{1}{2}} \widehat{\lambda}_{t'}^{-\frac{1}{2}} \left\langle \widehat{f}_t, \widehat{\mu}_1 - \widehat{\mu}_2 \right\rangle_{\mathcal{H}} \left\langle \widehat{f}_{t'}, \widehat{\mu}_1 - \widehat{\mu}_2 \right\rangle_{\mathcal{H}} \left\langle \widehat{f}_t, \widehat{f}_{t'} \right\rangle_{\mathcal{H}}.
\end{aligned}$$

As $(\widehat{f}_1, \ldots, \widehat{f}_n)$ form an orthnormal set of functions, for $t \neq t'$, we have $\left\langle \widehat{f}_t, \widehat{f}_{t'} \right\rangle_{\mathcal{H}} = 0$. We conclude that:

$$\widehat{\mathrm{D}_T}^2 = \sum_{t=1}^{T} \frac{n_1 n_2}{n\widehat{\lambda}_t} \left\langle \widehat{f}_t, \widehat{\mu}_1 - \widehat{\mu}_2 \right\rangle_{\mathcal{H}}^2 . \tag{1.21}$$

Then for $t \in \{1, \ldots, T\}$, we have the following kernel trick:

$$\frac{n_1 n_2}{n} \widehat{\lambda}_t^{-1} \left\langle \widehat{f}_t, \widehat{\mu}_1 - \widehat{\mu}_2 \right\rangle_{\mathcal{H}}^2 = \frac{n_1 n_2}{(n\widehat{\lambda})^2} \left(u'_{W,t} \mathbf{\Pi}_W \mathbf{K}_{\mathbf{Y}} \omega \right)^2 .$$

1.5 Conclusion

1.5.1 Related Work

Other Kernel Tests

In addition to the MMD test and the KFDA test, two other kernel two-sample tests have been proposed and are reviewed in [63]. The Kernel Density Ratio (KDR) test statistic follows the idea of estimating a f-divergence such as the Kullback-Leibler divergence between $\mathbb{P}_1$ and $\mathbb{P}_2$ using kernels [72]. The kernel change detection test is based on a MMD-like statistic computed on a weighted version of the kernel mean embeddings [33]. In [79], they suggest that the MMD witness function is sufficient to define a test and propose several tests based on witness functions. The authors of [28] apply the same normalization as KFDA at finite many locations for $O(n)$ operations, this statistic has worse performances than quadratic-time test procedures such as the permutation MMD procedure or the asymptotic MMD procedure. However, it can be of interest when ressource is limited. In this procedure, the location choice is random but it can be optimized to a proxy of the power while preserving the $O(n)$ complexity [70]. The choice of the norm can also vary, in [70], they use the Euclidean norm, a variant with the ℓ_1 norm exists [121]. Generalizations from two-sample tests to k-sample tests have been proposed for the KFDA test [15]. In Chapter 3, we propose a generalization of the KFDA test to any experimental design.

Kernel Choice

The MMD test remains the most studied kernel test and several variants have been proposed. One major research area around the MMD test is the kernel choice. The

most common choice is the Gaussian kernel with the bandwidth set as the median of the distances between the observations [43]. The authors of [123] propose to aggregate Gaussian kernels with different bandwidth sizes. Some others define the kernel as the output of a deep neural network [77]. We did not investigate this issue but it could be of interest to determine the best kernel for scRNA-Seq data.

Kernel Testing for High-Dimensional Data

Kernel tests are computationally of interest for high-dimensional data, as the computational complexity of these tests scales in n instead of scaling in the dimension. However, kernel tests do not tackle the curse of dimensionality [115, 8]. In particular, the power of a kernel test decreases when the dimension increases when comparing fair alternatives in the sense of the authors of [115]. It would be of interest to investigate more precisely the sensibility of kernel testing to the dimension.

1.5.2 The Truncated KFDA Statistic

The ridge KFDA test (Equation (1.16)) has raised more attention than the truncated KFDA test (Equation (1.17) because it inherits from the same property than the MMD test to be theoretically able to detect any difference between two distributions with a characteristic kernel. That is not the case for the truncated KFDA test with fixed truncation parameter. But the truncated KFDA statistic can be more efficiently implemented and our simulations studies and applications showed that the truncated KFDA asymptotic test has satisfying performance when at least hundreds of observations are available. Moreover, we consider that it is more suited to practical applications. Indeed, the truncation regularization can be interpreted as a dimension reduction adapted to the discrimination problem. Representing the embeddings projected in this finite dimensional space dedicated to the discrimination allows to visualize a possible difference spotted by the test. Put together, the truncated KFDA test associated to this representation make a complete framework to test and explore the differences between several datasets. This framework is presented and applied to biological data in the next Chapter.

Chapter 2

The Truncated Kernel Fisher Discriminant Analysis Test in Practice

Despite a tremendous success in the machine learning literature, there is currently no package or software that provides a dedicated implementation of kernel tests based on the MMD or on KFDA. The main contribution of this Chapter is to propose a turnkey solution for pair-wise multivariate comparisons in single-cell data analysis based on the KFDA framework (for testing and for discrimination). Thus, every method presented in this chapter is implemented in a package and can be run with a few lines of code. This package, called `ktest`[1], was implemented in both *Python* and *R* and is designed to correspond to the standards in use in the single-cell community. Future work will be dedicated to the introduction of the package to the `scverse` consortium of tools dedicated to single-cell data analysis. However, `ktest` is a general implementation of these kernel methods that may be applied to any type of data.

The methods presented here strongly rely on the KFDA framework. The truncated KFDA two-sample test is efficiently implemented and the Fisher linear discriminant axis in the feature space that is a side-product of this implementation is used as a visualization tool to explore the cell-wise differences that would be detected by the test. To the best of our knowledge, our package is the first user-oriented implementation of the KFDA approach. For completeness, we also implemented the MMD test. Single-cell datasets can be very large (from hundreds to tens of thousands of observations), which can be computationally prohibitive for some algorithms. This is the case of our kernel test that can take hours to compare two datasets with more than 5000 cells. We thus implemented a matrix factorisation method based on a Nyström approximation that is suited to drastically

[1]The package `ktest` is available on my Github page `https://github.com/AnthoOzier/ktest`

accelerate kernel methods without reducing too much the performance [147].

The first section of this chapter is an adapted version of the article we submitted jointly with the `ktest` package that describes how the KFDA framework can be practically applied on single-cell datasets (Section 2.1). It contains the description of the framework, our comparisons to other existing methods and our conclusions on the analyses done in collaboration with biologists from the Laboratoire de Biologie et Modelisation de la Cellule (LBMC) and the Institut Curie. The second section details the application of the Nyström method to accelerate the computation of the test statistic and allow the analysis of large datasets (Section 2.2).

2.1 Kernel-Based Testing for Single-Cell Differential Analysis

2.1.1 Abstract

Single-cell technologies have provided valuable insights into the distribution of molecular features, such as gene expression and epigenomic modifications. However, comparing these complex distributions in a controlled and powerful manner poses methodological challenges. Here we propose to benefit from the kernel-testing framework to compare the complex cell-wise distributions of molecular features in a non-linear manner based on their kernel embedding. Our framework not only allows for feature-wise analyses but also enables global comparisons of transcriptomes or epigenomes, considering their intricate dependencies. By using a classifier to discriminate cells based on the variability of their embedding, our method uncovers heterogeneities in cell populations that would otherwise go undetected. We show that kernel testing overcomes the limitations of differential analysis methods dedicated to single-cell. Kernel testing is applied to investigate the reversion process of differentiating cells, successfully identifying cells in transition between reversion and differentiation stages. Additionally, we analyze single-cell ChIP-Seq data and identify a subpopulation of untreated breast cancer cells that exhibit an epigenomic profile similar to persister cells.

2.1.2 Introduction

Thanks to the convergence of single-cell biology and massive parallel sequencing, it is now possible to create high dimensional molecular portraits of cell populations. This technological breakthrough allows for the measurement of gene expression [89, 68, 152], chromatin states [120], and genomic variations [45] at the single-cell resolution. These advances have resulted in the production of complex high dimensional data and revolutionized our understanding of the complexity of living tissues, both in normal and pathological states. Then, the field of single-cell data science has emerged, and new methodological challenges have arisen to fully exploit the potentialities of single-cell data, among which the statistical comparison of single-cell RNA sequencing (scRNA-Seq) datasets between conditions or tissues. This step has remained a prerequisite in the process to discriminate biological from technical variabilities and to assert meaningful expression differences. While most differential analysis methods primarily focus on expression data, similar methodological challenges have arisen in the comparative analysis of single cell epigenomic datasets, based for example on single cell chromatin accessibility assays (scATAC-Seq [111]) or single cell histone modifications profiling (e.g single-cell ChIP-Seq (scChIP-seq) [57], scCUT&Tag [18]). These approaches enable the mapping of chromatin states throughout the genome and their cell-to-cell variations at an unprecedented resolution [130, 23]. These single-cell epigenomic assays offer a quantitative perspective on regulatory processes, wherein cellular heterogeneity could drive cancer progression or the development of drug resistance for instance [92]. The identification of epigenomic features of interest by differential analysis in disease and complex eco-systems, will be key to understand regulatory principles of gene expression and identify potential drivers of tumor progression. Altogether, comparative analysis of single cell data sets, whatever their type, are an essential component of single cell data science, providing biological insights as well as opening therapeutic perspectives with the identification of biomarkers and therapeutic targets.

Differential Expression Analysis (DEA) is classically addressed by gene-wise two-sample tests designed to detect Differentially Expressed Genes (DEG) [32]. The generalized linear model (GLM) has been a powerful framework for linear parametric testing based on gene-expression summaries [86, 119, 118]. However, this parametric approach does not fully utilize the entire distribution of gene-expression that characterizes multiple transcriptional states. To achieve the full potential of differential analysis of scRNA-Seq data, DEA has been restated as a comparison between distributions. Distributional hypotheses were proposed to capture biologically relevant differences in univariate gene-

expressions [78]. Initially, these tests were performed using Gaussian-based clustering, that was further challenged by distribution-free methods based on ranks or cumulative distribution functions [122, 44, 139]. While distribution-free approaches are flexible enough to capture the numerous complex alternatives encountered in DEA, their fully agnostic point of view does not benefit from the significant progress made in modeling scRNA-Seq distributions, which leads to a loss of statistical power. As a trade-off, we propose a distribution-free test based on a representation of the data that can take advantage of finely-tuned probabilistic modeling of scRNA-Seq data.

Single-cell technologies provide a unique opportunity to obtain a quantitative snapshot of the entire transcriptome, which contains crucial information about between-gene dependencies and underlying regulatory networks and pathways. Therefore, univariate DEA only captures a part of the biological differences and is unable to detect complex global modifications in the joint expression of groups of genes. To fully exploit the complexity of scRNA-Seq data, joint multivariate testing or differential transcriptome analysis should be performed, allowing for cell-wise comparisons. This strategy can be complementary to gene-wise approaches, as the detection of DEG should be interpreted in the context of global differences between conditions. The joint multivariate testing strategy seems also particularly suited to compare epigenomic data since it is well established that chromatin conformation can induce complex dependencies between sites occupancy [91]. From a distributional perspective, this involves complementing joint distribution-based analyses with analyses based on marginals. Another significant advantage of differential transcriptome analysis is that it can be restricted to targeted GRNs or pathways, allowing for differential network or pathway analyses [101]. So far, global approaches were mainly developed for differential abundance testing [27, 31, 25], or for the comparison of cell-type compositions. Graph-based methods have been proposed to address differential transcriptome analysis [101, 16], but they only derive a global p-value without any representation or diagnostic tool.

In recent years, there have been significant advancements in statistical hypothesis testing, alongside the emergence of single-cell technologies. One important breakthrough in hypothesis testing was achieved by Gretton et al. [52], who combined kernel methods with statistical testing. Kernel methods are widely used in supervised learning [129] and are based on the concept of embedding data in a feature space, allowing for non-linear data analysis in the input space. Popular dimension reduction techniques, such as tSNE and UMAP [88, 96], also use kernel-based embedding [142]. The distribution of the embedded

data can be described using classical statistics such as means and variances, which can be applied in the feature space. Then the central concept of kernel-based testing is to rely on the Maximum Mean Discrepancy (MMD) test that compares the distance between mean embeddings of two conditions [100], allowing for non-linear comparison of two gene-expression distributions. Despite the significant potential of kernel-based testing, this approach has not yet been developed in single-cell data science.

In this work, we propose a new kernel-based framework for the exploration and comparison of single-cell data based on Differential Transcriptome/Epigenome Analysis. Our method relies on the Kernel Fisher Discriminant Analysis (KFDA) approach introduced by [62]. KFDA is a normalized version of the Maximum Mean Discrepancy to account for the variability of the datasets. This results in a test statistic that can be interpreted as the distance between mean embeddings projected onto the kernel-Fisher discriminant axis. Although KFDA was initially introduced as a non-linear classifier [99], it is a great example of how classifiers can be used for hypothesis testing [63, 85], and recent developments have demonstrated its optimality [60]. Here we show that the KFDA-witness function, which is the Fisher discriminant axis [79], can further be used for data exploration of scRNA-Seq and scChIP-seq data. Our method is implemented in a package called `ktest`[2] available in both *R* and *Python*, which offers many visualization tools based on the geometrical concepts from the Fisher Discriminant Analysis (FDA) to aid comparisons. We show the calibration and the power of our method compared with others on simulated [44] and multiple scRNA-Seq datasets [134]. Then we illustrate the power of the classification-based testing approach, that identifies sub-populations of cells based on expression and epigenomic data, that would not be detected otherwise. When applied to scRNA-Seq data, `ktest` reveals the heterogeneity in differentiating cell populations induced to revert toward an undifferentiated phenotype [153]. Our method also uncovers the epigenomic heterogeneity of breast cancer cells, revealing the pre-existence - prior to cancer treatment - of cells epigenomically identical to drug-persister cells, i.e the rare cells that can survive treatment.

As single-cell datasets grow larger and more complex, traditional testing methods may fail to capture subtle variations and accurately identify meaningful differences in molecular patterns. Here we show that kernel testing emerges as a promising approach to overcome these challenges, offering a robust and flexible framework. Kernel testing techniques are

[2] `https://github.com/AnthoOzier/ktest`

less sensitive to assumptions on data distribution than traditional methods, and can handle complex dependencies within and across cells. This capability is particularly relevant in the context of single-cell data, where inherent noise, sparsity, and heterogeneity pose unique challenges to accurate statistical inference. Overall, kernel testing represents a powerful tool for the differential analysis of single-cell data, enabling to uncover hidden patterns, and gain deeper insights into the intricate heterogeneities of cell populations.

2.1.3 Results

In the following we denote by $\mathbf{Y}_1 = (Y_{1,1}, \ldots, Y_{1,n_1})$ and $\mathbf{Y}_2 = (Y_{2,1}, \ldots, Y_{2,n_2})$ the gene expression measurements of G genes with distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ in conditions 1 and 2 on n_1 and n_2 cells respectively, $n = n_1 + n_2$. In the following, we will derive our method for expression data, but it can be generalized to any single-cell data. Then we suppose that

$$Y_{i,j} \sim \mathbb{P}_i, \quad i = 1, 2 \quad j = 1, \ldots n_i.$$

Two-sample testing between distributions consists in challenging the null hypothesis $H_0 : \mathbb{P}_1 = \mathbb{P}_2$ by the alternative hypothesis $H_1 : \mathbb{P}_1 \neq \mathbb{P}_2$. To construct a non-linear test we consider the embeddings of the original data denoted by $\Big(\phi(Y_{i,1}), \ldots, \phi(Y_{i,n_i})\Big)$ $(i = 1, 2)$, obtained using the feature map $\phi(\cdot)$ that maps the data into the so-called feature space $\mathcal{H}$ that is a reproducing kernel Hilbert space. The kernel provides a measure of the similarity between the observations, that turns out to be the inner product between the embeddings:

$$k(Y_{i,j}, Y_{i',j'}) = \Big\langle \phi(Y_{i,j}), \phi(Y_{i',j'}) \Big\rangle_{\mathcal{H}}.$$

Thanks to this relation, kernel methods are non-linear for the original data, but linear with respect to the embeddings in the feature space. They provide a non-linear dissimilarity between cells based either on the whole transcriptome or on univariate gene distributions. Kernel-based tests consist in the comparison of kernel mean embeddings of distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ [100], defined such that:

$$\forall i \in \{1, 2\}, \quad \mu_i = \mathbb{E}_{Y \sim \mathbb{P}_i}\Big(\phi(Y)\Big).$$

The initial contribution to kernel testing involved calculating the distance between kernel mean embeddings with the MMD statistic [54]. However, it is difficult to determine its null distribution, and since the MMD does not account for the variance of embedding, it

has recently been show to lack optimality [60]. By utilizing a Mahalanobis distance to standardize the difference between mean embeddings, we can not only obtain an asymptotic chi-square distribution for the resulting statistic [63], but we can also take advantage of the kernel Fisher Discriminant Analysis (KFDA) framework that is typically used for non-linear classification. Therefore, we present two complementary perspectives on the KFDA testing framework: one based on a distance-based construction of the statistic and the other on the kernel FDA, which offers several visualization tools to highlight the main cell-wise differences between the two tested conditions.

Testing with a Mahalanobis Distance Between Gene-Expression Embeddings

The squared distance between the kernel mean embeddings constitutes the so-called Maximum Mean Discrepancy statistic, such that:

$$\begin{aligned}\mathrm{MMD}^2(\mu_1,\mu_2) &= \|\mu_1-\mu_2\|^2_{\mathcal{H}} \\ &= \mathbb{E}_{Y_1\sim\mathbb{P}_1,Y_1'\sim\mathbb{P}_1}\left[k(Y_1,Y_1')\right]+\mathbb{E}_{Y_2\sim\mathbb{P}_2,Y_2'\sim\mathbb{P}_2}\left[k(Y_2,Y_2')\right] \\ &\quad -2\mathbb{E}_{Y_1\sim\mathbb{P}_1,Y_2\sim\mathbb{P}_2}\left[k(Y_1,Y_2)\right].\end{aligned}$$

This statistic tests the between-class separation by comparing expected pairwise similarities between and within conditions 1 and 2. To account for the residual variability, we introduce the weighted Mahalanobis distance between mean embeddings,

$$\mathrm{D}^2(\mu_1,\mu_2)=\frac{n_1n_2}{n}\|\Sigma_W^{-1/2}\left(\mu_1-\mu_2\right)\|^2_{\mathcal{H}},$$

where Σ_W is the homogeneous within-group covariance of embeddings:

$$\Sigma_W=\frac{n_1}{n}\Sigma_1+\frac{n_2}{n}\Sigma_2,$$

with:

$$\forall i\in\{1,2\},\quad \Sigma_i=\mathbb{E}_{Y\sim\mathbb{P}_i}\left[(\phi(Y)-\mu_i)^{\otimes 2}\right],$$

the covariance operator within each condition ($\otimes$ stands for the tensor product in the feature space). To avoid the singularity of Σ_W, we consider a regularized version of the kernel-based Mahalanobis distance, by approximating the within-covariance by its first T principal directions. This resumes to a kernel-PCA dimension-reduction step based on $\Sigma_{W,T}$ which catches the residual variability of expression data centered by condition.

Then the corresponding regularized statistic is based on the estimated mean embeddings and covariances:

$$\forall i \in \{1,2\}, \quad \widehat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\phi(Y_{i,j}), \quad \widehat{\Sigma}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}(\phi(Y_{i,j}) - \widehat{\mu}_i))^{\otimes 2}.$$

The main computational complexity comes from the eigen-decomposition of $\widehat{\Sigma}_W = (n_1\widehat{\Sigma}_1 + n_2\widehat{\Sigma}_2)/n$ which requires $O(n^3)$ operations and results in the truncated covariance $\widehat{\Sigma}_{W,T} = \sum_{t=1}^{T}\widehat{\lambda}_t(\widehat{f}_t \otimes \widehat{f}_t)$, where $(\widehat{\lambda}_t)_{t=1:T}$ are the decreasing eigenvalues of $\widehat{\Sigma}_{W,T}$ and $(\widehat{f}_t)_{t=1:T}$ are the associated eigenfunctions referred by extension in the following as principal components. Then the empirical weighted Mahalanobis distance between the two mean-embeddings is:

$$\widehat{\mathrm{D}}_T^2(\widehat{\mu}_1, \widehat{\mu}_2) = \frac{n_1 n_2}{n}\left\|\widehat{\Sigma}_{W,T}^{-\frac{1}{2}}(\widehat{\mu}_2 - \widehat{\mu}_1)\right\|_{\mathcal{H}}^2.$$

This statistic follows a $\chi^2(T)$ asymptotically under the null hypothesis [62], which resumes to the Hotelling's test in the feature space. Using the asymptotic distribution for testing seems reasonable for scRNA-Seq data for which $n \geq 100$, otherwise, it is possible to test with a permutation procedure for small sample sizes. Our implementation runs in ~ 5 minutes for $n \sim 4000$, and the package proposes a sampling-based Nyström approximation for larger sample sizes [147].

The Kernel Fisher Discriminant Analysis, a Powerful Tool for Non-Linear DEA

A major advantage of using the Mahalanobis distance between distributions is that the test statistic can be reinterpreted as a classification problem, thanks to its connection with the Fisher Discriminant Analysis (FDA). This framework induces a powerful cell-wise visualization tool that allows to explore and understand the nature of the differences between transcriptomes. FDA is a linear classification method that consists in finding the linear axis that optimizes the discrimination between the two distributions. Intuitively, a direction is discriminant if the observations projected on it (i) do not overlap and (ii) are far from each other. Hence the best discriminant axis is found by maximizing the Fisher Discriminant Ratio, that models a trade-off between minimizing the overlap while maximizing the distance between the means of the two groups. By finding this linear axis in the feature space to classify the embeddings, we obtain a non-linear function that makes the two distributions linearly separable. Thus, in the feature space we denote by

$h_T^\star$ the optimal axis that maximizes the truncated Fisher Discriminant Ratio:

$$h_T^\star = n \operatorname*{argmax}_{h \in \mathcal{H}} \frac{\langle h, \Sigma_B h \rangle_{\mathcal{H}}}{\langle h, \Sigma_{W,T} h \rangle_{\mathcal{H}}}.$$

where Σ_B is the between-group covariance capturing the part of the variance of the embeddings due to the difference between the two groups:

$$\Sigma_B = \frac{n_1 n_2}{n^2} (\mu_1 - \mu_2)^{\otimes 2}.$$

The numerator of the Fisher Discriminant Ratio captures the distance between the two mean embeddings on a given direction, to be maximized, and the denominator captures the variability of the embeddings projected on this direction, standing for a measure of the overlap, to be minimized. The discriminant axis $h_T^\star$ can be found in closed form from an analytical reasoning. The Mahalanobis distance then appears to be the maximal value of the ratio, which is the distance between the mean embeddings projected on $h_T^\star$:

$$\mathrm{D}_T^2 = n \frac{\langle h_T^\star, \Sigma_B h_T^\star \rangle_{\mathcal{H}}}{\langle h_T^\star, \Sigma_{W,T} h_T^\star \rangle_{\mathcal{H}}} = \frac{n_1 n_2}{n} \| \Sigma_{W,T}^{-1/2} (\mu_1 - \mu_2) \|_{\mathcal{H}}^2,$$

By relying on both the within-group and the between-group covariances, the FDA approach encompasses the total variability of the embeddings. We can interpret the projection of the embeddings on $h_T^\star$ in terms of similarity between the two groups. The extreme values of projected embeddings on the discriminant axis correspond to cells that contain the most significant information for distinguishing between conditions. Conversely, the central values of projected embeddings correspond to cells that do not contribute to the discrimination and hold less informative value. We will propose an illustration to show how this representation can be used to identify outliers or sub-populations.

Then non-linear testing turns out to be very powerful to detect complex alternatives, like the ones proposed in the context of distribution-based DEA [78]. We illustrate the discriminant axis by representing the four standard alternative hypotheses: differential mean (DE), differential proportions (DP), differential modality (DM) and differential both proportion and modality (DB) [78]. The DE, DP and DM alternatives are somehow easy to discriminate even with summary statistics because the distributions have different means, projecting the embeddings on the discriminant axis easily discriminates the two conditions. On the contrary, the DB alternative is the most difficult alternative to detect

with many DEA approaches, because the two conditions share the same mean expression [44]. The discriminant axis acts as a powerful non-linear transformation of the expression data to make the two distributions easily separable (Fig. 2.1).

Kernel Choice

The design of appropriate kernels is an active field of research [13, 123]. In kernel-based testing, choosing an appropriate kernel has many objectives like capturing important data characteristics and showing sufficient power to distinguish between different alternatives. To this extent, the conclusions drawn in the feature space from the mean embeddings should apply to the initial distributions. In other words, it should be equivalent to test $\mu_1 = \mu_2$ for $\mathbb{P}_1 = \mathbb{P}_2$ which is not true in general. However, both are equivalent for a particular class of kernels called universal kernels, which has led to theoretical and computational developments [131, 56, 123]. Fortunately, the Gaussian kernel fulfills this universality property. For two cells $\{(i,j),(i',j')\}$ and genes $g = 1,\ldots,G$, our developments will be based on k_{Gauss} defined such that:

$$k_{\text{Gauss}}(Y_{i,j}, Y_{i',j'}) = \exp\left(-\frac{1}{2\sigma^2}\sum_{g=1}^{G}(Y_{i,j}^g - Y_{i',j'}^g)^2\right).$$

This kernel can be used in both multivariate and univariate contexts. Once the Gaussian kernel has been chosen, the remaining question concerns the calibration of its bandwidth σ, which is done using the median heuristic [56, 123, 43]. We also propose to account for zero-inflation which is another important characteristic of scRNA-Seq data. This can be achieved by employing probability product kernels, which adapt kernel methods to specific probabilistic generative models [69]. Considering a zero-inflated Gaussian distribution with π the proportion of additional zeros and $f_{\mu,\sigma}$ the Gaussian probability function, we show that the probability product kernel between two zero-inflated Gaussian distributions of parameters (μ,σ,π) and (μ',σ,π') is (as detailed in the Methods section):

$$\begin{aligned} k_{\text{ZI-Gauss}}(Y_{i,j}, Y_{i',j'}) &= \pi\pi' + \pi(1-\pi')f_{\mu',\sigma}(0) + (1-\pi)\pi' f_{\mu,\sigma}(0) \\ &+ (1-\pi)(1-\pi')k_{\text{Gauss}}(Y_{i,j}, Y_{i',j'}). \end{aligned}$$

Kernel Testing is Calibrated and Powerful on Simulated Data

Simulations are required to compare the empirical performance of DE methods on controlled designs, to check their type-I error control and compare their power on targeted alternatives. Thanks to a very fruitful collaboration/data sharing with colleagues having developed a competing method [44], we challenged our kernel-based test with others on mixtures of zero-inflated negative binomial data reproducing the four Korthauer's alternatives [44] (as detailed in Material and Methods). Kernel testing was performed on the raw data using both the Gauss and the ZI-Gauss kernels. The type-I errors of the kernel test are controlled at the nominal levels $\alpha = 5\%$ and the performance increases with n (the asymptotic regime of the test is reached for $n \geq 100$). The kernel test is the best method for detecting the DB alternative, considered as the most difficult to detect, and it outperforms every other method in terms of global power excepted SigEMD. This gain in power can be explained by the non-linear nature of our method: despite the equality of means, the kernel-based transform of the data onto the discriminant axis allows a clear separation between distributions (Fig. 2.1). Our method shows its worst performances on the DP alternative, which is the only alternative for which all the values are covered by both conditions with different proportions. It shows that our method is particularly sensitive to alternatives where some values are occupied by one condition only (Fig. 2.2). Note that the probability product kernel did not improve the global performance, which indicates that the Gaussian kernel-based test is robust to zero inflation. This could also be due to the equality of the zero-inflation proportions between conditions. Finally, results on log-normalized data are similar.

Challenging DEA Methods on Experimental scRNA-Seq Data

Differential analysis methods require validation through experimental data, typically by using a ground truth list of differentially expressed (DE) genes and an accuracy criterion. In this study, we examine the framework proposed by Squair et al. [134], which compared 14 DE analysis methods on 18 scRNA-Seq datasets. The authors proposed three main conclusions: *i*) replicate variability needs to be corrected, *ii*) single-cell DE methods are susceptible to false discoveries, and *iii*) pseudo-bulk methods are the most powerful. Pseudo-bulk methods involve applying DEA methods dedicated to bulk-RNA-Seq to averaged scRNA-Seq. However, these conclusions are based on the use of bulk-RNA-Seq DE genes as the ground truth, which inevitably favors pseudo-bulk methods designed to

detect significant mean differences only. Hence, the study ignores genes with differential expression based on other characteristics, as shown in Korthauer's DB scenario [78]. Therefore, we propose to broaden the scope of this comparative study by comparing the outputs of different DE methods in a pairwise comparative manner, without relying on a reference ground truth list of DE genes. Based on pair-wise accuracies, Differential Analysis methods cluster into three groups of concordant groups that correspond to bulk, pseudo-bulk and single-cell based methods respectively (Fig. 2.3, top). As expected, bulk-based methods are separated from others, pseudo-bulk and single-cell methods performed similarly, scRNA-Seq data being more similar. Kernel testing shows performance close to single-cell methods.

We demonstrate that kernel testing does not show the same bias as other single-cell DEA methods that tend to over-detect highly-expressed genes as mentioned in the original study (Fig. 2.3, bottom). By inspecting the distributional changes associated to genes considered as false-positive in the original study (with bulk-RNA-Seq genes as the ground truth), we show that they can in fact be interpreted as true positives. Many of them belong to the DB alternative (difference in both modalities and proportions, [78]), and were thus undetectable from bulk-RNA-Seq data and pseudo-bulk methods (Fig. 2.7, left). Their classification in terms of false positives is then questionable, and kernel testing is clearly powerful to detect those alternatives on experimental data. Others present slight shifts in distribution and low zero proportions, these genes are correctly detected by the probability product kernel adapted to zero-inflation (examples of such distribution shapes are shown in Fig. 2.7, right).

Kernel Testing Reveals the Heterogeneity of Reverting Cells

Single-cell transcriptomics has been widely used to investigate the molecular bases of cell differentiation, and has highlighted the stochasticity and dynamics of the underlying gene regulatory networks. The stochasticity of GRNs allows plasticity between cell states, and is a source of heterogeneity between cells along the differentiation path, which calls for multivariate differential analysis methods. We focus on the differentiation path of chicken primary erythroid progenitor cells (T2EC). A first study highlighted the existence of plasticity, i.e. the ability of cells induced into differentiation to reacquire the phenotypic characteristics of undifferentiated cells (e.g. starting self-renewing again), until a differentiation point of commitment (around 24 hours after differentiation induction) after which this phenotype was lost [117]. A second study investigated the molecular mechanisms

underlying cell differentiation and reversion by measuring cell transcriptomes at four time points: undifferentiated T2EC maintained in a self-renewal medium (0H), then put in a differentiation-inducing medium for 24h (24H). The population was then split into a first population maintained in the same medium for 24h to achieve differentiation (48HDIFF), the second population was put back in the self-renewal medium to investigate potential reversion (48HREV) [153]. Cell transcriptomes were measured using scRT-qPCR on 83 genes selected to be involved in the differentiation process, as well as scRNA-Seq to complement the study by a non-targeted approach. Despite the strong global transcriptomic similarity between 0H and 48HREV cells, four DE genes were identified in the study (*RSFR*, *HBBA*, *TBC1D7*, *HSP90AA1*), interpreted as either a delay or as traces of engagement into differentiation of the 48HREV population, before returning to the self-renewal state. Hence, these first analyses suggested some heterogeneities between undifferentiated cells and reverted cells.

Our kernel-based test confirmed this heterogeneity by detecting a significant difference between undifferentiated cells (0H) and reverted cells (48HREV), both in scRT-qPCR and scRNA-Seq data (p-values $6.15\ 10^{-24}$ and $5.05\ 10^{-05}$ respectively), however considering the test statistic as a distance also confirmed the close proximity between these two conditions (Fig. 2.4.b and 2.8.a). We assumed that population 48HREV was heterogeneous and contained reverted cells and non-reverted cells. A k-means clustering was unable to detect any particular cell cluster (Fig. 2.9, middle). As the discriminant axis provided by our framework represents a synthetic summary of the global transcriptomic differences between two cell populations, it allowed to highlight the existence of a sub-population of 48HREV cells (denoted 48HREV-1) that overlaps the distribution summary of 48HDIFF-cells (48HREV vs. 48HDIFF, Fig. 2.4.c). Interestingly, these cells also matched the distribution summary of 24H-cells (48HREV vs. 24H, Fig. 2.4.c), and were separated from the undifferentiated cells (48HREV vs 0H, Fig. 2.4.c). A similar sub-population was detected using scRNA-Seq data (48HREV vs. 48HDIFF Fig. 2.8.b). According to our test, populations 48HDIFF and 48HREV-1 were very slightly different on scRT-qPCR data and similar on scRNA-Seq data (p-values $2.51\ 10^{-3}$ and 0.88 respectively). This slight difference may be explained by the targeted approach of scRT-qPCR that was based on a selection of 83 genes involved in differentiation and on the higher precision of the scRT-qPCR technology [153]. 48HREV-2 cells (48HREV cells after removing 48HREV-1 cells) were closer but still significantly different from 0H cells in both technologies (p-values $4.36\ 10^{-16}$ and $3.8\ 10^{-05}$ respectively). To describe these two sub-populations in

terms of genes, we performed a k-means clustering on the averaged expression of genes over cells in populations 0H, 24H, 48HDIFF, 48HREV-1, 48HREV-2. We identified three and five gene clusters on the scRT-qPCR and the scRNA-Seq data respectively. These clusters can be separated in three groups (Fig 2.4.d and 2.8.c): (*i*) genes activated during differentiation (scRT-qPCR cluster 2, scRNA-Seq clusters 0 and 2), e.g. hemoglobin related genes such as *HBA1* and *HBAD* (shown in Fig. 2.8.d), (*ii*) genes deactivated during differentiation (scRT-qPCR cluster 1, scRNA-Seq cluster 3) e.g. genes involved in metabolism of self-renewing cells such as *LDHA* and *LY6E* (shown in Fig. 2.8.d), and (*iii*) genes with no clear function pattern for which the expression levels did not change much during differentiation and reversion (scRT-qPCR cluster 0 and scRNA-Seq clusters 1 and 4). The p-value tables associated to each pair-wise univariate DE analysis with respect to each gene cluster are available online.[3]

To conclude, our differential transcriptome framework showed that population 48HREV is composed of two sub-populations, which sheds light on new putative mechanisms driving differentiation and reversion processes. Whereas a population is only slightly different to undifferentiated cells (48HREV-2), a sub-population (48HREV-1) has remained engaged in differentiation. This difference could be either due to a delay in engaging the reversion process for some cells, or to cells having crossed the irreversible point of commitment. Furthermore, our method has identified cellular pathways which could be important either for cell plasticity or cell differentiation, and can guide design of further experiments. Overall, it could enhance our comprehension of how gene regulatory networks react to differentiation and reversion signals.

2.1.4 Towards a New Testing Framework for Differential Binding Analysis in Single-cell ChIP-Seq Data

There is currently no dedicated method for the comparison of single-cell epigenomic profiles, existing studies often use non-parametric testing to compare epigenomic states and retrieve differentially enriched loci. The joint multivariate testing strategy seems particularly suited to compare epigenomic data since it is well established that chromatin conformation and natural spreading of histone modifications (in particular H3K27me3 [91]) can induce complex dependencies between sites occupancy. A recent study [92] has shown that the repressive histone mark H3K27me3 (trimethylation of histone H3 at ly-

[3] `https://github.com/AnthoOzier/kernel_testsDA.git`

sine 27) is involved in the emergence of drug persistence in breast cancer cells. Drug persistence occurs when only a subset of cells, known as persister cells, survives the initial drug treatment, thereby creating a reservoir of cells from which resistant cells will emerge. The study identified a persister expression program involving genes such as *TGFB1* and *FOXQ1*, with H3K27me3 as a lock to its activation. Changes in H3K27me3 modifications at the single-cell level showed a consistent pattern in persister cells compared to untreated cells, in particular cells display recurrent losses of repressive histone methylation at a subset of genes of the persister expression program. However, this pattern was not necessarily maintained in cells that developed full resistance, suggesting that part of the epigenomic features of persister cells might be transient. Moreover, analysis of untreated cells revealed heterogeneity within epigenomic profiles. Part of the population exhibited shared epigenomic features with persister cells, yet remaining distinguishable from them. This initial analysis suggested that a pool of untreated cells could contribute to the persister cell population later upon exposure to chemotherapy. However, unsupervised analyses were unable to clearly identify this pool of precursor cells.

We compared H3K27me3 scChIP-seq profiles between untreated and persister cells using kernel testing. Thanks to the discriminative approach, our framework offers a synthetic representation of the distributional differences between cell populations (Fig 2.5). Projecting cells on the kernelized discriminant axis reveals 3 sub-populations within the untreated cell population: Persister-Like (109 cells; 5% of untreated cells), Intermediate (1124 cells; 57%), Naive (744 cells; 38%), with increasing distance to persister cells (Fig. 2.5). We then performed a differential analysis of H3K27me3 enrichment between persister cells and the $n = 109$ untreated cells that were the most similar to persister cells on the discriminant axis. Over the 6,376 tested regions, only 14 were significantly differentially enriched (p-value$< 10^{-3}$, Table 2.1), suggesting that this sub-population of untreated cells is epigenomically very close to persister cells (with persister cells being hypo-methylated on these significant regions compared to persister-like cells). We then studied the differences between the three populations present in the untreated cell population, prior to any treatment. We performed differential analysis between the most distant untreated cells (*naive* vs *intermediate*), and between *intermediate* cells and *persister-like* cells. We detected significant changes in repressive epigenomic enrichments, both losses and gains, that will need further functional testing to understand their potential role in drug-persistence (Table 2.2). Altogether, our new kernel analytical framework shows that persister-like cells could exist prior to any treatment, and provides a novel level of appre-

ciation of epigenomic heterogeneity by revealing three sub-populations within treatment naive cell population. In addition, our method identifies small quantitative variations that are not detected by other methods and will need to be related to gene expression and other molecular features for further interpretation.

2.1.5 Conclusion

In this work we introduced the framework of kernel testing to perform differential analysis in a non-linear setting. This method compares the distribution of gene expression or epigenomic profiles through global or feature-wise comparisons, but can be extended to any measured single-cell features. Kernel testing has focused much attention in the machine learning community since it has the advantage of being non-linear, computationally tractable, and provides visualization combining dimension reduction and statistical testing. Its application to single-cell data is particularly promising, as it allows distributional comparisons without any assumptions about their shape. Moreover, using a classifier to perform discrimination-based testing has become popular [75], and allows powerful detection of population heterogeneities in both expression and epigenomics single-cell data. Our simulations show the power of this approach on specifically designed alternatives [78]. Furthermore, comparing kernel testing with other methods based on multiple scRNA-Seq data reveals its superior capability to identify distributional changes that go undetected by other approaches. Finally, the application of kernel testing to scRNA-Seq and scChIP-seq data uncovers biologically meaningful heterogeneities in cell populations that were not identified by standard procedures.

Perspectives of this work are numerous: we will first generalize the approach beyond the two-sample case and extend it to multiple sample comparisons. In particular, this will make it possible to have more general design that considers multiple factors such as batch effects. The adaptability of kernel methods makes them particularly well-suited for spatial data, so we plan to extend the framework to include spatial data analysis. To bridge the gap between global and feature-wise approaches, we are actively developing sensitivity analysis methods. These methods will help us identify influential features that contribute to the rejection of the global hypothesis. By combining these findings, we can create a joint approach that considers the complex dependencies inherent in single-cell data, while still providing interpretable outputs based on feature-wise information.

More than ever, single-cell data science appears at the convergence of many cutting-edge methodological developments in machine learning. As a result, these advancements

will have significant implications for the old-tale of differential analysis, offering new avenues for progress and improvement.

2.1.6 Materials and Methods

Simulation Settings

The comparison study on data simulated was performed on data following different mixtures of zero inflated negative binomial (ZINB) distributions [44]. The distribution parameters were chosen to reproduce the four Korthauer alternatives and two types of H_0 distributions. The performances were computed on 500 repetitions of a dataset composed of 1000 DE genes and 9000 non-DE genes. The DE genes are equaly separated in the four alternatives DE,DM, DP and DB. The non-DE genes are equally separated into a unimodal ZINB and a bimodal mixture of ZINB. The DE methods were applied on the raw data, type-I errors and powers were computed on the raw p-values while false discovery and true discovery rates were computed on the adjusted p-values, with the Benjamini-Hochberg correction [19]. The authors also shared their p-values tables with us for their methods (cicdf-asymp and citcdf-perm) [44], MAST [38], scDD [78], SigEMD [145], DESingle [98] and SCDE [73].

Comparison of Methods on Published scRNA-Seq Datasets

The eighteen comparison datasets were downloaded from the Zenodo repository (`https://doi.org/10.5281/zenodo.5048449`) compiled by Squair and coauthors [134]. They consists of six comparisons of bone marrow mononuclear phagocytes from mouse, rat, pig and rabbit in different conditions [59], eight comparisons of naive and memory T cells in different conditions [26] and four comparisons of alveolar macrophages and type II pneumocytes between young and old mouses [7] and between patients with pulmonary fibrosis and control individuals [116]. More details on the datasets are in [134] or in the original studies. The preprocessing step consisted in filtering genes present in less than three cells and normalizing data with the Seurat function *NormalizeData*, as in the original comparative study [134]. This not very restrictive preprocessing was chosen in order to not introduce biaises in the analyses, and many genes would have been ignored form the analysis in real conditions. The Area Under the Concordance Curves (AUCC) scores were computed with the original scripts [134].

Probability Product Kernel for Zero-Inflated Data

To derive the zero-inflated kernel, we consider a zero-inflated Gaussian distribution with π the proportion of additional zeros:

$$Y \sim \pi\delta_0(\bullet) + (1-\pi)f_{\mu,\sigma}(\bullet).$$

with $f_{\mu,\sigma}$ the Gaussian probability density function. This distribution has a mixture representation, with Z standing for the binary variable of distribution $\mathcal{B}(\pi)$, such that:

$$f_{\mu,\sigma,\pi}(x) = \mathbb{P}(Z=1)\delta_0(y) + \mathbb{P}(Z=0)f_{\mu,\sigma}(y).$$

We know the probability kernels for the Gaussian distributions:

$$k_{\text{Gauss}}(\mu,\mu') = \frac{1}{4\pi\sigma^2}\exp(-(\mu-\mu')^2/4\sigma^2),$$

and for the Bernoulli distribution:

$$k_{\mathcal{B}}\Big(\pi,\pi'\Big) = \pi\pi' + (1-\pi)(1-\pi').$$

To get the ZI-Gauss kernel, we compute the probability kernel $f_{\mu,\sigma,\pi}$ and $f_{\mu',\sigma,\pi'}$:

$$\begin{aligned} k_{\text{ZI-Gauss}}\Big(f_{\mu,\sigma,\pi}, f_{\mu',\sigma,\pi'}\Big) &= \int_y f_{\mu,\sigma,\pi}(y) f_{\mu',\sigma,\pi'}(y)dy \\ &= \pi\pi' + \pi(1-\pi')f_{\mu',\sigma}(0) + (1-\pi)\pi' f_{\mu,\sigma}(0) \\ &\quad +(1-\pi)(1-\pi')K_{\text{Gauss}}(\mu,\mu'). \end{aligned}$$

In the simulations, the probability product kernel was computed using the parameters of the Binomial distributions used to determine the drop-out rates of the simulated data (drawn uniformly in $[0.7, 0.9]$), the variance parameter σ was set as the median distance between the non-zero observations and the Gaussian means μ were set as the observed values.

Reversion Data

Details on the experiment and on the data can be found in the original paper [153]. The kernel-based testing framework was performed on the $\log(x+1)$ normalized RT-qPCR data and on the Pearson residuals of the 2000 most variable genes of the scRNA-Seq data

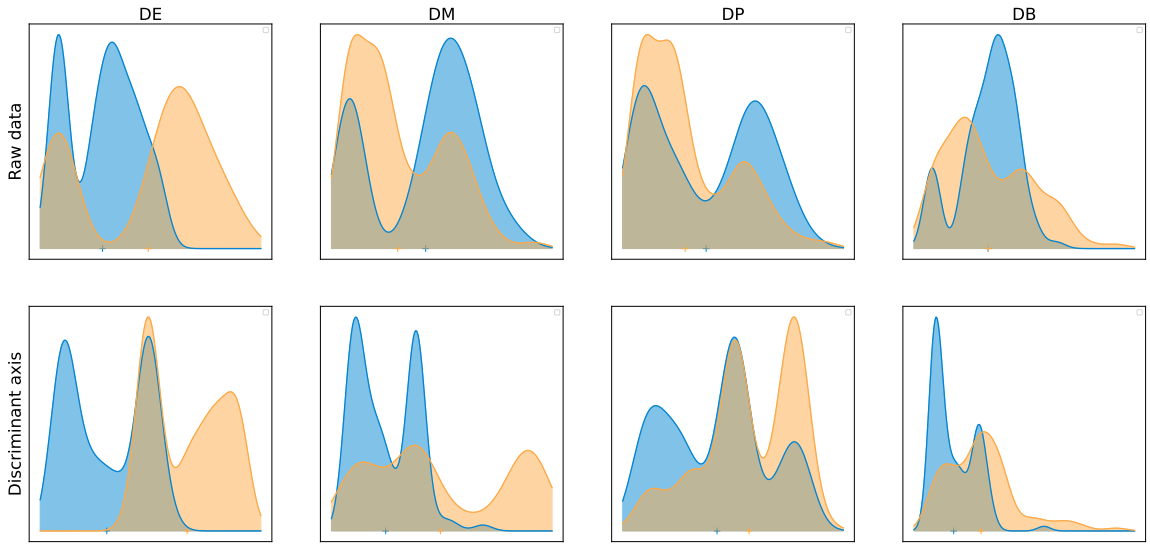

Figure 2.1: Top: Examples of distributions of the simulated data, DE: classical difference in expression, DM: difference in modalities, DP: difference in proportions, DB: difference in both modalities and proportions with equal means. Bottom: projection of cells on the discriminant axis ($T = 4$) for each alternative. The non-linear transform allows the separation of distributions on the discriminant axis.

obtained through the R package `sctransform` [58]. The truncation parameter of the multivariate comparisons ($T = 10$ for both technologies) was chosen to be large enough for the discriminant analysis to capture enough of the multivariate information and to maximize the discriminant ratio. The truncation parameter retained for univariate testing ($T = 4$) was chosen according to the simulation study.

2.1.7 Acknowledgments

The research was supported by a grant from the Agence Nationale de la Recherche ANR-18-CE45-0023 SingleStatOmics, by the projects AI4scMed, France 2030 ANR-22-PESN-0002, and SIRIC ILIAD (INCA-DGOS-INSERM-12558). The authors would like to thank Boris Hejblum for sharing the simulated data, François Gindraud for helping on the implementation of the kernel method, Stéphane Minvielle and Zaid Harchaoui for fruitful scientific discussions. This work was performed using HPC resources from GLiCID computing center.

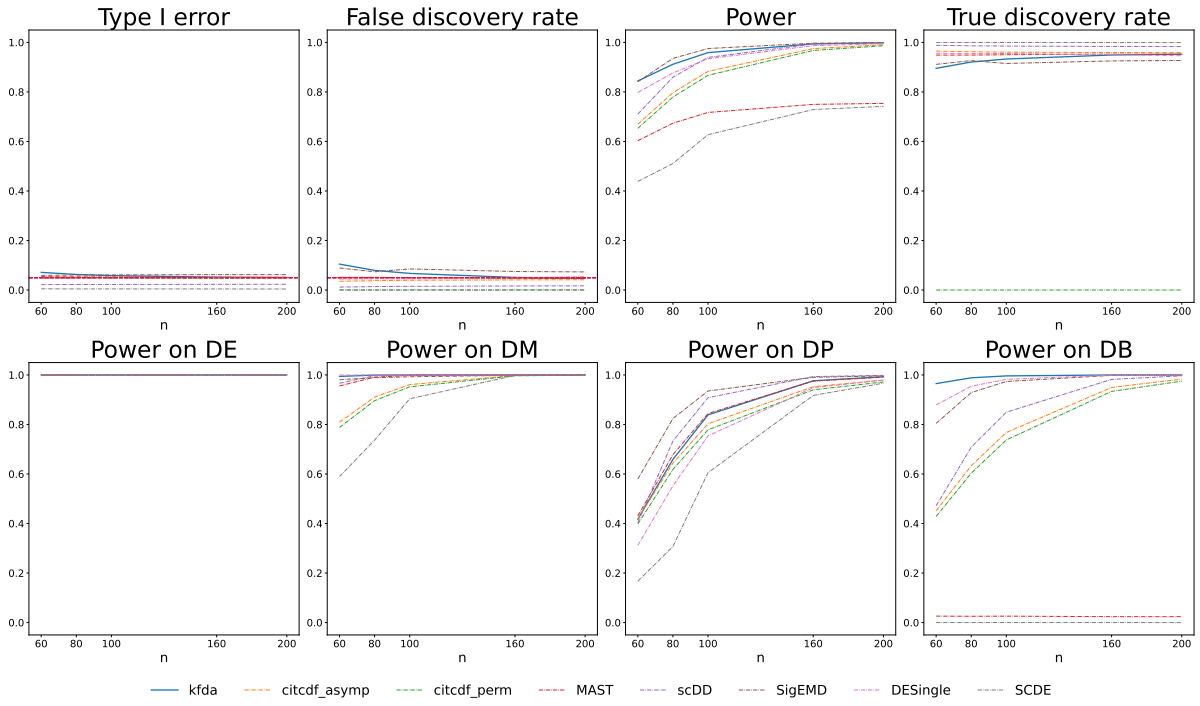

Figure 2.2: Comparison of DEA methods with respect to type-I errors and power. Top: Type-I errors are computed on raw p-values under H_0. False discovery Rate computed on Benjamini-Hochberg adjusted p-values. Power computed on raw p-values under H_1. True Discovery Rate computed on Benjamini-Hochberg adjusted p-values. Simulated data consists of 100 cells, 10000 genes (1000 DE, 9000 non-DE). Alternatives are simulated using DE: classical difference in expression (250 genes), DM: difference in modalities (250 genes), DP: difference in proportions (250 genes), DB: difference in both modalities and proportions with equal means (250 genes). Error rates are computed over 500 replicates.

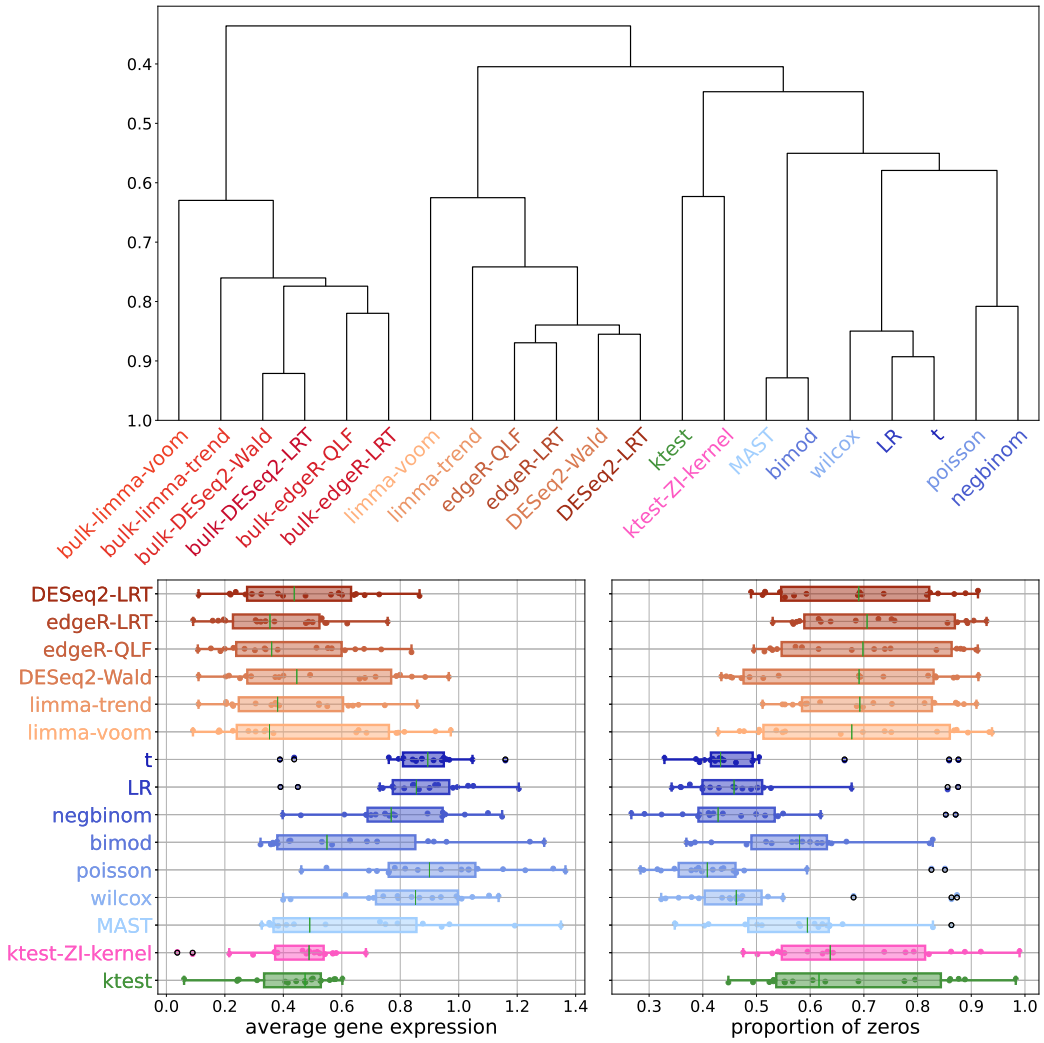

Figure 2.3: Top: Hierarchical clustering based on average AUCC scores computed between pairs of methods (over 18 datasets [134]). Bottom: Boxplot of the average expression (left) and proportion of zeros (right) of the top 500 DE genes for different DE methods (over 18 datasets [134]). Red: bulk methods, orange: pseudobulk methods, blue: single-cell methods. The truncation parameter is set to $T = 4$ for ktest.

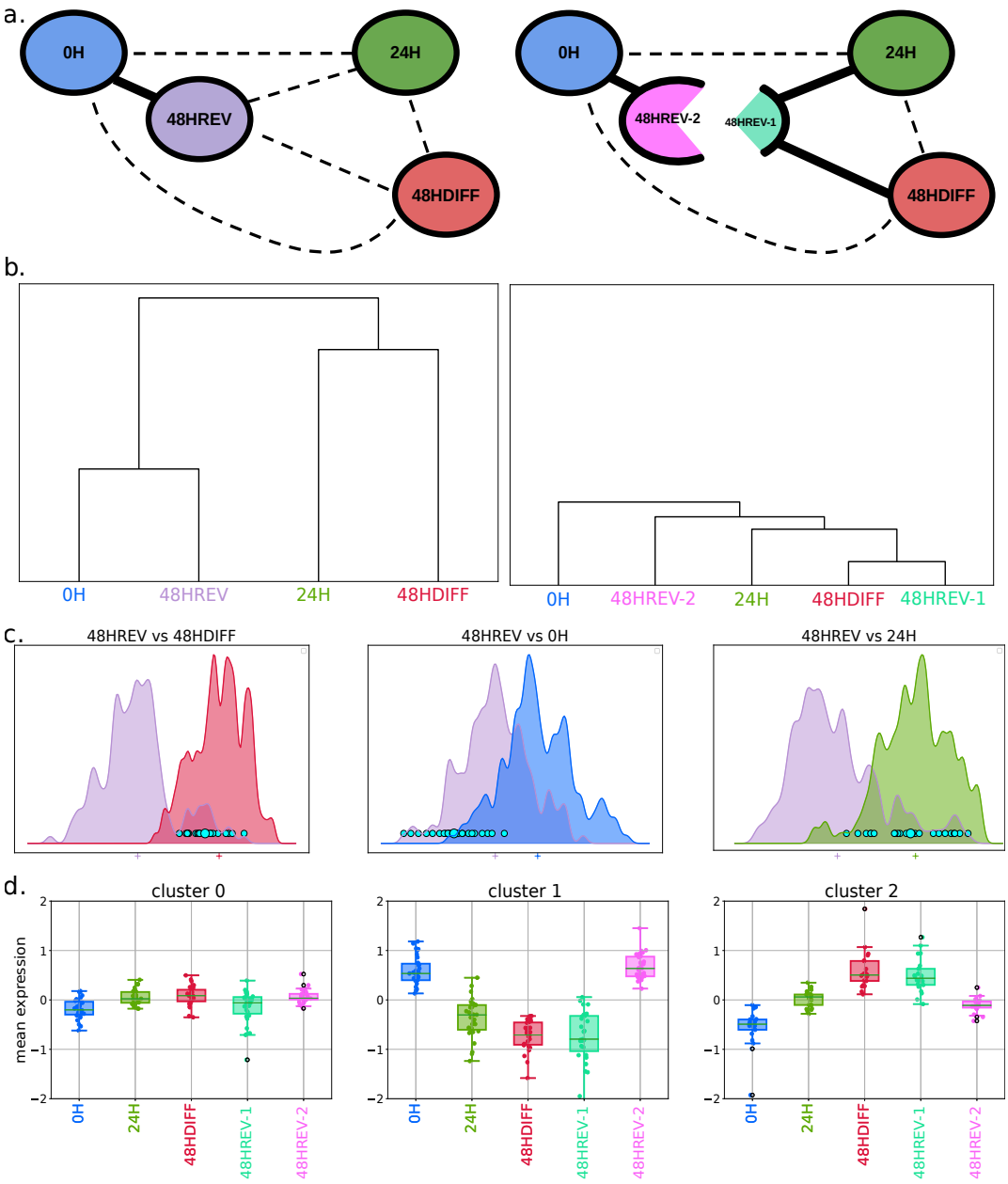

Figure 2.4: a: Summarized distance graphs between conditions before (left) and after (right) splitting condition 48HREV into populations 48HREV-1 and 48HREV-2. b: Trees from pairwise distances using our test statistic between conditions before (left) and after (right) splitting condition 48HREV into populations 48HREV-1 and 48HREV-2. c: Cell densities of compared conditions projected on the discriminant axis between conditions 48HREV and 48HDIFF (left), 48HREV and 0H (middle) and 48HREV and 24H (right) with highlighted population 48HREV-1. d: Boxplots of the mean expressions of the five populations 0H, 24H, 48HDIFF, 48HREV-1 and 48HREV-2 for the three genes clusters. a,b,c and d are obtained from scRT-qPCR data.

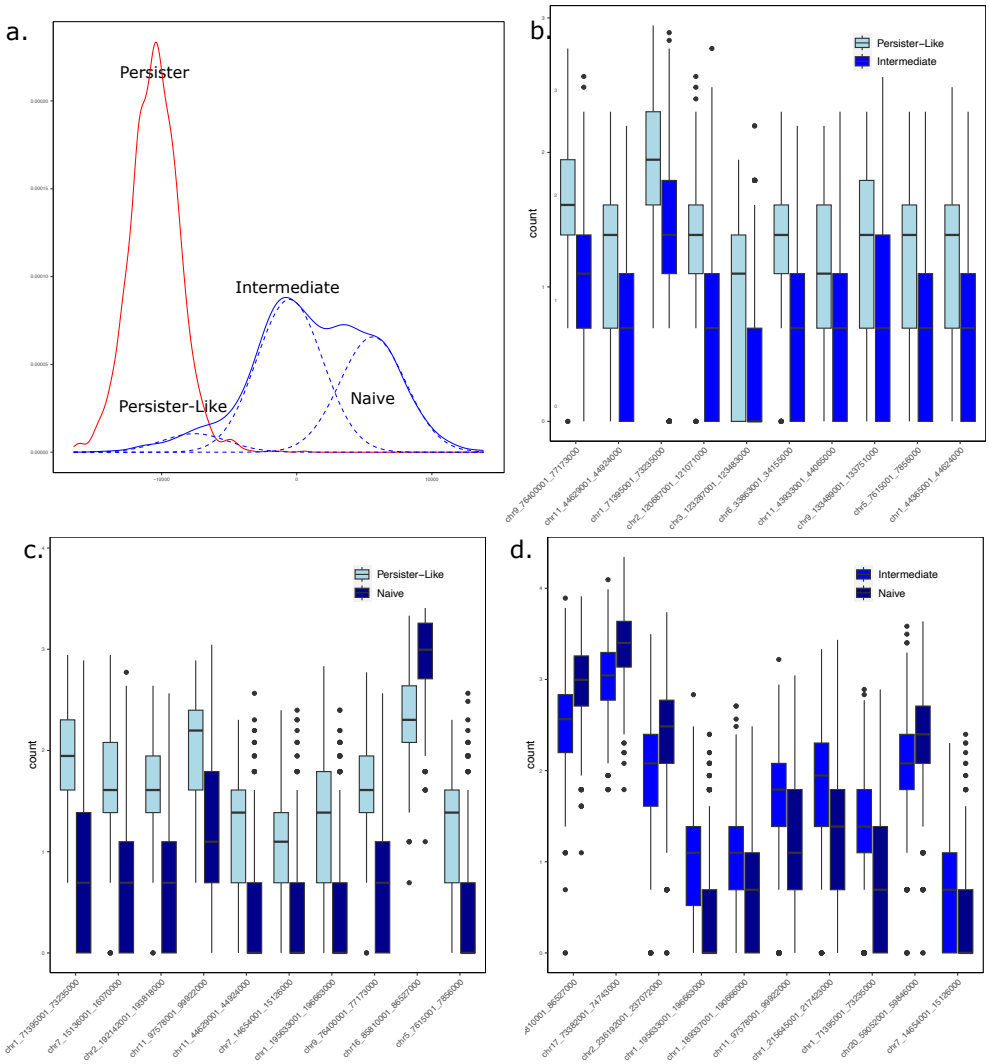

Figure 2.5: Differential analysis of scChIP-seq data on breast cancer cells. a. Cell densities of persister cells vs. untreated cells. Sub-populations of untreated cells were identified using 3-component mixture model, that revealed persister-like cells, intermediate and naive cells. b-c-d: boxplots of the top-10 differentially enriched H3K27me3 loci between the 3 sub-populations. Features are designated by the genomic coordinates of the ChIP-seq peaks. Corresponding overlapping genes are provided in Table 2.1.

2.1.8 Supplementary Material

Tuning the Truncation Hyperparameter

We use the simulation data to calibrate the hyperparameter of our method, i.e. the number T of principal directions of the within-covariance operator to retain to regularize the kernel-based Mahalanobis distance. The theoretical calibration of this hyperparameter still requires heavy mathematical developments, as shown by recent work [60]. However, these simulations provide a simple rule of thumb to choose it. Indeed, since T can be interpreted as the quantity of within-variance information used to describe the residual expression, increasing T will increase power in the detection of complex alternatives, at the price of increased type-I errors. In the simulations, Type-I errors of the kernel test remains at the nominal level $\alpha = 5\%$ until $T \leq 6$. with maximal power for $T = 4$ (Fig 2.6). Interestingly, the test was completely unable to detect the DB alternative when $T = 1$. These results confirm that the truncation hyperparameter should be chosen as a trade-off between maximizing testing power while keeping the type-I errors controlled at the nominal level to ensure calibration. This motivates the choice of $T = 4$ for all the univariate DE analyses in the paper.

For multivariate analyses, we assumed that the meaningful information was contained in more than four principal directions of the within-covariance operator and chose to take a larger truncation parameter in order to take into account more of the multivariate information available. We then chose the truncation parameter $T = 10$ that maximized the discriminant ratio while being not too large to still ensure the calibration.

Kernel Trick for the Effective Computation of the Test Statistic

In this section, we describe how to compute the test statistic $\widehat{D}_T^2(\widehat{\mu}_1, \widehat{\mu}_2)$ and the vector of projections of the embeddings onto the discriminant axis V, with $i \in \{1, 2\}$, $j \in \{1, \ldots, n_i\}$, and $V = (\langle h_T^\star, \phi(Y_{i,j})\rangle_{\mathcal{H}})_{i,j}$ for $T \in \{1, \ldots, n\}$. This computation relies on the kernel trick that consists in expressing every quantity of interest with respect to the gram matrix $\mathbf{K_Y}$ containing every pair-wise evaluation of the kernel function, such that for $i, i' \in \{1, 2\}$, $\mathbf{K_Y} = (\mathbf{K_Y}_{i,i'})_{i,i'}$, where for $j \in \{1, \ldots, n_i\}$, $j' \in \{1, \ldots, n_{i'}\}$, $\mathbf{K_Y}_{i,i'} = (k(Y_{i,j}, Y_{i',j'}))_{j,j'}$. The computation has two steps. First, we determine a matrix $\mathbf{K}_W$ that has the same eigenvalues as the operator $\widehat{\Sigma}_W$, then we compute the quantities of interest with respect to $\mathbf{K_Y}$, the T first eigenvalues $(\widehat{\lambda}_t)$, $t \in \{1, \ldots, T\}$ and the associated unit eigenvectors $(u_t)_t$ of $\mathbf{K}_W$. Let's denote by $\mathbf{I}_n$ the identity matrix of size n, $\mathbf{J}_n$ the

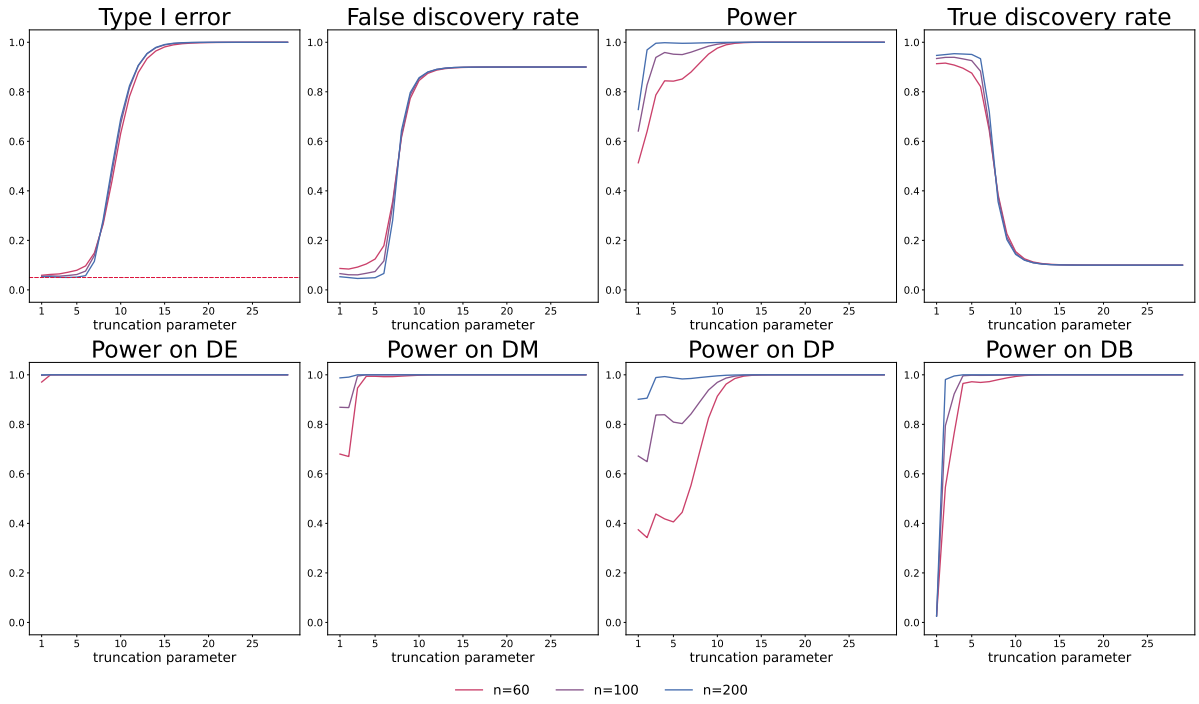

Figure 2.6: Calibration of the truncation with respect to type-I errors and power. Top: Type-I errors are computed on raw p-values under H_0. False discovery Rate computed on Benjamini-Hochberg adjusted p-values. Power computed on raw p-values under H_1. True Discovery Rate computed on Benjamini-Hochberg adjusted p-values. Simulated data consists of 10000 genes (1000 DE, 9000 non-DE). Alternatives are simulated using DE: classical difference in expression (250 genes), DM: difference in modalities (250 genes), DP: difference in proportions (250 genes), DB: difference in both modalities and proportions with equal means (250 genes). Error rates are computed over 500 replicates.

matrix of size n full of 1, and $\mathbb{1}_n$ the vector of size n full of 1. Then for $i \in \{1, 2\}$, let $\mathbf{\Pi}_{n_i} = \mathbf{I}_{n_i} - n_i^{-1}\mathbf{J}_{n_i}$, $\mathbf{\Pi}_W = \text{diag}(\mathbf{\Pi}_{n_1}, \mathbf{\Pi}_{n_2})$ and $\omega = (n_1^{-1}\mathbb{1}_{n_1}, -n_2^{-1}\mathbb{1}_{n_2})' \in \mathbb{R}^n$. We can show that the matrix $\mathbf{K}_W$ is equal to $\mathbf{K}_W = n^{-1}\mathbf{\Pi}_W\mathbf{K}_\mathbf{Y}\mathbf{\Pi}_W$. Then we have:

$$\widehat{D}_T^2(\widehat{\mu}_1, \widehat{\mu}_2) = \frac{n_1 n_2}{n^2}\sum_{t=1}^{T}\widehat{\lambda}_t^{-2}(u_t'\mathbf{\Pi}_W\mathbf{K}_\mathbf{Y}\omega)^2, \quad \text{and} \quad V = \frac{n_1 n_2}{n^2}\sum_{t=1}^{T}\widehat{\lambda}_t^{-2}(u_t'\mathbf{\Pi}_W\mathbf{K}_\mathbf{Y}\omega)\mathbf{K}_\mathbf{Y}\mathbf{\Pi}_W u_t.$$

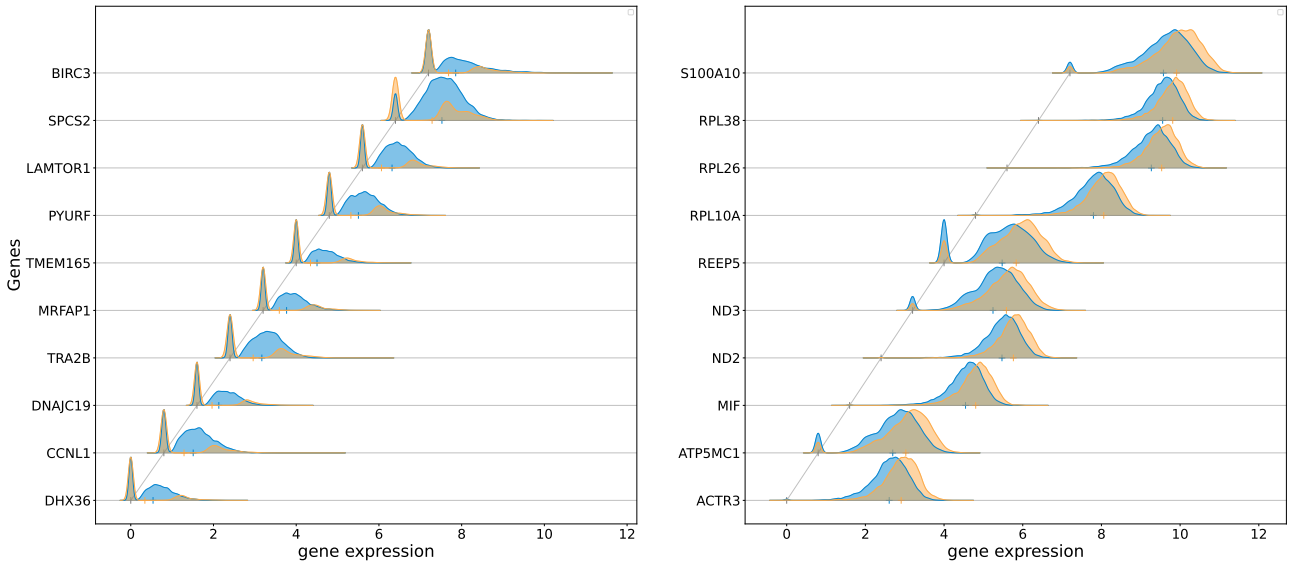

Figure 2.7: Expression densities of the two compared conditions for genes considered as DE by ktest-ZI-kernel and the other single-cell DE methods and considered as non-DE by pseudo-bulk methods. Left: stimulated memory Th0 cells (blue, 4766 cells) vs control memory Th0 cells (orange, 3110 cells) from [26]. Right: pig cells stimulated with lipopolysaccharide (blue, 6605 cells) vs control pig cells (orange, 6148 cells) from [59].

chr	start - end	$\widehat{D}_T^2$	average Persist. / Persist.-Like	average log2FC	gene
chr9	133489001 - 133751000	123.60	0.77 / 1.25	-0.51	ADAMTSL2, DBH, SARDH
chr5	1832001 - 2740000	104.70	2.45 / 2.80	-0.43	IRX4
chr9	135509001 - 135802000	79.40	0.83 / 1.22	-0.43	PAEP, LCN1, OBP2A, SOHLH1, KCNT1, LCN9
chr9	134445001 - 135458000	72.50	1.67 / 2.12	-0.49	OLFM1, FCN2, FCN1, COL5A1
chr9	76400001 - 77173000	51.00	1.21 / 1.59	-0.41	GCNT1
chr14	23331001 - 23355000	50.20	0.05 / 0.12	-0.09	SLC22A17
chr9	136123001 - 136206000	49.10	0.22 / 0.42	-0.22	LHX3
chr12	129982001 - 130786000	48.10	1.05 / 1.41	-0.37	RIMBP2, PIWIL1
chr3	123287001 - 123483000	45.20	0.69 / 0.91	-0.26	ADCY5
chr22	47950001 - 49760000	43.30	2.48 / 2.78	-0.38	FAM19A5

Table 2.1: Differential analysis of sc-chIPseq data: top-10 differential regions for pairwise comparisons between persister cells and persister-like cells. Adjusted p-values are $< 10^{-3}$ (Bonferroni correction). The last Gene column corresponds to the genes overlapping the regions.

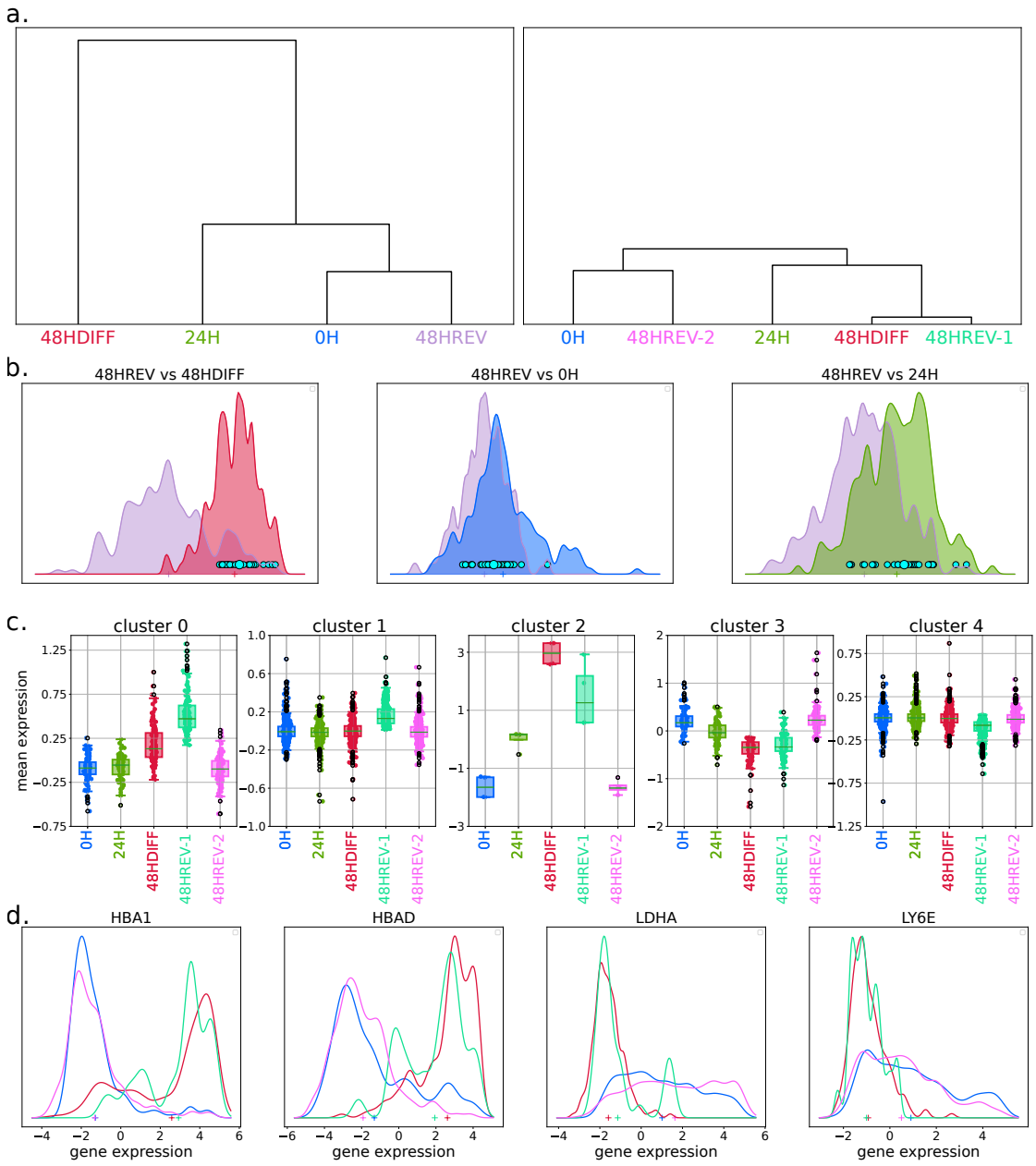

Figure 2.8: a: Trees from pairwise distances using our test statistic between conditions before (left) and after (right) splitting condition 48HREV into populations 48HREV-1 and 48HREV-2. b: Cell densities of compared conditions projected on the discriminant axis between conditions 48HREV and 48HDIFF (left), 48HREV and 0H (middle) and 48HREV and 24H (right) with highlighted population 48HREV-1. c: Boxplots of the mean expressions of the five populations 0H, 24H, 48HDIFF, 48HREV-1 and 48HREV-2 for the five identified genes clusters. d: Examples of gene expression distributions in populations 48HREV-1 (turquoise) and 48HREV-2 (pink) compared to populations 0H (blue) and 48HDIFF (red). a,b,c and d are obtained from scRNA-Seq data

chr	start - end	$\widehat{D}_T^2$	average Persist.-Like / Interm.	average log2FC	gene
chr9	76400001 - 77173000	151.10	1.59 / 0.93	-0.64	GCNT1
chr11	44629001 - 44924000	107.70	1.19 / 0.65	-0.50	TSPAN18
chr1	71395001 - 73235000	102.10	1.94 / 1.37	-0.51	NEGR1
chr2	120687001 - 121071000	99.80	1.31 / 0.76	-0.48	GLI2
chr3	123287001 - 123483000	94.80	0.91 / 0.47	-0.43	ADCY5
chr6	33863001 - 34155000	90.30	1.24 / 0.72	-0.48	GRM4
chr11	43933001 - 44065000	79.80	1.04 / 0.61	-0.40	ACCSL
chr9	133489001 - 133751000	78.50	1.25 / 0.79	-0.41	SARDH, DBH, ADAMTSL2
chr5	7615001 - 7856000	78.20	1.20 / 0.69	-0.42	C5orf49
chr1	44365001 - 44624000	73.70	1.20 / 0.77	-0.40	RNF220
chr	start - end	$\widehat{D}_T^2$	average Interm. / Naive	average log2FC	gene
chr16	85810001 - 86527000	466.50	2.52 / 2.95	0.50	FOXF1-IRF8
chr17	73382001 - 74743000	337.20	3.03 / 3.38	0.41	GPRC5C,CD300A,TTYH2, DNAI2, SDK2, RPL38, GPR142, CD300C, CD300LD, CD300LB, RAB37, KIF19, BTBD17, CD300LF, CD300E
chr2	236192001 - 237072000	314.50	1.99 / 2.41	0.48	IQCA1,ASB18
chr1	195633001 - 196663000	249.90	0.92 / 0.47	-0.51	KCNT2,CFH
chr1	189337001 - 190666000	235.80	1.03 / 0.58	-0.49	BRINP3
chr11	97578001 - 99922000	229.30	1.65 / 1.18	-0.55	CNTN5
chr1	215645001 - 217423000	221.10	1.79 / 1.30	-0.59	ESRRG,USH2A
chr1	71395001 - 73235000	215.20	1.37 / 0.91	-0.52	NEGR1
chr20	59052001 - 59846000	213.30	2.05 / 2.36	0.35	EDN3,PHACTR3
chr7	14654001 - 15126000	209.30	0.70 / 0.34	-0.37	DGKB
chr	start - end	$\widehat{D}_T^2$	average Persist.-Like / Naive	average log2FC	gene
chr1	71395001 - 73235000	292.20	1.94 / 0.91	-1.05	NEGR1
chr7	15136001 - 16070000	250.40	1.65 / 0.70	-0.95	MEOX2,AGMO
chr2	192142001 - 193818000	237.00	1.64 / 0.74	-0.90	TMEFF2
chr11	97578001 - 99922000	230.30	2.04 / 1.18	-0.86	CNTN5
chr11	44629001 - 44924000	217.10	1.19 / 0.39	-0.77	TSPAN18
chr7	14654001 - 15126000	203.80	1.09 / 0.34	-0.72	DGKB
chr1	195633001 - 196663000	201.50	1.31 / 0.47	-0.83	KCNT2,CFH
chr9	76400001 - 77173000	199.90	1.59 / 0.68	-0.91	GCNT1
chr16	85810001 - 86527000	199.30	2.29 / 2.95	0.93	FOXF1,IRF8
chr5	7615001 - 7856000	188.20	1.20 / 0.43	-0.73	C5orf49

Table 2.2: Differential analysis of sc-chIPseq data: top-10 differential regions for pairwise comparisons between the three sub-populations of untreated cells. Adjusted p-values are $< 10^{-3}$ (Bonferroni correction). The last Gene column corresponds to the genes overlapping the regions.

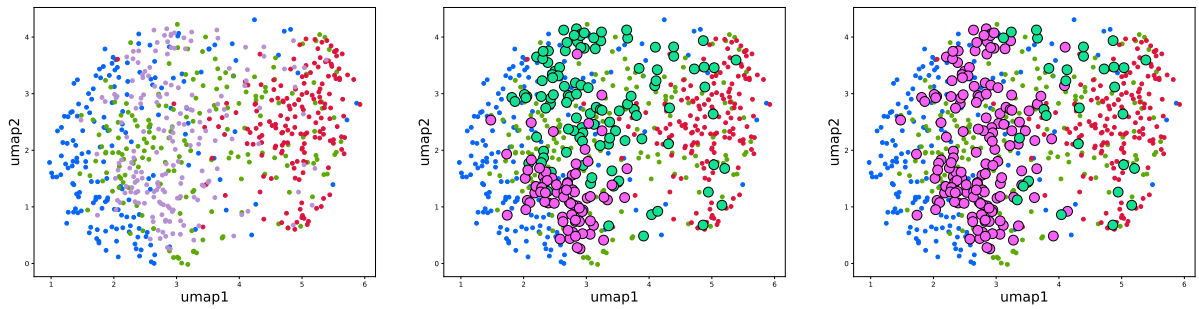

Figure 2.9: Left: Umap representation of the four conditions from scRNA-Seq data (0H (blue), 24H (green) 48HDIFF (red) and 48HREV (purple)). Middle: highlight of the 2 groups of 48HREV identified through a k-means algorithm. Right: The two groups 48HREV-1 (turquoise) and 48HREV-2 (pink) identified on the discriminant axis associated to the truncation parameter $t = 10$.

2.2 The Nyström Method

The computational cost of most kernel methods is polynomial in the number of observations n. When the degree of the polynomial is high, the computational cost may become prohibitive with n increasing. The computational cost of the KFDA statistic is $O(n^3)$, essentially because of the diagonalization of the within-group covariance operator. When datasets have more than 5000 observations, this diagonalization can take hours. The need for low-rank approximations that drastically reduces the computational burden is apparent. Nyström methods are one way of performing it. There exists others approaches, such as spectral clustering, as in [40] and [110], incomplete Cholesky decompositions [39], [11], [12] or random Fourier features-based approximations in the specific case of translation invariant kernels [113] [114] [81]. We will focus on Nyström methods since it has been demonstrated that they show very good performances in [83], [50] and [141].

The Nyström method applied to kernel methods is due to [147] who adapted the Nyström method introduced in [14] to compute integrals. The idea of the Nyström method consists in approximating the Gram matrix of $\mathcal{M}_n(\mathbb{R})$ by a product of three matrices of size $n \times r$, $r \times r$ and $r \times n$ respectively. Theoretically, it resumes to approximating the embeddings of the observations in the feature space $\mathcal{H}$ by their projection in an r-dimensional subspace of interest. The functions that form an orthonormal basis of this subspace are called the anchors. The anchors are determined through a subset of q observations called the landmarks. By working in the subspace spanned by the anchors, the

n-dimensional observations become r-dimensional and it reduces the computational cost of the KFDA statistic from $O(n^3)$ to $O(n^2)+O(r^3)$. The landmarks are generally selected with a random sampling.

2.2.1 The Nyström Landmarks

Let $(\mathcal{Y}, \mathfrak{Y})$ be a measurable space and $k(\cdot,\cdot) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a p.d. kernel associated to the RKHS $\mathcal{H}$ and the feature map $\phi(\cdot)$. We describe the Nyström method for two-sample testing. Let $\mathbf{Y}_1 = (Y_{1,1}, \ldots, Y_{1,n_1})$ and $\mathbf{Y}_2 = (Y_{2,1}, \ldots, Y_{2,n_2})$ two sets of n_1 and n_2 observations from $\mathcal{Y}$ respectively, with $n_1 + n_2 = n$ and $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2) = (Y_{1,1}, \ldots, Y_{1,n_1}, Y_{2,1}, \ldots, Y_{2,n_2})$. Their embeddings in $\mathcal{H}$ through the feature map $\phi(\cdot)$ are denoted $\Phi(\mathbf{Y}_1) = (\phi(Y_{1,1}), \ldots, \phi(Y_{1,n_1}))$ and $\Phi(\mathbf{Y}_2) = (\phi(Y_{2,1}), \ldots, \phi(Y_{2,n_2}))$ respectively, with $\Phi(\mathbf{Y}) = (\Phi(\mathbf{Y}_1), \Phi(\mathbf{Y}_2))$.

Let $\mathbf{Z} = (Z_1, \ldots, Z_q)$ be a set of $q < n$ landmarks in $\mathcal{Y}$. Let $\Phi(\mathbf{Z}) = (\phi(Z_1), \ldots, \phi(Z_n))$ be the embeddings of the landmarks in $\mathcal{H}$, $\mathbf{K}_\mathbf{Z} = (k(Z_i, Z_j))_{i,j\in\{1,\ldots,q\}}$ their associated Gram matrix, and $\mathcal{H}_\mathbf{Z}$ the subspace of $\mathcal{H}$ spanned by the landmarks. The landmarks should be chosen to be representative of the whole sample $\mathbf{Y}$. Common approaches are to define the landmarks $\mathbf{Z}$ as the centroids of a k-means algorithm, or to sample the landmarks in $\mathbf{Y}$, through a score-based sampling or a random sampling. In the case of two-sample testing, we propose to obtain a balanced set of landmarks by selecting $q_1 = \left\lfloor \frac{n_1}{n} \right\rfloor q$ landmarks $\mathbf{Z}_1 = (Z_{1,1}, \ldots, Z_{1,q_1})$ in $\mathbf{Y}_1$ and $q_2 = q - q_1$ landmarks $\mathbf{Z}_2 = (Z_{2,1}, \ldots, Z_{2,q_2})$ in $\mathbf{Y}_2$.

2.2.2 The Nyström Anchors

Our objective is to determine an r-dimensional subspace $\mathcal{H}_a$ of $\mathcal{H}_\mathbf{Z}$ and an orthonormal basis $\mathrm{a} = (a_1, \ldots, a_r)'$ of $\mathcal{H}_a$ composed by the anchors. The space $\mathcal{H}_a$ should be defined so that the embeddings of $\mathbf{Y}$ are well approximated by their orthogonal projections in $\mathcal{H}_a$. Note that when $r = q$, then $\mathcal{H}_a = \mathcal{H}_\mathbf{Z}$ and it resumes to the determination of an orthonormal basis of $\mathcal{H}_\mathbf{Z}$.

The anchors are often defined as the r eigenfunctions associated to the r highest eigenvalues $(\widehat{\lambda}_1^\mathbf{Z}, \ldots, \widehat{\lambda}_r^\mathbf{Z})$ of the empirical kernel covariance operator $\widehat{\Sigma}^\mathbf{Z}$ of $\mathbf{Z}$. Recall that $\widehat{\Sigma}^\mathbf{Z}$

is defined by:

$$\widehat{\Sigma}^{\mathbf{Z}} = \frac{1}{q}\sum_{j=1}^{q}\left(\phi(Z_j) - \widehat{\mu}^{\mathbf{Z}}\right)^{\otimes 2},$$

where $\widehat{\mu}^{\mathbf{Z}} = q^{-1}\sum_{j=1}^{q}\phi(Z_j)$ is the empirical kernel mean embedding associated to $\mathbf{Z}$. According to the kernel trick of the previous chapter (Section 1.2.2), we know that for $i \in \{1, \ldots, r\}$ the i^{th} anchor is such that:

$$a_i = \frac{1}{\sqrt{q\widehat{\lambda}_i^{\mathbf{Z}}}} u_i^{\mathbf{Z}\prime}\mathbf{\Pi}_q\Phi(\mathbf{Z}),$$

where $u_i^{\mathbf{Z}}$ is the eigenvector of $\mathbf{K}^{\mathbf{Z}} = \mathbf{\Pi}_q\mathbf{K}_{\mathbf{Z}}\mathbf{\Pi}_q$ associated to the eigenvalue $\widehat{\lambda}_i^{\mathbf{Z}}$ and $\mathbf{\Pi}_q = \mathbf{I}_m - q^{-1}\mathbf{J}_q$. We also have a matrix formulation for the vector of anchors $\mathrm{a} \in \mathcal{H}^r$:

$$\mathrm{a} = \frac{1}{q}\widehat{\mathbf{\Lambda}_r^{\mathbf{Z}}}^{-\frac{1}{2}}\mathbf{U}_r^{\mathbf{Z}\prime}\mathbf{\Pi}_q\Phi(\mathbf{Z}),$$

where $\widehat{\mathbf{\Lambda}_r^{\mathbf{Z}}} = \mathrm{Diag}(\widehat{\lambda}_1^{\mathbf{Z}}, \ldots, \widehat{\lambda}_r^{\mathbf{Z}}) \in \mathcal{M}_r(\mathbb{R})$ and $\mathbf{U}_r^{\mathbf{Z}} = (u_1^{\mathbf{Z}}, \ldots, u_r^{\mathbf{Z}}) \in \mathcal{M}_{q,r}(\mathbb{R})$ are the diagonal matrix containing the r highest eigenvalues of $\widehat{\Sigma}^{\mathbf{Z}}$ in decreasing order and the matrix containing the r associated eigenvectors of $\mathbf{K}^{\mathbf{Z}}$ as its columns respectively.

In the case of two samples, the embeddings of each group may be better approximated if the anchors are defined to be the orthonormal eigenfunctions associated to the r highest eigenvalues $(\widehat{\lambda}_{W,1}^{\mathbf{Z}}, \ldots, \widehat{\lambda}_{W,r}^{\mathbf{Z}})$ of the empirical within group covariance operator $\widehat{\Sigma}_W^{\mathbf{Z}}$ of $(\mathbf{Z}_1, \mathbf{Z}_2)$ defined such that:

$$\widehat{\Sigma}_W^{\mathbf{Z}} = \frac{q_1}{q}\widehat{\Sigma}^{\mathbf{Z}_1} + \frac{q_2}{q}\widehat{\Sigma}^{\mathbf{Z}_2},$$

where for $i \in \{1, 2\}$, $\widehat{\Sigma}^{\mathbf{Z}_i} = q_i^{-1}\sum_{j=1}^{q_i}\left(\phi(Z_{i,j}) - \widehat{\mu}^{\mathbf{Z}_i}\right)^{\otimes 2}$ and $\widehat{\mu}^{\mathbf{Z}_i} = q_i^{-1}\sum_{j=1}^{q_i}\phi(Z_{i,j})$ are the empirical kernel covariance operator and the empirical kernel mean embedding associated to $\mathbf{Z}_i$ respectively. We know from the diagonalization of the empirical within-group kernel covariance operator of the previous chapter (Section 1.4.3, Equation (1.18)), that for $i \in \{1, \ldots, r\}$, the i^{th} anchor is such that:

$$a_i = \frac{1}{\sqrt{q\widehat{\lambda}_{W,i}^{\mathbf{Z}}}} u_{W,i}^{\mathbf{Z}}{}'\mathbf{\Pi}_W^q\Phi(\mathbf{Z}),$$

where $\mathbf{\Pi}_W^q = \begin{pmatrix} \mathbf{\Pi}_{q_1} & 0 \\ 0 & \mathbf{\Pi}_{q_2} \end{pmatrix}$ and $u_{W,i}^{\mathbf{Z}}$ is the eigenvector of $\mathbf{K}_W^{\mathbf{Z}} = \mathbf{\Pi}_W^q \mathbf{K}_{\mathbf{Z}} \mathbf{\Pi}_W^q$ associated to the eigenvalue $\widehat{\lambda}_{W,i}^{\mathbf{Z}}$. Then the matrix formulation of the resulting vector of anchors $\mathrm{a} \in \mathcal{H}^r$ is such that:

$$\mathrm{a} = \frac{1}{q}\Big(\widehat{\mathbf{\Lambda}}_{W,r}^{\mathbf{Z}}\Big)^{-\frac{1}{2}} \mathbf{U}_{W,r}^{\mathbf{Z}\prime} \mathbf{\Pi}_W^q \Phi(\mathbf{Z}),$$

where $\widehat{\mathbf{\Lambda}}_{W,r}^{\mathbf{Z}} = \mathrm{Diag}(\widehat{\lambda}_{W,1}^{\mathbf{Z}}, \ldots, \widehat{\lambda}_{W,r}^{\mathbf{Z}}) \in \mathcal{M}_r(\mathbb{R})$ and $\mathbf{U}_{W,r}^{\mathbf{Z}} = (u_{W,1}^{\mathbf{Z}}, \ldots, u_{W,r}^{\mathbf{Z}}) \in \mathcal{M}_{q,r}(\mathbb{R})$ are the diagonal matrix containing the r highest eigenvalues of $\widehat{\Sigma}_W^{\mathbf{Z}}$ in decreasing order and the matrix containing the r associated eigenvectors of $\mathbf{K}_W^{\mathbf{Z}}$ as its columns respectively.

Observe that in the two proposed definitions of the anchors, the vector a is defined with respect to a $r \times q$ matrix, so that:

$$\mathrm{a} = \mathrm{A}_r \, \Phi(\mathbf{Z}) \in \mathcal{H}^r, \tag{2.1}$$

where A_r is either defined with respect to the diagonalization of $\widehat{\Sigma}^{\mathbf{Z}}$ or $\widehat{\Sigma}_W^{\mathbf{Z}}$. As we only consider anchors defined like this, we use the notation $\mathrm{A}_r \in \mathcal{M}_{r,q}(\mathbb{R})$ as a generic notation to write the anchors with respect to the landmarks as in Equation (2.1). Thus, an orthogonal projector Π_{a} on $\mathcal{H}_a = \mathrm{Span}\Big(a_1, \ldots, a_r\Big)$ has the following expressions:

$$\Pi_{\mathrm{a}} = \sum_{i=1}^{r} a_i \otimes a_i = \mathrm{a}' \, \mathrm{a} = (\mathrm{A}_r \, \Phi(\mathbf{Z}))'(\mathrm{A}_r \, \Phi(\mathbf{Z})).$$

Moreover, as $a_1, \ldots, a_r \in \mathcal{H}$ is an orthonormal set of functions, we have:

$$\mathrm{a} \, \mathrm{a}' = (\mathrm{A}_r \, \Phi(\mathbf{Z}))(\mathrm{A}_r \, \Phi(\mathbf{Z}))' = \mathbf{I}_r.$$

For $h \in \mathcal{H}$, we denote h^{a} the orthogonal projection of h in $\mathcal{H}_a$ such that:

$$h^{\mathrm{a}} = \Pi_{\mathrm{a}} h = \sum_{j=1}^{r} \langle h, a_j \rangle_{\mathcal{H}} \, a_j.$$

For $y \in \mathcal{Y}$, we denote $\phi_{\mathrm{a}}(y) = \Pi_{\mathrm{a}} \phi(y) = \sum_{j=1}^{r} \langle \phi(y), a_j \rangle_{\mathcal{H}} \, a_j$ the orthogonal projection of $\phi(y)$ on $\mathcal{H}_a$. In the context of the Nyström method, $\phi_{\mathrm{a}}(y)$ is considered as the Nyström approximation of $\phi(y)$. Then for $i \in \{1, 2\}$, the $n_i \times r$ matrix of coordinates of the

projected embeddings $\Phi_{\mathrm{a}}(\mathbf{Y}_i)$ in $\mathcal{H}_a$ is such that:

$$\Phi(\mathbf{Y}_i)\,\mathrm{a}' = \Big(\langle \phi(Y_{i,j}), a_{j'} \rangle_{\mathcal{H}} \Big)_{j \in \{1,\dots,n_i\}, j' \in \{1,\dots,r\}} \in \mathcal{M}_{n_i,r}(\mathbb{R}).$$

When we substitute a by its expression with respect to the landmarks $\Phi(\mathbf{Z})$, we obtain:

$$\begin{aligned}\Phi(\mathbf{Y}_i)\,\mathrm{a}' =& \Phi(\mathbf{Y}_i)\Phi(\mathbf{Z})'\,\mathrm{A}_r' \\ =& \mathbf{K}_{\mathbf{Y},\mathbf{Z}}\,\mathrm{A}_r',\end{aligned}$$

where $\mathbf{K}_{\mathbf{Y}_i,\mathbf{Z}} = \Big(\langle \phi(Y_{i,j}), \phi(Z_{j'}) \rangle_{\mathcal{H}} \Big)_{j \in \{1,\dots,n_i\}, j' \in \{1,\dots,q\}} \in \mathcal{M}_{n_i,q}(\mathbb{R})$. Then, the vector of approximated embeddings $\Phi_{\mathrm{a}}(\mathbf{Y}_i) = \Big(\phi_{\mathrm{a}}(Y_{i,1}), \dots, \phi_{\mathrm{a}}(Y_{i,n_i})\Big) \in \mathcal{H}^{n_i}$ may be expressed in matrix form as:

$$\Phi_{\mathrm{a}}(\mathbf{Y}_i) = \Phi(\mathbf{Y}_i)\,\mathrm{a}'\,\mathrm{a} = \mathbf{K}_{\mathbf{Y}_i,\mathbf{Z}}\,\mathrm{A}_r'\,\mathrm{A}_r\,\Phi(\mathbf{Z}).$$

Then the vector containing the Nyström approximations of both groups is such that:

$$\Phi_{\mathrm{a}}(\mathbf{Y}) = \Phi(\mathbf{Y})\,\mathrm{a}'\,\mathrm{a} = \mathbf{K}_{\mathbf{Y},\mathbf{Z}}\,\mathrm{A}_r'\,\mathrm{A}_r\,\Phi(\mathbf{Z}),$$

where $\mathbf{K}_{\mathbf{Y},\mathbf{Z}} = \begin{pmatrix} \mathbf{K}_{\mathbf{Y}_1,\mathbf{Z}} \\ \mathbf{K}_{\mathbf{Y}_2,\mathbf{Z}} \end{pmatrix} \in \mathcal{M}_{n,q}(\mathbb{R})$. We also define $\mathbf{K}_{\mathbf{Z},\mathbf{Y}} = \mathbf{K}'_{\mathbf{Y},\mathbf{Z}}$. These two formulations highlight that the Nyström approximations of the embeddings can be seen both as linear combinations of the landmarks and linear combinations of the anchors.

2.2.3 Low-Rank Approximations of the MMD and the KFDA Statistics

Let $\mathrm{a} = (a_1, \dots, a_r)' \in \mathcal{H}^r$ be an orthonormal set of r anchors defined with respect to $\mathbf{Z}$ through $\mathrm{a} = \mathrm{A}_r\,\Phi(\mathbf{Z})$, where $\mathrm{A}_r \in \mathcal{M}_{r,q}(\mathbb{R})$. Let $\mathcal{H}_a = \mathrm{Span}(a_1, \dots, a_r)$ and $\Pi_{\mathrm{a}} : \mathcal{H} \to \mathcal{H}$ the orthogonal projector on $\mathcal{H}_a$ such that $\Pi_{\mathrm{a}} = \mathrm{a}'\,\mathrm{a} = \sum_{i=1}^r a_i \otimes a_i$. Let $\Phi_{\mathrm{a}}(\mathbf{Y}) = \Phi(\mathbf{Y})\,\mathrm{a}'\,\mathrm{a} = \mathbf{K}_{\mathbf{Y},\mathbf{Z}}\,\mathrm{A}_r'\,\mathrm{A}_r\,\Phi(\mathbf{Z}) \in \mathcal{H}^n$ the vector of Nyström approximations of the embeddings.

The Nyström MMD Statistic

The biased Nyström MMD statistic $\widehat{\mathrm{MMD}}_{\mathrm{a}}^2$ is defined such that:

$$\widehat{\mathrm{MMD}}_{\mathrm{a}}^2 = \|\widehat{\mu}_1^{\mathrm{a}} - \widehat{\mu}_2^{\mathrm{a}}\|_{\mathcal{H}}^2 ,$$

where for $i \in \{1, 2\}$, the Nyström approximations of the i^{th} empirical kernel mean embedding is such that $\widehat{\mu}_i^{\mathrm{a}} = n_i^{-1} \sum_{j=1}^{n_i} \phi_{\mathrm{a}}(Y_{i,j})$.

Lemma 1. *The biased Nyström MMD statistic has the following expression:*

$$\widehat{\mathrm{MMD}}_{\mathrm{a}}^2 = \omega' \mathbf{K}_{\mathbf{Y},\mathbf{Z}} \, \mathrm{A}_r' \, \mathrm{A}_r \, \mathbf{K}_{\mathbf{Z}} \, \mathrm{A}_r' \, \mathrm{A}_r \, \mathbf{K}_{\mathbf{Z},\mathbf{Y}} \omega, \tag{2.2}$$

where $\omega = (n_1^{-1} \mathbb{1}_{n_1}', -n_2^{-1} \mathbb{1}_{n_2}')' \in \mathbb{R}^n$.

Proof. Observe that we have $\Phi_{\mathrm{a}}(\mathbf{Y})'\omega = \widehat{\mu}_1^{\mathrm{a}} - \widehat{\mu}_2^{\mathrm{a}}$, thus we deduce that:

$$\begin{aligned}
\widehat{\mathrm{MMD}}_{\mathrm{a}}^2 =& \omega' \Phi_{\mathrm{a}}(\mathbf{Y}) \Phi_{\mathrm{a}}(\mathbf{Y})' \omega \\
=& \omega' \mathbf{K}_{\mathbf{Y},\mathbf{Z}} \, \mathrm{A}_r' \, \mathrm{A}_r \, \Phi(\mathbf{Z}) \Phi(\mathbf{Z})' \, \mathrm{A}_r' \, \mathrm{A}_r \, \mathbf{K}_{\mathbf{Z},\mathbf{Y}} \omega \\
=& \omega' \mathbf{K}_{\mathbf{Y},\mathbf{Z}} \, \mathrm{A}_r' \, \mathrm{A}_r \, \mathbf{K}_{\mathbf{Z}} \, \mathrm{A}_r' \, \mathrm{A}_r \, \mathbf{K}_{\mathbf{Z},\mathbf{Y}} \omega .
\end{aligned}$$

□

Note that we also have $\Phi(\mathbf{Y})'\omega = \widehat{\mu}_1 - \widehat{\mu}_2$ and the biased MMD statistic can also be computed as the matrix product:

$$\begin{aligned}
\widehat{\mathrm{MMD}}_b^2 =& \|\widehat{\mu}_1 - \widehat{\mu}_2\|_{\mathcal{H}}^2 \\
=& \omega' \Phi(\mathbf{Y}) \Phi(\mathbf{Y})' \omega \\
=& \omega' \mathbf{K}_{\mathbf{Y}} \omega .
\end{aligned}$$

Whereas this product seems simpler than the Nyström MMD statistic of Equation (2.2), it necessitates exactly $n^2 + n$ operations. Oppositely, the Nyström MMD statistic is such

that:

$$\widehat{\mathrm{MMD}}_{\mathrm{a}}^2 = \omega' \mathbf{K}_{\mathbf{Y},\mathbf{Z}} \mathrm{A}_r' \, \mathrm{A}_r \, \mathbf{K}_{\mathbf{Z}} \mathrm{A}_r' \, \mathrm{A}_r \underbrace{\mathbf{K}_{\mathbf{Z},\mathbf{Y}} \omega}_{q \times n}.$$

$$\underbrace{\qquad\qquad}_{+r \times q}$$

$$\underbrace{\qquad\qquad}_{+q \times r}$$

$$\underbrace{\qquad\qquad}_{+q^2}$$

$$\underbrace{\qquad\qquad}_{+r \times q}$$

$$\underbrace{\qquad\qquad}_{+q \times r}$$

$$\underbrace{\qquad\qquad}_{+n \times q}$$

$$\underbrace{\qquad\qquad}_{+n}$$

Thus the Nyström MMD statistic necessitates exactly $n + q(q + 2n + 4r)$ operations to be computed, to which we have to add the $O(q^3)$ operations needed to determine a set of orthonormal anchors defined as the eigenvectors of an operator of interest. We see that the computation of the MMD statistic is quadratic in the number of observations n while the computation of the Nyström MMD statistic is linear in the number of observations n and cubic in the number of landmarks q (or quadratic if we ignore the determination of the anchors). When $q \ll n$, the Nyström MMD statistic $\widehat{\mathrm{MMD}}_{\mathrm{a}}^2$ is computationally cheaper than the MMD statistic $\widehat{\mathrm{MMD}}$.

Diagonalization of the Nyström Approximation of the Within-Group Covariance Operator

To compute the Nyström truncated KFDA statistic, we need to diagonalize the Nyström within-group covariance operator $\widehat{\Sigma}_W^{\mathrm{a}}$ defined with respect to the Nyström approximations of the embeddings, such that:

$$\widehat{\Sigma}_W^{\mathrm{a}} = \frac{1}{n} \Big(\mathbf{\Pi}_W \Phi_{\mathrm{a}}(\mathbf{Y}) \Big)' \Big(\mathbf{\Pi}_W \Phi_{\mathrm{a}}(\mathbf{Y}) \Big).$$

Lemma 2. *The Nyström within-group covariance operator $\widehat{\Sigma}_W^{\mathrm{a}}$ has the same spectrum than the following matrix:*

$$\mathbf{K}_W^{\mathrm{a}} = \frac{1}{n} \Big(\mathrm{A}_r \, \mathbf{K}_{\mathbf{Z},\mathbf{Y}} \mathbf{\Pi}_W \Big) \Big(\mathrm{A}_r \, \mathbf{K}_{\mathbf{Z},\mathbf{Y}} \mathbf{\Pi}_W \Big)' \in \mathcal{M}_r(\mathbb{R}).$$

Moreover, if the columns of $\mathbf{U}^{\mathrm{a}} = \Big(u_1^{\mathrm{a}}, \ldots, u_r^{\mathrm{a}} \Big)$ are an orthonormal set of eigenvectors of

$\mathbf{K}_W^{\mathrm{a}}$ *associated to the eigenvalues* $\widehat{\lambda}_1^{\mathrm{a}}, \ldots, \widehat{\lambda}_r^{\mathrm{a}}$, *then* $\mathrm{f}^{\mathrm{a}} = \Big(\mathrm{a}' \, u_1^{\mathrm{a}}, \ldots, \mathrm{a}' \, u_r^{\mathrm{a}} \Big)$ *is an orthonormal set of eigenfunctions of* $\widehat{\Sigma}_W^{\mathrm{a}}$.

Proof. We know from the previous chapter (Section 1.4.3) that the operator $\widehat{\Sigma}_W^{\mathrm{a}}$ has the same spectrum than the matrix $\widetilde{\mathbf{K}}_W^{\mathrm{a}} \in \mathcal{M}_n(\mathbb{R})$ defined such that:

$$\begin{aligned}
\widetilde{\mathbf{K}}_W^{\mathrm{a}} =& \frac{1}{n} \Big(\mathbf{\Pi}_W \Phi_{\mathrm{a}}(\mathbf{Y}) \Big) \Big(\mathbf{\Pi}_W \Phi_{\mathrm{a}}(\mathbf{Y}) \Big)'. \\
=& \frac{1}{n} \mathbf{\Pi}_W \Phi(\mathbf{Y}) \, \mathrm{a}' \underbrace{\mathrm{a} \, \mathrm{a}'}_{\mathbf{I}_r} \mathrm{a} \, \Phi(\mathbf{Y})' \mathbf{\Pi}_W \\
=& \frac{1}{n} \mathbf{\Pi}_W \Phi(\mathbf{Y}) \, \mathrm{a}' \, \mathrm{a} \, \Phi(\mathbf{Y})' \mathbf{\Pi}_W \\
=& \frac{1}{n} \Big(\mathrm{a} \, \Phi(\mathbf{Y})' \mathbf{\Pi}_W \Big)' \Big(\mathrm{a} \, \Phi(\mathbf{Y})' \mathbf{\Pi}_W \Big).
\end{aligned}$$

Then trivial matrix algebra shows us that this matrix has the same spectrum than the matrix:

$$\begin{aligned}
\mathbf{K}_W^{\mathrm{a}} =& \frac{1}{n} \Big(\mathrm{a} \, \Phi(\mathbf{Y})' \mathbf{\Pi}_W \Big) \Big(\mathrm{a} \, \Phi(\mathbf{Y})' \mathbf{\Pi}_W \Big)' \\
=& \frac{1}{n} \Big(\mathrm{A}_r \, \mathbf{K}_{\mathbf{Z},\mathbf{Y}} \mathbf{\Pi}_W \Big) \Big(\mathrm{A}_r \, \mathbf{K}_{\mathbf{Z},\mathbf{Y}} \mathbf{\Pi}_W \Big)' \in \mathcal{M}_r(\mathbb{R}).
\end{aligned}$$

Thus the operator $\widehat{\Sigma}_W^{\mathrm{a}} \in \mathrm{HS}(\mathcal{H})$ and the matrix $\mathbf{K}_W^{\mathrm{a}} \in \mathcal{M}_r(\mathbb{R})$ share the same spectrum. Moreover, we deduce that $\widehat{\Sigma}_W^{\mathrm{a}}$ has a maximal rank r that corresponds to the maximal rank of an operator defined on $\mathcal{H}_a$ that has dimension r. We now determine the relation between the eigenfunctions of $\widehat{\Sigma}_W^{\mathrm{a}}$ and the eigenvectors of $\mathbf{K}_W^{\mathrm{a}}$. Let u^{a} an eigenvector of $\mathbf{K}_W^{\mathrm{a}}$ associated to the eigenvalue $\widehat{\lambda}^{\mathrm{a}}$, we have:

$$\mathbf{K}_W^{\mathrm{a}} u^{\mathrm{a}} = \widehat{\lambda}^{\mathrm{a}} u^{\mathrm{a}}.$$

As we have $\mathrm{a} \, \mathrm{a}' = \mathbf{I}_r$, we show that:

$$\begin{aligned}
& \mathrm{a} \, \mathrm{a}' \, \mathbf{K}_W^{\mathrm{a}} \, \mathrm{a} \, \mathrm{a}' \, u^{\mathrm{a}} = \widehat{\lambda}^{\mathrm{a}} \, \mathrm{a} \, \mathrm{a}' \, u^{\mathrm{a}} \\
\Leftrightarrow & \frac{1}{n} \mathrm{a} \Big(\mathrm{a}' \, \mathrm{a} \, \Phi(\mathbf{Y})' \mathbf{\Pi}_W \Big) \Big(\mathrm{a}' \, \mathrm{a} \, \Phi(\mathbf{Y})' \mathbf{\Pi}_W \Big)' \mathrm{a}' \, u^{\mathrm{a}} = \widehat{\lambda}^{\mathrm{a}} \, \mathrm{a} \, \mathrm{a}' \, u^{\mathrm{a}} \\
\Leftrightarrow & \mathrm{a} \, \widehat{\Sigma}_W^{\mathrm{a}} \, \mathrm{a}' \, u^{\mathrm{a}} = \widehat{\lambda}^{\mathrm{a}} \, \mathrm{a} \, \mathrm{a}' \, u^{\mathrm{a}}
\end{aligned}$$

Then we multiply both sides of the equation by $n^{-1}\Big(\mathrm{a}'\,\mathrm{a}\,\Phi(\mathbf{Y})'\mathbf{\Pi}_W\Big)\Big(\mathrm{a}\,\Phi(\mathbf{Y})'\mathbf{\Pi}_W\Big)'$:

$$\frac{1}{n}\Big(\mathrm{a}'\,\mathrm{a}\,\Phi(\mathbf{Y})'\mathbf{\Pi}_W\Big)\Big(\mathrm{a}\,\Phi(\mathbf{Y})'\mathbf{\Pi}_W\Big)'\,\mathrm{a}\,\widehat{\Sigma}_W^{\mathrm{a}}\,\mathrm{a}'\,u^{\mathrm{a}} = \widehat{\lambda}^{\mathrm{a}}\frac{1}{n}\Big(\mathrm{a}'\,\mathrm{a}\,\Phi(\mathbf{Y})'\mathbf{\Pi}_W\Big)\Big(\mathrm{a}\,\Phi(\mathbf{Y})'\mathbf{\Pi}_W\Big)'\,\mathrm{a}\,\mathrm{a}'\,u^{\mathrm{a}}$$
$$\Leftrightarrow \widehat{\Sigma}_W^{\mathrm{a}}\widehat{\Sigma}_W^{\mathrm{a}}\,\mathrm{a}'\,u^{\mathrm{a}} = \widehat{\lambda}^{\mathrm{a}}\widehat{\Sigma}_W^{\mathrm{a}}\,\mathrm{a}'\,u^{\mathrm{a}}$$

We conclude by multiplying both sides of the equation on the left by the pseudo-inverse

$$\widehat{\Sigma}_W^{\mathrm{a}\,-1} = \sum_{\substack{t=1\\ \widehat{\lambda}_t^{\mathrm{a}}>0}}^{r}\frac{1}{\widehat{\lambda}_t^{\mathrm{a}}}\Big(\widehat{f}_t^{\mathrm{a}}\otimes\widehat{f}_t^{\mathrm{a}}\Big)$$

of $\widehat{\Sigma}_W^{\mathrm{a}}$, where $\widehat{f}_1^{\mathrm{a}},\ldots,\widehat{f}_r^{\mathrm{a}}$ is an orthonormal set of eigenfunctions of $\widehat{\Sigma}_W^{\mathrm{a}}$ associated to the eigenvalues $\widehat{\lambda}_1^{\mathrm{a}},\ldots,\widehat{\lambda}_r^{\mathrm{a}}$. We then conclude that:

$$\widehat{\Sigma}_W^{\mathrm{a}}\,\mathrm{a}'\,u^{\mathrm{a}} = \widehat{\lambda}^{\mathrm{a}}\,\mathrm{a}'\,u^{\mathrm{a}}.$$

Note that $\|\mathrm{a}'\,u^{\mathrm{a}}\|_{\mathcal{H}} = 1$. Thus, if the columns of $\mathbf{U}^{\mathrm{a}} = \Big(u_1^{\mathrm{a}},\ldots,u_r^{\mathrm{a}}\Big)$ are an orthonormal set of eigenvectors of $\mathbf{K}_W^{\mathrm{a}}$ associated to the eigenvalues $\widehat{\lambda}_1^{\mathrm{a}},\ldots,\widehat{\lambda}_r^{\mathrm{a}}$, then $\mathrm{f}^{\mathrm{a}} = \Big(\mathrm{a}'\,u_1^{\mathrm{a}},\ldots,\mathrm{a}'\,u_r^{\mathrm{a}}\Big)$ is an orthonormal set of eigenfunctions of $\widehat{\Sigma}_W^{\mathrm{a}}$. □

The Nyström Truncated KFDA Statistic

To define the Nyström truncated KFDA statistic $\widehat{\mathrm{D}_T^{\mathrm{a}}}^2$ for $T \leq r$, we consider the sum formulation of the truncated KFDA statistic of Equation (1.21):

$$\widehat{\mathrm{D}_T}^2 = \sum_{t=1}^{T}\frac{n_1 n_2}{n\widehat{\lambda}_t}\left\langle\widehat{f}_t, \widehat{\mu}_1 - \widehat{\mu}_2\right\rangle_{\mathcal{H}}^2,$$

where $\widehat{f}_1,\ldots,\widehat{f}_T$ are the orthonormal eigenfunctions of $\widehat{\Sigma}_W$ associated to the decreasing eigenvalues $\widehat{\lambda}_1,\ldots,\widehat{\lambda}_T$. The Nyström truncated KFDA statistic is obtained by substituting $\widehat{f}_1,\ldots,\widehat{f}_T$ and $\widehat{\lambda}_1,\ldots,\widehat{\lambda}_T$ by $\widehat{f}_1^{\mathrm{a}},\ldots,\widehat{f}_T^{\mathrm{a}}$ and $\widehat{\lambda}_1^{\mathrm{a}},\ldots,\widehat{\lambda}_T^{\mathrm{a}}$ respectively. Note that it is

unnecessary to replace $\widehat{\mu}_1 - \widehat{\mu}_2$ by $\widehat{\mu}_1^{\mathrm{a}} - \widehat{\mu}_2^{\mathrm{a}}$ as for $t \in \{1, \ldots, T\}$, we have:

$$\begin{aligned}\left\langle \widehat{f}_t^{\mathrm{a}}, \widehat{\mu}_1 - \widehat{\mu}_2 \right\rangle_{\mathcal{H}} =& u_t^{\mathrm{a}\prime}\, \mathrm{a}\, \Phi(\mathbf{Y})'\omega \\ =& u_t^{\mathrm{a}\prime}\, \mathrm{a}\, \mathrm{a}'\, \mathrm{a}\, \Phi(\mathbf{Y})'\omega \\ =& u_t^{\mathrm{a}\prime}\, \mathrm{a}\, \Phi_{\mathrm{a}}(\mathbf{Y})'\omega \\ =& \left\langle \widehat{f}_t^{\mathrm{a}}, \widehat{\mu}_1^{\mathrm{a}} - \widehat{\mu}_2^{\mathrm{a}} \right\rangle_{\mathcal{H}}.\end{aligned}$$

This result is explained by the fact that it is equivalent to project $\widehat{\mu}_1 - \widehat{\mu}_2 \in \mathcal{H}$ onto $\mathcal{H}_a$ to obtain $\widehat{\mu}_1^{\mathrm{a}} - \widehat{\mu}_2^{\mathrm{a}} \in \mathcal{H}_a$, and then to project it on the axis supported by $\widehat{f}_t \in \mathcal{H}_a$, then to directly project $\widehat{\mu}_1 - \widehat{\mu}_2 \in \mathcal{H}$ onto the same axis.

Lemma 3. *The Nyström approximation of the truncated KFDA statistic is defined such that:*

$$\widehat{\mathrm{D}_T^{\mathrm{a}}}^2 = \sum_{t=1}^{T} \frac{n_1 n_2}{n\widehat{\lambda}_t^{\mathrm{a}}} \left\langle \widehat{f}_t^{\mathrm{a}}, \widehat{\mu}_1 - \widehat{\mu}_2 \right\rangle_{\mathcal{H}}^2.$$

Moreover, we have the following kernel trick:

$$\widehat{\mathrm{D}_T^{\mathrm{a}}}^2 = \sum_{t=1}^{T} \frac{n_1 n_2}{n\widehat{\lambda}_t^{\mathrm{a}}} \left(u_{t\prime}^{\mathrm{a}}\, \mathrm{A}_r\, \mathbf{K}_{\mathbf{Z},\mathbf{Y}}\omega \right)^2.$$

Proof. We have:

$$\begin{aligned}\left\langle \widehat{f}^{\mathrm{a}}, \widehat{\mu}_1 - \widehat{\mu}_2 \right\rangle_{\mathcal{H}} =& u_{t\prime}^{\mathrm{a}}\, \mathrm{a}\, \Phi(\mathbf{Y})'\omega \\ =& u_{t\prime}^{\mathrm{a}}\, \mathrm{A}_r\, \mathbf{K}_{\mathbf{Z},\mathbf{Y}}\omega.\end{aligned}$$

□

Computational Cost of the Nyström Truncated KFDA Statistic

To determine the computational cost of the procedure, we need to determine both the computational cost of determining $\mathbf{K}_W^{\mathrm{a}}$ and of the test statistic. We find that the com-

putation of $\mathbf{K}_W^{\mathrm{a}}$ can be done for $r(n^2 + 2qn + qr) = O(n^2 r)$ operations:

$$\widehat{\Sigma}_W^{\mathrm{a}} = \underbrace{\frac{1}{n}\mathrm{A}_r \underbrace{\mathbf{K}_{\mathbf{Z},\mathbf{Y}} \underbrace{\mathbf{\Pi}_W \underbrace{\mathbf{K}_{\mathbf{Y},\mathbf{Z}}\,\mathrm{A}_r'}_{nqr}}_{+n^2 r}}_{+qnr}}_{+rqr}.$$

A similar approach shows that the computation of $\widehat{\mathrm{D}_T^{\mathrm{a}}}^2$ may be done for $T(nq + r(q+1))$, that is linear in every parameter. We then add the cost of the determination of the anchors ($O(q^3)$ operations) and the diagonalization of $\mathbf{K}_W^{\mathrm{a}} \in \mathcal{M}_r(\mathbb{R})$ ($O(r^3)$ operations), to obtain the total cost of the procedure that is $O(q^3 + r^3 + n^2 r)$. As $r \leq q$, when $q \ll n$, this procedure that is quadratic in the number of observations n is cheaper than the computation of the truncated KFDA statistic $\widehat{\mathrm{D}_T}^2$ that is cubic in the number of observations because the diagonalization of $\mathbf{K}_W \in \mathcal{M}_n(\mathbb{R})$ necessitates $O(n^3)$ operations.

2.2.4 Data Simulations

Simulation Procedure

We aim at assessing the quality of the Chi-square approximation on the Nyström truncated KFDA statistic by computing the empirical level of the testing procedure on simulated data. We simulate couples of samples following the same Gaussian distribution. To mimic a typical high-dimensional problem, we simulate data in a high dimensional space of dimension d with only $p < d$ informative features, so that the data actually lie in a lower dimensional space than the ambiant space. We refer to d as the global dimension and to p as the intrinsic dimension. To do so, we consider a distribution $\mathbb{P}_{d,p}$ such that:

$$\mathbb{P}_{d,p} = \mathcal{N}\left(\begin{pmatrix} 0_p \\ 0_{d-p} \end{pmatrix}, \begin{pmatrix} (1+\sigma)\mathbf{I}_p & 0 \\ 0 & \sigma\mathbf{I}_{d-p} \end{pmatrix}\right),$$

where $(1+\sigma)\mathbf{I}_p$ represents the covariance of the data in the lower dimensional space and $\sigma \in \mathbb{R}$ is an isotropic noise present in the whole space. Let $(\mathbf{Y}_1^1, \mathbf{Y}_2^1), \ldots, (\mathbf{Y}_1^C, \mathbf{Y}_2^C)$ be $C \geq 1$ couples of i.i.d samples of size n_1 and n_2 following the same distribution $\mathbb{P}_{d,p}$ with $n_1 + n_2 = n$. The false positive rate or type I error rate for a given level α of an asymptotic testing procedure based on a Nyström truncated KFDA statistic $\widehat{\mathrm{D}_T^{\mathrm{a}}}^2$ for $T \geq 1$ is obtained

through the followin expression:

$$\mathrm{FP}^{a}_{asymp}(T, C) = \frac{1}{C}\sum_{c=1}^{C} \mathbb{1}_{\widehat{\mathrm{D}^{\mathrm{a}}_{T}}^{2}(\mathbf{Y}^{c}_{1}, \mathbf{Y}^{c}_{2}) > q_{\chi^2(T)}(\alpha)},$$

where $q_{\chi^2(T)}(\alpha)$ is the α quantile of the $\chi^2(T)$ distribution with T degrees of freedom.

For every simulation $\sigma = 0.001$, the two samples have the same number of observations and we use the Gaussian kernel with median bandwidth. The level α is set to 5%

Results

We first assess the influence of n, d and p on the testing procedure with the non-Nyström truncated KFDA test. The results are shown in Figure 2.10. We observe that the quality of the asymptotic approximation increases when n increases. The results show that the type I error rate depends on the truncation parameter choice. However, when the truncation parameter is lower than the intrinsic dimension and the number of observations is large enough, we observe that the test is well calibrated.

The False positive rate increases until $T = p$ and then starts to decrease. Precisely, the nominal level of α seems to be reached for truncation parameters close to the intrinsic dimension ($T \simeq p$). Investigating the spectrum of the within-group covariance operator shows that the eigenvalues of the within-group covariance associated to $t > T$ are a few orders smaller than the eigenvalues associated to $t \leq T$, suggesting that eigendirections associated to $T \geq p$ catch random noise and/or are difficult to estimate. Thus the change in the type-I error rates when T reaches p is expected. We have no explanation for the pattern observed with pics and falls after $T \geq p$. We assume that it is highly related to numerical precision issues as the associated eigenvalues are very close to zero. We observe that the False Positive rate explodes when T is too large. This confirms that T should not be chosen too large and that the number of observations should be large enough (hundreds of cells), which is compatible with single-cell datasets.

Then, we compare the performances of randomly sampled landmarks with kmeans centroids landmarks, using the same anchor definition for both. We compared the performances to a reference standard truncated KFDA statistic. Interestingly, when $T \leq p$, the false positive rates of the two Nyström approximations are not differentiable from the

standard truncated KFDA statistic. When $T > p$, a difference appears and the random sampling seems to be slightly closer to the reference. It motivates the random sampling choice as the default landmark choice in our package. The results are shown in Figure 2.11. Note that for the Nyström approximations, the number of anchors upper-bounds the possible truncation parameters. We also studied the effect of the number of landmarks on the test procedure. When the number of landmarks is large enough (≥ 200), using more landmarks has negligeable effect on the test performances. The results are not shown.

To conclude this simulation studies, we compared the False Positive rates of the testing procedure when the anchors are defined as the normalized eigenfunctions of the total covariance of the landmarks with the procedure when the anchors are defined as the normalized eigenfunctions of the within-group covariance of the landmarks. The results are shown in Figure 2.12. It seems that the anchors definition has no effect when the intrinsic dimension p is small. We observe that when $p = 30$, the anchors defined as the normalized eigenfunctions of the within-group covariance of the landmarks, the performances of the Nyström approximation are closer to the performance of the reference. It suggests that the sub-space of the feature space defined by these landmarks is closer to the sub-space spanned by the eigenfunctions of the empirical within-group covariance operator of the observations that the Nyström procedure is approximating. Then we studied the effect of the anchor definition and the number of anchors. According to the results in Figure 2.13, increasing the number of anchors increases the test performances.

Finally, this simulation studies show that the truncation parameter should not be chosen too large in general, and that the Nyström approximation is efficient and does not impact the performances too much. Moreover, it suggests to randomly sample the landmarks and to define the anchors as the eigenfunctions of the within-group covariance operator of the landmarks for the Nyström KFDA procedure.

2.2.5 Discussion

Improvements on the Nyström Methods

An important part of the research on the Nyström approach is focused on improving the choice of the landmarks. The greedy approaches select each new landmark in order to minimize a criterion, typically an error, as in [133] and [105]. One may also draw the

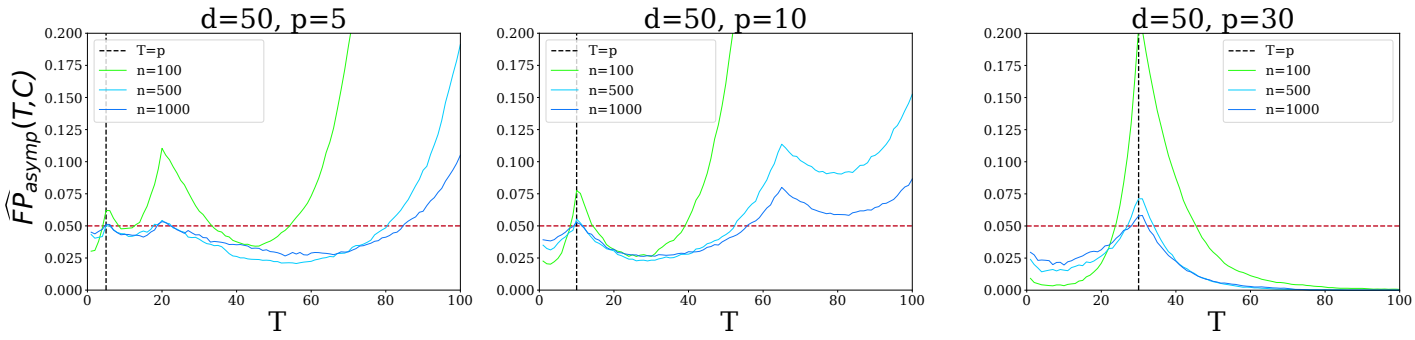

Figure 2.10: Asymptotic distribution-based empirical type-I error w.r.t. T for $n \in \{100, 500, 1000\}$ (green, sky blue and deep blue respectively), $d = 50$ and $p \in \{5, 10, 30\}$.

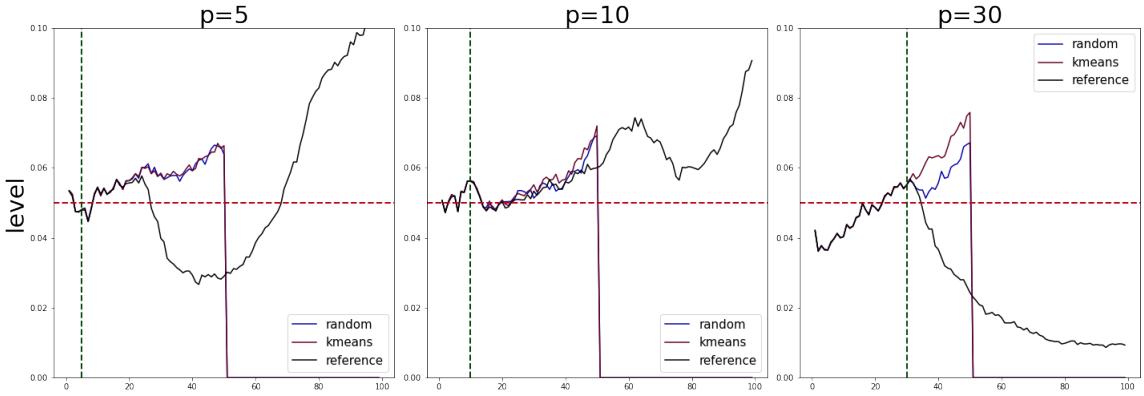

Figure 2.11: Asymptotic distribution-based empirical type-I error w.r.t. T for the truncated KFDA statistic and two Nyström approximations based on random sampled landmarks and kmeans centroïds landmarks (black, blue, purple respectively). Simulations are done for $n = 2000$ observations, $q = 200$ landmarks, $r = 50$ anchors and $C = 6000$ repetitions, $d = 50$ and $p \in \{5, 10, 30\}$.

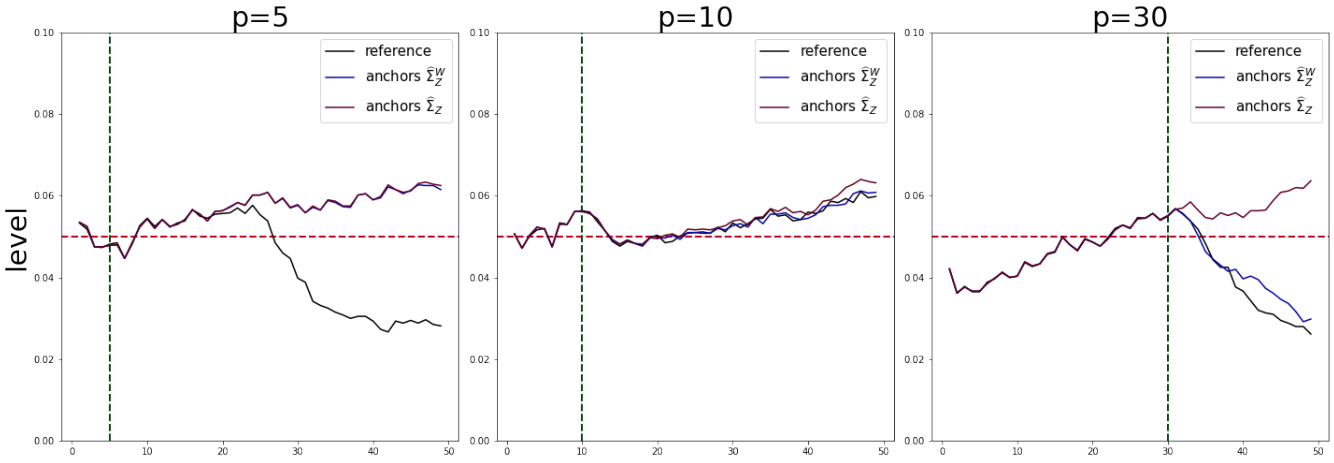

Figure 2.12: Asymptotic distribution-based empirical type-I error w.r.t. T for the truncated KFDA statistic and two Nyström approximations based on within-group covariance based anchors and total covariance base anchors (black, blue, purple respectively). Simulations are done for $n = 2000$ observations, $q = 200$ landmarks, $r = 50$ anchors and $C = 6000$ repetitions, $d = 50$ and $p \in \{5, 10, 30\}$.

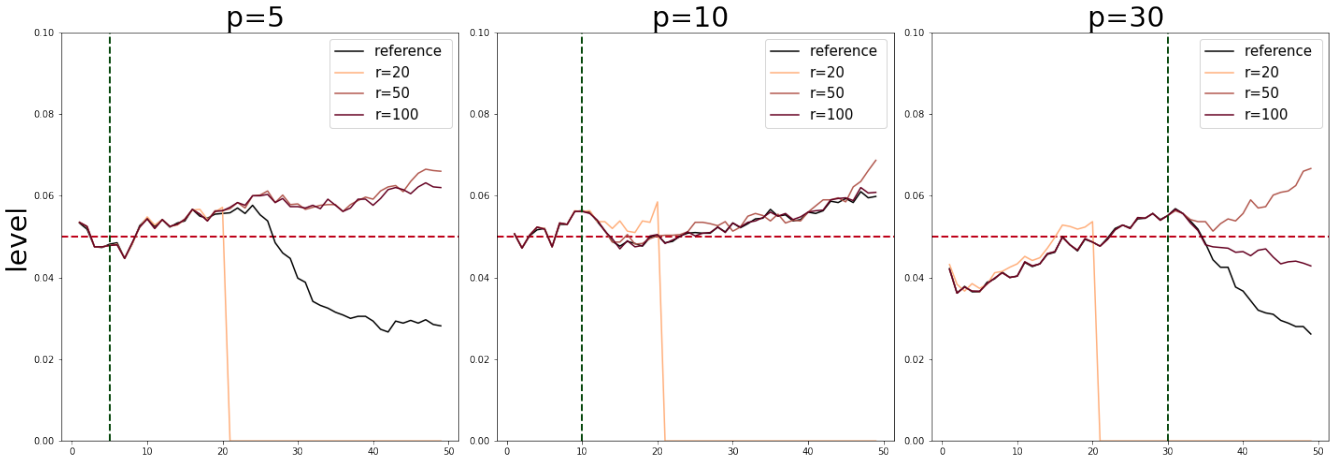

Figure 2.13: Asymptotic distribution-based empirical type-I error w.r.t. T for the truncated KFDA statistic and three Nyström approximations based on different number of anchors $r \in \{20, 50, 100\}$. Simulations are done for $n = 2000$ observations, $q = 200$ landmarks, $r = 50$ anchors and $C = 6000$ repetitions, $d = 50$ and $p \in \{5, 10, 30\}$.

landmarks according to a non-uniform distribution, in [35]. The distribution is based on the column norms. This technique takes roots in general low-rank approximation of matrices, hence it comes with bounds on the approximation error. When the observations are clusterized, it may be efficient to choose the landmarks as the centroids of a k-means

algorithm, as presented in [151] and improved in [103] with the k-means ++ algorithm. The leverage score based approaches have guaranties in terms of error approximation and statistical performances. These approaches take roots in the work of Mahoney and its collaborators [35, 50, 3]. They were at first only theoretical, too expensive to be implemented without strong hypotheses on K. The recursive algorithm proposed in [102] is low-cost and still has good performances according to the authors. It recursively approximate the sub-matrix by drawing approximatively half of the observations based on their ridge leverage score.

2.3 Conclusion

The comparison framework implemented in our package `ktest` and exploiting kernel testing shows promising results on several single-cell datasets and allows to detect the subpopulations that explain a difference between two conditions. The implementation of the Nyström method allows the analysis of large datasets. In the next Chapter, we propose a generalized test adapted to complex designs that allows to generalize from pair-wise comparisons to the global comparisons of several conditions.

Chapter 3

General Hypothesis Testing in the Feature Space

The KFDA framework may be applied to a wide variety of single-cell datasets in order to perform two-sample testing and discriminant-wise data exploration. However, it suffers from two main limitations. The first limitation is that it is restricted to two-sample comparisons. For more than two samples, one should perform pair-wise or one versus all analysis, and deal with multiple testing issues. This limitation has been overcome with a generalized KFDA k-sample testing framework [15]. The second limitation occurs when several explanatory variables may explain the data, and the others need to be "corrected" when testing for the influence of one particular variable. For instance, it may be of interest to compare more than two conditions, while taking into account other effects such as cell-types, donor or batch-effect. This issue may not always be dealt with a k-sample approach, as some explanatory variables may be non-categorical. In non-kernel contexts, these two limitations are usually tackled by linear models. Linear models naturally connect with the Fisher Discriminant Analysis because they both rely on the same quantities for simple designs.

In this chapter, we propose a statistical framework to generalize the KFDA approach to any general design. This framework is inspired from the multivariate linear model, we call it the kernel linear model. We basically fit a linear model on the embeddings in the feature space with respect to the explanatory factors. By doing so, the estimated model parameters are functions of the feature space that represent the contribution of each explanatory variable to the global kernel mean embedding of the observed sample. Then the linear model allows to test for any linear combination of the model parameters that are functions of the RKHS. Moreover, by stressing out the relation between the KFDA framework and the kernel linear model, we generalize the concept of discriminant directions to allow any hypothesis-based data exploration.

The first section of this chapter introduces or recalls some notations (Section 3.1). The second section introduces the kernel linear model as a generalization of the multivariate linear model (Section 3.2). Then, hypothesis testing based on the kernel linear model is introduced and we discuss some interpretations of the model (Section 3.3). We then propose a data exploration tool inspired by the kernel Fisher Discriminant direction and some diagnostic tools inspired from diagnostic tools on the multivariate linear model (Section 3.4). The kernel tricks to compute every introduced quantity are gathered in the next Section (Section 3.5), followed by some discussions (Section 3.6) and the proofs of our theorems (Section 3.7).

3.1 Notations

Norms in $\mathbb{R}^n$

Let $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$. The ℓ_1-norm $\|a\|_1$ and the euclidian norm $\|a\|$ of a are such that:

$$\|a\|_1 = \sum_{i=1}^{n} |a_i|,$$
$$\|a\| = \left(\sum_{i=1}^{n} a_i^2 \right)^{\frac{1}{2}}.$$

Hilbert-Schmidt Operators

To introduce the mathematical concepts we use in this chapter, we follow Section 2.1 *The Hilbert space of Hilbert-Schmidt operators* from [22]. Let $\mathcal{H}$ be a separable Hilbert space provided with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and the associated norm $\|\cdot\|_{\mathcal{H}}$. Let $(e_s)_{s \geq 1}$ be an orthonormal basis of $\mathcal{H}$. A Hilbert-Schmidt operator L from $\mathcal{H}$ to $\mathcal{H}$ is such that:

$$\sum_{s \geq 1} \|Le_s\|_{\mathcal{H}}^2 < +\infty.$$

The sum is independant from the orthonormal basis. The Hilbert-Schmidt operators form a separable Hilbert space $\mathrm{HS}(\mathcal{H})$ endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathrm{HS}(\mathcal{H})}$ such that for $L, N \in \mathrm{HS}(\mathcal{H})$, we have:

$$\langle L, N \rangle_{\mathrm{HS}(\mathcal{H})} = \sum_{s \geq 1} \langle Le_s, Ne_s \rangle_{\mathcal{H}}.$$

Hilbert-Schmidt operators are compact and have countable spectra. Every non-zero eigenvalue of a Hilbert-Schmidt operator is associated to an eigenspace of finite dimension. An operator L is called trace-class if $\sum_{s\geq 1}\langle e_s, Le_s\rangle_{\mathcal{H}}$ is a convergent series. The sum is independant from the orthonormal basis and is called trace of L. For a self-adjoint operator, we have $\mathrm{tr}(L) = \sum_{s\geq 1}\lambda_s(L)$. Note that if $L, N \in \mathrm{HS}(\mathcal{H})$, then the trace of their composition is equal to their inner product:

$$\mathrm{tr}(LN) = \langle L, N\rangle_{\mathrm{HS}(\mathcal{H})}\,. \tag{3.1}$$

We use the notation *tr* to refer to traces of trace-class operators from $\mathcal{H}$ to $\mathcal{H}$ and *trace* to refer to traces of matrices. If $f, g \in \mathcal{H}$ are non-zero, the tensor product $f \otimes g$ defines a rank one Hilbert-Schmidt operator. For $h \in \mathcal{H}$, we have:

$$\Big(f \otimes g\Big)h = \langle g, h\rangle_{\mathcal{H}}\, f.$$

We have the following identities:

$$\|f \otimes g\|_{\mathrm{HS}(\mathcal{H})} = \|f\|_{\mathcal{H}}\,\|g\|_{\mathcal{H}}\,, \tag{3.2}$$

$$\langle L, f \otimes g\rangle_{\mathrm{HS}(\mathcal{H})} = \langle f, Lg\rangle_{\mathcal{H}}\,. \tag{3.3}$$

Elements of $\mathcal{H}^n$

Let $f = (f_1, \ldots, f_n) \in \mathcal{H}^n$, we define a norm on $\mathcal{H}^n$ by:

$$\|f\|_{\mathcal{H}^n} = \left(\sum_{i=1}^{n}\|f_i\|_{\mathcal{H}}^2\right)^{\frac{1}{2}}.$$

If $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and $f = (f_1, \ldots, f_n) \in \mathcal{H}^n$, then $f \odot a \in \mathcal{H}$ stands for the linear combination of the elements of f with weights a:

$$f \odot a = \sum_{i=1}^{n} a_i f_i \in \mathcal{H}.$$

Then we can use the triangular inequality to immediatly get:

$$\|f \odot a\|_{\mathcal{H}} \leq \|a\|_1 \max_{i \in \{1,\dots,n\}} \|f_i\|_{\mathcal{H}} \tag{3.4}$$

Let $\mathbf{B} = (b_1, \dots, b_p)' \in \mathcal{M}_{p,n}(\mathbb{R})$ a $p \times n$ matrix and $b_j \in \mathbb{R}^n$ is the j^{th} row of $\mathbf{B}$. We classically consider $\mathbf{B}$ as the matrix of the linear operator from $\mathbb{R}^n$ to $\mathbb{R}^p$. Here, $\mathbf{B}$ is also used to represent the operator from $\mathcal{H}^n$ to $\mathcal{H}^p$, such that:

$$\mathbf{B}f = \begin{pmatrix} f \odot b_1 \\ \vdots \\ f \odot b_p \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n b_{1,i} f_i \\ \vdots \\ \sum_{i=1}^n b_{p,i} f_i \end{pmatrix} \in \mathcal{H}^p.$$

By an abuse of notation, if $\mathbf{B}$ is an $n \times n$ matrix, we denote by $f'\mathbf{B}f$ the Hilbert-Schmidt operator defined such that:

$$f'\mathbf{B}f = \sum_{i=1}^n f_i \otimes (f \odot b_i) \in \mathrm{HS}(\mathcal{H}). \tag{3.5}$$

The operator $f'\mathbf{B}f$ is a Hilbert-Schmidt operator as a linear combination of rank one operators $f_i \otimes f_j$:

$$f'\mathbf{B}f = \sum_{i,j=1}^n b_{i,j} f_i \otimes f_j.$$

3.2 The kernel Linear Model: a Linear Model in the Feature Space

In this section, we introduce a kernel generalization of the multivariate linear model and propose to perform kernel-based hypothesis testing on this model. Before introducing our model, we recall some generalities on the classical multivariate linear model.

Let $\mathbf{Y} = (Y_1, \dots, Y_n)$ a set of n observations from a measurable space $(\mathcal{Y}, \mathfrak{Y})$. Each Y_i is associated to a set of p explanatory variables $x_i = (x_{i,1}, \dots, x_{i,p})' \in \mathbb{R}^p$, $i \in \{1, \dots, n\}$. We define the design matrix as $\mathbf{X} = (x_1, \dots, x_n)' \in \mathcal{M}_{n,p}(\mathbb{R})$. We assume that $\mathbf{X}$ is full rank, i.e. $\dim(\mathrm{Im}(\mathbf{X})) = p$. We consider the case where the explanatory variables are deterministic and the observations are random and follow a probability distribution $\mathbb{P}$.

3.2.1 The Multivariate Linear Model

We adopt the definitions proposed in [104] for the classical multivariate linear model. Assume that $\mathcal{Y} = \mathbb{R}^m$ with $m \geq 1$. We denote by $Y^1, \ldots, Y^m$ in $\mathbb{R}^n$ the response variables such that $\mathbf{Y} = (Y_1, \ldots, Y_n)' = (Y^1, \ldots, Y^m) \in \mathcal{M}_{n,m}(\mathbb{R})$.

Definition of the Multivariate Linear Model

The multivariate linear model is a general approach to quantify the influence of the explanatory variables on the response variables. For $i \in \{1, \ldots, n\}$, the multivariate linear model is defined by:

$$Y_i = x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + \varepsilon_i \in \mathbb{R}^m,$$

where $\beta_1, \ldots, \beta_p \in \mathbb{R}^m$ are the model parameters and ε_i is the error of the model. The errors $\varepsilon_1, \ldots, \varepsilon_n \in \mathbb{R}^m$ are supposed to be n i.i.d. random variables with null expectations $\mathbb{E}\varepsilon_1 = 0_{\mathbb{R}^m}$ and covariance matrix $\mathrm{Cov}(\varepsilon_1) = \mathbf{\Sigma}_{\boldsymbol{\varepsilon}} \in \mathcal{M}_m(\mathbb{R})$. The model parameters $\beta_1, \ldots, \beta_p$ and the error covariance matrix $\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}$ are unknown and need to be inferred. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' \in \mathcal{M}_{p,m}(\mathbb{R})$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)' \in \mathcal{M}_{n,m}(\mathbb{R})$, then the model can be expressed in matrix form as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3.6}$$

An usual convention is to set the first column $\mathbf{X}^1$ of $\mathbf{X}$ as $\mathbb{1}_{\mathbb{R}^n}$, the vector full of ones. This explanatory variable is then called the trivial predictor and it is associated to the model parameter β_0. The multivariate linear model is a general framework that englobes different models. In models with a trivial predictor, if $m = 1$ and $p = 2$, the model belongs to the simple linear models. If $m = 1$ and $p \geq 3$, the model belongs to the multiple linear models. When $m \geq 2$ and $p \geq 2$, the model belongs to the multivariate linear models. If all the explanatory variables are categorical variables, the model is an ANalysis Of Variance (ANOVA) model when $m = 1$ and a Multivariate ANalysis Of Variance (MANOVA) otherwise. The models are summarized in Table 3.2.1.

Inference of the Model Parameters of the Multivariate Linear Model

We infer the model parameters $\beta_1, \ldots, \beta_p$ by fitting m independent multiple linear models with a least square estimator. Let $\beta^1, \ldots, \beta^m$ in $\mathbb{R}^p$ and $\varepsilon^1, \ldots, \varepsilon^m$ in $\mathbb{R}^n$ such that $\boldsymbol{\beta} = (\beta^1, \ldots, \beta^m)$ and $\boldsymbol{\varepsilon} = (\varepsilon^1, \ldots, \varepsilon^m)$. Let $s \in \{1, \ldots, m\}$, the s^{th} multiple linear model is such that:

$$Y^s = \mathbf{X}\beta^s + \varepsilon^s.$$

The least square estimator $\widehat{\beta}^s$ is the vector of $\mathbb{R}^p$ that minimizes the Mean Square Error:

$$\widehat{\beta}^s = \underset{\beta^s}{\operatorname{argmin}} \left\| Y^s - \mathbf{X}\beta^s \right\|_2^2.$$

If $\mathbf{X}$ is assumed to be full rank, the gradient of the MSE is equal to zero when:

$$\widehat{\beta}^s = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y^s \in \mathbb{R}^p.$$

By aggregating the m estimated parameters, we obtain $\widehat{\boldsymbol{\beta}}$ the least square estimator of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \in \mathcal{M}_{p,m}(\mathbb{R}).$$

Let $j \in \{1, \ldots, p\}$ and let $w_j = (w_{j,1}, \ldots, w_{j,n})' \in \mathbb{R}^n$ be the j^{th} column of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$. We have $\widehat{\beta}_j = \mathbf{Y}'w_j$. The predicted responses are such that:

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} \in \mathcal{M}_{n,m}(\mathbb{R}).$$

We define $\boldsymbol{\Pi} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' : \mathbb{R}^n \to \mathbb{R}^n$, the matrix of the orthogonal projection on $Im(\mathbf{X}) \subset \mathbb{R}^n$. For $i \in \{1, \ldots, n\}$, $\pi_i = (\pi_{i,1}, \ldots, \pi_{i,n})'$ is the i^{th} column of $\boldsymbol{\Pi}$. We have $\widehat{\mathbf{Y}} = \boldsymbol{\Pi}\mathbf{Y}$ and $\widehat{Y}_i = \mathbf{Y}'\pi_i \in \mathbb{R}^m$. The residual matrix is such that:

$$\widehat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}} \in \mathcal{M}_{n,m}(\mathbb{R}).$$

Then we determine the empirical covariance matrix:

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}} = \frac{1}{n}\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}} \in \mathcal{M}_m(\mathbb{R}).$$

	any explanatory variables		categorical explanatory variables only		
	$p = 2$	$p > 2$	$p = 2$	$p = 3$	$p \geq 2$
$m = 1$	simple linear model	multiple linear model	One-way ANOVA	Two-way ANOVA	ANOVA
$m \geq 2$	multivariate linear model		One-way MANOVA	Two-way MANOVA	MANOVA

Table 3.1: Summary of the different linear models.

3.2.2 The Kernel Linear Model

Now we propose a kernel generalization of the multivariate linear model for the relation between the embeddings $\phi(Y_i)$ and the corresponding vector of explanatory variables x_i, in order to perform hypothesis testing in this model. We call this model the kernel linear model.

Let $k(\cdot, \cdot) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a positive definite kernel associated to the Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$. For $i \in \{1, \ldots, n\}$, $\phi(Y_i) = k(Y_i, \cdot)$ is the embedding of Y_i in $\mathcal{H}$. The vector of $\mathcal{H}^n$ containing the embeddings of the observations is denoted $\Phi(\mathbf{Y}) = (\phi(Y_1), \ldots, \phi(Y_n))$.

Definition of the Kernel Linear Model

Let $\Theta = (\theta_1, \ldots, \theta_p)$ in $\mathcal{H}^p$ the model coefficients and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ in $\mathcal{H}^n$ a set of n i.i.d. random elements of $\mathcal{H}$ called errors. We assume that the errors have null expectations $\mathbb{E}(\varepsilon_1) = 0_{\mathcal{H}}$ and Hilbert-Schmidt covariance operator $\mathrm{Cov}(\varepsilon_1) = \Sigma_\varepsilon \in \mathrm{HS}(\mathcal{H})$. For $i \in \{1, \ldots, n\}$, the kernel linear model on $\phi(Y_i)$ is such that:

$$\phi(Y_i) = \Theta \odot x_i + \varepsilon_i.$$

The kernel linear model can also be written in matrix form as:

$$\Phi(\mathbf{Y}) = \mathbf{X}\Theta + \varepsilon. \tag{3.7}$$

This model can be interpreted as a multivariate linear model of infinite dimension. The main difference with the multivariate linear model is that the explanatory variables are used to explain the embeddings of the observations instead of the observations. Moreover, the errors and the model parameters are elements of $\mathcal{H}$. With this model, the kernel mean

embedding of the probability distribution underlying the data is expressed with respect to the explanatory variables, as we have $\mathbb{E}\Big(\Phi(\mathbf{Y})\Big) = \mathbf{X}\Theta$.

The feature space is a function space. Thus the kernel linear model is actually a functional linear model. Precisely, it belongs to the class of functional response models, that refers to the regression of functional responses with scalar explanatory variables and functional model parameters [65]. Our approach is nevertheless original as we restrict to functions of the feature space, and the observed functions are only convenient representations of our vectorial observations due to the RKHS embedding. The motivation to define this kernel linear model is that it allows to generalize the kernel two-sample tests in more general designs. We aim at performing hypothesis testing on the model parameters associated to the explanatory variables. Most situations we have in mind are models with categorical explanatory variables only. However, we develop the theoretical aspects of our asymptotic testing framework for general designs, allowing to test for the influence of non-categorical explanatory variables.

Inference of the Model Parameters of the Kernel Linear Model

As in the multivariate linear model, the estimation of the model parameters $\theta_1, \ldots, \theta_p$ is done through coordinate-wise least square estimation. To do so, we consider $(e_s)_{s\geq 1}$ an orthonormal basis of $\mathcal{H}$. Let $s \geq 1$, for $h \in \mathcal{H}$, we define $h^s = \langle h, e_s \rangle_{\mathcal{H}}$ such that $h = \sum_{s\geq 1} h^s e_s$. We obtain a multiple linear model by projecting the kernel linear model (3.7) onto span($\mathrm{e_s}$):

$$\Phi(\mathbf{Y})^s = \mathbf{X}\Theta^s + \varepsilon^s,$$

where $\Phi(\mathbf{Y})^s = (\phi(Y_1)^s, \ldots, \phi(Y_n)^s)' \in \mathbb{R}^n$, $\Theta^s = (\theta_1^s, \ldots, \theta_p^s)' \in \mathbb{R}^p$ and $\varepsilon^s = (\varepsilon_1^s, \ldots, \varepsilon_n^s)' \in \mathbb{R}^n$. As $\mathbf{X}$ is assumed to be full rank, the least square estimator of the model parameters of this multiple linear model is such that:

$$\widehat{\Theta}^s = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Phi(\mathbf{Y})^s.$$

Note that for $j \in \{1, \ldots, p\}$, $\widehat{\theta}_j = \sum_{s \geq 1} \widehat{\theta}_j^s e_s$. Then the least square estimator of $\Theta \in \mathcal{H}^p$ is given by:

$$\widehat{\Theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Phi(\mathbf{Y}). \tag{3.8}$$

If $w_j = (w_{j,1}, \ldots, w_{j,n})' \in \mathbb{R}^n$ is the j^{th} column of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \in \mathcal{M}_{n,p}(\mathbb{R})$, then $\widehat{\theta}_j = \Phi(\mathbf{Y}) \odot w_j$. We have $\widehat{\Theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_p)$ in $\mathcal{H}^p$. We can use $\widehat{\Theta}$ to predict $\Phi(\mathbf{Y})$:

$$\widehat{\Phi(\mathbf{Y})} = \mathbf{X}\widehat{\Theta}.$$

Now we consider $\mathbf{\Pi} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ the matrix of the orthogonal projection on the image space of $\mathbf{X}$ denoted $Im(\mathbf{X})$, also seen as an operator from $\mathcal{H}^n$ to $\mathcal{H}^n$. We have $\widehat{\Phi(\mathbf{Y})} = \mathbf{\Pi}\Phi(\mathbf{Y})$ and $\widehat{\phi(Y_i)} = \Phi(\mathbf{Y}) \odot \pi_i$. The design $\mathbf{X}$ is full rank, meaning that $rank(\mathbf{X}) = p$, let $(v_1, \ldots, v_p)$ be the p eigenvectors of $\mathbf{\Pi}$ such that:

$$\mathbf{\Pi} = \sum_{s=1}^{p} v_s v_s'. \tag{3.9}$$

The vector of residuals is defined by:

$$\widehat{\varepsilon} = \Phi(\mathbf{Y}) - \widehat{\Phi(\mathbf{Y})}.$$

We denote by $\mathbf{I}_n$ the identity matrix of $\mathbb{R}^n$, let $\mathbf{\Pi}^{\perp} = (\mathbf{I}_n - \mathbf{\Pi})$ be the matrix of the orthogonal projection on $\mathrm{Im}(\mathbf{X})^{\perp} \subset \mathbb{R}^n$. For $i \in \{1, \ldots, n\}$, we denote by $\pi_i^{\perp} = (\pi_{i,1}^{\perp}, \ldots, \pi_{i,n}^{\perp})' \in \mathbb{R}^n$ its i^{th} column and we have that $\widehat{\varepsilon}_i = \Phi(\mathbf{Y}) \odot \pi_i^{\perp}$, summarized in:

$$\widehat{\varepsilon} = \mathbf{\Pi}^{\perp}\Phi(\mathbf{Y}).$$

The residual covariance operator is defined such that:

$$\widehat{\Sigma}_{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\varepsilon}_i \otimes \widehat{\varepsilon}_i.$$

Note that in the case of a kernel function $k(\cdot, \cdot)$ associated to an infinite dimensional RKHS, the objects of $\mathcal{H}$ and $\mathrm{HS}(\mathcal{H})$ discussed here are not tractable. These are only defined theoretically and their expressions will be used further in this chapter to compute quantities of interest thanks to kernel tricks.

3.3 Testing Hypotheses

3.3.1 Testing Hypotheses on the Multivariate Linear Model

Let's go back to the multivariate linear model of Equation (3.6):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Our objective is to perform hypothesis testing on the model parameters $\beta_1, \ldots, \beta_p$. A general way to formulate the hypothesis test is to consider a surjective matrix $\mathbf{L} \in \mathcal{M}_{\mathcal{L},p}(\mathbb{R})$ such that the null and alternative hypotheses are:

$$H_0 : \mathbf{L}\boldsymbol{\beta} = 0_{\mathbb{R}^{\mathcal{L}}}, \quad \text{and} \quad H_1 : \mathbf{L}\boldsymbol{\beta} \neq 0_{\mathbb{R}^{\mathcal{L}}}.$$

It is usual to start by testing a model, i.e. choosing $\mathbf{L} \in \mathbb{R}^p$ such that the hypotheses are:

$$H_0 : \mathbb{E}(\mathbf{Y}) = \mu \quad \text{and} \quad H_1 : \mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}.$$

Then, if a the null hypothesis is rejected, we can test a contrast on the model parameters on the centered data by choosing a vector $\mathbf{L} \in \mathbb{R}^p$ such that $\mathbf{L}\mathbb{1}_p = 0$.

Definition of Several Test Statistics

There are several common tests for this kind of hypothesis, and most of them are defined with respect to the two matrices of $\mathcal{M}_m(\mathbb{R})$ (see [104]):

$$\begin{aligned}\widehat{\mathbf{H}_{\mathbf{L}}} =& (\mathbf{L}\widehat{\boldsymbol{\beta}})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}}) \\ \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}} =& \frac{1}{n}\sum_{i=1}^{n} \widehat{\varepsilon}_i \widehat{\varepsilon}_i'.\end{aligned}$$

The matrix $\widehat{\mathbf{H}_{\mathbf{L}}}$ can be interpreted as the norm of $\mathbf{L}\widehat{\boldsymbol{\beta}}$ normalized with respect to the design, that is the variance it would have if the covariance matrix of the error was identity. The matrix $n\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon}} \in \mathcal{M}_p(\mathbb{R})$ is called the residual sum of squares and it quantifies the empirical variance of the errors. These are different ways to take the observed residual covariance into account in the computation of the norm of the normalized $\mathbf{L}\widehat{\boldsymbol{\beta}}$. This leads to four popular hypothesis tests:

- The Roy's test statistic: $\text{Roy}(\mathbf{L}) = \max_{v \in \mathbb{R}^m, \|v\|_2 = 1} \left\| n^{-1} \widehat{\mathbf{\Sigma}}_{\varepsilon}^{-1} \widehat{\mathbf{H}_{\mathbf{L}}} v \right\|_2$.

- The Wilk's Lambda statistic: $\Lambda(\mathbf{L}) = |(\widehat{\mathbf{H}_{\mathbf{L}}} + n\widehat{\mathbf{\Sigma}}_{\varepsilon})^{-1} \widehat{\mathbf{H}_{\mathbf{L}}}|$.

- The Pillai's trace statistic: $\text{Pillai}(\mathbf{L}) = \text{trace}((\widehat{\mathbf{H}_{\mathbf{L}}} + n\widehat{\mathbf{\Sigma}}_{\varepsilon})^{-1} \widehat{\mathbf{H}_{\mathbf{L}}})$.

- The Hotelling-Lawley trace statistic: $\mathcal{F}(\mathbf{L}) = \text{trace}(n^{-1} \widehat{\mathbf{\Sigma}}_{\varepsilon}^{-1} \widehat{\mathbf{H}_{\mathbf{L}}})$.

Asymptotic Distribution of the Hotelling-Lawley Trace Statistic

When the errors are assumed to follow a Gaussian distribution, these statistics have different χ^2 distributions. These results are considered as robust since they hold asymptotically when the Gaussian assumption is relaxed. Here is a set of relaxed hypotheses used to prove the asymptotic distribution of the Hotelling-Lawley trace statistic, see for instance Theorems 12.7 and 12.8 in [104]:

$\mathtt{B}_1$ The i.i.d. errors of the model $\varepsilon_1, \ldots, \varepsilon_n$ have a fourth order moment: $\mathbb{E} \|\varepsilon_1\|^4 < +\infty$.

$\mathtt{B}_2$ We have $\max_i \pi_{i,i} \underset{n \to \infty}{\longrightarrow} 0$, where for $i \in \{1, \ldots, n\}$, $\pi_{i,i}$ is the i^{th} diagonal element of $\mathbf{\Pi} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

$\mathtt{B}_3$ The sampling gives a convergent design: $n(\mathbf{X}'\mathbf{X})^{-1} \underset{n \to \infty}{\to} W \in \mathbb{R}^{p \times p}$.

The asymptotic distribution of the Hotelling-Lawley trace statistic is given by the following theorem.

Theorem 10 (Theorem 12.8 from [104])**.** *Under Assumptions* $\mathtt{B}_1$*,* $\mathtt{B}_2$ *and* $\mathtt{B}_3$*, if* H_0 *is true, then the Hotelling-Lawley trace statistic asymptotically follows a* χ^2 *distribution with* $\mathcal{L} \times m$ *degrees of freedom:*

$$n\mathcal{F}(\mathbf{L}) \underset{n \to \infty}{\overset{\mathcal{D}}{\longrightarrow}} \chi^2_{\mathcal{L}m}$$

We choose to focus on the Hotelling-Lawley trace statistic for its connections with the Kernel Fisher Discriminant Analysis developed in Chapter 2, but similar results exist for the other test statistics in the multivariate setting and could be generalized to our kernelized linear model, see [104] and references cited therein.

3.3.2 Testing Hypotheses on the Kernelized Linear Model

The main contribution of this chapter is to generalize the Hotelling-Lawley trace statistic in the feature space. To do so, we define hypotheses on the model parameters of the kernel linear model and use the definition of the trace of Hilbert-Schmidt operators to define a kernel Hotelling-Lawley trace statistic. Then we determine the asymptotic distribution of a truncated regularization of this statistic.

Definition of the Truncated Kernel Hotelling-Lawley Trace Statistic

Let $\mathbf{L} = (l_{\ell,j})_{\ell\in\{1,\dots,\mathcal{L}\},j\in\{1,\dots,p\}} \in \mathbb{R}^{\mathcal{L}\times p}$ a surjective matrix such that each line encodes a linear combination of $(\theta_1,\dots,\theta_p)$ to be tested. We can formulate the null and alternative hypotheses as:

$$H_0 : \mathbf{L}\Theta = 0 \text{ vs. } H_1 : \mathbf{L}\Theta \neq 0.$$

We can generalize the Hotelling-Lawley trace statistic with a kernel Hotelling-Lawley trace statistic:

$$\mathcal{F} = \mathrm{tr}(n^{-1}\widehat{\Sigma}_\varepsilon^{-1}\widehat{H}_{\mathbf{L}}),$$

where $\widehat{H}_{\mathbf{L}} = (\mathbf{L}\widehat{\Theta})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\widehat{\Theta}) \in \mathrm{HS}(\mathcal{H})$ is referred as the test operator in the following. However, the residual covariance operator $\widehat{\Sigma}_\varepsilon$ may be singular and its inverse may not exist. To deal with this issue, we propose a truncated kernel Hotelling-Lawley trace statistic, inspired from the truncated Kernel Fisher Discriminant Analysis (KFDA) test statistic proposed by Harchaoui and coauthors for two-samples testing [62]. Let $\widehat{\Sigma}_{\varepsilon,T}$ be the spectral truncation of $\widehat{\Sigma}_\varepsilon$, defined by:

$$\widehat{\Sigma}_{\varepsilon,T} = \sum_{t=1}^{T} \widehat{\lambda}_t\Big(\widehat{f}_t \otimes \widehat{f}_t\Big), \tag{3.10}$$

where $(\widehat{\lambda}_t)_{t\geq 1}$ are the eigenvalues of $\widehat{\Sigma}_\varepsilon$ in non-increasing order and $(\widehat{f}_t)_{t\geq 1}$ are the associated eigenfunctions in $\mathcal{H}$. The pseudo-inverse of $\widehat{\Sigma}_{\varepsilon,T}$ is defined such that:

$$\widehat{\Sigma}_{\varepsilon,T}^{-1} = \sum_{\substack{t=1\\ \widehat{\lambda}_t>0}}^{T} \frac{1}{\widehat{\lambda}_t}\Big(\widehat{f}_t \otimes \widehat{f}_t\Big).$$

Then the truncated kernel Hotelling-Lawley trace statistic is such that:

$$\mathcal{F}_T = \mathrm{tr}(n^{-1}\widehat{\Sigma}_{\varepsilon,T}^{-1}\widehat{H}_{\mathbf{L}}),$$

There also exists a ridge regularization of the KFDA test statistic [64], thus we could also study a ridge regularized kernel Hotelling-Lawley trace statistic. Here we focus on the spectral truncation regularization because it has the main advantage of having a tractable χ^2 asymptotic distribution and the geometric interpretation of the eigenfunctions of $\widehat{\Sigma}_{\varepsilon,T}^{-1}\widehat{H}_{\mathbf{L}}$ are useful to develop data exploration tools.

Asymptotic Distribution of the Truncated Kernel Hotelling-Lawley Trace Statistic

As for the Hotelling-Lawley trace statistic, the truncated kernel Hotelling-Lawley trace statistic has a χ^2 asymptotic distribution. To establish this result, we make the following assumptions:

- $\mathtt{A}_1$ The kernel is bounded: $\|k(\cdot,\cdot)\|_\infty = M_k < +\infty$.
- $\mathtt{A}_2$ There exist $M_x \in \mathbb{R}$ such that for all $n \geq 1$, the design matrix $\mathbf{X} = \mathbf{X}(n) \in \mathcal{M}_{n,p}(\mathbb{R})$ is such that $\|\mathbf{X}\|_\infty \leq M_x$, where:

$$\|\mathbf{X}\|_\infty = \sup_{\substack{i\in\{1,\ldots,n\}\\ j\in\{1,\ldots,p\}}} |x_{i,j}|.$$

- $\mathtt{A}_3$ There exist $M_\pi \in \mathbb{R}$ such that for all $n \geq p$, the vectors $v_1, \ldots, v_p \in \mathbb{R}^n$ of an orthonormal basis of $\mathrm{Im}(\mathbf{X})$ are such that:

$$\sup_{\substack{s\in\{1,\ldots,p\}\\ i\in\{1,\ldots,n\}}} |v_{s,i}| \leq M_\pi/\sqrt{n}.$$

- $\mathtt{A}_4$ The sampling gives a convergent design: $n(\mathbf{X}'\mathbf{X})^{-1} \underset{n\to\infty}{\to} W \in \mathbb{R}^{p\times p}$.

The assumptions are discussed after the asymptotic result:

Theorem 11. *Under Assumptions* $\mathtt{A}_1$, $\mathtt{A}_2$, $\mathtt{A}_3$ *and* $\mathtt{A}_4$, *if* H_0 *is true and* $\lambda_T > 0$, *then:*

$$n\mathcal{F}_T \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2_{\mathcal{L}T}.$$

The advantage of this asymptotic distribution is that it is tractable and computation free. Thus performing hypothesis testing on the kernel linear model resumes to being able to compute the statistic. The computation of the statistic is detailed in the section dedicated to the kernel trick.

Note that the asymptotic distribution of the Hotelling-Lawley trace statistic derived from a multivariate linear model on m-dimensional observations follows a χ^2 distribution with $\mathcal{L} \times m$ degrees of freedom. Then the truncated kernel Hotelling-Lawley trace statistic that follows a χ^2 distribution with $\mathcal{L} \times T$ degrees of freedom seems to have the asymptotic distribution of the Hotelling-Lawley trace statistic from a T-dimensional problem. Indeed, using the truncated residual covariance in the statistic is equivalent to projecting the data in a random subspace of $\mathcal{H}$ of dimension T in order to perform the analysis. This property is one of the key points on which the proof is based.

Sketch of the proof. We define an alternative Hotelling-Lawley trace with the "true" error covariance operator Σ_ε instead of $\widehat{\Sigma}_\varepsilon$, such that $\tilde{\mathcal{F}}_T = \mathrm{tr}(\Sigma_{\varepsilon,T}^{-1}\widehat{H}_{\mathbf{L}})$, where $\Sigma_{\varepsilon,T}^{-1} = \sum_{t=1}^{T} \lambda_t^{-1}\Big(f_t \otimes f_t\Big)$ is well defined by assumption that $\lambda_T > 0$. We have:

$$
\begin{aligned}
|\mathcal{F}_T - \tilde{\mathcal{F}}_T| &= \left|\mathrm{tr}\left(\frac{1}{n}\Big(\widehat{\Sigma}_{\varepsilon,T}^{-1} - \Sigma_{\varepsilon,T}^{-1}\Big)\widehat{H}_{\mathbf{L}}\right)\right| \\
&= \left|\left\langle \Big(\widehat{\Sigma}_{\varepsilon,T}^{-1} - \Sigma_{\varepsilon,T}^{-1}\Big), \frac{1}{n}\widehat{H}_{\mathbf{L}} \right\rangle_{\mathrm{HS}(\mathcal{H})}\right| \\
&\leq \left\|\widehat{\Sigma}_{\varepsilon,T}^{-1} - \Sigma_{\varepsilon,T}^{-1}\right\|_{\mathrm{HS}(\mathcal{H})} \left\|\frac{1}{n}\widehat{H}_{\mathbf{L}}\right\|_{\mathrm{HS}(\mathcal{H})},
\end{aligned}
$$

where we used the linearity of the trace, Equation (3.1) and the Cauchy-Schwarz inequality. Then, in Proposition 3, we adapt the proofs of [128] and [154] by applying a bounded differences theorem to $\left\|\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}$ that outputs an exponential bound under Assumptions $\mathtt{A_1}$, $\mathtt{A_2}$ and $\mathtt{A_3}$. Then we make use of results from operator perturbation theory developed in [154] to derive an exponential bound on $\left\|\widehat{\Sigma}_{\varepsilon,T}^{-1} - \Sigma_{\varepsilon,T}^{-1}\right\|_{\mathrm{HS}(\mathcal{H})}$ in Lemma 11. As Lemma 9 shows that $n^{-1}\widehat{H}_{\mathbf{L}}$ is bounded under Assumptions $\mathtt{A_1}$ and $\mathtt{A_2}$, we can conclude that $\mathcal{F}_T \xrightarrow{P} \tilde{\mathcal{F}}_T$. We now need to derive the asymptotic distribution of $\tilde{\mathcal{F}}_T$. The derivation of the asymptotic distribution of $\tilde{\mathcal{F}}_T$ under Assumptions $\mathtt{A_1}$, $\mathtt{A_2}$, $\mathtt{A_3}$ and $\mathtt{A_4}$ is detailed in Lemma 10, where we show that it reduces to a standard Lawley-Hotelling test in $\mathbb{R}^T$. We then invoke a result from the literature (see for instance Theorem 12.8 from

[104]) to obtain its asymptotic distribution:

$$n\tilde{\mathcal{F}}_T \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2_{\mathcal{L}T}.$$

□

About Assumption $\mathtt{A_1}$ The proof of Theorem 11 essentially relies on the convergence in probability of $\widehat{\Sigma}_\varepsilon$ towards Σ_ε. We prove this convergence through a bounded difference theorem, adapting the approach of [128] and [154] on the covariance operator appearing in the Kernel Principal Component Analysis (KPCA) to the residual covariance operator. Assumption $\mathtt{A_1}$ is a strong but standard hypothesis for this type of result. We need this assumption to show that the errors are bounded and that their fourth order moment is finite.

About Assumption $\mathtt{A_2}$ The covariance operator of the KPCA may be considered as a special case of the residual covariance operator. It is the residual covariance operator obtained for a kernel linear model with a design reduced to the trivial explanatory variable only $\mathbf{X} = \mathbb{1}_n$, we call this model the KPCA linear model. For this model, the errors can be expressed explicitely with respect to the embeddings, thus Assumption $\mathtt{A_1}$ is sufficient to obtain that the errors are bounded and that their fourth order moment is finite. However, in more general models, the errors also depend on the design $\mathbf{X}$ and on the model parameters $\theta_1, \ldots, \theta_p$. For our general models, Assumption $\mathtt{A_2}$ is used in addition to $\mathtt{A_1}$ to obtain a bound on the errors and to show that their fourth moment is finite, as detailed in Lemma 4. Note that the fourth moment of the errors are also assumed to be finite in the linear Hotelling-Lawley test (Assumption $\mathtt{B_1}$) as it is used in Theorem 10. Thus, we do not need to reformulate this assumption under Assumptions $\mathtt{A_1}$ and $\mathtt{A_2}$.

About Assumption $\mathtt{A_3}$ The orthogonal projector $\mathbf{\Pi}$ associated to the KPCA linear model is $\mathbf{\Pi} = n^{-1}\mathbf{J}_n$, where $\mathbf{J}_n \in \mathcal{M}_n(\mathbb{R})$ is the matrix full of ones. This projector is characterized by its only norm 1 eigenvector $v = n^{-\frac{1}{2}}\mathbb{1}$ which is such that $\forall i \in \{1, \ldots, n\}, v_i = n^{-\frac{1}{2}} \underset{n\to\infty}{\longrightarrow} 0$. This convergence is explicit and it is used by the authors of [22] to prove the convergence of the empirical kernel covariance operator toward its population counterpart in the context of the KPCA. For a kernel linear model more complex than the KPCA linear model, we propose Assumption $\mathtt{A_3}$ as a possible generalization of the convergence property of the associated orthogonal projector. This

property is trivially verified for simple models such as the design matrix associated to a m-samples problem. Indeed, let $\mathbf{X}$ be the design matrix of a m-sample problem with asymptotically balanced effectives $n_1, \ldots, n_m$ such that:

$$\mathbf{X} = (w_1, \ldots, w_m) = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 1 & \ldots & 0 \\ 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix} \in \mathcal{M}_{n,m}(\mathbb{R}).$$

Then for $s \in \{1, \ldots, m\}$, the vector $v_s = \|w_s\|^{-1} w_s = n_s^{-\frac{1}{2}} w_s$ is an eigenvector of the orthogonal projector associated to $\mathbf{X}$ and we have for $i \in \{1, \ldots, n\}$ that $v_{s,i} \underset{n\to\infty}{\longrightarrow} 0$. Assumption $\mathtt{A_3}$ is stronger than Assumption $\mathtt{B_2}$ used to derive the asymptotic distribution of the linear Hotelling-Lawley statistic. We show how to derive Assumption $\mathtt{B_2}$ from Assumption $\mathtt{A_3}$ in Lemma 5.

About Assumption $\mathtt{A_4}$ Assumption $\mathtt{A_4}$ is exactly the same than Assumption $\mathtt{B_3}$.

About simplifying Assumption $\mathtt{A_2}$ and $\mathtt{A_3}$ We propose the following assumption:

$\mathtt{A_{2'}}$ There exists $M_x \in \mathbb{R}$ such that for all $n \geq 1$, $\mathbf{X} = \mathbf{X}(n) \in \mathcal{M}_{n,p}(\mathbb{R})$ is such that $\|\mathbf{X}\|_\infty \leq \frac{M_x}{\sqrt{n}}$.

This assumption can not replace assumption $\mathtt{A_2}$ and $\mathtt{A_3}$ in general. However, it can replace them when the columns of $\mathbf{X}$ have norm 1. First, note that $\mathtt{A_2}$ is a direct consequence of $\mathtt{A_2'}$. Then, we show that an orthonormal basis $(v_1, \ldots, v_p)$ of $\mathrm{Im}(\mathbf{X})$ obtained through a Gram-Schmidt orthogonalization of the columns of $\mathbf{X}$ verifies $\mathtt{A_3}$. Consider that the columns of $\mathbf{X}$ are ordered so that the first p columns $(x_1, \ldots, x_p)$ form a free set of vectors. The

Gram-Scmidt orthonormal basis $(v_1, \dots, v_p)$ obtained from them is defined by:

$$\begin{aligned} v_1 =& x_1, \\ \text{for } s \in \{1, \dots, p-1\}, \quad v_{s+1} =& x_{s+1} + \sum_{j=1}^{s} \langle x_{s+1}, v_j \rangle \, v_j. \end{aligned}$$

Assumption $\mathtt{A_3}$ is verified for v_1 as for $i \in \{1, \dots, n\}$, we have $|v_{1,i}| = |x_{1,i}| \leq M_x/\sqrt{n}$. Then, if for $s \in \{1, \dots, p-1\}$ and $i \in \{1, \dots, n\}$, we have $|v_{s,i}| \leq 2^{s-1} M_x/\sqrt{n}$, we deduce that:

$$\begin{aligned} |v_{s+1,i}| =& \left| x_{s+1,i} + \sum_{j=1}^{s} \langle x_{s+1}, v_j \rangle \, v_{j,i} \right| \\ \leq& |x_{s+1,i}| + \sum_{j=1}^{s} \|v_s\| \, \|x_{s+1,j}\| \, |v_{j,i}| \\ \leq& \frac{M_x}{\sqrt{n}} \left(1 + \sum_{j=1}^{s} 2^{s-1} \right) \\ \leq& \frac{M_x}{\sqrt{n}} 2^s \end{aligned}$$

By induction, we conclude that for $s \in \{1, \dots, p\}$ and $i \in \{1, \dots, n\}$:

$$|v_{s,i}| \leq \frac{2^{s-1} M_x}{\sqrt{n}} \leq \frac{2^{p-1} M_x}{\sqrt{n}}$$

Then, we obtain Assumption $\mathtt{A_3}$:

$$\sup_{\substack{s \in \{1, \dots, p\} \\ i \in \{1, \dots, n\}}} |v_{s,i}| \leq \frac{2^{p-1} M_x}{\sqrt{n}}.$$

3.3.3 Interpretations of the Model

Application to the Two-Ways MANOVA

A typical application of our kernel linear model would be a model with two additive effects, called two-way MANOVA Here is an example.

Consider a single-cell experiment where we expose cells to Treatment 1, Treatment 2 and Control (3). Each cell comes from patient 1 and patient 2, Let $Y_{i,j,k}$ be the measure of

the k^{th} cell of patient j treated with treatment i. A possible kernel linear model associated to this experiment could be:

$$\phi(Y_{i,j,k}) = \theta_0 + \theta_{\text{treatment } i} + \theta_{\text{patient } j} + \varepsilon_{i,j,k},$$

where θ_0 is the trivial predictor, $\theta_{\text{treatment } 1}, \theta_{\text{treatment } 2}, \theta_{\text{treatment } 3} \in \mathcal{H}$ are associated to treatment 1, treatment 2 and control 3 respectively, and $\theta_{\text{patient } 1}, \theta_{\text{patient } 2}$ stand for the effects of patients 1 and 2 respectively. The model parameter $\Theta = (\theta_0, \theta_{\text{treatment } 1}, \theta_{\text{treatment } 2}, \theta_{\text{treatment } 3}, \theta_{\text{patient } 1}, \theta_{\text{patient } 2}) \in \mathcal{H}^6$. A vector of explanatory variables $x_{i,j,k} \in \mathbb{R}^6$ associated to an observation $Y_{i,j,k}$ from Patient 1 treated with Treatment 2 would be $(1, 1, 0, 0, 1, 0)$. For instance, in the example discussed above with three treatments and two patients, the vector of model parameters was $\Theta = (\theta_0, \theta_{\text{treatment } 1}, \theta_{\text{treatment } 2}, \theta_{\text{treatment } 3}, \theta_{\text{patient } 1}, \theta_{\text{patient } 2}) \in \mathcal{H}^6$. A matrix $\mathbf{L}$ that would test if there is a difference between the treatments would be such that:

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \end{pmatrix}.$$

This matrix encodes for the hypotheses:

$$H_0 : \theta_{\text{treatment } 1} = \theta_{\text{treatment } 2} = \theta_{\text{treatment } 3} \text{ vs. } H_1 : \exists i, i' \in \{1, 2, 3\}, \theta_{\text{treatment } i} \neq \theta_{\text{treatment } i'}.$$

Explicitly, the first line of $\mathbf{L}$ encodes for the discrimination between Treatment 1 and Control, and the second line encodes for the discrimination between Treatment 2 and Control.

Links with the KFDA Test Statistic for the Two-Sample Test

Consider the setting of Chapter 2. We consider the measures of n_1 observations $\mathbf{Y}_1 = (Y_{1,1}, \ldots, Y_{1,n_1})$ from condition 1 and n_2 observations $\mathbf{Y}_2 = (Y_{2,1}, \ldots, Y_{2,n_2})$ from condition 2, with $n_1 + n_2 = n$. We associate each observation $Y_{i,j}$ to a vector of explanatory variables

$x_{i,j} = (x_{i,j,1}, x_{i,j,2})' \in \mathbb{R}^2$ defined such that:

$$x_{i,j} = \begin{cases} \begin{pmatrix} 1 \\ 0 \end{pmatrix} & \text{if } i = 1 \\ \begin{pmatrix} 0 \\ 1 \end{pmatrix} & \text{otherwise.} \end{cases}$$

We have $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2) = (Y_{1,1}, \ldots, Y_{1,n_1}, Y_{2,1}, \ldots, Y_{2,n_2})$ in $\mathcal{Y}^n$ and $\mathbf{X} = (x_{1,1}, \ldots, x_{1,n_1}, x_{2,1}, \ldots, x_{2,n_2}) \in \mathcal{M}_{n,2}(\mathbb{R})$. More specifically, we have:

$$\mathbf{X} = \begin{pmatrix} \mathbb{1}_{n_1} & 0_{n_1} \\ 0_{n_2} & \mathbb{1}_{n_2} \end{pmatrix},$$

where $\mathbb{1}_n \in \mathbb{R}^n$ is the vector full of 1 and $0_n \in \mathbb{R}^n$ is the vector full of zeros. For $i \in \{1, 2\}$ and $j \in \{1, \ldots, n_i\}$, the linear model on the embedded observation $\phi(Y_{i,j})$ is such that:

$$\phi(Y_{i,j}) = x_{i,j,1}\theta_1 + x_{i,j,2}\theta_2 + \varepsilon_{i,j},$$

where $\varepsilon_{i,j}$ is the j^{th} error of condition i. The errors are assumed to be i.i.d. with $\mathbb{E}\varepsilon_{1,1} = 0$ and $\mathrm{Cov}(\varepsilon_{1,1}) = \Sigma_\varepsilon$. Let $\Theta = (\theta_1, \theta_2)$ in $\mathcal{H}^2$ and $\varepsilon = (\varepsilon_{1,1}, \ldots, \varepsilon_{1,n_1}, \varepsilon_{2,1}, \ldots, \varepsilon_{2,n_2})$ in $\mathcal{H}^n$, the model in matrix form is such that:

$$\Phi(\mathbf{Y}) = \mathbf{X}\Theta + \varepsilon.$$

Note that for $i \in \{1, 2\}$, the model gives for $j \in \{1, \ldots, n_i\}$, $\mathbb{E}\phi(Y_{1,j}) = \theta_i$. Thus θ_i is actually the kernel mean embedding of the distribution underlying the observations in $\mathbf{Y}_i$. The least square estimator of Θ is $\widehat{\Theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Phi(\mathbf{Y})$. It can be computed explicitely and we obtain $\widehat{\theta}_1 = n_1^{-1}\sum_{j=1}^{n_1}\phi(Y_{1,j})$ and $\widehat{\theta}_2 = n_2^{-1}\sum_{j=1}^{n_2}\phi(Y_{2,j})$. We recognize the empirical kernel mean embeddings associated to $\mathbf{Y}_1$ and $\mathbf{Y}_2$ respectively. The predicted observation are $\widehat{\phi(Y_{i,j})} = \widehat{\theta}_i$ and the residuals are $\widehat{\varepsilon}_{i,j} = \phi(Y_{i,j}) - \widehat{\theta}_i$. The residual covariance operator is then equal to:

$$\widehat{\Sigma}_\varepsilon = \frac{n_1}{n}\widehat{\Sigma}_1 + \frac{n_2}{n}\widehat{\Sigma}_2,$$

where $\widehat{\Sigma}_i = n_i^{-1}\sum_{j=1}^{n_i}(\phi(Y_{i,j}) - \widehat{\theta}_i)^{\otimes 2}$ for $i \in \{1, 2\}$. If we define $\mathbf{L} = (-1, 1) \in \mathcal{M}_{1,2}(\mathbb{R})$, then the hypotheses $H_0 : \mathbf{L}\Theta = 0$ and $H_1 : \mathbf{L}\Theta \neq 0$ resumes to hypotheses $H_0 : \theta_1 = \theta_2$

and $H_1 : \theta_1 \neq \theta_2$. Then the test operator $\widehat{H}_{\mathbf{L}}$ is such that:

$$\widehat{H}_{\mathbf{L}} = \frac{n_1 n_2}{n}(\widehat{\theta}_2 - \widehat{\theta}_1)^{\otimes 2}.$$

Note that we recognize the within-group covariance operator and the between group covariance operator defined in chapter 2 for the two sample case in $\widehat{\Sigma}_\varepsilon$ and $n^{-1}\widehat{H}_{\mathbf{L}}$ respectively. Finally, the spectral regularization of the Hotelling-Lawley test statistic associated to this hypothesis is such that:

$$\begin{aligned}
\mathcal{F}_T =& \operatorname{tr}(\frac{1}{n}\Sigma_{\varepsilon,T}^{-1}\widehat{H}_{\mathbf{L}}) \\
=& \frac{1}{n}\left\langle \Sigma_{\varepsilon,T}^{-1}, \widehat{H}_{\mathbf{L}} \right\rangle_{\mathrm{HS}(\mathcal{H})} \\
=& \frac{n_1 n_2}{n^2}\left\langle \Sigma_{\varepsilon,T}^{-1}, (\widehat{\theta}_2 - \widehat{\theta}_1)^{\otimes 2} \right\rangle_{\mathrm{HS}(\mathcal{H})}.
\end{aligned}$$

Then we apply Equation (3.3) and use the fact that $\Sigma_{\varepsilon,T}$ is self-adjoint:

$$\begin{aligned}
=& \frac{n_1 n_2}{n^2}\left\langle \Sigma_{\varepsilon,T}^{-1}(\widehat{\theta}_2 - \widehat{\theta}_1), (\widehat{\theta}_2 - \widehat{\theta}_1) \right\rangle_{\mathcal{H}} \\
=& \frac{n_1 n_2}{n^2}\left\langle \Sigma_{\varepsilon,T}^{-\frac{1}{2}}(\widehat{\theta}_2 - \widehat{\theta}_1), \Sigma_{\varepsilon,T}^{-\frac{1}{2}}(\widehat{\theta}_2 - \widehat{\theta}_1) \right\rangle_{\mathcal{H}} \\
=& \frac{n_1 n_2}{n^2}\left\| \Sigma_{\varepsilon,T}^{-\frac{1}{2}}(\widehat{\theta}_2 - \widehat{\theta}_1) \right\|_{\mathcal{H}}^2.
\end{aligned}$$

Here, we recognize the truncated KFDA two-sample test statistic presented in Chapter 2. In fact we have $\mathcal{F}_T = n^{-1}\widehat{\mathrm{D}_T}^2$. Thus, the truncated KFDA statistic is a special case of the truncated Hotelling-Lawley test statistic on our general kernel linear model. We can indeed consider hypothesis testing on the kernel linear model as a generalization of KFDA two-sample testing. It is then easy to apply the framework to problems with $k > 2$ samples or more general designs. Moreover, it also motivates the generalization of the discriminant directions used for hypothesis-based data exploration.

3.4 Data Exploration and Diagnostics

3.4.1 Projection on the Discriminant Directions

The Kernel Fisher Discriminant Analysis (KFDA) was introduced in [99]. It is a classifier based on the determination of the discriminant directions in the feature space. It is usually used to discriminate between $p \geq 2$ groups of observations. Here we highlight the link between the KFDA approach for $p \geq 2$ groups and the kernel Hotelling-trace statistic in order to motivate the generalization of discriminant directions to general designs.

We consider p groups. For $i \in \{1, \ldots, p\}$, let $Y_{i,1}, \ldots, Y_{i,n_i}$ the n_i observations of group i, with $n = \sum_{i=1}^{p} n_i$. Each observation belong to one group only, for $i \in \{1, \ldots, p\}$ and $j \in \{1, \ldots, n_i\}$, the vector of explanatory variables $x_{i,j} \in \mathbb{R}^p$ is the vector full of zeros except in position i where it is one and let $\mathbf{X}$ be the associated design matrix. Let $\Theta = (\theta_1, \ldots, \theta_p)$. Here the kernel linear model takes the form:

$$\phi(Y_{i,j}) = \theta_i + \varepsilon_{i,j},$$

where $(\varepsilon_{i,j})_{i \in \{1,\ldots,p\}, \{1,\ldots,n_i\}}$ are i.i.d. with null expectation and common covariance operator Σ_ε. The least square estimator of the model parameters is such that for $i \in \{1, \ldots, p\}$, $\widehat{\theta}_i = n_i^{-1} \sum_{j=1}^{n_i} \phi(Y_{i,j})$. We recognize the empirical kernel mean embedding of the distribution underlying group i. Let $\mu = n^{-1} \sum_{i=1}^{p} n_i \theta_i$ and $\widehat{\mu} = n^{-1} \sum_{i=1}^{p} n_i \widehat{\theta}_i$ be the global kernel mean embedding and its empirical estimator respectively. For the KFDA, the within group covariance operator Σ_W and the between group covariance operator Σ_B are defined such that:

$$\widehat{\Sigma}_W = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n_i} (\phi(Y_{i,j}) - \widehat{\theta}_i)^{\otimes 2},$$
$$\widehat{\Sigma}_B = \frac{1}{n} \sum_{i=1}^{p} n_i (\widehat{\theta}_i - \widehat{\mu})^{\otimes 2}.$$

The discriminant directions computed in the KFDA approach are the eigenfunctions of the operator $\widehat{\Sigma}_W^{-1} \widehat{\Sigma}_B$, that has a maximal rank of $p - 1$.

In Chapter 2, we discussed how Harchaoui and coauthors [62] highlighted the link between the KFDA for two groups and the Hotelling T^2 test statistic in the feature space.

Actually, for any number of groups p, the kernelized Hotelling-Lawley test statistic has a link with the KFDA. Indeed, the matrix

$$\mathbf{L} = \begin{pmatrix} 1-\frac{n_1}{n} & -\frac{n_2}{n} & \cdots & -\frac{n_p}{n} \\ -\frac{n_1}{n} & 1-\frac{n_2}{n} & \cdots & -\frac{n_p}{n} \\ \vdots & \ddots & \ddots & \vdots \\ -\frac{n_1}{n} & \cdots & 1-\frac{n_{p-1}}{n} & -\frac{n_p}{n} \end{pmatrix} \in \mathcal{M}_{p-1,p}(\mathbb{R}), \tag{3.11}$$

encodes for the hypotheses $H_0 : \theta_1 = \theta_2 = \cdots = \theta_p = \mu$ versus $H_1 : \exists j \in \{1, \ldots, p\}, \theta_j \neq \mu$. The test operator $\widehat{H}_{\mathbf{L}}$ obtained from the matrix $\mathbf{L}$ is colinear to the empirical between-group covariance operator $\widehat{\Sigma}_B$ and the residual covariance operator $\widehat{\Sigma}_\varepsilon$ obtained from the mean square estimation of the model parameters is equal to the within-group covariance operator $\widehat{\Sigma}_W$. Thus, the KFDA discriminant directions that can be used to discriminate the p groups are in fact the eigenfunctions of the Hilbert-Schmidt operator in the kernel Hotelling-Lawley trace statistic $\Sigma_\varepsilon^{-1}\widehat{H}_{\mathbf{L}}$.

We propose to generalize this approach by considering the eigenfunctions of the operator $\Sigma_\varepsilon^{-1}\widehat{H}_{\mathbf{L}}$ with respect to any kernel linear model and to any matrix $\mathbf{L}$ as directions of interest for the problem. These directions can be used to explain the response of the test or be used as exploratory tools in order to detect cell populations of interest.

3.4.2 Diagnostic Plots

A multivariate linear model or a kernel linear model both rely on assumptions on the model parameters and the errors. The response plot and the residual plot are two plots that allow to visually assess the veracity of the hypotheses, and to eventually detect outliers or model errors in the multivariate linear model [71, 104]. Here we extend these diagnostic plots to the kernel linear model.

Consider the m dimensional multivariate linear model of Equation (3.6). Let $s \in \{1, \ldots, m\}$, the s^{th} response plot represents the s^{th} coordinate of observed data with respect to the s^{th} predicted responses, i.e. the couples $((Y_1^s, \widehat{Y}_1^s), \ldots, (Y_n^s, \widehat{Y}_n^s))$. The s^{th} residual plot represents the s^{th} coordinate of the residuals with respect to the s^{th} coordinate of the predicted responses, i.e. the couples $((\widehat{\varepsilon}_1^s, \widehat{Y}_1^s), \ldots, (\widehat{\varepsilon}_n^s, \widehat{Y}_n^s))$. If the heteroscedasticity assumption is true, then the points in the response plot should be

uniformly spread around the identity line, and they should be uniformly spread around the zero horizontal axis in the residual plot. Some visual outlier may be detected in the response plot as observations placed far from the identity line compared to the others. We can detect a dependency between the predictions and the residuals in the residual plot, that would suggest that an effect has not been taken into account in the model. We can also detect heteroscedasticity if the vertical spread is not uniform along the horizontal axis in the residual plot. Heteroscedasticity may be overcome with data normalization.

Now we consider the kernel linear model of Equation (3.7). For $i \in \{1, \ldots, n\}$, the embeddings $\phi(Y_i)$, predictions $\widehat{\phi(Y_i)}$ and residuals $\widehat{\varepsilon}_i$, are elements of $\mathrm{Span}(\phi(Y_1), \ldots, \phi(Y_n)) \subset \mathcal{H}$ that is an n-dimensional subspace of $\mathcal{H}$. As n-dimensional objects, we would need n diagnostic plots to draw an exhaustive picture of the diagnostics. Instead, we propose to select a few directions of interest of $\mathcal{H}$ in order to compute informative but non-exhaustive diagnostic plots. Let $h \in \mathcal{H}$ be a direction of interest, we propose to represent the diagnostic plots with respect to $\langle \phi(Y_i), h \rangle_{\mathcal{H}}$, $\left\langle \widehat{\phi(Y_i)}, h \right\rangle_{\mathcal{H}}$ and $\langle \widehat{\varepsilon}_i, h \rangle_{\mathcal{H}}$ respectively. For T such that $\widehat{\lambda}_T > 0$, where $\widehat{f}_1, \ldots, \widehat{f}_T$ are the first T eigenfunctions of the residual covariance operator $\widehat{\Sigma}_\varepsilon$ associated to the eigenvalues $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_T$, we propose to use these directions as directions of interest to compute the diagnostic plots. We define the following matrices:

$$\begin{aligned}
\mathcal{D}(\Phi(\mathbf{Y}), \widehat{\Sigma}_\varepsilon, T) =& (\left\langle \phi(Y_i), \widehat{f}_t \right\rangle_{\mathcal{H}})_{i \in \{1,\ldots,n\}, t \in \{1,\ldots,T\}}, \\
\mathcal{D}(\widehat{\Phi(\mathbf{Y})}, \widehat{\Sigma}_\varepsilon, T) =& (\left\langle \widehat{\phi(Y_i)}, \widehat{f}_t \right\rangle_{\mathcal{H}})_{i \in \{1,\ldots,n\}, t \in \{1,\ldots,T\}}, \\
\mathcal{D}(\widehat{\varepsilon}, \widehat{\Sigma}_\varepsilon, T) =& (\left\langle \widehat{\varepsilon}_i, \widehat{f}_t \right\rangle_{\mathcal{H}})_{i \in \{1,\ldots,n\}, t \in \{1,\ldots,T\}},
\end{aligned}$$

to be explicitely computed in the kernel trick section. As the eigenfunctions $\widehat{f}_1, \ldots, \widehat{f}_T$ are ordered with respect to decreasing variability of the residuals, these are especially suitable for the diagnostic plots. In fact, it is more important that the model hypotheses are true on directions with high variability. Oppositely, it should not have a huge impact on the results if model assumptions are not verified on weakly variable directions, namely $\widehat{f}_{T+1}, \ldots, \widehat{f}_n$. It would also be possible to use the eigenfunctions of $\widehat{\Sigma}_{\varepsilon,T}^{-1} \widehat{H}_{\mathbf{L}}$ as directions of interest for the diagnostic plots. These directions also depend on the choice of T and they are directly associated to the testing hypotheses used to construct $\widehat{H}_{\mathbf{L}}$. According to the relation between the KFDA test and the kernel linear model, these diagnostic plots could also be applied for KFDA testing.

3.4.3 Influence of the Observations

We are interested in the detection of outliers in our kernel linear model. The n diagonal elements $(\pi_{i,i})_{i\in\{1,\dots,n\}}$ of $\mathbf{\Pi}$ are usually called leverage. A large leverage on the i^{th} observation compared to the other may be a clue of observation i having a large influence on the estimated model parameters.

Another way to study the influence of individuals or group of observations is to compute the Cook's distance. Several multivariate generalizations of the Cook's distance have been proposed, see for instance [36] and references therein. These Cook's distances have trace formulations. The Cook's distance for the i^{th} observation associated to the multivariate linear model of Equation (3.6) is such that:

$$\mathcal{D}_{Cook} = \frac{1}{p}\operatorname{trace}\left((\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})\operatorname{Cov}(\widehat{\boldsymbol{\beta}})^{-1}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})\right),$$

where $\widehat{\boldsymbol{\beta}}_{(i)}$ is the vector of model parameters obtained when the observation i is ignored. Using a matrix inversion lemma, the authors also show that:

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)} = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i\widehat{\varepsilon}_i'}{1-\pi_{i,i}} \in \mathcal{M}_{p,m}(\mathbb{R}).$$

In order to generalize the Cook's distance to our kernel linear model, we need to define the expectation and covariance of vectors of $\mathcal{H}^p$ in order to compute $\mathbb{E}(\widehat{\Theta})$ and $\operatorname{Cov}(\widehat{\Theta})$. Let $h = (h_1, \dots, h_p)$ a random vector of $\mathcal{H}^p$. We define the expectation of h as $\mathbb{E}(h) = (\mathbb{E}(h_1), \dots, \mathbb{E}(h_p)) \in \mathcal{H}^p$. We define the covariance of h as a $p \times p$ matrix where the coordinates are elements of $\mathrm{HS}(\mathcal{H})$, such that for $i, j \in \{1, \dots, p\}$, we have $\operatorname{Cov}(h)_{i,j} = \operatorname{Cov}(h_i, h_j) = \mathbb{E}\left((h_i - \mathbb{E}(h_i)) \otimes (h_j - \mathbb{E}(h_j))\right)$. Now we compute the expectation $\mathbb{E}\widehat{\Theta}$ and covariance $\operatorname{Cov}(\widehat{\Theta})$ of $\widehat{\Theta}$. Note that the combination of Equations (3.8) and (3.7) gives $\widehat{\Theta} = \Theta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon$. Let $j \in \{1, \dots, p\}$, we have $\widehat{\theta}_j = \theta_j + \varepsilon \odot w_j$, where $w_j \in \mathbb{R}^n$ is the j^{th} column of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$. Then we have $\mathbb{E}\widehat{\theta}_j = \theta_j$ and $\mathbb{E}\widehat{\Theta} = (\theta_1, \dots, \theta_p) \in \mathcal{H}^p$. Let

$i, j \in \{1, \dots, p\}$, we have:

$$\begin{aligned}
\mathrm{Cov}(\widehat{\theta}_i, \widehat{\theta}_j) =& \mathbb{E}\left((\widehat{\theta}_i - \theta_i) \otimes (\widehat{\theta}_j - \theta_j)\right) \\
=& \mathbb{E}\left(\varepsilon \odot w_i \otimes \varepsilon w_j\right) \\
=& \sum_{k,l=1}^{n} w_{i,k} w_{j,l} \mathbb{E}\left(\varepsilon_k \otimes \varepsilon_l\right) \\
=& \sum_{k=1}^{n} w_{i,k} w_{j,k} \Sigma_\varepsilon \\
=& (\mathbf{X}'\mathbf{X})^{-1}_{i,j} \Sigma_\varepsilon.
\end{aligned}$$

Thus we have $\mathrm{Cov}(\widehat{\Theta}) = \left((\mathbf{X}'\mathbf{X})^{-1}_{i,j}\Sigma_\varepsilon\right)_{i,j\in\{1,\dots,p\}}$. We use the notation $\boxtimes$, such that $\Sigma_\varepsilon \boxtimes (\mathbf{X}'\mathbf{X})^{-1}$ represents the matrix where the element (i, j) is $(\mathbf{X}'\mathbf{X})^{-1}_{i,j} \in \mathbb{R}$ multiplied with the Hilbert-Schmidt operator Σ_ε. Then we have $\mathrm{Cov}(\Theta) = \Sigma_\varepsilon \boxtimes (\mathbf{X}'\mathbf{X})^{-1}$. We use $\widehat{\Sigma}_{\varepsilon,T} \boxtimes (\mathbf{X}'\mathbf{X})^{-1}$ as an empirical estimator of $\mathrm{Cov}(\Theta)$. Let $\widehat{\Theta}_{(i)}$ be the estimated model parameters obtained when ignoring the i^{th} observation, we have the kernel Cook's distance:

$$\mathcal{D}_{Cook} = \frac{1}{p} \mathrm{tr}\left((\widehat{\Theta} - \widehat{\Theta}_{(i)}) \mathrm{Cov}(\widehat{\Theta})^{-1} (\widehat{\Theta} - \widehat{\Theta}_{(i)})\right).$$

It is possible to extend the definition of the kernel Cook distance to any linear combination of Θ [36]:

$$\mathcal{D}_{Cook,L} = \frac{1}{\mathcal{L}} \mathrm{tr}\left((\mathbf{L}(\widehat{\Theta} - \widehat{\Theta}_{(i)}))' \mathrm{Cov}(\mathbf{L}\widehat{\Theta})^{-1} (\mathbf{L}(\widehat{\Theta} - \widehat{\Theta}_{(i)}))\right),$$

where we can show that $\mathrm{Cov}(\mathbf{L}\Theta) = \Sigma_\varepsilon \boxtimes \mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'$. The computation of the kernel Cook distance is developed in the next section dedicated to the kernel trick.

3.4.4 Application to the Reversion Dataset

We propose to illustrate the data exploration and diagnostic tools of this section on the Reversion RTqPCR dataset presented in Chapter 2 and first published in [153]. The Reversion RTqPCR dataset contains four conditions 0H, 24H, 48HDIFF and 48HREV and each conditions is divided in eight batches corresponding to manipulation repetitions. In Chapter 2, we compared the four comparisons in a pair-wise approach using the truncated KFDA two-sample test. Here we propose to model the condition and batch effect with a kernel linear model and then to apply the kernel Hotelling-Lawley test to

test if there is an effect of the condition and if there is an effect of the batch. The main differences between these two approaches are that we do not test for the same hypothesis, the noise is better estimated with the Hotelling-Lawley test, and the batch effect is automatically corrected. We represent each observation of the dataset as $Y_{i,j,k}$ where $i \in \{0\text{H},24\text{H},48\text{HDIFF},48\text{HREV}\}$ stands for the condition, $j \in \{1,\dots,8\}$ stands for the batch effect and $k \in \{1,\dots,n_{i,j}\}$ represents the k^{th} cell of condition i in batch j. We propose the following kernel linear model:

$$\phi(Y_{i,j,k}) = \theta_i + \theta_j + \varepsilon_{i,j,k},$$

where $\theta_{0H}, \theta_{24H}, \theta_{48HDIFF}, \theta_{48HREV} \in \mathcal{H}$ are the model parameters associated to the condition and $\theta_1,\dots,\theta_8 \in \mathcal{H}$ are the model parameters associated to the batch effect. The design matrix associated to this model is in $\mathcal{M}_{n_{rev},12}(\mathbb{R})$, where n_{rev} is the total number of observed cells.

The kernelized response plots and the kernelized residual plots give a general idea on the relevance of our model. We represented these diagnostic plots with respect to the first three eigendirections of the residual covariance operator in Figure (3.1). As expected, the response plots follows the identity line and the residual plots follow the horizontal null axis. We also observe that the effect of the condition (different colors) is very important compared to the effect of the batch (eight parallel lines per color).

Then, we apply twice the Hotelling-Lawley test. First, we assess the presence of a significant batch effect with a one-versus-all test matrix similar to the matrix of Equation (3.11). The test does not reject the null hypothesis, meaning that the experience is reproducible and there is no biais induced by the batch. Then we test for an effect of the condition with a one-versus-all approach and reject the null hypothesis, that confirms the results of Chapter 2 about the existence of significant differences for at least one condition compared to the others. As we have four conditions, the matrix $\mathbf{L}$ encoding for the one-versus-all hypothesis has three rows and the resulting test operator $\Sigma_{\varepsilon,T}^{-1}\widehat{H}_{\mathbf{L}}$ has rank 3 when $T \geq 3$. Thus, there are three generalized discriminant directions. The densities of the four conditions projected on each generalized discriminant direction are represented in Figure 3.2. we observe that the first generalized discriminant direction, to interpret as encoding the most meaningful differences, orders the four conditions as 48HDIFF, 48HREV, 0H, 24H. It seems consistent with the analyses of Chapter 2, where

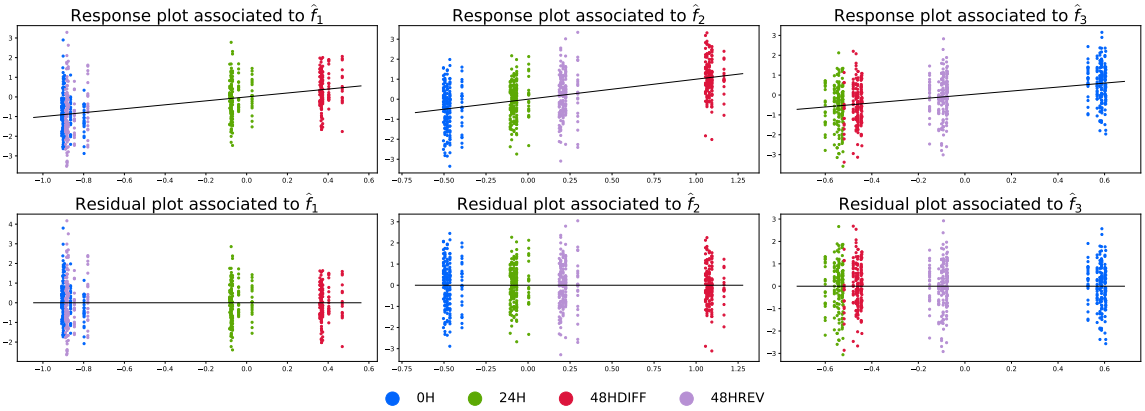

Figure 3.1: Diagnostic plots associated to the densities of the four conditions of the Reversion dataset projected on the first (left), second (middle) and third (right) eigendirections of the residual covariance operator.

we showed that the population 48HREV contained a sub-population close to condition 48HDIFF and the remaining cells were close to condition 0H. We thus focus on the first generalized discriminant direction and observe the Cook influence of the observations on the model with respect to their position on this axis on Figure (3.3). Interestingly, the most influential observations are in condition 48HREV.

This application highlights how starting the comparative analysis of a complex dataset with several meta-information (here, batch effect and condition) can be very visual and intuitive with our approach, as each test response comes with a range of representations. Further analyses would be needed to complete the data analysis of the Reversion RTqPCR dataset and eventually enhance the conclusions we reached with the first analyses performed with the truncated KFDA framework.

3.5 Kernel Trick : the Effective Computation of the Statistic

All the quantities presented in this chapter are implemented in our package *ktest*. In this section, we apply kernel tricks to obtain the explicit expressions associated to them. We start with the diagonalization of the residual covariance operator. The obtained quantites are used to compute the truncated kernel Hotelling-Lawley trace, the informative

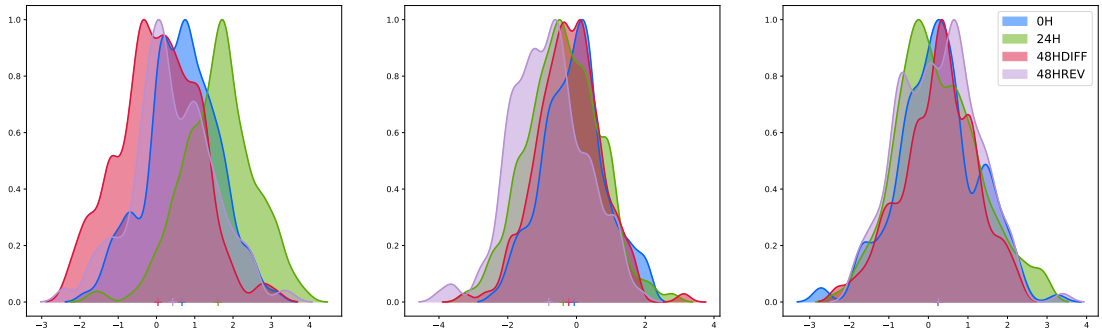

Figure 3.2: Densities of the four conditions of the Reversion dataset projected on the first (left), second (middle) and third (right) generalized discriminant directions associated to the Hotelling-Lawley operator used to test the effect of the condition with $T = 12$.

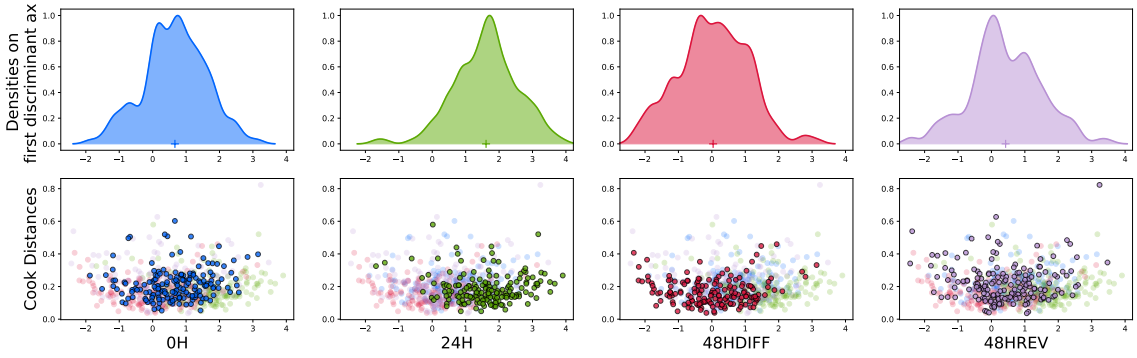

Figure 3.3: Influence of the observations with respect to the position of their projection on the first generalized discriminant directions associated to the Hotelling-Lawley operator used to test the effect of the condition with $T = 12$.

directions for data exploration, the quantities needed to obtain the diagnostic plots, and the Cook distance.

Diagonalization of the Residual Covariance Operator

Theorem 12. *The Hilbert-Schmidt operator $\widehat{\Sigma}_\varepsilon$ has the same spectrum than the matrix $\mathbf{K}_\varepsilon$ defined such that:*

$$\mathbf{K}_\varepsilon = \frac{1}{n}\mathbf{\Pi}^\perp \mathbf{K}_\mathbf{Y} \mathbf{\Pi}^\perp$$

where $\mathbf{K}_\mathbf{Y} = (k(Y_i, Y_j))_{i,j\in\{1,\dots,n\}}$ is the Gram matrix of $\mathbf{Y}$ with respect to $k(\cdot,\cdot)$.

Moreover, if u is a unit eigenvector of $\mathbf{K}_\varepsilon$ associated to the eigenvalue $\widehat{\lambda}$, then a unit eigenfunction $\widehat{f}$ of $\widehat{\Sigma}_\varepsilon$ associated to $\widehat{\lambda}$ is obtained by:

$$\widehat{f} = \frac{1}{\sqrt{n\widehat{\lambda}}}\widehat{\varepsilon} \odot u \tag{3.12}$$

Sketch of proof. We first show that for each eigenfunction $\widehat{f}$ of the Hilbert-Schmidt operator $\widehat{\Sigma}_\varepsilon$ associated to the eigenvalue $\widehat{\lambda}$, the vector $\left(\left\langle \widehat{f}, \widehat{\varepsilon}_1 \right\rangle_\mathcal{H}, \dots, \left\langle \widehat{f}, \widehat{\varepsilon}_n \right\rangle_\mathcal{H}\right)$ of $\mathbb{R}^n$ is an eigenvector of $\mathbf{K}_\varepsilon \in \mathcal{M}_n(\mathbb{R})$ associated to the eigenvalue $\widehat{\lambda}$. It shows that the spectrum of $\widehat{\Sigma}_\varepsilon$ is a subset of the spectrum of $\mathbf{K}_\varepsilon$. Then, we show that for each eigenvector u of $\mathbf{K}_\varepsilon$, the function $\widehat{\varepsilon} \odot u \in \mathcal{H}$ is an eigenfunction of $\widehat{\Sigma}_\varepsilon$. It shows that the spectrum of $\mathbf{K}_\varepsilon$ is a subset of the spectrum of $\widehat{\Sigma}_\varepsilon$. We then conclude that $\mathbf{K}_\varepsilon$ and $\widehat{\Sigma}_\varepsilon$ have the same spectrum. Then we consider u, a unit eigenvector of $\mathbf{K}_\varepsilon$ associated to the eigenvalue $\widehat{\lambda}$. We know that $\widehat{\varepsilon} \odot u$ is an eigenfunction of $\widehat{\Sigma}_\varepsilon$ associated to the eigenvalue $\widehat{\lambda}$. We can determine that $\|\widehat{\varepsilon} \odot u\|_\mathcal{H} = \sqrt{n\widehat{\lambda}}$, thus the vector $\widehat{f} = (n\widehat{\lambda})^{-\frac{1}{2}}\widehat{\varepsilon} \odot u$ is a unit eigenfunction of $\widehat{\Sigma}_\varepsilon$. □

Computation of the Truncated Kernel Hotelling-Lawley Trace Statistic

To compute the test statistic $\mathcal{F}_T$ in practice we need to rewrite it with respect to some tractable quantities. Let $\mathbf{A} = \mathbf{L}'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}\mathbf{L} = (\alpha_{i,j})_{i,j\in\{1,\dots,p\}} \in \mathcal{M}_p(\mathbb{R})$. The matrix $\mathbf{A}$ is related to the model as it is defined with respect to $\mathbf{L}$ and $\mathbf{X}$, we compute it for $O(p^3 + \mathcal{L}^3)$. Then, according Equation (3.5), we have:

$$\begin{aligned}\widehat{H}_\mathbf{L} =& \widehat{\Theta}'\mathbf{L}'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}\mathbf{L}\widehat{\Theta} \\ =& \sum_{i,j=1}^{p} \alpha_{i,j}\widehat{\theta}_i \otimes \widehat{\theta}_j.\end{aligned}$$

Applying Equation (3.1) and replacing the residual covariance operator by its spectral decomposition in the truncated kernel Hotelling-Lawley trace, we have:

$$\begin{aligned}\mathcal{F}_T &= \operatorname{tr}(\widehat{\Sigma}_{\varepsilon,T}^{-1}\widehat{H}_{\mathbf{L}}) \\ &= \left\langle \widehat{\Sigma}_{\varepsilon,T}^{-1}, \widehat{H}_{\mathbf{L}} \right\rangle_{\mathrm{HS}(\mathcal{H})} \\ &= \sum_{i,j=1}^{p}\sum_{t=1}^{T} \alpha_{i,j}\widehat{\lambda}_t^{-1} \left\langle \widehat{f}_t \otimes \widehat{f}_t, \widehat{\theta}_i \otimes \widehat{\theta}_j \right\rangle_{\mathrm{HS}(\mathcal{H})} \\ &= \sum_{i,j=1}^{p}\sum_{t=1}^{T} \alpha_{i,j}\widehat{\lambda}_t^{-1} \left\langle \widehat{f}_t, \widehat{\theta}_i \right\rangle_{\mathcal{H}} \left\langle \widehat{f}_t, \widehat{\theta}_j \right\rangle_{\mathcal{H}}.\end{aligned}$$

The kernel trick to compute this expression is to rewrite it with respect to $\mathbf{U}_T$ and $\widehat{\mathbf{\Lambda}}_T$, the matrices of eigenvectors and eigenfunctions of $\mathbf{K}_\varepsilon$. Let $\mathbf{K}_{\widehat{\Theta}}$ be such that:

$$\mathbf{K}_{\widehat{\Theta}} = \left(\frac{1}{\sqrt{\widehat{\lambda}_t}} \left\langle \widehat{\theta}_j, \widehat{f}_t \right\rangle_{\mathcal{H}} \right)_{t\in\{1,\dots,T\}, j\in\{1,\dots,p\}}.$$

Theorem 13. *We have:*

$$\mathbf{K}_{\widehat{\Theta}} = \left(\frac{1}{\sqrt{\widehat{\lambda}_t}} \left\langle \widehat{\theta}_j, \widehat{f}_t \right\rangle_{\mathcal{H}} \right)_{j\in\{1,\dots,p\}, t\in\{1,\dots,T\}} = \frac{1}{\sqrt{n}} (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}_{\mathbf{Y}}\mathbf{\Pi}^{\perp}\mathbf{U}_T'\widehat{\mathbf{\Lambda}}_T^{-1},$$

where $\mathbf{U}_T = (u_1, \dots, u_T) \in \mathcal{M}_{n,T}(\mathbb{R})$ *and* $\widehat{\mathbf{\Lambda}}_T = \operatorname{diag}(\widehat{\lambda}_1, \dots, \widehat{\lambda}_T) \in \mathcal{M}_T(\mathbb{R})$ *are the matrix containing the* T *first eigenvectors and eigenvalues of* $\mathbf{K}_\varepsilon$ *respectively.*

Proof. Let $\widehat{f}_{\varepsilon,T} = (\widehat{f}_1, \dots, \widehat{f}_T)$ in $\mathcal{H}^T$ be the vector of $\mathcal{H}^T$ containing the T eigenfunctions of $\widehat{\Sigma}_\varepsilon$ associated to the eigenvalues $\widehat{\lambda}_1, \dots, \widehat{\lambda}_T$. From Equation (3.12), we have that:

$$\widehat{f}_{\varepsilon,T} = \frac{1}{\sqrt{n}} \widehat{\mathbf{\Lambda}}_T^{-\frac{1}{2}} \mathbf{U}_T \widehat{\varepsilon}.$$

As $\widehat{\varepsilon} = \mathbf{\Pi}^{\perp}\Phi(\mathbf{Y})$, we have:

$$\widehat{f}_{\varepsilon,T} = \frac{1}{\sqrt{n}} \widehat{\mathbf{\Lambda}}_T^{-\frac{1}{2}} \mathbf{U}_T \mathbf{\Pi}^{\perp}\Phi(\mathbf{Y}).$$

Note that $\widehat{\mathbf{\Lambda}}_T^{-\frac{1}{2}}\widehat{f}_{\varepsilon,T} = (\widehat{\lambda}_1^{-\frac{1}{2}}\widehat{f}_1, \dots, \lambda_T^{-\frac{1}{2}}\widehat{f}_T)$. Let $\widetilde{\mathbf{U}}_T = \frac{1}{\sqrt{n}}\widehat{\mathbf{\Lambda}}_T^{-1}\mathbf{U}_T\mathbf{\Pi}^{\perp} \in \mathcal{M}_{T,n}(\mathbb{R})$ and let $\tilde{u}_1, \dots, \tilde{u}_T$ be the T columns of $\widetilde{\mathbf{U}}_T'$, such that $\widehat{\mathbf{\Lambda}}_T^{-\frac{1}{2}}\widehat{f}_{\varepsilon,T} = \widetilde{\mathbf{U}}_T\Phi(\mathbf{Y})$ and for $t \in \{1, \dots, T\}$,

$\widehat{\lambda}_t^{-\frac{1}{2}}\widehat{f}_t = \Phi(\mathbf{Y}) \odot \tilde{u}_t$. Let $j \in \{1, \ldots, p\}$, we recall that $w_j = (w_{j,1}, \ldots, w_{j,n})$ in $\mathbb{R}^n$ is the j^{th} column of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ and that $\widehat{\theta}_j = \Phi(\mathbf{Y}) \odot w_j$. Then we have:

$$\begin{aligned}\frac{1}{\sqrt{\widehat{\lambda}_t}}\left\langle \widehat{\theta}_j, \widehat{f}_t \right\rangle_{\mathcal{H}} &= \sum_{i,i'=1}^{n} w_{j,i}\tilde{u}_{t,i'} \left\langle \phi(Y_i), \phi(Y_{i'}) \right\rangle_{\mathcal{H}} \\ &= \sum_{i,i'=1}^{n} w_{j,i}\tilde{u}_{t,i'} k(Y_i, Y_{i'}) \\ &= w_j' \mathbf{K}_{\mathbf{Y}} \tilde{u}_t.\end{aligned}$$

Hence, we replace w_j' by $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\tilde{u}_t$ by $\widetilde{\mathbf{U}}_T$ to obtain $\mathbf{K}_{\widehat{\Theta}}$:

$$\mathbf{K}_{\widehat{\Theta}} = \frac{1}{\sqrt{n}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{K}_{\mathbf{Y}}\mathbf{\Pi}^{\perp}\mathbf{U}_T'\widehat{\mathbf{\Lambda}}_T^{-1},$$

□

The truncated kernel Hotelling-Lawley test statistic is simply obtained by computing:

$$\mathcal{F}_T = \text{trace}(\mathbf{K}_{\widehat{\Theta}}'\mathbf{A}\mathbf{K}_{\widehat{\Theta}})$$

Computation of the Informative Directions

Now we determine the informative directions defined as the eigenfunctions of $\widehat{\Sigma}_{\varepsilon,T}^{-1}\widehat{H}_{\mathbf{L}}$ of rank $\kappa \leq \mathcal{L} \wedge T$.

Theorem 14. *Let* $\widetilde{\mathbf{U}}_T = \frac{1}{\sqrt{n}}\widehat{\mathbf{\Lambda}}_T^{-1}\mathbf{U}_T\mathbf{\Pi}^{\perp} \in \mathcal{M}_{T,n}(\mathbb{R})$, $\Psi = \widetilde{\mathbf{U}}_T'\widetilde{\mathbf{U}}_T$ *and* $\mathbf{R} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L})^{-1}\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ *in* $\mathcal{M}_n(\mathbb{R})$. *Let* $\widehat{g}_1, \ldots, \widehat{g}_\kappa$ *the* κ *eigenfunctions of* $\widehat{\Sigma}_{\varepsilon,T}^{-1}\widehat{H}_{\mathbf{L}}$ *associated to the eigenvalues* $\xi_1 \geq \xi_2 \geq \cdots \geq \xi_\kappa$. *Then the matrix* $\mathbf{K}_{\varepsilon,L,T}$ *defined by:*

$$\mathbf{K}_{\varepsilon,L,T} = \mathbf{K}_{\mathbf{Y}}\Psi\mathbf{K}_{\mathbf{Y}}\mathbf{R},$$

has the same eigenvalues than $\widehat{\Sigma}_{\varepsilon,T}^{-1}\widehat{H}_{\mathbf{L}}$. *Moreover, for* $k \in \{1, \ldots, \kappa\}$, *if* v_k *is an eigenvector of* $\mathbf{K}_{\varepsilon,L,T}$ *associated to the eigenvalue* ξ_k, *then we have:*

$$\widehat{g}_k = \frac{1}{\sqrt{\xi_k v_k'\mathbf{R}\mathbf{K}_{\mathbf{Y}}\Psi v_k}}\Phi(\mathbf{Y}) \odot (\Psi\mathbf{K}_{\mathbf{Y}}\mathbf{R}v_k).$$

Sketch of proof. As for Theorem 12, the proof consists in showing that $\mathbf{K}_{\varepsilon,L,T}$ and $\widehat{\Sigma}_{\varepsilon,T}^{-1}\widehat{H}_{\mathbf{L}}$ have the same spectrums. To do so, we show that for an eigenfunction $\widehat{g}$ of $\widehat{\Sigma}_{\varepsilon,T}^{-1}\widehat{H}_{\mathbf{L}}$ associated to the eigenvalue ξ, the vector $(\langle\phi(Y_1),\widehat{g}\rangle_{\mathcal{H}},\ldots,\langle\phi(Y_n),\widehat{g}\rangle_{\mathcal{H}})$ of $\mathbb{R}^n$ is an eigenvector of $\mathbf{K}_{\varepsilon,L,T}$ associated to the eigenvalue ξ. Also, for an eigenvector v of $\mathbf{K}_{\varepsilon,L,T}$ associated to the eigenvalue ξ, the function $\Phi(\mathbf{Y})\odot\Psi\mathbf{K_Y}\mathbf{R}v\in\mathcal{H}$ is an eigenfunction of $\widehat{\Sigma}_{\varepsilon,T}^{-1}\widehat{H}_{\mathbf{L}}$ associated to the eigenvalue ξ. Then we determine that the function $\Phi(\mathbf{Y})\odot\Psi\mathbf{K_Y}\mathbf{R}v$ has norm $\sqrt{\xi v'\mathbf{R}\mathbf{K_Y}\Psi v}$ and normalize it. □

The exploration of the data with respect to the hypothesis is done by computing the orthogonal projections of the embeddings on $\widehat{g}_1,\ldots,\widehat{g}_\kappa$. The matrix $\mathbf{D}=(\langle\phi(Y_i),\widehat{g}_k\rangle_{\mathcal{H}})_{i\in\{1,\ldots,n\},k\in\{1,\ldots,\kappa\}}$ of coordinates of the embbedings in $\text{span}(\widehat{g}_1,\ldots,\widehat{g}_\kappa)$ is such that:

$$\mathbf{D}=\mathbf{K_Y}\Psi\mathbf{K_Y}\mathbf{R}\mathbf{V}\widetilde{\mathbf{\Lambda}}_\xi,$$

where the columns of $\mathbf{V}$ are the eigenvectors of $\mathbf{K}_{\varepsilon,L,T}$ and $\widetilde{\mathbf{\Lambda}}_\xi=\text{diag}(\frac{1}{\sqrt{\xi_1 v_1'\mathbf{R}\mathbf{K_Y}\Psi v_1}},\ldots,\frac{1}{\sqrt{\xi_\kappa v_\kappa'\mathbf{R}\mathbf{K_Y}\Psi v_\kappa}})$

All these quantities may vary with respect to the truncation parameter T. We discuss the choice of T in the Conclusion chapter but the dependance on T of these quantities can be used as a heuristic to choose the value of T in practice.

How to Obtain the Diagnostic Plots

The diagnostic plot relies on the following quantities:

$$\begin{aligned}\mathcal{D}(\Phi(\mathbf{Y}),\widehat{\Sigma}_\varepsilon,T)=&(\left\langle\phi(Y_i),\widehat{f}_t\right\rangle_{\mathcal{H}})_{i\in\{1,\ldots,n\},t\in\{1,\ldots,T\}},\\ \mathcal{D}(\widehat{\Phi(\mathbf{Y})},\widehat{\Sigma}_\varepsilon,T)=&(\left\langle\widehat{\phi(Y_i)},\widehat{f}_t\right\rangle_{\mathcal{H}})_{i\in\{1,\ldots,n\},t\in\{1,\ldots,T\}},\\ \mathcal{D}(\widehat{\varepsilon},\widehat{\Sigma}_\varepsilon,T)=&(\left\langle\widehat{\varepsilon}_i,\widehat{f}_t\right\rangle_{\mathcal{H}})_{i\in\{1,\ldots,n\},t\in\{1,\ldots,T\}},\end{aligned}$$

The determination of these matrix is similar to the determination of the matrix $\mathbf{K}_{\widehat{\Theta}}$, we obtain:

$$\begin{aligned}\mathcal{D}(\Phi(\mathbf{Y}), \widehat{\Sigma}_\varepsilon, T) =& \frac{1}{\sqrt{n}} \mathbf{K_Y \Pi U}_T \widehat{\mathbf{\Lambda}}_T^{-\frac{1}{2}},\\ \mathcal{D}(\widehat{\Phi(\mathbf{Y})}, \widehat{\Sigma}_\varepsilon, T) =& \frac{1}{\sqrt{n}} \mathbf{\Pi K_Y \Pi U}_T \widehat{\mathbf{\Lambda}}_T^{-\frac{1}{2}},\\ \mathcal{D}(\widehat{\varepsilon}, \widehat{\Sigma}_\varepsilon, T) =& \frac{1}{\sqrt{n}} \mathbf{\Pi}^\perp \mathbf{K_Y \Pi U}_T \widehat{\mathbf{\Lambda}}_T^{-\frac{1}{2}}.\end{aligned}$$

Computation of the Cook Distance

Let $i \in \{1, \ldots, n\}$, the RKHS version of the expression expression of $(\widehat{\Theta} - \widehat{\Theta}_{(i)})$ is such that (see [36] for the multivariate version):

$$\widehat{\Theta} - \widehat{\Theta}_{(i)} = \begin{pmatrix} \frac{w_{1,i}}{1-\pi_{i,i}} \widehat{\varepsilon}_i \\ \vdots \\ \frac{w_{p,i}}{1-\pi_{i,i}} \widehat{\varepsilon}_i \end{pmatrix} \in \mathcal{H}^p$$

Then we have that:

$$\begin{aligned}\mathcal{D}_{Cook} =& \frac{1}{p} \operatorname{tr} \left(\sum_{j,k=1}^{p} \frac{w_{j,i}}{1-\pi_{i,i}} \frac{w_{k,i}}{1-\pi_{i,i}} (\mathbf{X}'\mathbf{X})_{j,k} \widehat{\varepsilon}_i \otimes \widehat{\Sigma}_\varepsilon \widehat{\varepsilon}_i \right)\\ =& \frac{1}{p} \sum_{j,k=1}^{p} \frac{w_{j,i}}{1-\pi_{i,i}} \frac{w_{k,i}}{1-\pi_{i,i}} (\mathbf{X}'\mathbf{X})_{j,k} \operatorname{tr} \left(\widehat{\varepsilon}_i \otimes \widehat{\Sigma}_\varepsilon \widehat{\varepsilon}_i \right)\end{aligned}$$

We remark that:

$$\sum_{j,k=1}^{p} \frac{w_{j,i}}{1-\pi_{i,i}} \frac{w_{k,i}}{1-\pi_{i,i}} (\mathbf{X}'\mathbf{X})_{j,k} = \frac{w_i' \mathbf{X}'\mathbf{X} w_i}{(1-\pi_{i,i})^2}$$

And we have:

$$\operatorname{tr} \left(\widehat{\varepsilon}_i \otimes \widehat{\Sigma}_\varepsilon \widehat{\varepsilon}_i \right) = \sum_{t=1}^{T} \widehat{\lambda}_t^{-1} \left\langle \widehat{f}_t, \widehat{\varepsilon}_i \right\rangle_{\mathcal{H}}^2.$$

We recognize the elements of the i^{th} line of $\mathcal{D}(\widehat{\varepsilon}, \widehat{\Sigma}_\varepsilon, T)$, thus we have:

$$\begin{aligned}\operatorname{tr}\left(\widehat{\varepsilon}_i \otimes \widehat{\Sigma}_\varepsilon \widehat{\varepsilon}_i\right) =& (\mathcal{D}(\widehat{\varepsilon}, \widehat{\Sigma}_\varepsilon, T)\widehat{\mathbf{\Lambda}}^{-1}\mathcal{D}(\widehat{\varepsilon}, \widehat{\Sigma}_\varepsilon, T)')_{i,i} \\ =& \frac{1}{n}(\mathbf{\Pi}^\perp \mathbf{K_Y} \mathbf{\Pi} \mathbf{U}_T \widehat{\mathbf{\Lambda}}_T^{-2} \mathbf{U}_T' \mathbf{\Pi} \mathbf{K_Y} \mathbf{\Pi}^\perp)_{i,i}\end{aligned}$$

Finally, we have an expression for the Cook distance $\mathcal{D}_{Cook}$ associated to the i^{th} observation:

$$\frac{w_i' \mathbf{X}' \mathbf{X} w_i}{pn(1-\pi_{i,i})^2}(\mathbf{\Pi}^\perp \mathbf{K_Y} \mathbf{\Pi} \mathbf{U}_T \widehat{\mathbf{\Lambda}}_T^{-2} \mathbf{U}_T' \mathbf{\Pi} \mathbf{K_Y} \mathbf{\Pi}^\perp)_{i,i}.$$

3.6 Discussions

In this Chapter, we introduced a framework to model the embeddings in the feature space with a linear model, this approach has never been done as far as we know. We conjecture that this absence is due to the very bad performance such a model would have for a prediction task. However, in the context of testing, this is a natural generalization of kernel two-sample tests. With this connection between kernel testing and multivariate linear models, kernel testing is enriched with well established diagnostic tools, and the exploration tools derived from the discriminant analysis are generalized to any design.

Computational Aspects

As for kernel two-sample testing, testing with the kernel linear model simplifies the analysis of high-dimensional data, as the more expensive operation is the diagonalisation of the residual covariance operator that depends on the number of observations and not on the number of variables. Moreover, a Nyström approximation of the residual covariance operator can drastically reduce the computational cost of the statistic. Such a Nyström approximation raises issues on balanced landmark sampling adapted to the design, and on the anchors definition, probably as the eigenfunctions of the residual covariance defined with respect to the landmarks.

Variants of the Kernel Hotelling-Lawley Test

We could also define a ridge kernel Hotelling-Lawley statistic through a ridge regularization of the residual covariance operator. Similarly to the KFDA test, the ridge approach

has theoretically more power than the truncated approach, but lacks of interpretability and practicality. The truncated approach relies on the choice of a truncation hyperparameter. The issues associated to this choice discussed in the Conclusion chapter for the truncated KFDA test can be generalized to the truncated kernel Hotelling-Lawley test.

The kernel linear model encompasses kernelization of the ANOVA model for univariate data and a MANOVA model for multivariate data as particular cases.

Our approach on the kernel linear model could also enhance the set of approches for the multivariate linear model. More precisely, the truncated kernel Hotelling-Lawley test is not the exact counterpart of the linear Hotelling-Lawley test as no truncation regularization is needed for the later. It would be possible to define a spectral truncation of the residual covariance matrix in the linear case, that would correspond to a dimension reduction based on the residual variability, and obtain a truncated linear Hotelling-Lawley statistic.

3.7 Proofs

3.7.1 Non-Asymptotic Results on the Residual Covariance Operator.

Proposition 3. *If Assumptions* $\mathtt{A_1}$, $\mathtt{A_2}$ *and* $\mathtt{A_3}$ *are verified, then we have with probability* $1 - e^{-\xi}$*:*

$$\left\|\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})} \leq \zeta(n, \xi), \tag{3.13}$$

where

$$\begin{aligned}\zeta(n,\xi) =& \frac{1}{n}\left(2M_\varepsilon^4 p + (n-p)\left(\mathbb{E}\left\|\varepsilon_1\right\|_{\mathcal{H}}^4 - \left\|\Sigma_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}^2\right)\right)^{1/2} \\ &+ \frac{p}{n}\left\|\Sigma_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})} + \sqrt{\frac{2\xi}{n}} M_\varepsilon^2 (5 + 2p + 2p^2),\end{aligned}$$

and $M_\varepsilon = M_k^{\frac{1}{2}} + M_x \max_{j\in\{1,\dots,p\}} \|\theta_j\|_{\mathcal{H}}$.

Proof. We observe that:

$$\begin{aligned}\left\|\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})} &= \left\|\widehat{\Sigma}_\varepsilon - \mathbb{E}\widehat{\Sigma}_\varepsilon + \mathbb{E}\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})} \\ &\leq \left\|\widehat{\Sigma}_\varepsilon - \mathbb{E}\widehat{\Sigma}_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})} + \left\|\mathbb{E}\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}.\end{aligned}$$

We know from Lemma 6 that:

$$\begin{aligned}\left\|\mathbb{E}\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})} &= \left\|\frac{n-p}{n}\Sigma_\varepsilon - \Sigma_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})} \\ &= \frac{p}{n}\left\|\Sigma_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}.\end{aligned}$$

Now we apply the McDiarmid inequality 16 to the function:

$$\begin{aligned} f: \quad \begin{array}{ccc} \mathcal{H}^n & \longrightarrow & \mathbb{R} \\ \varepsilon & \longmapsto & \left\|\widehat{\Sigma}_\varepsilon - \mathbb{E}(\widehat{\Sigma}_\varepsilon)\right\|_{\mathrm{HS}(\mathcal{H})}. \end{array}\end{aligned}$$

Let $i_0 \in \{1, \ldots, n\}$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ and $\tilde{\varepsilon} = (\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n)$ in $\mathcal{H}^n$ such that $\varepsilon_{i_0} \neq \tilde{\varepsilon}_{i_0}$ and $\forall i \neq i_0, \varepsilon_i = \tilde{\varepsilon}_i$. Lemma 7 gives us that:

$$|f(\varepsilon) - f(\tilde{\varepsilon})| \leq \frac{2M_\varepsilon^2}{n}(5 + 2p + 2p^2).$$

We use this bound to apply the McDiarmid inequality to f. We have with probability $1 - e^{-\xi}$:

$$\left\|\widehat{\Sigma}_\varepsilon - \mathbb{E}\widehat{\Sigma}_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})} \leq \mathbb{E}\left\|\widehat{\Sigma}_\varepsilon - \mathbb{E}\widehat{\Sigma}_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})} + \sqrt{\frac{2\xi}{n}}M_\varepsilon^2(5 + 2p + 2p^2).$$

By Lemma 8, we bound the expectation term to obtain that with probability $1 - e^{-\xi}$:

$$\begin{aligned}\left\|\widehat{\Sigma}_\varepsilon - \mathbb{E}\widehat{\Sigma}_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})} \leq& \frac{1}{n}\left(2M_\varepsilon^4 p + (n-p)\left(\mathbb{E}\left\|\varepsilon_1\right\|_{\mathcal{H}}^4 - \left\|\Sigma_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}^2\right)\right)^{1/2} \\ &+ \sqrt{\frac{2\xi}{n}}M_\varepsilon^2(5 + 2p + 2p^2).\end{aligned}$$

□

A direct consequence of Proposition 3 is the convergence in probability of $\widehat{\Sigma}_\varepsilon$ towards

Σ_ε:

$$\widehat{\Sigma}_\varepsilon \xrightarrow[n\to\infty]{p} \Sigma_\varepsilon.$$

According to the exponential bound obtained, we also deduce that the rate of convergence is equal to $\sqrt{n}$.

Lemma 4. *If Assumption* $\mathtt{A_1}$ *and* $\mathtt{A_2}$ *are verified, then there exists* $M_\varepsilon \in \mathbb{R}$ *such that:*

$$\forall i \in \{1,\dots,n\}, \|\varepsilon_i\|_{\mathcal{H}} < M_\varepsilon.$$

Moreover, we have $M_\varepsilon = M_k^{\frac{1}{2}} + pM_x \max_{j\in\{1,\dots,p\}} \|\theta_j\|_{\mathcal{H}}$

Proof. We have:

$$\begin{aligned}\|\varepsilon_i\|_{\mathcal{H}} &= \|\phi(Y_i) - (\mathbf{X}\Theta)_i\|_{\mathcal{H}} \\ &\leq \|\phi(Y_i)\|_{\mathcal{H}} + \|(\mathbf{X}\Theta)_i\|_{\mathcal{H}} \\ &\leq \|\phi(Y_i)\|_{\mathcal{H}} + \|\Theta \odot x_i\|_{\mathcal{H}}.\end{aligned}$$

From Assumption $\mathtt{A_1}$, we know that $\|\phi(Y_i)\|_{\mathcal{H}} \leq M_k^{\frac{1}{2}}$. We apply Equation (3.4) to obtain $\|\Theta \odot x_i\|_{\mathcal{H}} \leq \|x_i\|_1 \max_{j\in\{1,\dots,p\}} \|\theta_j\|_{\mathcal{H}}$, then Assumption $\mathtt{A_2}$ gives $\|x_i\|_1 = \sum_{j=1}^p |x_{i,j}| \leq pM_x$. Thus $\|\Theta \odot x_i\|_{\mathcal{H}} \leq pM_x \max_{j\in\{1,\dots,p\}} \|\theta_j\|_{\mathcal{H}}$. The sum of the two bounds gives the result. □

Lemma 5. *Under Assumption* $\mathtt{A_3}$*, we have for* $i \in \{1,\dots,n\}$*:*

$$\pi_{i,i} \xrightarrow[n\to\infty]{} 0$$

Proof. For $i,j \in \{1,\dots,n\}$, we know that $\pi_{i,j} = \sum_{s=1}^p v_{s,i}v_{s,j}$ from Lemma 16. Thus we have:

$$|\pi_{i,i}| \leq \sum_{s=1}^p v_{s,i}^2$$

From Assumption $\mathtt{A_3}$, $v_{s,i} \leq M_\pi/\sqrt{n}$, thus $|\pi_{i,i}| \leq \frac{pM_\pi^2}{n}$. As $\frac{pM_\pi^2}{n} \xrightarrow[n\to\infty]{} 0$, we have that $\pi_{i,i} \xrightarrow[n\to\infty]{} 0$. □

Lemma 6. *We have that:*

$$\mathbb{E}(\widehat{\Sigma}_\varepsilon) = \frac{n-p}{n}\Sigma_\varepsilon.$$

Proof. Denote $\mathbf{\Pi}^\perp = (I_\mathcal{H} - \mathbf{\Pi})$ and $\pi_i^\perp = (\pi_{i,1}^\perp, \dots, \pi_{i,n}^\perp)' \in \mathbb{R}^n$ the i^{th} column of $\mathbf{\Pi}^\perp$. We have:

$$\widehat{\varepsilon}_i = \varepsilon \odot \pi_i^\perp = \sum_{j=1}^n \pi_{i,j}^\perp \varepsilon_j.$$

Consequently, the expectation is such that:

$$\begin{aligned}
\mathbb{E}(\widehat{\Sigma}_\varepsilon) =& \frac{1}{n}\mathbb{E}(\sum_{i=1}^n \widehat{\varepsilon}_i \otimes \widehat{\varepsilon}_i) \\
=& \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \pi_{i,j}^\perp \pi_{i,k}^\perp \mathbb{E}(\varepsilon_j \otimes \varepsilon_k) \\
=& \frac{1}{n}\sum_{i=1}^n \left(\sum_{j=1}^n {\pi_{i,j}^\perp}^2 \mathbb{E}(\varepsilon_j \otimes \varepsilon_j) + \sum_{j\neq k} \pi_{i,j}^\perp \pi_{i,k}^\perp \mathbb{E}(\varepsilon_j) \otimes \mathbb{E}(\varepsilon_k) \right) \\
=& \frac{1}{n}(\sum_{i,j=1}^n {\pi_{i,j}^\perp}^2)\Sigma_\varepsilon \\
\overset{(\star)}{=}& \frac{n-p}{n}\Sigma_\varepsilon,
\end{aligned}$$

where we applied Lemma 14 on $\mathbf{\Pi}^\perp$ with $\operatorname{rank}(\mathbf{\Pi}^\perp) = n - p$ to obtain $(\star)$. □

Lemma 7. *Assume that Assumptions* $\mathtt{A_1}$, $\mathtt{A_2}$ *and* $\mathtt{A_3}$ *are true. Let* $f(\varepsilon) = \left\|\widehat{\Sigma}_\varepsilon - \mathbb{E}(\widehat{\Sigma}_\varepsilon)\right\|_{\mathrm{HS}(\mathcal{H})} = \left\|\widehat{\Sigma}_\varepsilon - (n-p)/n\Sigma_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}$. *Let* $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ *and* $\tilde{\varepsilon} = (\tilde{\varepsilon}_1, \dots, \tilde{\varepsilon}_n)$ *in* $\mathcal{H}^n$ *such that* $\varepsilon_{i_0} \neq \tilde{\varepsilon}_{i_0}$ *and* $\forall i \neq i_0, \varepsilon_i = \tilde{\varepsilon}_i$. *We have:*

$$|f(\varepsilon) - f(\tilde{\varepsilon})| \leq \frac{2M_\varepsilon^2}{n}(1 + p(1 + M_\pi + M_\pi^2)),$$

where p *is the rank of the design matrix* $\mathbf{X}$, M_ε *is defined in Lemma 4, and* M_π *comes from Assumption* $\mathtt{A_3}$.

Proof. Denote $\widetilde{\Sigma_\varepsilon} = \frac{1}{n}\sum_{i=1}^n \tilde{\varepsilon}_i \otimes \tilde{\varepsilon}_i$, we have:

$$\begin{aligned}
|f(\varepsilon) - f(\tilde{\varepsilon})| &= \left| \left\| \widehat{\Sigma}_\varepsilon - \mathbb{E}(\widehat{\Sigma}_\varepsilon) \right\|_{\mathrm{HS}(\mathcal{H})} - \left\| \widetilde{\Sigma_\varepsilon} - \mathbb{E}(\widehat{\Sigma}_\varepsilon) \right\|_{\mathrm{HS}(\mathcal{H})} \right| \\
&\leq \left\| \widehat{\Sigma}_\varepsilon - \mathbb{E}(\widehat{\Sigma}_\varepsilon) - (\widetilde{\Sigma_\varepsilon} - \mathbb{E}(\widehat{\Sigma}_\varepsilon)) \right\|_{\mathrm{HS}(\mathcal{H})} \\
&\leq \left\| \widehat{\Sigma}_\varepsilon - \widetilde{\Sigma_\varepsilon} \right\|_{\mathrm{HS}(\mathcal{H})}.
\end{aligned}$$

As $\widehat{\varepsilon}_i = \varepsilon_i - \varepsilon \odot \pi_i$, we have:

$$\widehat{\Sigma}_\varepsilon = \frac{1}{n}\sum_{i=1}^n \varepsilon_i \otimes \varepsilon_i - \frac{1}{n}\sum_{i=1}^n \varepsilon_i \otimes (\varepsilon \odot \pi_i) - \frac{1}{n}\sum_{i=1}^n (\varepsilon \odot \pi_i) \otimes \varepsilon_i + \frac{1}{n}\sum_{i=1}^n (\varepsilon \odot \pi_i) \otimes (\varepsilon \odot \pi_i).$$

As $\mathbf{\Pi}$ is an orthogonal projector, we have that $\mathbf{\Pi} = \mathbf{\Pi}' = \mathbf{\Pi}^2$. Thus for $i, j \in \{1, \ldots, n\}$, we have $\pi_{i,j} = \pi_{j,i}$ and $\sum_{k=1}^n \pi_{i,k}\pi_{j,k} = \pi_{i,j}$. Then by developing each $\varepsilon \odot \pi_i = \sum_{j=1}^n \pi_{i,j}\varepsilon_j$, we obtain that:

$$\frac{1}{n}\sum_{i=1}^n \varepsilon_i \otimes (\varepsilon \odot \pi_i) = \frac{1}{n}\sum_{i=1}^n (\varepsilon \odot \pi_i) \otimes \varepsilon_i = \frac{1}{n}\sum_{i=1}^n (\varepsilon \odot \pi_i) \otimes (\varepsilon \odot \pi_i).$$

It leads to:

$$\widehat{\Sigma}_\varepsilon = \frac{1}{n}\sum_{i=1}^n \varepsilon_i \otimes \varepsilon_i - \frac{1}{n}\sum_{i=1}^n \varepsilon_i \otimes (\varepsilon \odot \pi_i).$$

We replace $\widehat{\Sigma}_\varepsilon$ and $\widetilde{\Sigma_\varepsilon}$ by this expression and use the triangular inequality to obtain:

$$\begin{aligned}
|f(\varepsilon) - f(\tilde{\varepsilon})| \leq & \frac{1}{n} \left\| \sum_{i=1}^n \varepsilon_i \otimes \varepsilon_i - \tilde{\varepsilon}_i \otimes \tilde{\varepsilon}_i \right\|_{\mathrm{HS}(\mathcal{H})} \\
& + \underbrace{\frac{1}{n} \left\| \sum_{i=1}^n \tilde{\varepsilon}_i \otimes (\tilde{\varepsilon} \odot \pi_i) - \varepsilon_i \otimes (\varepsilon \odot \pi_i) \right\|_{\mathrm{HS}(\mathcal{H})}}_{=A},
\end{aligned}$$

The first term is such that:

$$\begin{aligned}
\frac{1}{n} \left\| \sum_{i=1}^n \varepsilon_i \otimes \varepsilon_i - \tilde{\varepsilon}_i \otimes \tilde{\varepsilon}_i \right\|_{\mathrm{HS}(\mathcal{H})} &= \frac{1}{n} \left\| \varepsilon_{i_0} \otimes \varepsilon_{i_0} - \tilde{\varepsilon}_{i_0} \otimes \tilde{\varepsilon}_{i_0} \right\|_{\mathrm{HS}(\mathcal{H})} \\
&\leq \frac{1}{n} \left\| \varepsilon_{i_0} \right\|_{\mathcal{H}}^2 + \left\| \tilde{\varepsilon}_{i_0} \right\|_{\mathcal{H}}^2 \\
&\leq \frac{2M_\varepsilon^2}{n}.
\end{aligned}$$

The second term can be decomposed as:

$$\begin{aligned}
A =& \frac{1}{n}\left\| \sum_{i=1, i\neq i_0}^{n} \varepsilon_i \otimes \Big((\widetilde{\varepsilon} - \varepsilon) \odot \pi_i \Big) + \widetilde{\varepsilon}_{i_0} \otimes (\widetilde{\varepsilon} \odot \pi_{i_0}) - \varepsilon_{i_0} \otimes (\varepsilon \odot \pi_{i_0}) \right\|_{\mathrm{HS}(\mathcal{H})} \\
=& \frac{1}{n}\left\| \sum_{i=1, i\neq i_0}^{n} \varepsilon_i \otimes \Big((\widetilde{\varepsilon} - \varepsilon) \odot \pi_i \Big) + \widetilde{\varepsilon}_{i_0} \otimes \Big((\widetilde{\varepsilon} - \varepsilon) \odot \pi_{i_0} \Big) + (\widetilde{\varepsilon} - \varepsilon)_{i_0} \otimes (\varepsilon \odot \pi_{i_0}) \right\|_{\mathrm{HS}(\mathcal{H})} \\
\leq& \underbrace{\frac{1}{n}\left\| \sum_{i=1, i\neq i_0}^{n} \varepsilon_i \otimes \Big((\widetilde{\varepsilon} - \varepsilon) \odot \pi_i \Big) \right\|_{\mathrm{HS}(\mathcal{H})}}_{A_1} + \underbrace{\frac{1}{n}\left\| \widetilde{\varepsilon}_{i_0} \otimes \Big((\widetilde{\varepsilon} - \varepsilon) \odot \pi_{i_0} \Big) \right\|_{\mathrm{HS}(\mathcal{H})}}_{A_2} \\
& + \underbrace{\frac{1}{n} \| (\widetilde{\varepsilon} - \varepsilon)_{i_0} \otimes (\varepsilon \odot \pi_{i_0}) \|_{\mathrm{HS}(\mathcal{H})}}_{A_3}.
\end{aligned}$$

We now bound each term. Remark that according to Lemma 16, for $h \in \mathcal{H}^n$ and $i \in \{1, \ldots, n\}$, we have

$$\begin{aligned}
h \odot \pi_i &= \sum_{s=1}^{p} v_{s,i} h \odot v_s \\
&= \sum_{s=1}^{p} \sum_{j=1}^{n} v_{s,i} v_{s,j} h_j.
\end{aligned}$$

In particular, we have:

$$(\widetilde{\varepsilon} - \varepsilon) \odot \pi_i = \sum_{s=1}^{p} v_{s,i} v_{s,i_0} (\widetilde{\varepsilon}_{i_0} - \varepsilon_{i_0}).$$

Thus:

$$\begin{aligned}
A_1 \leq & \frac{1}{n} \sum_{i=1, i\neq i_0}^{n} \left\| \varepsilon_i \otimes (\widetilde{\varepsilon} - \varepsilon) \odot \pi_i \right\|_{\mathrm{HS}(\mathcal{H})} \\
\leq & \frac{M_\varepsilon}{n} \sum_{i=1, i\neq i_0}^{n} \left\| (\widetilde{\varepsilon} - \varepsilon) \odot \pi_i \right\|_{\mathcal{H}} \\
\leq & \frac{M_\varepsilon}{n} \sum_{i=1, i\neq i_0}^{n} \sum_{s=1}^{p} |v_{s,i}||v_{s,i_0}| \left\| (\widetilde{\varepsilon}_{i_0} - \varepsilon_{i_0}) \right\|_{\mathcal{H}} \\
\leq & \frac{2M_\varepsilon^2}{n} \frac{M_\pi^2}{n} \sum_{i=1, i\neq i_0}^{n} \sum_{s=1}^{p} 1 \\
\leq & \frac{2M_\varepsilon^2}{n} M_\pi^2 p.
\end{aligned}$$

We also have:

$$\begin{aligned}
A_2 \leq & \frac{M_\varepsilon}{n} \left\| (\widetilde{\varepsilon} - \varepsilon) \odot \pi_{i_0} \right\|_{\mathcal{H}} \\
\leq & \frac{M_\varepsilon}{n} \sum_{s=1}^{p} |v_{s,i_0}||v_{s,i_0}| \left\| (\widetilde{\varepsilon}_{i_0} - \varepsilon_{i_0}) \right\|_{\mathcal{H}} \\
\leq & \frac{2M_\varepsilon^2}{n} \sum_{s=1}^{p} |v_{s,i_0}|^2 \\
\leq & \frac{2M_\varepsilon^2}{n} p,
\end{aligned}$$

where we used that for $i \in \{1, \ldots, n\}$ and $s \in \{1, \ldots, p\}$,$|v_{s,i_0}| \leq 1$. We also have:

$$\begin{aligned}
A_3 \leq & \frac{2M_\varepsilon}{n} \left\| \varepsilon \odot \pi_{i_0} \right\|_{\mathcal{H}} \\
\leq & \frac{2M_\varepsilon}{n} \sum_{s=1}^{p} \sum_{i=1}^{n} |v_{s,i_0}||v_{s,i}| \left\| \varepsilon_i \right\|_{\mathcal{H}} \\
\leq & \frac{2M_\varepsilon^2}{n} \sum_{s=1}^{p} |v_{s,i_0}| \sum_{i=1}^{n} |v_{s,i}|.
\end{aligned}$$

Note that $\sum_{i=1}^{n} |v_{s,i}| = \|v_s\|_1 \leq \sqrt{n} \|v_s\| \leq \sqrt{n}$ and that $\sum_{s=1}^{p} |v_{s,i_0}| \leq \frac{pM_\pi}{\sqrt{n}}$. Injecting these two results in the bound on A_3 leads to:

$$A_3 \leq \frac{2M_\varepsilon^2}{n} p M_\pi.$$

Finally, we have that:

$$\begin{aligned}|f(\varepsilon) - f(\tilde{\varepsilon})| \leq & \frac{2M_\varepsilon^2}{n} + \frac{2M_\varepsilon^2}{n} M_\pi^2 p + \frac{2M_\varepsilon^2}{n} p + \frac{2M_\varepsilon^2}{n} p M_\pi \\ \leq & \frac{2M_\varepsilon^2}{n}(1 + p(1 + M_\pi + M_\pi^2)).\end{aligned}$$

□

Lemma 8. *Under Assumptions* $\mathtt{A}_1$ *and* $\mathtt{A}_2$*, we have:*

$$\mathbb{E}\left\|\widehat{\Sigma}_\varepsilon - \mathbb{E}\widehat{\Sigma}_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})} \leq \frac{1}{n}\left(2M_\varepsilon^4 p + (n-p)\left(\mathbb{E}\left\|\varepsilon_1\right\|_{\mathcal{H}}^4 - \left\|\Sigma_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}^2\right)\right)^{1/2},$$

where M_ε *is defined in Lemma 4.*

Proof. Note that:

$$\begin{aligned}\mathbb{E}\left\|\widehat{\Sigma}_\varepsilon - \mathbb{E}\widehat{\Sigma}_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}^2 =& \mathbb{E}\left\|\widehat{\Sigma}_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}^2 - 2\mathbb{E}\left\langle\widehat{\Sigma}_\varepsilon, \mathbb{E}\widehat{\Sigma}_\varepsilon\right\rangle_{\mathrm{HS}(\mathcal{H})} + \left\|\mathbb{E}\widehat{\Sigma}_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}^2 \\ =& \mathbb{E}\left\|\widehat{\Sigma}_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}^2 - 2\left\langle\mathbb{E}\widehat{\Sigma}_\varepsilon, \mathbb{E}\widehat{\Sigma}_\varepsilon\right\rangle_{\mathrm{HS}(\mathcal{H})} + \left\|\mathbb{E}\widehat{\Sigma}_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}^2 \\ =& \mathbb{E}\left\|\widehat{\Sigma}_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}^2 - \left\|\mathbb{E}\widehat{\Sigma}_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}^2\end{aligned}$$

By Jensen's inequality, we have that:

$$\begin{aligned}\mathbb{E}\left\|\widehat{\Sigma}_\varepsilon - \mathbb{E}\widehat{\Sigma}_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})} \leq & \left[\mathbb{E}\left\|\widehat{\Sigma}_\varepsilon - \mathbb{E}\widehat{\Sigma}_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}^2\right]^{\frac{1}{2}} \\ \leq & \left[\mathbb{E}\left\|\widehat{\Sigma}_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}^2 - \left\|\mathbb{E}\widehat{\Sigma}_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}^2\right]^{\frac{1}{2}}.\end{aligned}$$

We can develop $\widehat{\Sigma}_\varepsilon$ to obtain that:

$$\widehat{\Sigma}_\varepsilon = \frac{1}{n}\sum_{i=1}^n (1 - \pi_{i,i})\varepsilon_i \otimes \varepsilon_i + \frac{1}{n}\sum_{i,j=1, i\neq j}^n \pi_{i,j}\varepsilon_i \otimes \varepsilon_j. \tag{3.14}$$

Thus:

$$\left\|\widehat{\Sigma}_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}^2 = A - 2B + C,$$

where

$$
\begin{aligned}
A =& \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(1-\pi_{i,i})(1-\pi_{j,j})\left\langle \varepsilon_i\otimes\varepsilon_i, \varepsilon_j\otimes\varepsilon_j\right\rangle_{\mathrm{HS}(\mathcal{H})},\\
B =& \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1,k\neq j}^{n}(1-\pi_{i,i})\pi_{j,k}\left\langle \varepsilon_i\otimes\varepsilon_i, \varepsilon_j\otimes\varepsilon_k\right\rangle_{\mathrm{HS}(\mathcal{H})},\\
C =& \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\sum_{k=1}^{n}\sum_{l=1,l\neq k}^{n}\pi_{i,j}\pi_{k,l}\left\langle \varepsilon_i\otimes\varepsilon_j, \varepsilon_k\otimes\varepsilon_l\right\rangle_{\mathrm{HS}(\mathcal{H})}.
\end{aligned}
$$

We now compute the expectation of each term:

$$
\begin{aligned}
\mathbb{E}(A) =& \frac{1}{n^2}\sum_{i=1}^{n}(1-\pi_{i,i})^2\mathbb{E}\left(\left\langle \varepsilon_i\otimes\varepsilon_i, \varepsilon_i\otimes\varepsilon_i\right\rangle_{\mathrm{HS}(\mathcal{H})}\right)\\
&+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}(1-\pi_{i,i})(1-\pi_{j,j})\left\langle \mathbb{E}\left(\varepsilon_i\otimes\varepsilon_i\right), \mathbb{E}\left(\varepsilon_j\otimes\varepsilon_j\right)\right\rangle_{\mathrm{HS}(\mathcal{H})}\\
=& \frac{1}{n^2}\sum_{i=1}^{n}(1-\pi_{i,i})^2\mathbb{E}\left(\left\|\varepsilon_1\right\|_{\mathcal{H}}^4\right)+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}(1-\pi_{i,i})(1-\pi_{j,j})\left\|\Sigma_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}^2.
\end{aligned}
$$

We can directly see that every term of B contains at least one indice j or k different from the others, involving that the expectation of each term of B is null, which give that $\mathbb{E}(B)=0$. The same happens for every term of C where indices k and l are different from indices i and j. Consequently we have:

$$
\begin{aligned}
\mathbb{E}(C) =& \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}{\pi_{i,j}}^2\left(\mathbb{E}\left\langle \varepsilon_i\otimes\varepsilon_j, \varepsilon_i\otimes\varepsilon_j\right\rangle_{\mathrm{HS}(\mathcal{H})}+\mathbb{E}\left\langle \varepsilon_i\otimes\varepsilon_j, \varepsilon_j\otimes\varepsilon_i\right\rangle_{\mathrm{HS}(\mathcal{H})}\right)\\
=& \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}{\pi_{i,j}}^2\mathbb{E}\left\|\varepsilon_i\right\|_{\mathcal{H}}^2\mathbb{E}\left\|\varepsilon_j\right\|_{\mathcal{H}}^2.
\end{aligned}
$$

According to Equation (3.14), $\mathbb{E}\widehat{\Sigma}_\varepsilon$ is such that:

$$
\begin{aligned}
\mathbb{E}\widehat{\Sigma}_\varepsilon =& \frac{1}{n}\sum_{i=1}^{n}(1-\pi_{i,i})\mathbb{E}(\varepsilon_i\otimes\varepsilon_i)+\frac{1}{n}\sum_{i,j=1,i\neq j}^{n}\sum_{j=1,j\neq i}^{n}\pi_{i,j}\mathbb{E}\varepsilon_i\otimes\mathbb{E}\varepsilon_j\\
=& \frac{1}{n}\sum_{i=1}^{n}(1-\pi_{i,i})\Sigma_\varepsilon.
\end{aligned}
$$

Then:

$$\begin{aligned}\left\|\mathbb{E}\widehat{\Sigma}_\varepsilon\right\|^2_{\mathrm{HS}(\mathcal{H})} =& \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(1-\pi_{i,i})(1-\pi_{j,j})\left\|\Sigma_\varepsilon\right\|^2_{\mathrm{HS}(\mathcal{H})} \\ =& \frac{1}{n^2}\sum_{i=1}^{n}(1-\pi_{i,i})^2\left\|\Sigma_\varepsilon\right\|^2_{\mathrm{HS}(\mathcal{H})} + \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}(1-\pi_{i,i})(1-\pi_{j,j})\left\|\Sigma_\varepsilon\right\|^2_{\mathrm{HS}(\mathcal{H})}.\end{aligned}$$

Now we can sum $\mathbb{E}(A), \mathbb{E}(C)$ and $\left\|\mathbb{E}\widehat{\Sigma}_\varepsilon\right\|^2_{\mathrm{HS}(\mathcal{H})}$:

$$\begin{aligned}\mathbb{E}\left\|\widehat{\Sigma}_\varepsilon\right\|^2_{\mathrm{HS}(\mathcal{H})} - \left\|\mathbb{E}\widehat{\Sigma}_\varepsilon\right\|^2_{\mathrm{HS}(\mathcal{H})} =& \mathbb{E}(A) + \mathbb{E}(C) - \left\|\mathbb{E}\widehat{\Sigma}_\varepsilon\right\|^2_{\mathrm{HS}(\mathcal{H})} \\ =& \frac{1}{n^2}\sum_{i=1}^{n}(1-\pi_{i,i})^2\mathbb{E}\left\|\varepsilon_1\right\|^4_{\mathcal{H}} + \frac{2}{n^2}\sum_{i,j=1,i\neq j}^{n}{\pi_{i,j}}^2\mathbb{E}\left\|\varepsilon_i\right\|^2_{\mathcal{H}}\mathbb{E}\left\|\varepsilon_j\right\|^2_{\mathcal{H}} \\ & - \frac{1}{n^2}\sum_{i=1}^{n}(1-\pi_{i,i})^2\left\|\Sigma_\varepsilon\right\|^2_{\mathrm{HS}(\mathcal{H})} \\ =& \frac{1}{n^2}\sum_{i=1}^{n}(1-\pi_{i,i})^2\left(\mathbb{E}\left\|\varepsilon_1\right\|^4_{\mathcal{H}} - \left\|\Sigma_\varepsilon\right\|^2_{\mathrm{HS}(\mathcal{H})}\right) \\ & + \frac{2}{n^2}\sum_{i,j=1,i\neq j}^{n}{\pi_{i,j}}^2\mathbb{E}\left\|\varepsilon_i\right\|^2_{\mathcal{H}}\mathbb{E}\left\|\varepsilon_j\right\|^2_{\mathcal{H}} \\ \leq& \frac{n-p}{n^2}\left(\mathbb{E}\left\|\varepsilon_1\right\|^4_{\mathcal{H}} - \left\|\Sigma_\varepsilon\right\|^2_{\mathrm{HS}(\mathcal{H})}\right) + \frac{2M_\varepsilon^4 p}{n^2}.\end{aligned}$$

where the bound comes from applying Lemma 14. Finally:

$$\mathbb{E}\left\|\widehat{\Sigma}_\varepsilon - \mathbb{E}\widehat{\Sigma}_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})} \leq \frac{1}{n}\left(2M_\varepsilon^4 p + (n-p)\left(\mathbb{E}\left\|\varepsilon_1\right\|^4_{\mathcal{H}} - \left\|\Sigma_\varepsilon\right\|^2_{\mathrm{HS}(\mathcal{H})}\right)\right)^{1/2}.$$

□

3.7.2 Proof of the Asymptotic Distribution of the Hotelling-Lawley Trace Test Statistic.

In this section, we give the details that were not given on the sketch of the proof of Theorem 11.

Lemma 9. *Under Assumptions* $\mathtt{A}_1$ *and* $\mathtt{A}_2$*, if* H_0 *is true, then we have:*

$$\left\|\frac{1}{n}\widehat{H}_{\mathbf{L}}\right\|_{\mathrm{HS}(\mathcal{H})} \leq M_\varepsilon^2,$$

where M_ε is defined in Lemma 4.

Proof. We inject the expression of $\widehat{\Theta}$ in the expression of $\widehat{H}_{\mathbf{L}}$:

$$\widehat{H}_{\mathbf{L}} = (\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Phi(\mathbf{Y}))'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Phi(\mathbf{Y}))$$

□

According to the linear model on the embeddings, we have

$$(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Phi(\mathbf{Y})) = (\mathbf{L}\Theta + \mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon).$$

The term $\mathbf{L}\Theta$ is null under H_0. Then we can write:

$$\widehat{H}_{\mathbf{L}} = \varepsilon\mathbf{R}\varepsilon$$

where $\mathbf{R} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \in \mathcal{M}_n(\mathbb{R})$ is an orthogonal projector, as $\mathbf{R}' = \mathbf{R}$ and $\mathbf{R}^2 = \mathbf{R}$. Since $\mathbf{L}$ and $\mathbf{X}$ are both full rank, we have that $\text{rank}(\mathbf{R}) \leq \min(p, \mathcal{L})$. Let $R_i = (r_{i,1}, \ldots, r_{i,n}) \in \mathbb{R}^n$ be the i^{th} column of $\mathbf{R}$. We have:

$$\begin{aligned}
\left\|\widehat{H}_{\mathbf{L}}\right\|_{\text{HS}(\mathcal{H})} &= \|(\mathbf{R}\varepsilon)'(\mathbf{R}\varepsilon)\|_{\text{HS}(\mathcal{H})} \\
&= \sum_{i=1}^{n} \|(\mathbf{R}\varepsilon)_i \otimes (\mathbf{R}\varepsilon)_i\|_{\text{HS}(\mathcal{H})} \\
&= \sum_{i=1}^{n} \|(\mathbf{R}\varepsilon)_i\|_{\mathcal{H}}^2 \\
&= \|\mathbf{R}\varepsilon\|_{\mathcal{H}^n}^2 .
\end{aligned}$$

By Lemma 15, we have $\|\mathbf{R}\varepsilon\|_{\mathcal{H}^n} \leq \|\varepsilon\|_{\mathcal{H}^n}$, thus:

$$\begin{aligned}
\left\|\widehat{H}_{\mathbf{L}}\right\|_{\text{HS}(\mathcal{H})} &\leq \|\varepsilon\|_{\mathcal{H}^n}^2 \\
&\leq \sum_{i=1}^{n} \|\varepsilon_i\|_{\mathcal{H}}^2 \\
&\leq n M_\varepsilon^2
\end{aligned}$$

Thus we have $\left\|n^{-1}\widehat{H}_{\mathbf{L}}\right\|_{\text{HS}(\mathcal{H})} \leq M_\varepsilon^2$.

Lemma 10. *Under Assumptions* $\mathtt{A}_1$, $\mathtt{A}_2$, $\mathtt{A}_3$ *and* $\mathtt{A}_4$, *we have:*

$$n\widetilde{\mathcal{F}}_T = \mathrm{tr}(\Sigma_{\varepsilon,T}^{-1}\widehat{H}_{\mathbf{L}}) \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2_{\mathcal{L}T}.$$

Proof. We recall that the i^{th} equation of the kernel linear model on $\Phi(\mathbf{Y})$ is:

$$\phi(Y_i) = \Theta \odot x_i + \varepsilon_i.$$

We know that the eigenfunctions $(f_s)_{s\geq 1}$ of Σ_ε form an orthonormal basis of $\mathcal{H}$. For $s \geq 1$ and $h \in \mathcal{H}$, we define $h^s = \langle h, f_s\rangle_{\mathcal{H}}$, such that $h = \sum_{s\geq 1} h^s f_s$. For $i \in \{1,\dots,n\}$, let $Z_i = (\phi(Y_i)^1,\dots,\phi(Y_i)^T)$ in $\mathbb{R}^T$, $\beta_i = (\theta_i^1,\dots,\theta_i^T)$ in $\mathbb{R}^T$ and $\boldsymbol{\beta} = (\beta_1,\dots,\beta_p)' \in \mathcal{M}_{p,T}(\mathbb{R})$. We also define $\widetilde{\varepsilon}_i = (\varepsilon_i^1,\dots,\varepsilon_i^T)$ in $\mathbb{R}^T$. The projection of the i^{th} equation of the kernel linear model on $\mathrm{Span}(f_1,\dots,f_T) \subset \mathcal{H}$ is:

$$Z_i = \boldsymbol{\beta}' x_i + \widetilde{\varepsilon}_i.$$

We recognize a multivariate linear model in $\mathbb{R}^T$ that has the matrix form:

$$\mathbf{Z} = X\boldsymbol{\beta} + \widetilde{\boldsymbol{\varepsilon}},$$

where $\mathbf{Z} = (Z_1,\dots,Z_n) \in \mathcal{M}_{n,T}(\mathbb{R})$ and $\widetilde{\boldsymbol{\varepsilon}} = (\widetilde{\varepsilon}_1,\dots,\widetilde{\varepsilon}_n)' \in \mathcal{M}_{n,T}(\mathbb{R})$. Remark that the errors $\widetilde{\varepsilon}_1,\dots,\widetilde{\varepsilon}_n$ are i.i.d. with $\mathbb{E}\widetilde{\varepsilon}_1 = 0$ and covariance matrix $\Sigma_T = (\sigma_{s,s'})_{s,s'\in\{1,\dots,T\}} \in \mathcal{M}_T(\mathbb{R})$. For $s, s' \in \{1,\dots,T\}$, we have:

$$\begin{aligned}
\sigma_{s,s'} =& \mathbb{E}(\varepsilon_1^s \varepsilon_1^{s'}) \\
=& \mathbb{E}\,\langle \varepsilon_1, f_s\rangle_{\mathcal{H}} \langle \varepsilon_1, f_{s'}\rangle_{\mathcal{H}} \\
=& \langle f_s, \mathbb{E}(\varepsilon_1 \otimes \varepsilon_1) f_{s'}\rangle_{\mathcal{H}} \\
=& \langle f_s, \Sigma_\varepsilon f_{s'}\rangle_{\mathcal{H}} \\
=& \sum_{t\geq 1} \lambda_t \langle f_s, (f_t \otimes f_t) f_{s'}\rangle_{\mathcal{H}} \\
=& \sum_{t\geq 1} \lambda_t \langle f_s, f_t\rangle_{\mathcal{H}} \langle f_{s'}, f_t\rangle_{\mathcal{H}}. \\
=& \begin{cases} \lambda_s & \text{if } s = s' \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

Thus, we have $\Sigma_T = \text{diag}(\lambda_1, \ldots, \lambda_T)$.

The Hotelling-Lawley trace associated to hypotheses $\tilde{H}_0 : \mathbf{L}\boldsymbol{\beta} = 0_{\mathbb{R}^{\mathcal{L}}}$ versus $\tilde{H}_1 : \mathbf{L}\boldsymbol{\beta} \neq 0_{\mathbb{R}^{\mathcal{L}}}$ is equal to:

$$\mathcal{G}_n = \text{trace}\left(\frac{1}{n}\widehat{\boldsymbol{\Sigma}}_T^{-1}\widehat{H}_{\mathbf{L}T}\right),$$

where $\widehat{\boldsymbol{\Sigma}}_T = \text{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_T)$, $\widehat{H}_{\mathbf{L}T} = (\mathbf{L}\widehat{\boldsymbol{\beta}})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}})$, and $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)' \in \mathcal{M}_{p,T}(\mathbb{R})$ is the least square estimator of $\boldsymbol{\beta}$, defined such that:

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}.$$

We can develop $\mathbf{Z}$ in this equation to show that for $i \in \{1, \ldots, p\}$, $\widehat{\beta}_i = (\widehat{\theta}_i^1, \ldots, \widehat{\theta}_i^T)' \in \mathbb{R}^T$. According to Theorem 12.8 from [104], if Assumptions $\mathtt{B_1}$, $\mathtt{B_2}$ and $\mathtt{B_3}$ (implied by assumptions $\mathtt{A_1}$, $\mathtt{A_2}$, $\mathtt{A_3}$ and $\mathtt{A_4}$), we have that:

$$n\mathcal{G}_n \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2_{\mathcal{L}T}.$$

In particular, in the proof of the above result, they also show that:

$$n\tilde{\mathcal{G}}_n = \text{trace}(\boldsymbol{\Sigma}_T^{-1}\widehat{H}_{\mathbf{L}T}) \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2_{\mathcal{L}T}.$$

Now we show that $\tilde{\mathcal{F}}_T = \tilde{\mathcal{G}}_n$ to conclude the proof. We denote $\mathbf{L}'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}\mathbf{L} = (\alpha_{i,j})_{i,j\in\{1,\ldots,p\}} \in \mathcal{M}_p(\mathbb{R})$. Then, with a computation similar from Equation (3.5), we have:

$$\widehat{H}_{\mathbf{L}T} = \sum_{i,j=1}^{p} \alpha_{i,j}\widehat{\beta}_i\widehat{\beta}_j'.$$

Note that for $i, j \in \{1, \ldots, p\}$, we have:

$$\boldsymbol{\Sigma}_T^{-1}\widehat{\beta}_i\widehat{\beta}_j' = \begin{pmatrix} \lambda_1^{-1}\widehat{\theta}_i^1\widehat{\theta}_j^1 & \ldots & \lambda_1^{-1}\widehat{\theta}_i^1\widehat{\theta}_j^T \\ \vdots & \ddots & \vdots \\ \lambda_T^{-1}\widehat{\theta}_i^T\widehat{\theta}_j^1 & \ldots & \lambda_T^{-1}\widehat{\theta}_i^T\widehat{\theta}_j^T \end{pmatrix}$$

Then,

$$\begin{aligned} n\tilde{\mathcal{G}}_n &= \operatorname{trace}\left(\mathbf{\Sigma}_T^{-1}\widehat{H}_{\mathbf{L}T}\right) \\ &= \sum_{i,j=1}^{p} \alpha_{i,j} \sum_{t=1}^{T} \lambda_t^{-1} \widehat{\theta}_i^t \widehat{\theta}_j^t. \end{aligned}$$

On the other hand, we have:

$$\begin{aligned} n\tilde{\mathcal{F}}_T =& \operatorname{tr}\left(\Sigma_{\varepsilon,T}^{-1}\widehat{H}_{\mathbf{L}}\right) \\ =& \left\langle \Sigma_{\varepsilon,T}^{-1}, \widehat{H}_{\mathbf{L}} \right\rangle_{\mathrm{HS}(\mathcal{H})} \\ =& \sum_{i,j=1}^{p} \sum_{t=1}^{T} \alpha_{i,j} \lambda_t^{-1} \left\langle f_t \otimes f_t, \widehat{\theta}_i \otimes \widehat{\theta}_j \right\rangle_{\mathrm{HS}(\mathcal{H})} \\ =& \sum_{i,j=1}^{p} \sum_{t=1}^{T} \alpha_{i,j} \lambda_t^{-1} \left\langle f_t, \widehat{\theta}_i \right\rangle_{\mathcal{H}} \left\langle f_t, \widehat{\theta}_j \right\rangle_{\mathcal{H}} \\ =& \sum_{i,j=1}^{p} \sum_{t=1}^{T} \alpha_{i,j} \lambda_t^{-1} \widehat{\theta}_i^t \widehat{\theta}_j^t \\ =& n\tilde{\mathcal{G}}_n. \end{aligned}$$

Finally we conclude that:

$$n\tilde{\mathcal{F}}_T \xrightarrow[n\to\infty]{\mathcal{D}} \chi^2_{\mathcal{L}T}.$$

□

3.7.3 Results from Operator Perturbations Theory

In this section, we invoke some results from operator perturbations theory to obtain an exponential bound on $\left\|\widehat{\Sigma}_\varepsilon^{-1} - \Sigma_\varepsilon^{-1}\right\|_{\mathrm{HS}(\mathcal{H})}$, that is used in the proof of Theorem 11.

We want to apply Theorem 2 from Zwald and Blanchard [154].

Theorem 15 (Theorem 2 from [154])**.** *Let $A \in \mathrm{HS}(\mathcal{H})$ be a symmetric positive operator with eigenvalues $\lambda_1 > \lambda_2 > \dots$. For an integer r such that $\lambda_r > 0$, let $\delta_r = (\lambda_r - \lambda_{r+1})/2$ and $\widetilde{\delta}_r = \delta_r \wedge \delta_{r-1}$.*

Let $B \in \mathrm{HS}(\mathcal{H})$ a symmetric operator such that $A + B$ is a positive operator with simple non-zero eigenvalues and $\|B\|_{\mathrm{HS}(\mathcal{H})} \leq \widetilde{\delta}_r/2$.

Let $\Pi_r(A)$ denote the orthogonal projector onto the one-dimentionnal subspace of $\mathcal{H}$ spanned by the r^{th} eigenfunction of A. Then we have:

$$\|\Pi_r(A) - \Pi_r(A+B)\|_{\mathrm{HS}(\mathcal{H})} \leq \frac{2\|B\|_{\mathrm{HS}(\mathcal{H})}}{\tilde{\delta}_r}.$$

This theorem allows us to obtain the following result.

Lemma 11. *Assume that $\mathtt{A}_1$, $\mathtt{A}_2$ and $\mathtt{A}_3$ are verified and that $\lambda_1, \ldots, \lambda_T$ and $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_T$ are simple non-zero eigenvalues of Σ_ε (resp. $\widehat{\Sigma}_\varepsilon$). There exist $\tilde{\eta}_T$ such that for $n \geq \tilde{\eta}_T$, we have with probability $1 - e^{-\xi}$:*

$$\left\|\widehat{\Sigma}_\varepsilon^{-1} - \Sigma_\varepsilon^{-1}\right\|_{\mathrm{HS}(\mathcal{H})} \leq \frac{\zeta(n,\xi)}{\lambda_T - \zeta(n,\xi)} \left(\sum_{t=1}^{T} \frac{2}{\tilde{\delta}_t} + \frac{T}{\lambda_T}\right)$$

Proof. We can use the spectral decomposition of the inverse operators and apply the triangular inequality to obtain:

$$\begin{aligned}
\left\|\widehat{\Sigma}_\varepsilon^{-1} - \Sigma_\varepsilon^{-1}\right\|_{\mathrm{HS}(\mathcal{H})} &= \left\|\sum_{t=1}^{T} \widehat{\lambda}_t^{-1} \widehat{f}_t \otimes \widehat{f}_t - \lambda_t^{-1} f_t \otimes f_t\right\|_{\mathrm{HS}(\mathcal{H})} \\
&= \left\|\sum_{t=1}^{T} \widehat{\lambda}_t^{-1} \Pi_{\widehat{f}_t} - \lambda_t^{-1} \Pi_{f_t}\right\|_{\mathrm{HS}(\mathcal{H})} \\
&= \left\|\sum_{t=1}^{T} \widehat{\lambda}_t^{-1} (\Pi_{\widehat{f}_t} - \Pi_{f_t}) + (\widehat{\lambda}_t^{-1} - \lambda_t^{-1}) \Pi_{f_t}\right\|_{\mathrm{HS}(\mathcal{H})} \\
&\leq \sum_{t=1}^{T} \widehat{\lambda}_t^{-1} \left\|\Pi_{\widehat{f}_t} - \Pi_{f_t}\right\|_{\mathrm{HS}(\mathcal{H})} + \sum_{t=1}^{T} |\widehat{\lambda}_t^{-1} - \lambda_t^{-1}| \, \|\Pi_{f_t}\|_{\mathrm{HS}(\mathcal{H})}
\end{aligned}$$

According to Lemma 13, we have $|\widehat{\lambda}_t - \lambda_t| \leq \zeta(n,\xi)$. Thus $(\lambda_T - \zeta(n,\xi)) \leq \widehat{\lambda}_T \leq (\lambda_T + \zeta(n,\xi))$ with probability $1 - e^{-\xi}$. As $\zeta(n,\xi) \underset{n\to\infty}{\longrightarrow} 0$, there exists an integer $N(\xi, \lambda_T) \geq 1$ such that for $n \geq N(\xi, \lambda_T)$, we have $\lambda_T - \zeta(n,\xi) > 0$. For such n, we have with probability $1 - e^{-\xi}$:

$$\frac{1}{\widehat{\lambda}_T} \leq \frac{1}{\lambda_T - \zeta(n,\xi)}.$$

Then, note that for $t \in \{1, \dots, T\}$ we have:

$$\begin{aligned}|\widehat{\lambda}_t^{-1} - \lambda_t^{-1}| =& \widehat{\lambda}_t^{-1}\lambda_t^{-1}|\widehat{\lambda}_t - \lambda_t| \\ \leq& \lambda_T^{-1}\widehat{\lambda}_T^{-1}\zeta(n,\xi), \\ \leq& \frac{1}{\lambda_T}\frac{\zeta(n,\xi)}{\lambda_T - \zeta(n,\xi)}.\end{aligned}$$

As we have:

$$\|\Pi_{f_t}\|_{\mathrm{HS}(\mathcal{H})} = \|f_t\|_{\mathcal{H}}^2 = 1,$$

we obtain that:

$$\sum_{t=1}^{T} |\widehat{\lambda}_t^{-1} - \lambda_t^{-1}| \, \|\Pi_{f_t}\|_{\mathrm{HS}(\mathcal{H})} \leq \frac{T}{\lambda_T}\frac{\zeta(n,\xi)}{\lambda_T - \zeta(n,\xi)}$$

By applying Lemma 12, for $n \geq \eta_T$ and with probability $1 - e^{-\xi}$, we have:

$$\begin{aligned}\sum_{t=1}^{T} \widehat{\lambda}_t^{-1} \left\|\Pi_{\widehat{f}_t} - \Pi_{f_t}\right\|_{\mathrm{HS}(\mathcal{H})} \leq& \frac{\zeta(n,\xi)}{\widehat{\lambda}_T}\sum_{t=1}^{T}\frac{2}{\tilde{\delta}_t} \\ \leq& \frac{\zeta(n,\xi)}{\lambda_T - \zeta(n,\xi)}\sum_{t=1}^{T}\frac{2}{\tilde{\delta}_t}\end{aligned} \tag{3.15}$$

Thus, for $n \geq N(\xi, \lambda_T) \vee \eta_T$, we have with probability $1 - e^{-\xi}$:

$$\left\|\widehat{\Sigma}_\varepsilon^{-1} - \Sigma_\varepsilon^{-1}\right\|_{\mathrm{HS}(\mathcal{H})} \leq \frac{\zeta(n,\xi)}{\lambda_T - \zeta(n,\xi)}\left(\sum_{t=1}^{T}\frac{2}{\tilde{\delta}_t} + \frac{T}{\lambda_T}\right)$$

□

Lemma 12. *Assume that* $\mathtt{A}_1$, $\mathtt{A}_2$ *and* $\mathtt{A}_3$ *are verified and that* $\lambda_1, \dots, \lambda_T$ *and* $\widehat{\lambda}_1, \dots, \widehat{\lambda}_T$ *are simple non-zero eigenvalues of* Σ_ε *(resp.* $\widehat{\Sigma}_\varepsilon$*). Let* $t \in \{1, \dots, T\}$*, denote* $\delta_t = (\lambda_t - \lambda_{t+1})/2$*,* $\tilde{\delta}_t = \min(\delta_t, \delta_{t-1})$*. Let* $\Pi_{f_t} = f_t \otimes f_t$ *and* $\Pi_{\widehat{f}_t} = \widehat{f}_t \otimes \widehat{f}_t$ *the orthogonal projector on the axis spanned by* f_t *(resp.* $\widehat{f}_t$*). Then there exists* $\eta_t > 0$ *such that for* $n \geq \eta_t$*, we have with probability* $1 - e^{-\xi}$*:*

$$\left\|\Pi_{f_t} - \Pi_{\widehat{f}_t}\right\|_{\mathrm{HS}(\mathcal{H})} \leq \frac{2\zeta(n,\xi)}{\tilde{\delta}_t}.$$

Proof. Let $t \in \{1, \dots, T\}$. As $\zeta(n, \xi) \underset{n\to\infty}{\longrightarrow} 0$, according to Proposition 3, there exists η_t such that for $n \geq \eta_t$, we have with probability $1 - e^{-\xi}$:

$$\left\|\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})} \leq \frac{\tilde{\delta}_T}{2}.$$

Then, every hypothesis of Theorem 15 is verified, and for $n \geq \eta_t$, we have with probability $1 - e^{-\xi}$:

$$\left\|\Pi_{f_t} - \Pi_{\widehat{f}_t}\right\|_{\mathrm{HS}(\mathcal{H})} \leq \frac{2\left\|\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}}{\tilde{\delta}_t} \leq \frac{2\zeta(n, \xi)}{\tilde{\delta}_t}$$

□

Lemma 13. *If Assumptions* $\mathtt{A_1}$, $\mathtt{A_2}$ *and* $\mathtt{A_3}$ *are verified, for* $t \in \{1, \dots, T\}$, *we have with probability* $1 - e^{-\xi}$*:*

$$|\widehat{\lambda}_t - \lambda_t| \leq \zeta(n, \xi).$$

This lemma is an intermediate result in the proof of Theorem 3 from [154].

Proof. The result is a direct application of the Hoffman-Wielandt inequality in infinite dimensional setting [20] and Theorem 3. Let $t \in \{1, \dots, T\}$, according to the Hoffman-Wielandt inequality, we have:

$$|\widehat{\lambda}_t - \lambda_t| \leq \left\|\widehat{\Sigma}_\varepsilon - \Sigma_\varepsilon\right\|_{\mathrm{HS}(\mathcal{H})}.$$

We conclude the proof by applying Theorem 3. □

3.7.4 Proofs for the Kernel Trick

Proof of Theorem 12. We define the matrix:

$$\mathbf{K}_\varepsilon = \frac{1}{n}\begin{pmatrix} \langle \widehat{\varepsilon}_1, \widehat{\varepsilon}_1 \rangle_{\mathcal{H}} & \dots & \langle \widehat{\varepsilon}_1, \widehat{\varepsilon}_n \rangle_{\mathcal{H}} \\ \vdots & \ddots & \vdots \\ \langle \widehat{\varepsilon}_n, \widehat{\varepsilon}_1 \rangle_{\mathcal{H}} & \dots & \langle \widehat{\varepsilon}_n, \widehat{\varepsilon}_n \rangle_{\mathcal{H}} \end{pmatrix}$$

Let $i, j \in \{1, \ldots, n\}$, remind that $\widehat{\varepsilon}_i = \Phi(\mathbf{Y}) \odot \pi_i^{\perp}$, where $\pi_i^{\perp}$ is the i^{th} column of $\mathbf{\Pi}^{\perp} = \mathbf{I}_n - \mathbf{\Pi}$ we have that:

$$\begin{aligned}\langle \widehat{\varepsilon}_i, \widehat{\varepsilon}_j \rangle_{\mathcal{H}} =& \left\langle \Phi(\mathbf{Y}) \odot \pi_i^{\perp}, \Phi(\mathbf{Y}) \odot \pi_j^{\perp} \right\rangle_{\mathcal{H}} \\ =& \sum_{k,l=1}^{n} \pi_{i,k}^{\perp} \pi_{j,l}^{\perp} \langle \phi(Y_k), \phi(Y_l) \rangle_{\mathcal{H}} \\ =& \pi_i^{\perp\prime} \mathbf{K}_{\mathbf{Y}} \pi_j^{\perp}.\end{aligned}$$

Thus we have:

$$\mathbf{K}_{\varepsilon} = \frac{1}{n} \mathbf{\Pi}^{\perp} \mathbf{K}_{\mathbf{Y}} \mathbf{\Pi}^{\perp}.$$

Let $\widehat{f}$ be an eigenfunction of $\widehat{\Sigma}_{\varepsilon}$ associated to the non-zero eigenvalue $\widehat{\lambda} \in \mathbb{R}^+$. We have the eigen relation:

$$\widehat{\Sigma}_{\varepsilon} \widehat{f} = \widehat{\lambda} \widehat{f}$$

Let $j \in \{1, \ldots, n\}$, we compute the inner product of both side of the eigen equation with $\widehat{\varepsilon}_j$:

$$\begin{array}{rcl} \frac{1}{n} \sum_{i=1}^{n} \left\langle \widehat{\varepsilon}_i \otimes \widehat{\varepsilon}_i \widehat{f}, \widehat{\varepsilon}_j \right\rangle_{\mathcal{H}} & = & \widehat{\lambda} \left\langle \widehat{f}, \widehat{\varepsilon}_j \right\rangle_{\mathcal{H}} \\ \iff \frac{1}{n} \sum_{i=1}^{n} \langle \widehat{\varepsilon}_i, \widehat{\varepsilon}_j \rangle_{\mathcal{H}} \left\langle \widehat{\varepsilon}_i, \widehat{f} \right\rangle_{\mathcal{H}} & = & \widehat{\lambda} \left\langle \widehat{f}, \widehat{\varepsilon}_j \right\rangle_{\mathcal{H}} \end{array}$$

Let $u = \left(\left\langle \widehat{f}, \widehat{\varepsilon}_1 \right\rangle_{\mathcal{H}}, \ldots, \left\langle \widehat{f}, \widehat{\varepsilon}_n \right\rangle_{\mathcal{H}} \right)$ in $\mathbb{R}^n$, we have $u_j = \left\langle \widehat{f}, \widehat{\varepsilon}_j \right\rangle_{\mathcal{H}}$. We remark that:

$$(\mathbf{K}_{\varepsilon} u)_j = \frac{1}{n} \sum_{i=1}^{n} \langle \widehat{\varepsilon}_i, \widehat{\varepsilon}_j \rangle_{\mathcal{H}} \left\langle \widehat{\varepsilon}_i, \widehat{f} \right\rangle_{\mathcal{H}}.$$

Thus, we have:

$$\mathbf{K}_{\varepsilon} u = \widehat{\lambda} u.$$

It proves that each eigenvalue of $\mathbf{K}_{\varepsilon}$ is an eigenvalue of $\widehat{\Sigma}_{\varepsilon}$.

On the other hand, let $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$ be a unit eigenvector of $\mathbf{K}_{\varepsilon}$ associated

to the non-zero eigenvalue $\nu \in \mathbb{R}^+$, we know that:

$$\mathbf{K}_\varepsilon u = \nu u,$$

and we deduce from this equality that for $i \in \{1, \dots, n\}$, we have:

$$\begin{aligned} u_i =& \frac{1}{n\nu} \sum_{j=1}^{n} \langle \widehat{\varepsilon}_i, \widehat{\varepsilon}_j \rangle_{\mathcal{H}} u_j \\ =& \frac{1}{n\nu} \langle \widehat{\varepsilon}_i, \widehat{\varepsilon} \odot u \rangle_{\mathcal{H}} . \end{aligned}$$

Then we observe that:

$$\begin{aligned} \widehat{\varepsilon} \odot u =& \sum_{i=1}^{n} \frac{1}{n\nu} \langle \widehat{\varepsilon}_i, \widehat{\varepsilon} \odot u \rangle_{\mathcal{H}} \widehat{\varepsilon}_i \\ =& \frac{1}{n\nu} \sum_{i=1}^{n} \widehat{\varepsilon}_i \otimes \widehat{\varepsilon}_i (\widehat{\varepsilon} \odot u) \\ \widehat{\varepsilon} \odot u =& \frac{1}{\nu} \widehat{\Sigma}_\varepsilon (\widehat{\varepsilon} \odot u). \end{aligned}$$

This equality shows that each eigenvalue ν of $\mathbf{K}_\varepsilon$ is an eigenvalue of $\widehat{\Sigma}_\varepsilon$. We can conclude that $\widehat{\Sigma}_\varepsilon$ and $\mathbf{K}_\varepsilon$ have the same spectra.

Now we consider $\widehat{\varepsilon} \odot u \in \mathcal{H}$, an eigenfunction of $\widehat{\Sigma}_\varepsilon$ associated to the eigenvalue $\widehat{\lambda}$. Then a unit eigenfunction $\widehat{f}$ of $\widehat{\Sigma}_\varepsilon$ is such that:

$$\widehat{f} = \frac{1}{\|\widehat{\varepsilon} \odot u\|_{\mathcal{H}}} \widehat{\varepsilon} \odot u,$$

where:

$$\begin{aligned} \|\widehat{\varepsilon} \odot u\|_{\mathcal{H}}^2 =& \left\langle \sum_{i=1}^{n} u_i \widehat{\varepsilon}_i, \sum_{j=1}^{n} u_j \widehat{\varepsilon}_j \right\rangle_{\mathcal{H}} \\ =& \sum_{i=1}^{n} u_i \left[\sum_{j=1}^{n} \langle \widehat{\varepsilon}_i, \widehat{\varepsilon}_j \rangle_{\mathcal{H}} u_j \right] \\ =& \sum_{i=1}^{n} u_i [n\widehat{\lambda} u_i] \\ =& n\widehat{\lambda} \sum_{i=1}^{n} u_i^2 \\ =& n\widehat{\lambda} \end{aligned}$$

□

Proof of Theorem 14. Note that:

$$\begin{aligned}\psi =& \widetilde{\mathbf{U}}_T' \widetilde{\mathbf{U}}_T \\ =& \sum_{t=1}^{T} \tilde{u}_t \tilde{u}_t' \\ =& \left(\sum_{t=1}^{T} \tilde{u}_{t,i} \tilde{u}_{t,j} \right)_{i,j \in \{1,\dots,n\}}\end{aligned}$$

Thus we have:

$$\begin{aligned}\Psi \mathbf{K_Y} &= \left(\sum_{t=1}^{T} \sum_{j=1}^{n} \tilde{u}_{t,i} \tilde{u}_{t,j} \left\langle \phi(Y_j), \phi(Y_k) \right\rangle_{\mathcal{H}} \right)_{i,k \in \{1,\dots,n\}} \\ &= \left(\sum_{t=1}^{T} \tilde{u}_{t,i} \left\langle \sum_{j=1}^{n} \tilde{u}_{t,j} \phi(Y_j), \phi(Y_k) \right\rangle_{\mathcal{H}} \right)_{i,k \in \{1,\dots,n\}} \\ &= \left(\sum_{t=1}^{T} \tilde{u}_{t,i} \left\langle (\Phi(\mathbf{Y}) \odot \tilde{u}_t), \phi(Y_k) \right\rangle_{\mathcal{H}} \right)_{i,k \in \{1,\dots,n\}} \\ &= \left(\sum_{t=1}^{T} \widehat{\lambda}_t^{-\frac{1}{2}} \tilde{u}_{t,i} \left\langle \widehat{f}_t, \phi(Y_k) \right\rangle_{\mathcal{H}} \right)_{i,k \in \{1,\dots,n\}}.\end{aligned}$$

We also have:

$$\mathbf{K_Y} \Psi \mathbf{K_Y} = \left(\sum_{t=1}^{T} \widehat{\lambda}_t^{-1} \left\langle \widehat{f}_t, \phi(Y_i) \right\rangle_{\mathcal{H}} \left\langle \widehat{f}_t, \phi(Y_j) \right\rangle_{\mathcal{H}} \right)_{i,j \in \{1,\dots,n\}} \in \mathcal{M}_n(\mathbb{R}).$$

If we denote $R = (r_{i,j})_{i,j \in \{1,\dots,n\}}$, we have:

$$\mathbf{K_Y} \Psi \mathbf{K_Y} R = \left(\sum_{t=1}^{T} \sum_{j=1}^{n} r_{j,k} \widehat{\lambda}_t^{-1} \left\langle \widehat{f}_t, \phi(Y_i) \right\rangle_{\mathcal{H}} \left\langle \widehat{f}_t, \phi(Y_j) \right\rangle_{\mathcal{H}} \right)_{i,k \in \{1,\dots,n\}} \in \mathcal{M}_n(\mathbb{R}).$$

If v is an eigenvector of $\mathbf{K}_{\varepsilon,L,T}$ associated to the eigenvalue ξ, we have:

$$\mathbf{K_Y} \Psi \mathbf{K_Y} R v = \xi v.$$

Thus, for $i \in \{1, \dots, n\}$, we have the relation:

$$\sum_{t=1}^{T} \sum_{j,k=1}^{n} r_{j,k} \widehat{\lambda}_t^{-1} \left\langle \widehat{f}_t, \phi(Y_i) \right\rangle_{\mathcal{H}} \left\langle \widehat{f}_t, \phi(Y_j) \right\rangle_{\mathcal{H}} v_k = \xi v_i$$

Now we show that $\Phi(\mathbf{Y}) \odot \Psi \mathbf{K}_{\mathbf{Y}} R v$ is an eigenfunction of $\widehat{\Sigma}_{\varepsilon}^{-1} \widehat{H}_{\mathbf{L}}$. Note that:

$$\Psi \mathbf{K}_{\mathbf{Y}} R v = \begin{pmatrix} \sum_{t=1}^{T} \sum_{j=1}^{n} \sum_{k=1}^{n} \widehat{\lambda}_t^{-\frac{1}{2}} \tilde{u}_{t,1} \left\langle \widehat{f}_t, \phi(Y_j) \right\rangle_{\mathcal{H}} r_{j,k} v_k \\ \vdots \\ \sum_{t=1}^{T} \sum_{j=1}^{n} \sum_{k=1}^{n} \widehat{\lambda}_t^{-\frac{1}{2}} \tilde{u}_{t,n} \left\langle \widehat{f}_t, \phi(Y_j) \right\rangle_{\mathcal{H}} r_{j,k} v_k \end{pmatrix}.$$

Thus,

$$\begin{aligned} \Phi(\mathbf{Y}) \odot \Psi \mathbf{K}_{\mathbf{Y}} R v =& \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{j=1}^{n} \sum_{k=1}^{n} \widehat{\lambda}_t^{-\frac{1}{2}} \tilde{u}_{t,i} \left\langle \widehat{f}_t, \phi(Y_j) \right\rangle_{\mathcal{H}} r_{j,k} v_k \phi(Y_i) \\ =& \sum_{t=1}^{T} \sum_{j=1}^{n} \sum_{k=1}^{n} \widehat{\lambda}_t^{-\frac{1}{2}} \left\langle \widehat{f}_t, \phi(Y_j) \right\rangle_{\mathcal{H}} r_{j,k} v_k \sum_{i=1}^{n} \tilde{u}_{t,i} \phi(Y_i) \\ =& \sum_{t=1}^{T} \sum_{j,k=1}^{n} \frac{\left\langle \widehat{f}_t, \phi(Y_j) \right\rangle_{\mathcal{H}} r_{j,k}}{\widehat{\lambda}_t} v_k \widehat{f}_t. \end{aligned}$$

Remark that for $h \in \mathcal{H}$, $t \in \{1, \dots, T\}$ and $i, j \in \{1, \dots, n\}$, we have:

$$\widehat{f}_t \otimes \widehat{f}_t \phi(Y_i) \otimes \phi(Y_j) h = \langle h, \phi(Y_i) \rangle_{\mathcal{H}} \left\langle \widehat{f}_t, \phi(Y_j) \right\rangle_{\mathcal{H}} \widehat{f}_t.$$

Then,

$$\begin{aligned} \widehat{\Sigma}_{\varepsilon}^{-1} \widehat{H}_{\mathbf{L}} (\Phi(\mathbf{Y}) \odot \Psi \mathbf{K}_{\mathbf{Y}} R v) =& \sum_{t=1}^{T} \sum_{i,j=1}^{n} \frac{r_{i,j}}{\widehat{\lambda}_t} \widehat{f}_t \otimes \widehat{f}_t \phi(Y_i) \otimes \phi(Y_j) \sum_{s=1}^{T} \sum_{k,\ell=1}^{n} \frac{\left\langle \widehat{f}_s, \phi(Y_k) \right\rangle_{\mathcal{H}} r_{k,\ell}}{\widehat{\lambda}_s} v_\ell \widehat{f}_s \\ =& \sum_{t=1}^{T} \sum_{i,j=1}^{n} \frac{r_{i,j}}{\widehat{\lambda}_t} \left\langle \widehat{f}_t, \phi(Y_j) \right\rangle_{\mathcal{H}} \underbrace{\sum_{s=1}^{T} \sum_{k,\ell=1}^{n} \frac{\left\langle \widehat{f}_s, \phi(Y_k) \right\rangle_{\mathcal{H}} r_{k,\ell}}{\widehat{\lambda}_s} \left\langle \widehat{f}_s, \phi(Y_i) \right\rangle_{\mathcal{H}} v_\ell}_{=\xi v_i} \widehat{f}_t \\ =& \sum_{t=1}^{T} \sum_{i,j=1}^{n} \widehat{\lambda}_t^{-1} r_{i,j} \left\langle \widehat{f}_t, \phi(Y_j) \right\rangle_{\mathcal{H}} \xi v_i \widehat{f}_t \\ =& \xi (\Phi(\mathbf{Y}) \odot \Psi \mathbf{K}_{\mathbf{Y}} R v) \end{aligned}$$

Thus, $\Phi(\mathbf{Y}) \odot \Psi \mathbf{K_Y} R v$ is an eigenvector of $\widehat{\Sigma}_\varepsilon^{-1} \widehat{H}_\mathbf{L}$ associated to the eigenvalue ξ. We can conclude that every eigenvalue of $\mathbf{K}_{\varepsilon,L,T}$ is an eigenvalue of $\widehat{\Sigma}_\varepsilon^{-1} \widehat{H}_\mathbf{L}$. Now we consider $\widehat{g}$, an eigenvector of $\widehat{\Sigma}_\varepsilon^{-1} \widehat{H}_\mathbf{L}$ associated to the eigenvalue ξ. We have:

$$\widehat{\Sigma}_\varepsilon^{-1} \widehat{H}_\mathbf{L} \widehat{g} = \xi \widehat{g}.$$

We also have that:

$$\widehat{\Sigma}_\varepsilon^{-1} \widehat{H}_\mathbf{L} \widehat{g} = \sum_{t=1}^{T} \sum_{i,j=1}^{n} \widehat{\lambda}_t^{-1} r_{i,j} \left\langle \widehat{f}_t, \phi(Y_i) \right\rangle_\mathcal{H} \langle \phi(Y_j), \widehat{g} \rangle_\mathcal{H} \widehat{f}_t$$

We show that the vector $v = (\langle \phi(Y_1), \widehat{g} \rangle_\mathcal{H}, \ldots, \langle \phi(Y_n), \widehat{g} \rangle_\mathcal{H})$ in $\mathbb{R}^n$ is an eigenvector of $\mathbf{K}_{\varepsilon,L,T}$ associated to the eigenvalue ξ:

$$\begin{aligned}
\mathbf{K}_{\varepsilon,L,T} v =& \mathbf{K_Y} \Psi \mathbf{K_Y} R v \\
=& \left(\sum_{t=1}^{T} \sum_{j,k=1}^{n} r_{j,k} \widehat{\lambda}_t^{-1} \left\langle \widehat{f}_t, \phi(Y_i) \right\rangle_\mathcal{H} \left\langle \widehat{f}_t, \phi(Y_j) \right\rangle_\mathcal{H} \langle \phi(Y_i), \widehat{g} \rangle_\mathcal{H} \right)_{i \in \{1,\ldots,n\}} \\
=& \left(\left\langle \underbrace{\sum_{t=1}^{T} \sum_{i,j=1}^{n} \widehat{\lambda}_t^{-1} r_{i,j} \left\langle \widehat{f}_t, \phi(Y_i) \right\rangle_\mathcal{H} \langle \phi(Y_j), \widehat{g} \rangle_\mathcal{H} \widehat{f}_t}_{= \xi \widehat{g}}, \phi(Y_i) \right\rangle_\mathcal{H} \right)_{i \in \{1,\ldots,n\}} \\
=& \xi v.
\end{aligned}$$

We conclude that every eigenvalue of $\widehat{\Sigma}_\varepsilon^{-1} \widehat{H}_\mathbf{L}$ is an eigenvalue of $\mathbf{K}_{\varepsilon,L,T}$. As a result, the operator $\widehat{\Sigma}_\varepsilon^{-1} \widehat{H}_\mathbf{L}$ and the matrix $\mathbf{K}_{\varepsilon,L,T}$ share the same spectrum. Let v be a unit eigenvector of $\mathbf{K}_{\varepsilon,L,T}$, we know that $\Phi(\mathbf{Y}) \odot \Psi \mathbf{K_Y} R v \in \mathcal{H}$ is an eigenfunction of $\widehat{\Sigma}_\varepsilon^{-1} \widehat{H}_\mathbf{L}$ and we have:

$$\begin{aligned}
\| \Phi(\mathbf{Y}) \odot \Psi \mathbf{K_Y} R v \|_\mathcal{H}^2 =& v' R \mathbf{K_Y} \Psi (\mathbf{K_Y} \Psi \mathbf{K_Y} R v) \\
=& \xi v' R \mathbf{K_Y} \Psi v.
\end{aligned}$$

Thus, $\widehat{g} = \frac{1}{\sqrt{\xi v' R \mathbf{K_Y} \Psi v}} \Phi(\mathbf{Y}) \odot \Psi \mathbf{K_Y} R v$ is a unit eigenfunction of $\widehat{\Sigma}_\varepsilon^{-1} \widehat{H}_\mathbf{L}$. □

3.7.5 Results on Orthogonal Projectors

Lemma 14. *Let $\mathbf{\Pi} = (\pi_{i,j})_{i,j\in\{1,\dots,n\}} \in \mathcal{M}_n(\mathbb{R})$ be the matrix of an orthogonal projector of rank p, then we have:*

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{n} {\pi_{i,j}}^2 = p,\\
&\sum_{i=1}^{n} {\pi_{i,i}}^2 \leq p,\\
&\sum_{i=1}^{n}(1-\pi_{i,i})^2 \leq n-p.
\end{aligned}$$

Proof. As $\mathbf{\Pi}^2 = \mathbf{\Pi}$, we directly have $\sum_{j=1}^{n} {\pi_{i,j}}^2 = \pi_{i,i}$ and by computing the trace we find that:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} {\pi_{i,j}}^2 = p.$$

Then:

$$\begin{aligned}
\sum_{i=1}^{n} {\pi_{i,i}}^2 \leq& \sum_{i=1}^{n}\sum_{j=1}^{n} {\pi_{i,j}}^2\\
\leq& p,\\
\sum_{i=1}^{n}(1-\pi_{i,i})^2 =& \sum_{i=1}^{n}(1-2\pi_{i,i}+{\pi_{i,i}}^2)\\
\leq& n-p.
\end{aligned}$$

□

Lemma 15. *If $\mathbf{\Pi} = (\pi_{i,j})_{i,j\in\{1,\dots,n\}} \in \mathcal{M}_n(\mathbb{R})$ is the matrix of an orthogonal projector and $f = (f_1,\dots,f_n) \in \mathcal{H}^n$, then we have that:*

$$\|\mathbf{\Pi} f\|_{\mathcal{H}^n} \leq \|f\|_{\mathcal{H}^n}.$$

Proof. Let $(e_i)_{i\geq 1}$ an orthonormal basis of $\mathcal{H}$. For $s \geq 1$, let $f^s = (f_1^s,\dots,f_n^s)$ in $\mathbb{R}^n$ where for $i \in \{1,\dots,n\}$, $f_i^s = \langle f_i, e_s\rangle_{\mathcal{H}}$. Note that $f_i = \sum_{s\geq 1} f_i^s e_s$, $\|f_i\|_{\mathcal{H}}^2 = \sum_{s\geq 1} {f_i^s}^2$ and

$\|f^s\|^2 = \sum_{j=1}^n f_j^{s2}$. Then we have:

$$\begin{aligned}\|f\|_{\mathcal{H}^n}^2 &= \sum_{i=1}^n \sum_{s\geq 1} f_i^{s2} \\ &= \sum_{s\geq 1} \sum_{i=1}^n f_i^{s2} \\ &= \sum_{s\geq 1} \|f^s\|^2 .\end{aligned}$$

Considering $\mathbf{\Pi}$ as a projector from $\mathbb{R}^n$ to $\mathbb{R}^n$, we have $(\mathbf{\Pi} f^s)_i = \sum_{j=1}^n \pi_{i,j} f_j^s$ and $\|\mathbf{\Pi} f^s\|^2 = \sum_{i=1}^n (\sum_{j=1}^n \pi_{i,j} f_j^s)^2$. By property of an orthogonal projector, we have $\|\mathbf{\Pi} f^s\| \leq \|f^s\|$. Now we compute $\|\mathbf{\Pi} f\|_{\mathcal{H}}$:

$$\begin{aligned}\|\mathbf{\Pi} f\|_{\mathcal{H}} &= \sum_{i=1}^n \left\| \sum_{j=1}^n \pi_{i,j} f_j \right\|_{\mathcal{H}}^2 \\ &= \sum_{i=1}^n \left\| \sum_{s\geq 1} \sum_{j=1}^n \pi_{i,j} f_j^s e_s \right\|_{\mathcal{H}}^2 \\ &= \sum_{i=1}^n \sum_{s\geq 1} \left(\sum_{j=1}^n \pi_{i,j} f_j^s \right)^2 \\ &= \sum_{s\geq 1} \|\mathbf{\Pi} f^s\|^2 \\ &\leq \sum_{s\geq 1} \|f^s\|^2 \\ &\leq \|f\|_{\mathcal{H}^n}^2 .\end{aligned}$$

□

Lemma 16. *Let $\mathbf{\Pi} = (\pi_{i,j})_{i,j\in\{1,\dots,n\}} \in \mathcal{M}_n(\mathbb{R})$ be the matrix of an orthogonal projector of rank p, and let $v_1, \dots, v_p \in \mathbb{R}^n$ be an orthonormal basis of $\mathrm{Im}(\mathbf{\Pi}) \subset \mathbb{R}^n$. Then, for $i \in \{1, \dots, n\}$, the column $\pi_i = (\pi_{i,j})_{j\in\{1,\dots,n\}}$ is such that:*

$$\pi_i = \sum_{s=1}^p v_{s,i} v_s,$$

and for $i, j \in \{1, \dots, n\}$, we have:

$$\pi_{i,j} = \sum_{s=1}^{p} v_{s,i} v_{s,j}.$$

Proof. The results are direct from Equation (3.9). □

3.7.6 McDiarmid Inequality

Theorem 16 (McDiarmid inequality [94])**.** *If $Y_1, \dots, Y_n$ are n i.i.d. random variables in a measurable space $\mathcal{Y}$ and the function*

$$\begin{aligned} f : \quad \mathcal{Y}^n \quad &\longrightarrow \mathbb{R} \\ y_1, \dots, y_n \quad &\longmapsto f(y_1, \dots, y_n) \end{aligned}$$

is such that for all $i \in \{1, \dots, n\}$, we have:

$$\sup_{y_1, \dots, y_n, y_{i'} \in \mathcal{Y}} \Big| f(y_1, \dots, y_i, \dots, y_n) - f(y_1, \dots, y_{i'}, \dots, y_n) \Big| \leq c_i,$$

then we have with probability lower than $e^{-\xi}$ that:

$$f(y_1, \dots, y_n) - \mathbb{E}\Big(f(y_1, \dots, y_n)\Big) \geq \sqrt{\frac{\xi}{2} \sum_{i=1}^{n} c_i^2},$$

and we also have with probability lower than $e^{-\xi}$ that:

$$\mathbb{E}\Big(f(y_1, \dots, y_n)\Big) - f(y_1, \dots, y_n) \geq \sqrt{\frac{\xi}{2} \sum_{i=1}^{n} c_i^2}.$$

Chapter 4

Influence Functions for the KFDA Statistic

Identifying the sub-population driving a significant difference between several conditions is a recurrent task of single-cell data analysis but there is no dedicated tool yet. In Chapter 2, we suggested that our visualization tool based on the Kernel Fisher Discriminant Axis could allow to identify a sub-population of interest that supports a detected global difference. While efficient in practice, this approach is only qualitative, and the definition of such a detected sub-population yet remains subjective. A possible quantitative measure of the participation of a cell or a group of cell to the rejection of the null hypothesis could be handled by robust statistics. The field of robust statistics aims to avoid misleading results due to overly influential outlier observations when analysing a dataset. To this aim, influence functions are defined to quantify the influence of an observation on a statistic, and Gateaux derivatives are generalized influence functions to quantify the influence of a group of observations [61]. Thus, Gateaux derivatives and influence functions have the potential to quantify the influence of a cell or of a group of cells on the outcome of the test. In general, Gateaux derivatives are used as a way to solve an optimization problem by identifying the zeros of the Gateaux derivative of the statistical functional associated to the loss [2, 76]. Influence functions may be defined as a particular case of Gateaux derivative and are often used to assess the robustness of a statistic [34, 67] and to detect influential outliers, that has strong links with influences measured through Cook distances [30]. In the context of single-cell data analysis, isolated cells are often considered as outliers, and sub-populations may be considered as biologically significant, thus we are also interested in the influence of groups of observations and consider Gateaux derivatives as a practical tool, which is not usual. Gateaux derivatives and influence functions have been applied to various kernel methods, and to test procedures [48], but not to kernel tests yet. Beyond single-cell sub-population detection, studying the robustness of kernel tests in general is as such a question of interest. This chapter presents a work initiated

on Gateaux derivatives for kernel testing that still needs to be pursued.

In this chapter, we introduce Gateaux derivatives and influence functions (Section 4.1). In particular, we introduce partial Gateaux derivatives and partial influence functions to study statistics that are functions of two probability distributions. The second section deals with the application of these concepts on kernel tests (Section 4.2).

4.1 Gateaux Derivatives and Influence Functions

In this section, we introduce Gateaux derivatives and influence functions in Hilbert spaces. This introduction relies on the presentation proposed in [2], where they derive the Gateaux derivative and influence function of the kernel canonical correlation analysis. We consider a general measurable Hilbert space $(\mathcal{Y}, \mathfrak{Y})$ endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{Y}}$. Let $\mathcal{P}(\mathcal{Y})$ be the set of all the probability distributions on $\mathcal{Y}$ and $\widetilde{\mathcal{P}}(\mathcal{Y})$ be a convex subset of $\mathcal{P}(\mathcal{Y})$. For $i \geq 1$, we also denote $\widetilde{\mathcal{P}}_i(\mathcal{Y})$ the subspace of $\widetilde{\mathcal{P}}(\mathcal{Y})$ containing all the probability distributions having a finite i^{th}-order moment.

4.1.1 Introduction to Gateaux Derivatives and Influence Functions

Statistical Functionals

Gateaux derivatives and influence functions are defined as the generalization of the concept of derivation to statistics considered as functions of probability distributions. These functions taking probability distributions as inputs are called statistical functionals because they output statistics when applied to empirical probability distributions.

Definition 8 (Statistical Functional)**.** *A statistical functional is a function $T(\circledcirc)$ from $\mathcal{P}(\mathcal{Y})$ to any convenient space $\mathfrak{F}$.*

The notation $\circledcirc$ is used to highlight that the arguments of statistical functionals are probability distributions. For instance, the statistical functional of the mean $m(\circledcirc)$ is such that for $\mathbb{P} \in \widetilde{\mathcal{P}}_1(\mathcal{Y})$, we have:

$$m(\mathbb{P}) = \int_{\mathcal{Y}} y \, \mathrm{d}\, \mathbb{P}(y) = \mathbb{E}_{\mathbb{P}}(\mathrm{d}\, y).$$

Similarly, the statistical functional of the covariance $S(\circledcirc)$ is such that for $\mathbb{P} \in \widetilde{\mathcal{P}}_2(\mathcal{Y})$, we have:

$$S(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}(Y \otimes Y) - m(\mathbb{P}) \otimes m(\mathbb{P}).$$

Contaminated Distributions

To define the Gateaux derivative of a statistical functional, we first recall the definition of the derivative of a linear operator f of a normed space $\mathfrak{F}$. Let $u, v \in \mathfrak{F}$, when it exists, the derivative of f at u in the direction of v is defined as the following limit:

$$\lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} \frac{f(u + \epsilon v) - f(u)}{\epsilon}.$$

This definition relies on the fact that the function $\epsilon \in \mathbb{R} \mapsto u + \epsilon v \in \mathfrak{F}$ is continuous at 0. This is not the case in $\mathcal{P}(\mathcal{Y})$ where the function $\epsilon \mapsto \mathbb{P} + \epsilon\mathbb{Q}$ is not continuous at 0 in general. To define the derivative of a statistical functional, the idea of Gateaux is to use contaminated distributions

Definition 9 (Contaminated distribution)**.** *Let $\mathbb{P}, \mathbb{Q} \in \widetilde{\mathcal{P}}(\mathcal{Y})$ and $\epsilon > 0$. The probability distribution $\mathbb{P}$ contaminated by $\mathbb{Q}$ at the level ϵ, denoted $\mathbb{P}^{(\mathbb{Q},\epsilon)}$ is defined by:*

$$\mathbb{P}^{(\mathbb{Q},\epsilon)} = \mathbb{P} + \epsilon(\mathbb{Q} - \mathbb{P}).$$

When $\mathbb{P}, \mathbb{Q}$ are in $\widetilde{\mathcal{P}}(\mathcal{Y})$, the convexity of $\widetilde{\mathcal{P}}(\mathcal{Y})$ ensures that the mapping $\epsilon \mapsto \mathbb{P}^{(\mathbb{Q},\epsilon)}$ is continuous at 0. For $\epsilon \in [0, 1]$, the quantity $\mathbb{P}^{(\mathbb{Q},\epsilon)}$ is a mixture of $\mathbb{P}$ and $\mathbb{Q}$ weighted by $(1 - \epsilon)$ and ϵ. It can be interpreted as the probability distribution $\mathbb{P}$ where a portion ϵ of its mass has been 'contaminated' by $\mathbb{Q}$.

Let $\mathbb{P}, \mathbb{Q} \in \widetilde{\mathcal{P}}_2(\mathcal{Y})$ and $\epsilon > 0$, the statistical functional of the mean $m(\circledcirc)$ and of the covariance $S(\circledcirc)$ at the contaminated distribution $\mathbb{P}^{(\mathbb{Q},\epsilon)}$ are such that:

$$m(\mathbb{P}^{(\mathbb{Q},\epsilon)}) = m(\mathbb{P}) + \epsilon\Big(m(\mathbb{Q}) - m(\mathbb{P})\Big), \tag{4.1}$$

$$S(\mathbb{P}^{(\mathbb{Q},\epsilon)}) = S(\mathbb{P}) + \epsilon\Big(S(\mathbb{Q}) - S(\mathbb{P}) + \Big(m(\mathbb{Q}) - m(\mathbb{P})\Big)^{\otimes 2}\Big) + \epsilon^2\Big(m(\mathbb{Q}) - m(\mathbb{P})\Big)^{\otimes 2}. \tag{4.2}$$

Proof. For the statistical functional of the mean $m(\circledcirc)$, we have:

$$\begin{aligned} m(\mathbb{P}^{(\mathbb{Q},\epsilon)}) &= \int_{\mathcal{Y}} y \,\mathrm{d}\Big(\mathbb{P} + \epsilon(\mathbb{Q} - \mathbb{P})\Big)(y) \\ &= \int_{\mathcal{Y}} y \,\mathrm{d}\,\mathbb{P}(y) + \epsilon\left(\int_{\mathcal{Y}} y d\mathbb{Q}(y) - \int_{\mathcal{Y}} y \,\mathrm{d}\,\mathbb{P}(y)\right) \\ &= m(\mathbb{P}) + \epsilon\Big(m(\mathbb{Q}) - m(\mathbb{P})\Big). \end{aligned}$$

For the statistical functional of the covariance $S(\circledcirc)$, we have:

$$S(\mathbb{P}^{(\mathbb{Q},\epsilon)}) = \mathbb{E}_{\mathbb{P}^{(\mathbb{Q},\epsilon)}}(Y \otimes Y) - m(\mathbb{P}^{(\mathbb{Q},\epsilon)}) \otimes m(\mathbb{P}^{(\mathbb{Q},\epsilon)}),$$

where:

$$\begin{aligned} \mathbb{E}_{\mathbb{P}^{(\mathbb{Q},\epsilon)}}(Y \otimes Y) =& \mathbb{E}_{\mathbb{P}}(Y \otimes Y) + \epsilon\Big(\mathbb{E}_{\mathbb{Q}}(Y \otimes Y) - \mathbb{E}_{\mathbb{P}}(Y \otimes Y)\Big) \\ =& S(\mathbb{P}) + m(\mathbb{P})^{\otimes 2} + \epsilon\Big(S(\mathbb{Q}) - S(\mathbb{P}) + m(\mathbb{Q})^{\otimes 2} - m(\mathbb{P})^{\otimes 2}\Big), \end{aligned}$$

and

$$\begin{aligned} m(\mathbb{P}^{(\mathbb{Q},\epsilon)}) \otimes m(\mathbb{P}^{(\mathbb{Q},\epsilon)}) =& \Big(m(\mathbb{P}) + \epsilon\Big(m(\mathbb{Q}) - m(\mathbb{P})\Big)\Big)^{\otimes 2} \\ =& m(\mathbb{P})^{\otimes 2} + \epsilon\Big(m(\mathbb{P}) \otimes \Big(m(\mathbb{Q}) - m(\mathbb{P})\Big) \\ &+ \Big(m(\mathbb{Q}) - m(\mathbb{P})\Big) \otimes m(\mathbb{P})\Big) + \epsilon^2\Big(m(\mathbb{Q}) - m(\mathbb{P})\Big)^{\otimes 2} \\ =& m(\mathbb{P})^{\otimes 2} + \epsilon\Big(m(\mathbb{P}) \otimes m(\mathbb{Q}) + m(\mathbb{Q}) \otimes m(\mathbb{P}) - 2m(\mathbb{P})^{\otimes^2}\Big) \\ &+ \epsilon^2\Big(m(\mathbb{Q}) - m(\mathbb{P})\Big)^{\otimes 2} \end{aligned}$$

We sum the two terms and focus on the terms multiplied by ϵ, that are equal to:

$$\begin{aligned} &\Big(S(\mathbb{Q}) - S(\mathbb{P}) + m(\mathbb{Q})^{\otimes 2} - m(\mathbb{P})^{\otimes 2}\Big) - \Big(m(\mathbb{P}) \otimes m(\mathbb{Q}) + m(\mathbb{Q}) \otimes m(\mathbb{P}) - 2m(\mathbb{P})^{\otimes^2}\Big) \\ =& S(\mathbb{Q}) - S(\mathbb{P}) + \Big(m(\mathbb{Q}) - m(\mathbb{P})\Big)^{\otimes 2}. \end{aligned}$$

□

Gateaux Derivatives

The Gateaux derivative measures the impact of an infinitesimal contamination of a probability distribution on a statistical functional.

Definition 10 (Gateaux derivative)**.** *Let $\mathbb{P}, \mathbb{Q} \in \widetilde{\mathcal{P}}(\mathcal{Y})$. When it exists, the Gateaux derivative of a statistical functional $T(\circledcirc)$ at $\mathbb{P}$ in the direction of $\mathbb{Q}$ is denoted $T'\Big(\mathbb{P}, \mathbb{Q}\Big)$ and is defined as the following limit:*

$$T'\Big(\mathbb{P}, \mathbb{Q}\Big) = \lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} \frac{T\Big(\mathbb{P}^{(\mathbb{Q}, \epsilon)}\Big) - T(\mathbb{P})}{\epsilon}.$$

Observe that according to this definition, every classical formula on the derivative of products, quotients and compositions hold for Gateaux derivatives. The Gateaux derivative of the statistical functionals of the mean $m(\circledcirc)$ and covariance S follow directly from Equations (4.1) and (4.2):

$$m'\Big(\mathbb{P}, \mathbb{Q}\Big) = m(\mathbb{Q}) - m(\mathbb{P}),$$

$$S'\Big(\mathbb{P}, \mathbb{Q}\Big) = S(\mathbb{Q}) - S(\mathbb{P}) + \Big(m(\mathbb{Q}) - m(\mathbb{P})\Big)^{\otimes 2}.$$

Influence Functions

Influence functions are a particular case of Gateaux derivatives where the contaminating probability distribution $\mathbb{Q}$ is reduced to a Dirac δ_y, $y \in \mathcal{Y}$. The influence function of a statistical functional with respect to δ_y aims at capturing the influence of a single observation y of the input space $\mathcal{Y}$ on the statistical functional.

Definition 11 (Influence function)**.** *Let $\mathbb{P} \in \widetilde{\mathcal{P}}(\mathcal{Y})$ and $y \in \mathcal{Y}$. When it exists, the influence of y on a statistical functional $T(\circledcirc)$ at $\mathbb{P}$, denoted $\mathrm{I}_T(\mathbb{P}, y)$ is defined such that:*

$$\mathrm{I}_T(\mathbb{P}, y) = T'\Big(\mathbb{P}, \delta_y\Big).$$

According to [112], the influence function of a statistical functional $T(\circledcirc)$ can be interpreted as the first term of a power series expansion of $T(\circledcirc)$ at $\mathbb{P}^{(\delta_y, \epsilon)}$:

$$T(\mathbb{P}^{(\delta_y, \epsilon)}) = T + \epsilon\, \mathrm{I}_T(\mathbb{P}, y) + \epsilon\psi(\epsilon),$$

where $\lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} \psi(\epsilon) = 0$. Let $\mathbb{P} \in \widetilde{\mathcal{P}}_2(\mathcal{Y})$ and $y \in \mathcal{Y}$. The influence of y on the statistical functionals of the mean and covariance at $\mathbb{P}$ is such that:

$$\mathrm{I}_m(\mathbb{P}, y) = y - m(\mathbb{P}),$$

$$\mathrm{I}_S(\mathbb{P}, y) = \Big(y - m(\mathbb{P})\Big)^{\otimes 2} - S(\mathbb{P}).$$

Influence of a Sub-Group of Points and Gateaux Derivatives

Influence functions provide the influence of a single observation and are used to assess the robustness of an estimator. In single-cell applications, we may be interested in the measure of the influence of a group of observations. That can be done with empirical Gateaux derivatives.

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be a n-sample of i.i.d. observations such that $Y_1 \sim \mathbb{P} \in \widetilde{\mathcal{P}}(\mathcal{Y})$. We assume that $\widetilde{\mathcal{P}}(\mathcal{Y})$ is defined so that the empirical distribution $\mathbb{P}_n$ associated to $\mathbf{Y}$ is in $\widetilde{\mathcal{P}}(\mathcal{Y})$ almost-surely. Let $\mathbf{Z} = (Z_1, \ldots, Z_q)$ be a sub-sample of $\mathbf{Y}$. We would like to assess the influence of $\mathbf{Z}$ on a statistical functional computed with respect to $\mathbb{P}_n$. Let $\mathbb{P}_{n,q}$ be the discrete distribution associated to the points in $\mathbf{Z}$ and $T(\circledcirc)$ be a statistical functional. The influence of the group of observations $\mathbf{Z}$ on $T(\circledcirc)$ can be evaluated through the Gateaux derivative $T'\Big(\mathbb{P}_n, \mathbb{P}_{n,q}\Big)$ of $T(\circledcirc)$ at $\mathbb{P}_n$ in the direction of $\mathbb{P}_{n,q}$. Note that when $\mathbf{Z}$ is reduced to a single point y of $\mathbf{Y}$, this quantity corresponds to the influence function $\mathrm{I}_T(\mathbb{P}, y)$.

4.1.2 Advanced Concepts for Gateaux Derivative and Influence Function Adapted to Kernel Testing

Two generalizations of Gateaux derivatives and influence function are needed to define those of the kernel test approaches of the previous chapters. The first generalization is to define the Gateaux derivatives and influence functions of statistical functionals that output elements of a RKHS $\mathcal{H}$ associated to a p.d. kernel $k(\cdot, \cdot)$, or Hilbert-Schmidt operators from $\mathcal{H}$ to $\mathcal{H}$. To do so, we define $\widetilde{\mathcal{P}}_{\mathcal{H}}(\mathcal{Y}) = \{\mathbb{P} \in \widetilde{\mathcal{P}}(\mathcal{Y}) |\ \mathbb{E}_{\mathbb{P}}\Big(k(Y, Y)\Big) < \infty\}$, the subset of $\widetilde{\mathcal{P}}(\mathcal{Y})$ that contains all the probability distributions for which the kernel covariance operators are well defined. Since kernel testing is inherently based on quantities defined over several probability distributions to compare, the second generalization is to

define statistical functional defined on joint distributions. Here, we only focus on statistical functionals defined on two probability distribution, but the generalization to more than two probability distribution would be needed for the Gateaux derivative and influence function of the truncated Hotelling-Lawley trace statistic. The Gateaux derivatives and influence functions defined on such statistical functionals are called partial Gateaux derivatives and partial influence function and were introduced in [109]. A last tool of interest is the derivation of the Gateaux derivatives and influence functions of the statistical functionals associated to the eigenfunctions and eigenvalues of a statistical functional that outputs a self-adjoint Hilbert-Schmidt operator. This tool has been developed to detect the influential observations of the Kernel Principal Component Analysis [34, 67].

Let $(\mathcal{Y}, \mathfrak{Y})$ be a measurable space and $k(\cdot,\cdot)$ a p.d. kernel associated to the separable RKHS $\mathcal{H}$ and the feature map $\phi(\cdot)$. The statistical functional of the kernel mean embedding $\mu(\circledcirc)$ and the kernel covariance operator $\Sigma(\circledcirc)$ are defined such that for $\mathbb{P} \in \widetilde{\mathcal{P}}_{\mathcal{H}}(\mathcal{Y})$, we have $\mu(\mathbb{P}) = \mu$ and $\Sigma(\mathbb{P}) = \Sigma$, where μ and Σ are defined in Chapter 1. The Gateaux derivatives and influence functions of $\mu(\circledcirc)$ and $\Sigma(\circledcirc)$ are similar to those of $m(\circledcirc)$ and $S(\circledcirc)$ respectively. Let $\mathbb{P}, \mathbb{Q} \in \widetilde{\mathcal{P}}_{\mathcal{H}}(\mathcal{Y})$ and $y \in \mathcal{Y}$, we have the following Gateaux derivatives and influence functions:

$$
\begin{aligned}
&\mu'\Big(\mathbb{P}, \mathbb{Q}\Big) = \mu(\mathbb{Q}) - \mu(\mathbb{P}),\\
&\Sigma'\Big(\mathbb{P}, \mathbb{Q}\Big) = \Sigma(\mathbb{Q}) - \Sigma(\mathbb{P}) + \Big(\mu(\mathbb{Q}) - \mu(\mathbb{P})\Big)^{\otimes 2},\\
&\mathrm{I}_{\mu}(\mathbb{P}, y) = \phi(y) - \mu(\mathbb{P}),\\
&\mathrm{I}_{\Sigma}(\mathbb{P}, y) = \Big(\phi(y) - \mu(\mathbb{P})\Big)^{\otimes 2} - \Sigma(\mathbb{P}).
\end{aligned}
$$

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be a n-sample of i.i.d. observations such that $Y_1 \sim \mathbb{P}$ and $\mathbf{Z} = (Z_1, \ldots, Z_q)$ be a sub-sample of $\mathbf{Y}$. Let $\mathbb{P}_n$ be the empirical distribution associated to $\mathbf{Y}$ and $\mathbb{P}_{n,q}$ the discrete distribution associated to the points in $\mathbf{Z}$. The influence of $\mathbf{Z}$ on the statistical functionals of the kernel mean embedding and the kernel covariance operator are such that:

$$
\mu'\Big(\mathbb{P}_n, \mathbb{P}_{n,q}\Big) = \widehat{\mu}^{\mathbf{Z}} - \widehat{\mu},
$$

$$
\Sigma'\Big(\mathbb{P}_n, \mathbb{P}_{n,q}\Big) = \widehat{\Sigma}^{\mathbf{Z}} - \widehat{\Sigma} + (\widehat{\mu}^{\mathbf{Z}} - \widehat{\mu})^{\otimes 2}
$$

where $\widehat{\mu} = n^{-1}\sum_{j=1}^{n}\phi(Y_j)$ and $\widehat{\mu}^{\mathbf{Z}} = q^{-1}\sum_{j=1}^{q}\phi(Z_j)$ are the empirical kernel covariance associated to $\mathbf{Y}$ and $\mathbf{Z}$ respectively, and $\widehat{\Sigma} = n^{-1}\sum_{j=1}^{n}\Big(\phi(Y_j) - \widehat{\mu}\Big)^{\otimes 2}$ and $\widehat{\Sigma}^{\mathbf{Z}} = q^{-1}\sum_{j=1}^{q}\Big(\phi(Z_j) - \widehat{\mu}^{\mathbf{Z}}\Big)^{\otimes 2}$ are the empirical kernel covariance operators associated to $\mathbf{Y}$ and $\mathbf{Z}$ respectively.

Diagonalization of Symmetric Operators

Let $A(\circledcirc)$ a statistical functional such that for $\mathbb{P} \in \widetilde{\mathcal{P}}_{\mathcal{H}}(\mathcal{Y})$, $A(\mathbb{P})$ is a self-adjoint Hilbert-Schmidt operator with an orthonormal set of eigenvectors $(f_t^{A(\mathbb{P})})_{t\geq 1}$ associated to the distinct non-increasing eigenvalues $(\lambda_t^{A(\mathbb{P})})_{t\geq 1}$. Let $(f_t^A(\circledcirc))_{t\geq 1}$ and $(\lambda_t^A(\circledcirc))_{t\geq 1}$ the statistical functionals such that for $t \geq 1$, we have $f_t^A(\mathbb{P}) = f_t^{A(\mathbb{P})}$ and $f_t^A(\mathbb{P}) = \lambda_t^{A(\mathbb{P})}$. The influence functions of $f_t^A(\circledcirc)$ and $\lambda_t^A(\circledcirc)$ have been simultaneously proposed in [34, 67]. We generalize these results by proposing their Gateaux derivatives:

Theorem 17. *Let $\mathbb{P}, \mathbb{Q} \in \widetilde{\mathcal{P}}_{\mathcal{H}}(\mathcal{Y})$. For $t \geq 1$, the Gateaux derivatives of $f_t^A(\circledcirc)$ and $\lambda_t^A(\circledcirc)$ at $\mathbb{P}$ in the direction of $\mathbb{Q}$ are such that:*

$$\lambda_t^{A'}\Big(\mathbb{P}, \mathbb{Q}\Big) = \Big\langle A'\Big(\mathbb{P}, \mathbb{Q}\Big) f_t^A(\mathbb{P}), f_t^A(\mathbb{P})\Big\rangle_{\mathcal{H}},$$

$$f_t^{A'}\Big(\mathbb{P}, \mathbb{Q}\Big) = \sum_{\substack{t'\geq 1\\ t'\neq t}} \frac{\Big\langle A'\Big(\mathbb{P}, \mathbb{Q}\Big) f_t^A(\mathbb{P}), f_{t'}^A(\mathbb{P})\Big\rangle_{\mathcal{H}}}{\lambda_t^A(\mathbb{P}) - \lambda_{t'}^A(\mathbb{P})} f_{t'}^A(\mathbb{P}).$$

The previous theorem generalizes the influence functions defined in [34] given in the following corollary.

Corollary 1. *Let $\mathbb{P} \in \widetilde{\mathcal{P}}_{\mathcal{H}}(\mathcal{Y})$ and $y \in \mathcal{Y}$. The influence of y on $f_t^A(\circledcirc)$ and $\lambda_t^A(\circledcirc)$ at $\mathbb{P}$ are such that:*

$$\mathrm{I}_{\lambda_t^A}(\mathbb{P}, y) = \Big\langle \mathrm{I}_A(\mathbb{P}, y) f_t^A(\mathbb{P}), f_t^A(\mathbb{P})\Big\rangle_{\mathcal{H}},$$

$$\mathrm{I}_{f_t^A}(\mathbb{P}, y) = \sum_{\substack{t'\geq 1\\ t'\neq t}} \frac{\Big\langle \mathrm{I}_A(\mathbb{P}, y) f_t^A(\mathbb{P}), f_{t'}^A(\mathbb{P})\Big\rangle_{\mathcal{H}}}{\lambda_t^A(\mathbb{P}) - \lambda_{t'}^A(\mathbb{P})} f_t^A(\mathbb{P}).$$

Proof. We adapt the proof of [34] for the influence functions to the Gateaux derivatives.

Note that for any statistical functional $T(\circledcirc)$, we have the following power serie expansion:

$$T(\mathbb{P}^{(\mathbb{Q},\epsilon)}) = T(\mathbb{P}) + \epsilon T'\Big(\mathbb{P}, \mathbb{Q}\Big) + \epsilon\psi(\epsilon),$$

where $\lim\limits_{\substack{\epsilon\to 0\\ \epsilon>0}}\psi(\epsilon) = 0$. Thus, we have:

$$\begin{aligned} A\Big(\mathbb{P}^{(\mathbb{Q},\epsilon)}\Big) &= A(\mathbb{P}) + \epsilon A'\Big(\mathbb{P}, \mathbb{Q}\Big) + \epsilon\psi_{\mathrm{HS}(\mathcal{H})}(\epsilon), \\ f_t^A\Big(\mathbb{P}^{(\mathbb{Q},\epsilon)}\Big) &= f_t^A(\mathbb{P}) + \epsilon f_t^{A'}\Big(\mathbb{P}, \mathbb{Q}\Big) + \epsilon\psi_{\mathcal{H}}(\epsilon), \\ \lambda_t^A\Big(\mathbb{P}^{(\mathbb{Q},\epsilon)}\Big) &= \lambda_t^A(\mathbb{P}) + \epsilon \lambda_t^{A'}\Big(\mathbb{P}, \mathbb{Q}\Big) + \epsilon\psi_{\mathbb{R}}(\epsilon), \end{aligned} \tag{4.3}$$

where $\psi_{\mathrm{HS}(\mathcal{H})}(\epsilon) \in \mathrm{HS}(\mathcal{H})$, $\psi_{\mathcal{H}}(\epsilon) \in \mathcal{H}$ and $\psi_{\mathbb{R}}(\epsilon) \in \mathbb{R}$ with $\lim\limits_{\substack{\epsilon\to 0\\ \epsilon>0}}\Big\|\psi_{\mathrm{HS}(\mathcal{H})}(\epsilon)\Big\|_{\mathrm{HS}(\mathcal{H})} = 0$, $\lim\limits_{\substack{\epsilon\to 0\\ \epsilon>0}}\|\psi_{\mathcal{H}}(\epsilon)\|_{\mathcal{H}} = 0$ and $\lim\limits_{\substack{\epsilon\to 0\\ \epsilon>0}}\psi_{\mathbb{R}}(\epsilon) = 0$. Let $t \geq 1$. We have the following eigen-equation for the statistical functionals:

$$A(\circledcirc) f_t^A(\circledcirc) = \lambda_t^A(\circledcirc) f_t^A(\circledcirc).$$

If we substitute every element of this eigen-equation applied on $\mathbb{P}^{(\mathbb{Q},\epsilon)}$ by the associated power serie expansion of Equations (4.3), we obtain:

$$\begin{aligned} &\Big(A(\mathbb{P}) + \epsilon A'\Big(\mathbb{P}, \mathbb{Q}\Big) + \epsilon\psi_{\mathrm{HS}(\mathcal{H})}(\epsilon)\Big)\Big(f_t^A(\mathbb{P}) + \epsilon f_t^{A'}\Big(\mathbb{P}, \mathbb{Q}\Big) + \epsilon\psi_{\mathcal{H}}(\epsilon)\Big) \\ &= \Big(\lambda_t^A(\mathbb{P}) + \epsilon\lambda_t^{A'}\Big(\mathbb{P}, \mathbb{Q}\Big) + \epsilon\psi_{\mathbb{R}}(\epsilon)\Big)\Big(f_t^A(\mathbb{P}) + \epsilon f_t^{A'}\Big(\mathbb{P}, \mathbb{Q}\Big) + \epsilon\psi_{\mathcal{H}}(\epsilon)\Big). \end{aligned}$$

From now, we omit the probability distribution $\mathbb{P}$ on which are evaluated the statistical functional so that $A(\mathbb{P})$ is denoted A and we omit the probability distributions $\mathbb{P}$ and $\mathbb{Q}$ on which are evaluated the Gateaux derivatives so that $A'\Big(\mathbb{P}, \mathbb{Q}\Big)$ becomes A', so that the previous equation becomes:

$$\Big(A + \epsilon A' + \epsilon\psi_{\mathrm{HS}(\mathcal{H})}(\epsilon)\Big)\Big(f_t^A + \epsilon f_t^{A'} + \epsilon\psi_{\mathcal{H}}(\epsilon)\Big) = \Big(\lambda_t^A + \epsilon\lambda_t^{A'} + \epsilon\psi_{\mathbb{R}}(\epsilon)\Big)\Big(f_t^A + \epsilon f_t^{A'} + \epsilon\psi_{\mathcal{H}}(\epsilon)\Big).$$

Then we develop this equation and use that $Af_t^A = \lambda_t^A f_t^A$ to obtain a power serie

expansion on each side of the equation. By identifying the ϵ-terms, we find that:

$$A' f_t^A + A f_t^{A'} = \lambda_t^{A'} f_t^A + \lambda_t^A f_t^{A'}. \tag{4.4}$$

Now consider the norm of $f_t^A(\circledcirc)$, we have:

$$\begin{aligned}\left\| f_t^A(\mathbb{P}^{(\mathbb{Q},\epsilon)}) \right\|_{\mathcal{H}}^2 &= \left\| f_t^A + \epsilon f_t^{A'} + \epsilon \psi_{\mathcal{H}}(\epsilon) \right\|_{\mathcal{H}}^2 \\ &= \left\| f_t^A \right\|_{\mathcal{H}}^2 + 2\epsilon \left\langle f_t^A, f_t^{A'} + \psi_{\mathcal{H}}(\epsilon) \right\rangle_{\mathcal{H}} + \epsilon^2 \left\| f_t^{A'} + \psi_{\mathcal{H}}(\epsilon) \right\|_{\mathcal{H}}^2.\end{aligned}$$

That leads to:

$$N' = 2\left\langle f_t^A, f_t^{A'} \right\rangle_{\mathcal{H}},$$

where $N(\circledcirc)$ is the statistical functional defined such that $N(\mathbb{P}) = \left\| f_t^A(\mathbb{P}) \right\|_{\mathcal{H}}$. As we know that $\left(f_t^A\right)_{t\geq 1}$ is an orthonormal set of eigenfunctions, we also know that $N(\circledcirc)$ is actually a constant equal to 1, and its Gateaux derivative is thus equal to 0. We deduce that:

$$\left\langle f_t^A, f_t^{A'} \right\rangle_{\mathcal{H}} = 0.$$

We take advantage of this orthogonality by computing the inner product between the terms of Equation (4.4) and $f_t^A(\mathbb{P})$ on a first hand, and $f_{t'}^A$, for $t' \neq t$ on a second hand. First, we have:

$$\left\langle A' f_t^A + A f_t^{A'}, f_t^A \right\rangle_{\mathcal{H}} = \left\langle \lambda_t^{A'} f_t^A + \lambda_t^A f_t^{A'}, f_t^A \right\rangle_{\mathcal{H}}.$$

As A is self-adjoint and $\left\langle f_t^{A'}, A f_t^A \right\rangle_{\mathcal{H}} = \lambda_t^A \left\langle f_t^{A'}, f_t^A \right\rangle_{\mathcal{H}}$, we have:

$$\begin{aligned}&\left\langle A' f_t^A, f_t^A \right\rangle_{\mathcal{H}} + \underbrace{\left\langle f_t^{A'}, A f_t^A \right\rangle_{\mathcal{H}}}_{=0} = \lambda_t^{A'} \underbrace{\left\langle f_t^A, f_t^A \right\rangle_{\mathcal{H}}}_{=1} + \lambda_t^A \underbrace{\left\langle f_t^{A'}, f_t^A \right\rangle_{\mathcal{H}}}_{=0}. \\ \iff &\left\langle A' f_t^A, f_t^A \right\rangle_{\mathcal{H}} = \lambda_t^{A'}.\end{aligned}$$

That proves the expression of the Gateaux derivative of $\lambda_t^A(\circledcirc)$. Secondly, for $t' \neq t$, we

have:

$$\begin{aligned}&\left\langle A' f_t^A + A f_t^{A'}, f_{t'}^A \right\rangle_{\mathcal{H}} = \left\langle \lambda_t^{A'} f_t^A + \lambda_t^A f_t^{A'}, f_{t'}^A \right\rangle_{\mathcal{H}} \\ \iff &\left\langle A' f_t^A, f_{t'}^A \right\rangle_{\mathcal{H}} + \lambda_{t'}^A \left\langle f_t^{A'}, f_{t'}^A \right\rangle_{\mathcal{H}} = \lambda_t^{A'} \underbrace{\left\langle f_t^A, f_{t'}^A \right\rangle_{\mathcal{H}}}_{=0} + \lambda_t^A \left\langle f_t^{A'}, f_{t'}^A \right\rangle_{\mathcal{H}}.\end{aligned}$$

By reordering the equation, we obtain that:

$$\left\langle f_t^{A'}, f_{t'}^A \right\rangle_{\mathcal{H}} = \frac{\left\langle A' f_t^A, f_{t'}^A \right\rangle_{\mathcal{H}}}{\lambda_t^A - \lambda_{t'}^A}.$$

Since $(f_t^A)_{t \geq 1}$ is an orthonormal basis of $\mathcal{H}$ and $\left\langle f_t^A, f_t^{A'} \right\rangle_{\mathcal{H}} = 0$, we conclude that:

$$f_t^{A'} = \sum_{\substack{t' \geq 1 \\ t' \neq t}} \frac{\left\langle A' f_t^A, f_{t'}^A \right\rangle_{\mathcal{H}}}{\lambda_t^A - \lambda_{t'}^A} f_{t'}^A.$$

□

Partial Gateaux Derivatives and Partial Influence Functions

When a statistic is defined with respect to several samples, the associated statistical functional should be defined accordingly. For instance, let $\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{P}(\mathcal{Y})$, the statistical functional associated to the within-group covariance $\Sigma_W(\circledcirc, \circledcirc)$ and the between-group covariance $\Sigma_B(\circledcirc, \circledcirc)$ are such that:

$$\Sigma_W(\mathbb{P}_1, \mathbb{P}_2) = \frac{n_1}{n}\Sigma(\mathbb{P}_1) + \frac{n_2}{n}\Sigma(\mathbb{P}_2),$$

$$\Sigma_B(\mathbb{P}_1, \mathbb{P}_2) = \frac{n_1 n_2}{n^2}\Big(\mu(\mathbb{P}_1) - \mu(\mathbb{P}_2)\Big)^{\otimes 2},$$

where $n_1, n_2 \geq 1$ and $n_1 + n_2 = n$. To properly define the Gateaux derivative and influence function of such statistical functionals, the partial Gateaux derivative and partial influence function have been introduced [109]. The partial Gateaux derivative and partial influence function rely on the generalization of contaminated distribution to joint distributions.

Definition 12 (Joint contaminated distribution)**.** *Let $\mathbb{P}_1, \mathbb{P}_2, \mathbb{Q}_1, \mathbb{Q}_2 \in \widetilde{\mathcal{P}}(\mathcal{Y})$, and $\epsilon > 0$. The joint probability distribution $(\mathbb{P}_1, \mathbb{P}_2)$ contaminated by $(\mathbb{Q}_1, \mathbb{Q}_2)$ at the level ϵ, denoted*

$(\mathbb{P}_1, \mathbb{P}_2)^{(\mathbb{Q}_1,\mathbb{Q}_2),\epsilon}$, *is defined such that:*

$$(\mathbb{P}_1, \mathbb{P}_2)^{(\mathbb{Q}_1,\mathbb{Q}_2),\epsilon} = \Big(\mathbb{P}_1^{(\mathbb{Q}_1,\epsilon)}, \mathbb{P}_2^{(\mathbb{Q}_2,\epsilon)}\Big)$$

Then the partial Gateaux derivative and partial influence function are defined similarly to the Gateaux derivative and influence function.

Definition 13 (Partial Gateaux derivative)**.** *Let* $\mathbb{P}_1, \mathbb{P}_2, \mathbb{Q}_1, \mathbb{Q}_2 \in \widetilde{\mathcal{P}}(\mathcal{Y})$*. When it exists, the partial Gateaux derivative of a statistical functional* $T(\circledcirc, \circledcirc)$ *at* $(\mathbb{P}_1, \mathbb{P}_2)$ *in the direction of* $(\mathbb{Q}_1, \mathbb{Q}_2)$ *is denoted* $T'\Big((\mathbb{P}_1, \mathbb{P}_2), (\mathbb{Q}_1, \mathbb{Q}_2)\Big)$ *and is defined as the following limit:*

$$T'\Big((\mathbb{P}_1, \mathbb{P}_2), (\mathbb{Q}_1, \mathbb{Q}_2)\Big) = \lim_{\substack{\epsilon \to 0 \\ \epsilon > 0}} \frac{T\Big((\mathbb{P}_1, \mathbb{P}_2)^{(\mathbb{Q}_1,\mathbb{Q}_2),\epsilon}\Big) - T(\mathbb{P}_1, \mathbb{P}_2)}{\epsilon}.$$

Definition 14 (Partial influence function)**.** *Let* $\mathbb{P}_1, \mathbb{P}_2 \in \widetilde{\mathcal{P}}(\mathcal{Y})$ *and* $y \in \mathcal{Y}$*. When it exists, the partial influence of* y *contaminating* $\mathbb{P}_1$ *and letting* $\mathbb{P}_2$ *unchanged on a statistical functional* $T(\circledcirc, \circledcirc)$*, denoted* $\mathrm{I}_{T,\mathbb{P}_2}(\mathbb{P}_1, y)$*, is defined such that:*

$$\mathrm{I}_{T,\mathbb{P}_2}(\mathbb{P}_1, y) = T'\Big((\mathbb{P}_1, \mathbb{P}_2), (\delta_y, \mathbb{P}_2)\Big).$$

Similarly, the partial influence of y *contaminating* $\mathbb{P}_2$ *and letting* $\mathbb{P}_1$ *unchanged on a statistical functional* $T(\circledcirc, \circledcirc)$*, denoted* $\mathrm{I}_{T,\mathbb{P}_1}(\mathbb{P}_2, y)$*, is defined such that:*

$$\mathrm{I}_{T,\mathbb{P}_1}(\mathbb{P}_2, y) = T'\Big((\mathbb{P}_1, \mathbb{P}_2), (\mathbb{P}_1, \delta_y)\Big).$$

Let $\mathbb{Q}_1, \mathbb{Q}_2 \in \widetilde{\mathcal{P}}_2(\mathcal{Y})$, we find the following results:

$$\begin{aligned}\Sigma_B{}'\Big((\mathbb{P}_1, \mathbb{P}_2), (\mathbb{Q}_1, \mathbb{Q}_2)\Big) = \frac{n_1 n_2}{n}\Bigg(&\Big(\mu(\mathbb{P}_1) - \mu(\mathbb{P}_2)\Big) \otimes \Big(\mu'\Big(\mathbb{P}_1, \mathbb{Q}_1\Big) - \mu'\Big(\mathbb{P}_2, \mathbb{Q}_2\Big)\Big) \\ &+ \Big(\mu'\Big(\mathbb{P}_1, \mathbb{Q}_1\Big) - \mu'\Big(\mathbb{P}_2, \mathbb{Q}_2\Big)\Big) \otimes \Big(\mu(\mathbb{P}_1) - \mu(\mathbb{P}_2)\Big)\Bigg),\end{aligned}$$

$$\Sigma_W{}'\Big((\mathbb{P}_1, \mathbb{P}_2), (\mathbb{Q}_1, \mathbb{Q}_2)\Big) = \frac{n_1}{n}\Sigma'\Big(\mathbb{P}_1, \mathbb{Q}_1\Big) + \frac{n_2}{n}\Sigma'\Big(\mathbb{P}_2, \mathbb{Q}_2\Big).$$

Let $(f_t^{\Sigma_W}(\circledcirc, \circledcirc))_{t\geq 1}$ be the statistical functionals of the orthonormal eigenfunctions of $\Sigma_W(\circledcirc, \circledcirc)$ and $(\lambda_t^{\Sigma_W}(\circledcirc, \circledcirc))_{t\geq 1}$ be the sequence of statistical functionals of associated

eigenvalues. A direct application of Theorem 17 gives that:

$$\lambda_t^{\Sigma_W}{}'\Big((\mathbb{P}_1,\mathbb{P}_2),(\mathbb{Q}_1,\mathbb{Q}_2)\Big) = \Big\langle \Sigma_W{}'\Big((\mathbb{P}_1,\mathbb{P}_2),(\mathbb{Q}_1,\mathbb{Q}_2)\Big) f_t^{\Sigma_W}(\mathbb{P}_1,\mathbb{P}_2), f_t^{\Sigma_W}(\mathbb{P}_1,\mathbb{P}_2)\Big\rangle_{\mathcal{H}},$$

$$\begin{aligned} &f_t^{\Sigma_W}{}'\Big((\mathbb{P}_1,\mathbb{P}_2),(\mathbb{Q}_1,\mathbb{Q}_2)\Big) \\ &= \sum_{\substack{t'\geq 1\\ t'\neq t}} \frac{\Big\langle \Sigma_W{}'\Big((\mathbb{P}_1,\mathbb{P}_2),(\mathbb{Q}_1,\mathbb{Q}_2)\Big) f_t^{\Sigma_W}(\mathbb{P}_1,\mathbb{P}_2), f_{t'}^{\Sigma_W}(\mathbb{P}_1,\mathbb{P}_2)\Big\rangle_{\mathcal{H}}}{\lambda_t^{\Sigma_W}(\mathbb{P}_1,\mathbb{P}_2) - \lambda_{t'}^{\Sigma_W}(\mathbb{P}_1,\mathbb{P}_2)} f_{t'}^{\Sigma_W}(\mathbb{P}_1,\mathbb{P}_2). \end{aligned}$$

4.2 Application of Gateaux Derivative and Influence Function to Kernel Tests

4.2.1 Gateaux Derivatives of Kernel Tests Statistics

Gateaux Derivative of the MMD Statistic

Let $\mathrm{MMD}^2(\circledcirc,\circledcirc)$ be the statistical functional of the biased squared MMD statistic such that for $\mathbb{P}_1,\mathbb{P}_2 \in \widetilde{\mathcal{P}}_{\mathcal{H}}(\mathcal{Y})$, we have:

$$\mathrm{MMD}^2(\circledcirc,\circledcirc) = \|\Delta(\circledcirc,\circledcirc)\|_{\mathcal{H}}^2,$$

where $\Delta(\circledcirc,\circledcirc)$ is such that for $\mathbb{P}_1,\mathbb{P}_2 \in \widetilde{\mathcal{P}}_{\mathcal{H}}(\mathcal{Y})$ we have $\Delta(\mathbb{P}_1,\mathbb{P}_2) = \mu(\mathbb{P}_1) - \mu(\mathbb{P}_2)$. According to the expression of $\Delta(\circledcirc,\circledcirc)$, for $\mathbb{P}_1,\mathbb{P}_2,\mathbb{Q}_1,\mathbb{Q}_2 \in \widetilde{\mathcal{P}}_{\mathcal{H}}(\mathcal{Y})$, we have the following Gateaux derivative:

$$\begin{aligned} \Delta'\Big((\mathbb{P}_1,\mathbb{P}_2),(\mathbb{Q}_1,\mathbb{Q}_2)\Big) =& \Delta(\mathbb{Q}_1,\mathbb{Q}_2) - \Delta(\mathbb{P}_1,\mathbb{P}_2) \\ =& \mu(\mathbb{Q}_1) - \mu(\mathbb{Q}_2) - \Big(\mu(\mathbb{P}_1) - \mu(\mathbb{P}_2)\Big). \end{aligned}$$

Then, the following theorem introduces the Gateaux derivative of $\mathrm{MMD}^2(\circledcirc,\circledcirc)$.

Theorem 18. *Let $\mathbb{P}_1,\mathbb{P}_2,\mathbb{Q}_1,\mathbb{Q}_2 \in \widetilde{\mathcal{P}}_{\mathcal{H}}(\mathcal{Y})$, the Gateaux derivative of $\mathrm{MMD}^2(\circledcirc,\circledcirc)$ at $(\mathbb{P}_1,\mathbb{P}_2)$ in the direction of $(\mathbb{Q}_1,\mathbb{Q}_2)$ is such that:*

$$\mathrm{MMD}^{2'} = 2\Big\langle \Delta', \Delta\Big\rangle_{\mathcal{H}},$$

where we omitted to mention $(\mathbb{P}_1,\mathbb{P}_2)$ for the statistical functionals and $\Big((\mathbb{P}_1,\mathbb{P}_2),(\mathbb{Q}_1,\mathbb{Q}_2)\Big)$

for the partial Gateaux derivative to have a readable expression.

Proof. The result follows directly as the Gateaux derivative of a composition of statistical functionals follows the same formula than the composition of classical derivatives. □

Gateaux Derivative of the Truncated KFDA Statistic

Let $T \geq 1$ and $\mathrm{D}_T^2(\circledcirc, \circledcirc)$ be the statistical functional of the truncated KFDA statistic defined in Chapter 1. We assume that T is chosen so that for $\mathbb{P}_1, \mathbb{P}_2 \in \widetilde{\mathcal{P}}_{\mathcal{H}}(\mathcal{Y})$, $\mathrm{D}_T^2(\mathbb{P}_1, \mathbb{P}_2)$ is well defined. The statistical functional $\mathrm{D}_T^2(\circledcirc, \circledcirc)$ can be expressed as:

$$\mathrm{D}_T^2(\circledcirc, \circledcirc) = \sum_{t=1}^{T} \frac{n_1 n_2}{n \lambda_t^{\Sigma_W}(\circledcirc, \circledcirc)} \left\langle f_t^{\Sigma_W}(\circledcirc, \circledcirc), \Delta(\circledcirc, \circledcirc) \right\rangle_{\mathcal{H}}^2,$$

The next theorem presents the Gateaux derivative of $\mathrm{D}_T^2(\circledcirc, \circledcirc)$.

Theorem 19. *Let $\mathbb{P}_1, \mathbb{P}_2, \mathbb{Q}_1, \mathbb{Q}_2 \in \widetilde{\mathcal{P}}_{\mathcal{H}}(\mathcal{Y})$, the Gateaux derivative of $\mathrm{D}_T^2(\circledcirc, \circledcirc)$ at $(\mathbb{P}_1, \mathbb{P}_2)$ in the direction of $(\mathbb{Q}_1, \mathbb{Q}_2)$ is such that:*

$$\mathrm{D}_T^{2\prime}\Big((\mathbb{P}_1, \mathbb{P}_2), (\mathbb{Q}_1, \mathbb{Q}_2)\Big) = \sum_{t=1}^{T} \frac{n_1 n_2 \left\langle f_t^{\Sigma_W}, \Delta \right\rangle_{\mathcal{H}}}{n \lambda_t^{\Sigma_W}} \left(2 \left\langle f_t^{\Sigma_W}, \Delta' \right\rangle_{\mathcal{H}} + 2 \left\langle f_t^{\Sigma_W\prime}, \Delta \right\rangle_{\mathcal{H}} - \frac{\lambda_t^{\Sigma_W\prime}}{\lambda_t^{\Sigma_W}} \left\langle f_t^{\Sigma_W}, \Delta \right\rangle_{\mathcal{H}} \right),$$

where we omitted to mention $(\mathbb{P}_1, \mathbb{P}_2)$ for the statistical functionals and $\Big((\mathbb{P}_1, \mathbb{P}_2), (\mathbb{Q}_1, \mathbb{Q}_2)\Big)$ for the partial Gateaux derivative to have a readable expression.

Proof. The result follows directly from applying classical operations to obtain the Gateaux derivatives of products, compositions and quotients. □

4.2.2 Kernel Tricks

The kernel trick has been applied several times in the previous chapters. As there is no originality in the derivation of the explicit formulas for the Gateaux derivatives of the kernel tests, we directly give their expressions.

Let $\mathbf{Y}_1 = (Y_{1,1}, \ldots, Y_{1,n_1})$ $\mathbf{Y}_2 = (Y_{2,1}, \ldots, Y_{2,n_2})$ two i.i.d. samples drawn from $\mathbb{P}_1$ and $\mathbb{P}_2$ with $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$ and $n = n_1 + n_2$. We consider the general situation of the influence of a sub-sample $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$ that contains observations from $\mathbf{Y}_1$ and $\mathbf{Y}_2$ such that $\mathbf{Z}_1 = (Z_{1,1}, \ldots, Z_{1,q_1})$ is a sub-sample of $\mathbf{Y}_1$ and $\mathbf{Z}_2 = (Z_{2,1}, \ldots, Z_{2,q_2})$ is a sub-sample

of $\mathbf{Y}_2$, with $q = q_1 + q_2$. We denote $\mathbb{P}_{n_1}$ and $\mathbb{P}_{n_2}$ the empirical probability distributions associated to $\mathbf{Y}_1$ and $\mathbf{Y}_2$ respectively and $\mathbb{P}_{n_1,q_1}$ and $\mathbb{P}_{n_2,q_2}$ the discreted distribution associated to the points in $\mathbf{Z}_1$ and $\mathbf{Z}_2$ respectively. The embeddings of these samples and sub-samples are denoted $\Phi(\mathbf{Y}), \Phi(\mathbf{Z}), \Phi(\mathbf{Y}_1), \Phi(\mathbf{Y}_2), \Phi(\mathbf{Z}_1)$ and $\Phi(\mathbf{Z}_2)$ respectively. The Gram matrix of samples with themself are denoted as $\mathbf{K}_\mathbf{Y}, \mathbf{K}_\mathbf{Z}$ and for $i \in \{1,2\}$, $\mathbf{K}_{\mathbf{Y}_i}, \mathbf{K}_{\mathbf{Z}_i}$. The rectangular Gram matrices associated to two samples are denoted with two subscripts, for instance $\mathbf{K}_{\mathbf{Y},\mathbf{Z}} = \Big(k(Y_i, Z_j)\Big)_{i\in\{1,\dots,n\}, j\in\{1,\dots,q\}}$. Let $\omega = (n_1^{-1}\mathbb{1}'_{n_1}, -n_2^{-1}\mathbb{1}'_{n_2})' \in \mathbb{R}^n$ and $\omega_\mathbf{Z} = (q_1^{-1}\mathbb{1}'_{q_1}, -q_2^{-1}\mathbb{1}'_{q_2})' \in \mathbb{R}^q$.

Computation of the Gateaux Derivative of the MMD Statistic

We have:

$$\mathrm{MMD}^{2'}\Big((\mathbb{P}_{n_1}, \mathbb{P}_{n_2}), (\mathbb{P}_{n_1,q_1}, \mathbb{P}_{n_2,q_2})\Big) = 2\Big(\omega'_\mathbf{Z}\mathbf{K}_{\mathbf{Z},\mathbf{Y}}\omega - \omega'\mathbf{K}_\mathbf{Y}\omega\Big)$$

Computation of the Gateaux Derivative of the Truncated KFDA Statistic

For $t \in \{1,\dots,n\}$ the statistical functionals $f_t^{\Sigma_W}(\circledcirc,\circledcirc)$ and $\lambda_t^{\Sigma_W}(\circledcirc,\circledcirc)$ evaluated on $(\mathbb{P}_{n_1}, \mathbb{P}_{n_2})$ are equal to the t^{th} eigenvector and eigenvalue of the empirical within group covariance operator $\widehat{\Sigma}_W = \Sigma_W(\mathbb{P}_{n_1}, \mathbb{P}_{n_2})$. Thus, we define $\mathbf{K}_W = \mathbf{\Pi}_W\mathbf{K}_\mathbf{Y}\mathbf{\Pi}_W$, where $\mathbf{\Pi}_W$ is the bi-centering matrix, and denote $\widehat{\lambda}_1,\dots,\widehat{\lambda}_n$ and $u_{W,1},\dots,u_{W,n}$ the eigenvalues and associated orthonormal eigenvectors of $\mathbf{K}_W$. We then have $\lambda_t^{\Sigma_W}(\mathbb{P}_{n_1}, \mathbb{P}_{n_2}) = \widehat{\lambda}_t$ and $f_t^{\Sigma_W}(\mathbb{P}_{n_1}, \mathbb{P}_{n_2}) = (n\widehat{\lambda}_t)^{-\frac{1}{2}}u'_{W,t}\mathbf{\Pi}_W\Phi(\mathbf{Y})$. For $T \in \{1,\dots,n\}$, we denote $\mathbf{U}_{W,T} = \Big(u_{W,1},\dots,u_{W,T}\Big) \in \mathcal{M}_{n,T}(\mathbb{R})$ and $\mathbf{\Lambda}_{W,T} = \mathrm{Diag}\Big(\widehat{\lambda}_1,\dots,\widehat{\lambda}_T\Big) \in \mathcal{M}_T(\mathbb{R})$, such that $\widetilde{\mathbf{U}}_{W,T} = n^{-\frac{1}{2}}\mathbf{\Lambda}_{W,T}^{-\frac{1}{2}}\mathbf{U}'_{W,T}\mathbf{\Pi}_W$.

For $i \in \{1,2\}$, we define the Gram Gateaux matrix of sample i such that:

$$\begin{aligned}\mathbf{G}_i =& \frac{1}{q_i}\mathbf{K}_{\mathbf{Y},\mathbf{Z}_i}\mathbf{\Pi}_{q_i}\mathbf{K}_{\mathbf{Z}_i,\mathbf{Y}} - \frac{1}{n}\mathbf{K}_{\mathbf{Y},\mathbf{Y}_i}\mathbf{\Pi}_{n_i}\mathbf{K}_{\mathbf{Y}_i,\mathbf{Y}} \\ &- \frac{1}{q_i^2}\mathbf{K}_{\mathbf{Y},\mathbf{Z}_i}\mathbf{J}_{q_i}\mathbf{K}_{\mathbf{Z}_i,\mathbf{Y}} + \frac{1}{nq}\mathbf{K}_{\mathbf{Y},\mathbf{Y}_i}\mathbf{J}_{n_i,q_i}\mathbf{K}_{\mathbf{Z}_i,\mathbf{Y}} + \frac{1}{nq}\mathbf{K}_{\mathbf{Y},\mathbf{Z}_i}\mathbf{J}_{q_i,n_i}\mathbf{K}_{\mathbf{Y}_i,\mathbf{Y}} - \frac{1}{n^2}\mathbf{K}_{\mathbf{Y},\mathbf{Y}_i}\mathbf{J}_{n_i}\mathbf{K}_{\mathbf{Y}_i,\mathbf{Y}},\end{aligned}$$

where $\mathbf{J}_{n,m} \in \mathcal{M}_{n,m}(\mathbb{R})$ is the matrix full of ones and $\mathbf{J}_{n,n}$ is denoted $\mathbf{J}_n$. Then, we have:

$$\left(\Big\langle\Sigma'\Big(\mathbb{P}_{n_i}, \mathbb{P}_{n_i,q_i}\Big)f_t^{\Sigma_W}(\mathbb{P}_{n_1}, \mathbb{P}_{n_2}), f_{t'}^{\Sigma_W}(\mathbb{P}_{n_1}, \mathbb{P}_{n_2})\Big\rangle_\mathcal{H}\right)_{t,t'\in\{1,\dots,T\}} = \widetilde{\mathbf{U}}'_{W,T}\mathbf{G}_i\widetilde{\mathbf{U}}_{W,T} \in \mathcal{M}_T(\mathbb{R}).$$

Therefore, we define $\mathbf{G}_W = \frac{n_1}{n}\mathbf{G}_1 + \frac{n_2}{n}\mathbf{G}_2$, such that:

$$\left(\left\langle {\Sigma_W}'\Big((\mathbb{P}_{n_1}, \mathbb{P}_{n_2}), (\mathbb{P}_{n_1,q_1}, \mathbb{P}_{n_2,q_2})\Big) f_t^{\Sigma_W}(\mathbb{P}_{n_1}, \mathbb{P}_{n_2}), f_{t'}^{\Sigma_W}(\mathbb{P}_{n_1}, \mathbb{P}_{n_2}) \right\rangle_{\mathcal{H}} \right)_{t,t' \in \{1,\dots,T\}} = \widetilde{\mathbf{U}}'_{W,T} \mathbf{G}_W \widetilde{\mathbf{U}}_{W,T} \in \mathcal{M}_T(\mathbb{R}),$$

and $\Delta_{\mathbf{\Lambda}_{W,T}} \in \mathcal{M}_T(\mathbb{R})$ such that for $t, t' \in \{1, \dots, T\}$:

$$\left(\Delta_{\mathbf{\Lambda}_{W,T}}\right)_{t,t'} = \begin{cases} 0 & \text{if } t = t', \\ \frac{1}{\widehat{\lambda}_t - \widehat{\lambda}_{t'}} & \text{otherwise.} \end{cases}$$

Then, for $t \in \{1, \dots, T\}$, we have:

$$\begin{aligned}
&{\lambda_t^{\Sigma_W}}'\Big((\mathbb{P}_{n_1}, \mathbb{P}_{n_2}), (\mathbb{P}_{n_1,q_1}, \mathbb{P}_{n_2,q_2})\Big) = \left(\widetilde{\mathbf{U}}'_{W,T} \mathbf{G}_W \widetilde{\mathbf{U}}_{W,T}\right)_{t,t} \\
&\left\langle f_t^{\Sigma_W}(\mathbb{P}_{n_1}, \mathbb{P}_{n_2}), \Delta'\Big((\mathbb{P}_{n_1}, \mathbb{P}_{n_2}), (\mathbb{P}_{n_1,q_1}, \mathbb{P}_{n_2,q_2})\Big) \right\rangle_{\mathcal{H}} = \left(\widetilde{\mathbf{U}}'_{W,T}(\mathbf{K}_{\mathbf{Y},\mathbf{Z}}\omega_{\mathbf{Z}} - \mathbf{K}_{\mathbf{Y}}\omega)\right)_t \\
&\left\langle {f_t^{\Sigma_W}}'\Big((\mathbb{P}_{n_1}, \mathbb{P}_{n_2}), (\mathbb{P}_{n_1,q_1}, \mathbb{P}_{n_2,q_2})\Big), \Delta(\mathbb{P}_{n_1}, \mathbb{P}_{n_2}) \right\rangle_{\mathcal{H}} = \left(\widetilde{\mathbf{U}}'_{W,T} \mathbf{G}_W \widetilde{\mathbf{U}}_{W,T} \Delta_{\mathbf{\Lambda}_{W,T}} \widetilde{\mathbf{U}}'_{W,T} \mathbf{K}_{\mathbf{Y}} \omega\right)_t.
\end{aligned}$$

We then use these quantities to compute the vector of partial Gateaux derivatives associated to each contribution of the KFDA statistic according to the formula in Theorem 19.

4.2.3 Illustration on the Reversion Dataset

We illustrate an application of influence functions by computing the influence of each observation to the pairwise comparisons of the Reversion dataset described in Chapters 2 and 3. Recall that the Reversion RTqPCR dataset contains four conditions 0H, 24H, 48HDIFF and 48HREV and that each conditions is divided in eight batches corresponding to manipulation repetitions. For a pair-wise comparison and a given value of the truncation parameter, we propose to visualize the influence of the observations on the contribution of the associated eigendirection of the within-group covariance operator with respect to their position on the discriminant axis associated to this truncation value. The results are shown in Figure 4.1.

We observe that important drops of the log p-value are associated to directions on which many observations have a large influence on the statistic. Oppositely, when the influence of all the cells is very low, the p-value tends to increase. The influence is a measure of

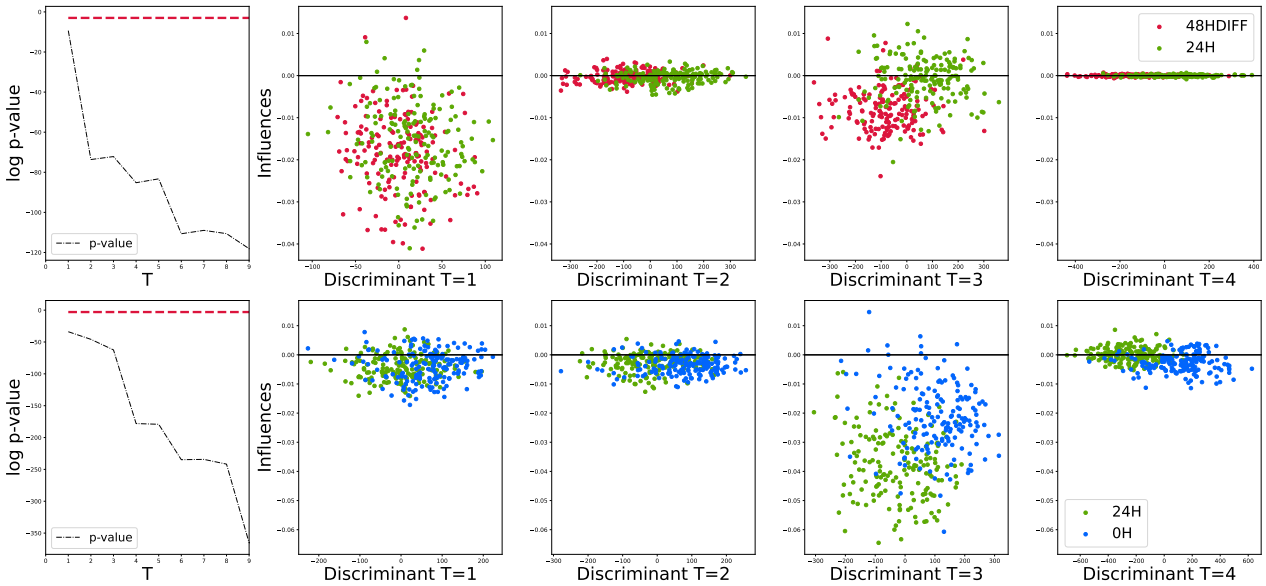

Figure 4.1: Pairwise comparison of conditions 48HDIFF versus 24H (top) and conditions 24H versus 0H (bottom). Left: log p-values with respect to T. Right: Influence of the observations on the contribution of the T^{th} eigendirection with respect to the position of their projection on the discriminant axis associated to a truncation parameter T for $T \in \{1, 2, 3, 4\}$, the null horizontal axis is represented in black.

how the absence of an observation would modify a statistic. Thus, if the influence of an observation on the test statistic is negative, it means that this observation advocates for a difference between the two samples, oppositely, if the influence of an observation is positive, it means that this observation participates to the similarity of the two datasets.

4.3 Conclusion

Interpretation of the Results on Kernel Testing

Applied to a kernel testing, Gateaux derivatives and influence functions can be used to identify the observations that advocate for the similarity of the two samples and those that advocate for a difference between the two samples. To do so, we propose to compute the influence of each observation on the value of the test statistic. As the MMD and KFDA statistics are based on the distance between both mean embeddings, each observation can influence the test to increase or decrease this distance. As single-cell data analysis often focuses on sub-groups of observations called sub-populations, we developed the concept of

Gateaux derivative in order to allow the investigation of the influence of a sub-population that could have been detected through an independant statistical analysis.

Future Work

Future work should be dedicated to the analysis of the level and power influence function and Gateaux derivative associated to kernel tests. These concepts have been introduced in [61] and recently applied for Wald-type tests in [48]. Another aspect that could be explored is the possible definitions of contaminating distributions for Gateaux derivative with a biological signification, such as the absence or presence of some sub-population, or different positions in a differentiation trajectory.

Conclusion

The work presented in this manuscript is mainly motivated by the advanced application of kernel testing to single-cell data science. We presented and implemented a complete framework for KFDA-based hypothesis testing, and proposed practical solutions to issues encountered, like reducing the computational cost of the procedure, interpreting the results or generalizing the test. This work opens many exciting research directions to deepen our understanding of unanswered questions, such as the asymptotic properties of the Nyström approximations or the interpretation of the model parameters in the kernel linear model. We propose a discussion in this last chapter on a selection of subjects that we consider as being future research directions.

Before discussing possible improvements, we introduce a discussion on a central aspect of kernel testing based on spectral regularizations of Hilbert-Schmidt operators that we only hinted so far: the theoretical importance of the truncation parameter and practical results or heuristics to fix its value.

Tuning the Truncation Parameter

We observed from numerical simulations shared in Chapter 2 that the asymptotic regime is reached for hundreds of observations, that is suited for applications on single-cell data. However, we also observed that the level of the test depends on the truncation parameter. One issue is to describe the expected risk of the test in general. The issues raised by the mathematical formulation of the task of choosing a truncation parameter that optimize the test performances can be formulated in terms of optimizing a trade-off between level and power. Some work has been done to describe the performances of the ridge regularized KFDA test performances in [60] but it remains to be done for the truncated KFDA test. Such an exploration could answer to the issue of having a controled type I error delivers an algorithm for the data-driven calibration of the truncation parameter.

Until then, a preliminary heuristic can be derived from some geometrical intuition on the effect of the truncation parameter. Here, we propose a geometrical interpretation of the difference captured by the kernel Fisher Discriminant axis with respect to the truncation parameter used to regularize the within-group covariance operator in the two-sample case. This description motivates the introduction of a last diagnostic tool that describes the geometry of the discrimination problem with respect to the truncation parameter that can be used to choose the truncation parameter in practice. While being focused on the two-sample case, this approach could be generalized to the Hotelling-Lawley test for general designs.

Geometric Considerations to Define Proper Alternatives

In truncated KFDA testing, the choice of the truncation parameter T in the regularization of the within-group covariance is critical since it underlie the definition of what we consider as a meaningful difference in complex situations. This parameter is representative of the complexity allowed to the non-linear transformation of the kernel Fisher Discriminant, as it stands for a delimitation between signal against noise.

First, it is clear that too large truncations may result in badly calibrated tests. By restricting the truncation to moderate values (less than 15 in the context of sinlge-cell data analysis), we can guarantee that the test is well calibrated, according to the several simulation studies performed in Chapter 2. Then, we observed that the evolution of the p-value with respect to the truncation parameter is not monotonous, as it highly depends on the geometry of the problem.

The kernel Fisher discriminant axis associated to the truncation parameter $T \geq 1$ is a sum of the contributions of the first T eigendirections of the empirical within-group covariance operator. Each contribution is colinear to the inner product between the eigenvector and the direction supported by the difference of the two empirical kernel mean embedding $\widehat{\mu}_1 - \widehat{\mu}_2$. In other words, eigendirections that are close to the direction of interest $\widehat{\mu}_1 - \widehat{\mu}_2$ contribute more to the discrimination. This geometrical interpretation with respect to the direction of interest $\widehat{\mu}_1 - \widehat{\mu}_2$ highlight the existence of at least two types of alternative that act as two blind spots of the discriminant axis associated to a truncation parameter T, that we call the orthogonal case and the parallel case.

- **Orthogonal case** : when the directions supporting the variability of each group are orthogonal to each other, each eigendirection of the within-group covariance is dedicated to the variability of one of the two conditions only. Thus, the resulting discriminant axis is not oriented in a direction able to spot the difference and fails to reject the null hypothesis until the truncation parameter is large enough to allow for a non-linear transformation that captures this orthogonality. This example is illustrated in Figure 4.2 with a cross dataset (two orthogonal bivariate Gaussians, forming a cross).

- **Parallel case** : when the two conditions share the same covariance stucture, but the difference between them is supported by a direction capturing little variance from the embeddings. The test fails to reject the null hypothesis until the truncation parameter T is large enough for the kernel Fisher Discriminant to consider this low variability direction. This situation is ambiguous, since it may also be interpreted as a not so important difference on a direction supporting little variance of the data. We illustrate this situation in Figure 4.3 with two parallel lines (two parallel bivariate Gaussians).

A Diagnostic Graph to Monitor the Effects of the Truncation Parameter

These two types of geometry can be detected by some quantities that can be monitored with respect to T. For instance, each eigendirection of the within-group covariance supports a part of the total difference $\widehat{\mu}_1 - \widehat{\mu}_2$ that can be quantified and allow to detect for a parallel case. Also, each eigendirection captures a part of the variability of each group, quantifying this captured variability can allow to detect an orthogonal case. Finally, as the Fisher Discriminant Analysis is defined as the optimization of the ratio between the captured difference and the captured variability, a discrimination score can be measured for each eigendirection as the ratio between the capture difference and the captured within-group variability. Thus, we propose a diagnostic graph to lead the data exploration by giving insights on the directions containing meaningful differences to look at and to ensure that the comparison is well interpreted. This diagnostic graph allows to monitor the captured variability of each group, the captured within-group variability, the captured difference, and the discrimination score with respect to T. We also add the p-value of the test statistic associated to each truncation parameter. We show the diagnostic graphs

associated to the two examples of geometric situations in Figures 4.2 and 4.3. Note that we split the diagnostics into several graphs for readability, but they can be gathered in one graph, as all the diagnostics have values comprised between 0 and 1.

General Perspectives on Kernel Testing

Kernel testing has proved to be particularly suited for the analysis of single-cell data. The underlying geometric properties and the diagnostics derived from the kernel linear model allow a wide range of insightful interpretations. However, a lot of work remains to be done.

Approximations of Kernel Test Statistics

The reduction of the computational cost of the kernel test statistics is a central concern of practical kernel testing. We focused on one version of the Nyström approximation to reduce the computational cost of the test statistic. It would be of interest to assess the performances of the test obtained through different approximations such as the alternative Nyström landmark choices or other approaches such as the Random Fourier Features [113]. In addition, kernel testing is a playground suited for the systematic comparison of these different approaches. From a theoretical point of view, it would be of interest to ensure that these approximation do not modify the asymptotic distributions of the kernel test statistics. To go further, if the asymptotic distribution is conserved, it would be possible to precisely study the asymptotic properties when both the number of observations and the number of landmarks increase.

Computational Aspects

The package `ktest` we implemented aims at giving a handy implementation of kernel testing. Several features could be enhanced to facilitate its use, and the range of available tests and approximation methods could be completed with existing alternatives. Moreover, the release of several tutorial notebooks and the schedule of courses are planned to reach an audience of non-statisticians. A nice evolution for the package `ktest` would be allow exterior contributors that are developing insightful tools for kernel methods.

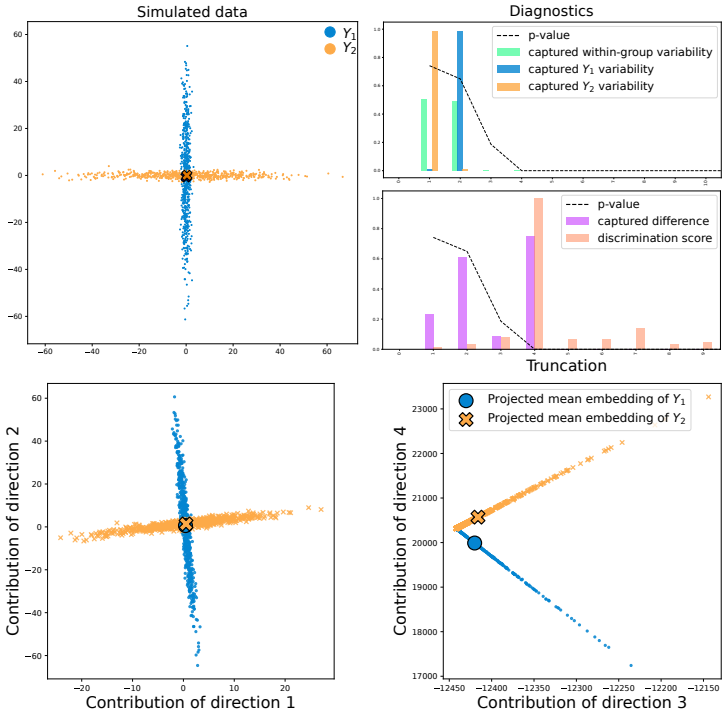

Figure 4.2: Comparison of two orthogonal bivariate Gaussians forming a cross. Upper-left : simulated data. Upper-right : diagnostic plots. Bottom: distributions of the embeddings on the first four eigendirections associated to the within-group covariance. Interpretation: observe on the diagnostic graphs that the truncation $t = 1$ captures almost all the variability of the second sample (captured Y_2 variability) and almost none of the variability of the first sample (captured Y_1 variability), and the truncation $t = 2$ captures the opposite. These directions have a low discrimination score as they do not allow to discriminate between the two positions, and we see on the graph of the contributions of directions 1 and 2 that these directions kind of reconstructed the cross. The direction having the highest discrimination score is associated to $t = 4$ and we see that this is for this value of t that the p-value falls below the $\alpha = 0.05$ threshold. We can see on the graph of contributions 3 and 4 that the direction associated to $t = 4$ is the first direction on which the two projected mean embeddings are separated.

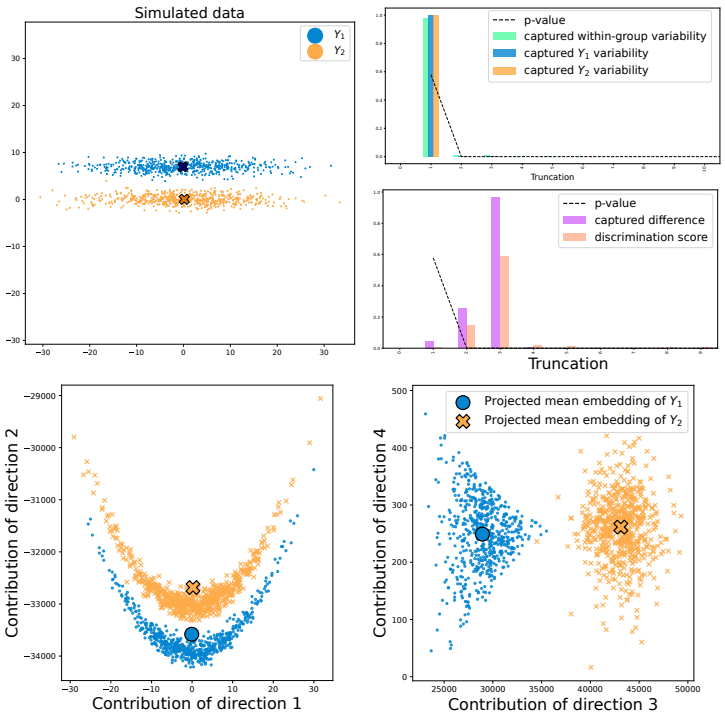

Figure 4.3: Comparison of two parallel bivariate Gaussians. Upper-left : simulated data. Upper-right : diagnostic plots. Bottom: distributions of the embeddings on the first four eigendirections associated to the within-group covariance. Interpretation: The direction associated to $t = 1$ captures almost all the variability of the samples but none of their difference and has thus a low discrimination score. The direction associated to $t = 2$ as a higher discrination score and it suffice to reject the null hypothesis, we see on its contribution that it is a non-linear square-like transformation of the direction $t = 1$. The direction associated to $t = 3$ confirms the difference with an even higher discrimination score.

Theoretical Enhancements

So far we only introduced the kernel linear model and an associated test. The test performances of the test on simulated data should be assessed in a near future. This model broadens the possible approaches to compare dataset and could be used complementary to the pair-wise comparisons allowed by the KFDA framework. The kernel linear model we defined is a favorable framework to introduce modelling aspects in kernel methods. In particular, the interpretation of the kernel linear model could be related to the notion of conditional embedding of distributions.

The kernel linear model associates each explanatory variable used in the model to an element of the RKHS that is the associated model parameter. These model parameters can be considered as the embeddings of the associated explanatory variables, and could be used in further analyses.

Influence of Observations and Variables

We introduced influence functions, that are a promising tool to understand the contribution of each observation to the results of a data analysis. The practical use of influence functions will necessitate to precisely define the notion of influential sub-populations and cells and the biological interpretations of the results. Especially when applied to kernel testing.

A drawback of kernel methods is that the dependance to the variables in lost with the embedding. It would be interesting to have a direct measure of which variables support a detected difference. Thus, complementary to influence functions, sensitivity analysis tools could be applied to kernel testing to measure the contribution of local differences to the global difference.

Specific Aspects of Single-Cell Data Analysis

Some issues specific to single-cell data analysis have not been examinated yet. For instance, kernel testing is non-parametric, but some probabilistic models have been proposed for single-cell data and could be used to improve the test performances. A way to introduce model priors on kernel testing consist in using kernels adapted to the model,

such as Fisher kernels that can be defined for any probability distribution. The Zero-Inflated Negative Binomial distribution is considered to be suited to model single-cell datasets, and it would be possible to determine a Fisher kernel adapted to this model.

We proposed kernel testing to perform global comparisons of single-cell datasets on the basis of all the variables. Transcriptomic activity is known to be related to Gene Regulatory Networks (GRN). As GRNs are groups of genes considered to have related acitvity, our approach allows to investigate the GRN-wise differences between conditions.

Spatial single-cell RNA sequencing datasets have the position of the cell in a tissue in addition to the transcriptomic information [136]. The position of a cell could be encoded in a kernel function or modeled in the kernel linear model to allow the analysis of such datasets with our comparison framework.

BIBLIOGRAPHY

[1] Divyansh Agarwal et al., "Distribution-Free Multisample Test Based on Optimal Matching with Applications to Single Cell Genomics", en, in: *arXiv* (June 2019), arXiv: 1906.04776.

[2] Md Ashad Alam, Kenji Fukumizu, and Yu-Ping Wang, "Influence function and robust variant of kernel canonical correlation analysis", in: *Neurocomputing* 304 (2018), Publisher: Elsevier, pp. 12–29.

[3] Ahmed El Alaoui and Michael W. Mahoney, "Fast Randomized Kernel Methods With Statistical Guarantees", in: *arXiv* (Nov. 2015), arXiv: 1411.0306.

[4] Robert A. Amezquita et al., "Orchestrating single-cell analysis with Bioconductor", in: *Nature methods* 17.*2* (2020), Publisher: Nature Publishing Group US New York, pp. 137–145.

[5] T. W. Anderson, "On the Distribution of the Two-Sample Cramer-von Mises Criterion", in: *The Annals of Mathematical Statistics* 33.*3* (Sept. 1962), Publisher: Institute of Mathematical Statistics, pp. 1148–1159.

[6] Tallulah S. Andrews and Martin Hemberg, "Identifying cell populations with scRNASeq", in: *Molecular aspects of medicine* 59 (2018), Publisher: Elsevier, pp. 114–122.

[7] Ilias Angelidis et al., "An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics", en, in: *Nature Communications* 10.*1* (Feb. 2019), Number: 1 Publisher: Nature Publishing Group, p. 963.

[8] Ery Arias-Castro, Bruno Pelletier, and Venkatesh Saligrama, "Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension", in: *Journal of Nonparametric Statistics* 30.*2* (2018), Publisher: Taylor & Francis, pp. 448–471.

[9] N. Aronszajn, "Theory of Reproducing Kernels", in: *Transactions of the American Mathematical Society* 68.*3* (1950), Publisher: American Mathematical Society, pp. 337–404.

[10] Francis Bach and Michael Jordan, "Learning spectral clustering", in: *Advances in neural information processing systems* 16 (2003).

[11] Francis R. Bach and Michael I. Jordan, "Kernel Independent Component Analysis", in: *Journal of Machine Learning Research* 3.*Jul* (2002), pp. 1–48.

[12] Francis R. Bach and Michael I. Jordan, "Predictive low-rank decomposition for kernel methods", en, in: *Proceedings of the 22nd international conference on Machine learning - ICML '05*, Bonn, Germany: ACM Press, 2005, pp. 33–40.

[13] Francis R. Bach, Gert R. G. Lanckriet, and Michael I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm", in: *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, New York, NY, USA: Association for Computing Machinery, July 2004, p. 6.

[14] Christopher T. H. Baker, *The Numerical Treatment of Integral Equations*, en, Google-Books-ID: B9sIAQAAIAAJ, Clarendon Press, 1977.

[15] Armando S. K. Balogoun, Guy M. Nkiet, and Carlos Ogouyandjou, "Kernel based method for the k -sample problem with functional data", en, in: *Communications in Statistics - Theory and Methods* (Dec. 2020), pp. 1–26.

[16] Trambak Banerjee, Bhaswar B. Bhattacharya, and Gourab Mukherjee, "A Nearest-Neighbor Based Nonparametric Test for Viral Remodeling in Heterogeneous Single-Cell Proteomic Data", en, in: *arXiv* (June 2020), arXiv: 2003.02937.

[17] Ludwig Baringhaus and Carsten Franz, "On a new multivariate two-sample test", in: *Journal of multivariate analysis* 88.*1* (2004), Publisher: Elsevier, pp. 190–206.

[18] Marek Bartosovic, Mukund Kabbe, and Gonçalo Castelo-Branco, "Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues", eng, in: *Nature Biotechnology* 39.*7* (July 2021), pp. 825–835.

[19] Benjamini et Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing on JSTOR*, 1995.

[20] Rajendra Bhatia and Ludwig Elsner, "The Hoffman-Wielandt inequality in infinite dimensions", en, in: *Proceedings of the Indian Academy of Sciences - Mathematical Sciences* 104.*3* (Aug. 1994), pp. 483–494.

[21] Peter J. Bickel, "A distribution free version of the Smirnov two sample test in the p-variate case", in: *The Annals of Mathematical Statistics* 40.*1* (1969), Publisher: JSTOR, pp. 1–23.

[22] Gilles Blanchard, Olivier Bousquet, and Laurent Zwald, "Statistical properties of kernel principal component analysis", en, in: *Machine Learning* 66.*2-3* (Mar. 2007), pp. 259–294.

[23] Jason D. Buenrostro et al., "Single-cell chromatin accessibility reveals principles of regulatory variation", eng, in: *Nature* 523.*7561* (July 2015), pp. 486–490.

[24] Andrew Butler et al., "Integrating single-cell transcriptomic data across different conditions, technologies, and species", in: *Nature biotechnology* 36.*5* (2018), Publisher: Nature Publishing Group US New York, pp. 411–420.

[25] M. Büttner et al., "scCODA is a Bayesian model for compositional single-cell data analysis", en, in: *Nature Communications* 12.*1* (Nov. 2021), Number: 1 Publisher: Nature Publishing Group, p. 6876.

[26] Eddie Cano-Gamez et al., "Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4+ T cells to cytokines", en, in: *Nature Communications* 11.*1* (Apr. 2020), p. 1801.

[27] Yue Cao et al., "scDC: single cell differential composition analysis", in: *BMC Bioinformatics* 20.*19* (Dec. 2019), p. 721.

[28] Kacper P Chwialkowski et al., "Fast Two-Sample Testing with Analytic Representations of Probability Measures", in: *Advances in Neural Information Processing Systems*, vol. 28, Curran Associates, Inc., 2015.

[29] Corinna Cortes and Vladimir Vapnik, "Support-vector networks", in: *Machine learning* 20 (1995), Publisher: Springer, pp. 273–297.

[30] Frank Critchley, "Influence in principal components analysis", in: *Biometrika* 72.*3* (Dec. 1985), pp. 627–636.

[31] Emma Dann et al., "Differential abundance testing on single-cell data using k-nearest neighbor graphs", en, in: *Nature Biotechnology* 40.*2* (Feb. 2022), pp. 245–253.

[32] Samarendra Das, Anil Rai, and Shesh N. Rai, "Differential Expression Analysis of Single-Cell RNA-Seq Data: Current Statistical Approaches and Outstanding Challenges", eng, in: *Entropy (Basel, Switzerland)* 24.*7* (July 2022), p. 995.

[33] Manuel Davy et al., "An online support vector machine for abnormal events detection", in: *Signal processing* 86.*8* (2006), Publisher: Elsevier, pp. 2009–2025.

[34] Michiel Debruyne, Mia Hubert, and Johan Van Horebeek, "Detecting influential observations in Kernel PCA", en, in: *Computational Statistics & Data Analysis* 54.*12* (Dec. 2010), pp. 3007–3019.

[35] Petros Drineas and Michael W. Mahoney, "On the Nystrom Method for Approximating a Gram Matrix for Improved Kernel-Based Learning", in: *Journal of Machine Learning Research* 6.*72* (2005), pp. 2153–2175.

[36] Jose A Dıaz-Garcıa and Graciela Gonzalez-Farıas, "Sensitivity analysis in linear regression", en, in: (2005).

[37] Mirjana Efremova et al., "CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes", en, in: *Nature Protocols* 15.*4* (Apr. 2020), Number: 4 Publisher: Nature Publishing Group, pp. 1484–1506.

[38] Greg Finak et al., "MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data", en, in: *Genome Biology* 16.*1* (Dec. 2015), p. 278.

[39] Shai Fine and Katya Scheinberg, "Efficient SVM Training Using Low-Rank Kernel Representations", in: *Journal of Machine Learning Research* 2.*Dec* (2001), pp. 243–264.

[40] C. Fowlkes et al., "Spectral grouping using the nystrom method", en, in: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.*2* (Feb. 2004), pp. 214–225.

[41] Jerome H. Friedman, "On multivariate goodness–of–fit and two–sample testing", in: *Statistical Problems in Particle Physics, Astrophysics, and Cosmology* 1 (2003), Publisher: Citeseer, p. 311.

[42] Jerome H. Friedman and Lawrence C. Rafsky, "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests", in: *The Annals of Statistics* (1979), Publisher: JSTOR, pp. 697–717.

[43] Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa, *Large sample analysis of the median heuristic*, arXiv:1707.07269 [math, stat], Oct. 2018.

[44] Marine Gauthier et al., *Distribution-free complex hypothesis testing for single-cell RNA-seq differential expression analysis*, en, bioRxiv, Nov. 2021.

[45] Charles Gawad, Winston Koh, and Stephen R. Quake, "Single-cell genome sequencing: current state of the science", en, in: *Nature Reviews Genetics* 17.*3* (Mar. 2016), pp. 175–188.

[46] Benyamin Ghojogh, Fakhri Karray, and Mark Crowley, "Fisher and Kernel Fisher Discriminant Analysis: Tutorial", en, in: *arXiv:1906.09436 [cs, stat]* (June 2019), arXiv: 1906.09436.

[47] Promit Ghosal and Bodhisattva Sen, "Multivariate ranks and quantiles using optimal transport: Consistency, rates and nonparametric testing", in: *The Annals of Statistics* 50.*2* (2022), Publisher: Institute of Mathematical Statistics, pp. 1012–1037.

[48] Abhik Ghosh et al., "Influence analysis of robust Wald-type tests", in: *Journal of Multivariate Analysis* 147 (2016), Publisher: Elsevier, pp. 102–126.

[49] Christophe Giraud, *Introduction to high-dimensional statistics*, CRC Press, 2021.

[50] Alex Gittens and Michael W. Mahoney, "Revisiting the Nystrom Method for Improved Large-Scale Machine Learning", in: *arXiv:1303.1849 [cs]* (June 2013), arXiv: 1303.1849.

[51] Arthur Gretton et al., "A Fast, Consistent Kernel Two-Sample Test", in: *Advances in Neural Information Processing Systems*, vol. 22, Curran Associates, Inc., 2009.

[52] Arthur Gretton et al., "A Kernel Method for the Two-Sample-Problem", in: *Advances in Neural Information Processing Systems*, vol. 19, MIT Press, 2006.

[53] Arthur Gretton et al., "A Kernel Statistical Test of Independence", in: *Advances in Neural Information Processing Systems*, vol. 20, Curran Associates, Inc., 2007.

[54] Arthur Gretton et al., "A Kernel Two-Sample Test", in: *Journal of Machine Learning Research* 13.*25* (2012), pp. 723–773.

[55] Arthur Gretton et al., "Measuring Statistical Dependence with Hilbert-Schmidt Norms", en, in: *Algorithmic Learning Theory*, ed. by Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita, Lecture Notes in Computer Science, Berlin, Heidelberg: Springer, 2005, pp. 63–77.

[56] Arthur Gretton et al., "Optimal kernel choice for large-scale two-sample tests", in: *Advances in Neural Information Processing Systems*, vol. 25, Curran Associates, Inc., 2012.

[57] Kevin Grosselin et al., “High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer”, eng, in: *Nature genetics* 51.*6* (June 2019), pp. 1060–1066.

[58] Christoph Hafemeister and Rahul Satija, “Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression”, in: *Genome Biology* 20.*1* (Dec. 2019), p. 296.

[59] Tzachi Hagai et al., “Gene expression variability across cells and species shapes innate immunity”, en, in: *Nature* 563.*7730* (Nov. 2018), pp. 197–202.

[60] Omar Hagrass, Bharath K. Sriperumbudur, and Bing Li, *Spectral Regularized Kernel Two-Sample Tests*, arXiv:2212.09201 [cs, math, stat], Dec. 2022.

[61] Frank R. Hampel et al., *Robust Statistics: The Approach Based on Influence Functions*, en, John Wiley & Sons, 1981.

[62] Zaid Harchaoui et al., “A regularized kernel-based approach to unsupervised audio segmentation”, en, in: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan: IEEE, Apr. 2009, pp. 1665–1668.

[63] Zaid Harchaoui et al., “Kernel-Based Methods for Hypothesis Testing: A Unified View”, en, in: *IEEE Signal Processing Magazine* 30.*4* (July 2013), pp. 87–97.

[64] Zaïd Harchaoui, Francis Bach, and Eric Moulines, “Testing for homogeneity with kernel Fisher discriminant analysis”, in: *Advances in Neural Information Processing Systems* 20 (2007).

[65] Lajos Horváth and Piotr Kokoszka, *Inference for Functional Data with Applications*, en, Google-Books-ID: OVezLB___ZpYC, Springer Science & Business Media, May 2012.

[66] Harold Hotelling, “The Generalization of Student’s Ratio”, in: *The Annals of Mathematical Statistics* 2.*3* (Aug. 1931), Publisher: Institute of Mathematical Statistics, pp. 360–378.

[67] Su-Yun Huang, Yi-Ren Yeh, and Shinto Eguchi, “Robust Kernel Principal Component Analysis”, en, in: *Neural Computation* 21.*11* (Nov. 2009), pp. 3179–3213.

[68] Diego Adhemar Jaitin et al., “Massively parallel single cell RNA-Seq for marker-free decomposition of tissues into cell types”, in: *Science (New York, N.Y.)* 343.*6172* (Feb. 2014), pp. 776–779.

[69] Tony Jebara, Risi Kondor, and Andrew Howard, “Probability Product Kernels”, in: *Journal of Machine Learning Research* 5.*Jul* (2004), pp. 819–844.

[70] Wittawat Jitkrittum et al., “Interpretable Distribution Features with Maximum Testing Power”, en, in: *arXiv:1605.06796 [cs, stat]* (Oct. 2016), arXiv: 1605.06796.

[71] Richard Arnold Johnson and Dean W. Wichern, “Applied multivariate statistical analysis”, in: (2002), Publisher: Prentice hall Upper Saddle River, NJ.

[72] Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama, “f-Divergence Estimation and Two-Sample Homogeneity Test Under Semiparametric Density-Ratio Models”, in: *IEEE Transactions on Information Theory* 58.*2* (Feb. 2012), Conference Name: IEEE Transactions on Information Theory, pp. 708–720.

[73] Peter V Kharchenko, Lev Silberstein, and David T Scadden, “Bayesian approach to single-cell differential expression analysis”, en, in: *Nature Methods* 11.*7* (July 2014), pp. 740–742.

[74] Ilmun Kim, Sivaraman Balakrishnan, and Larry Wasserman, “Robust multivariate nonparametric tests via projection averaging”, in: *The Annals of Statistics* 48.*6* (Dec. 2020), Publisher: Institute of Mathematical Statistics, pp. 3417–3441.

[75] Ilmun Kim et al., “Classification accuracy as a proxy for two-sample testing”, in: *The Annals of Statistics* 49.*1* (2021), pp. 411–434.

[76] JooSeuk Kim and Clayton D. Scott, “Robust kernel density estimation”, in: *The Journal of Machine Learning Research* 13.*1* (2012), Publisher: JMLR. org, pp. 2529–2565.

[77] Matthias Kirchler et al., “Two-sample testing using deep learning”, in: *International Conference on Artificial Intelligence and Statistics*, PMLR, 2020, pp. 1387–1398.

[78] Keegan D. Korthauer et al., “A statistical approach for identifying differential distributions in single-cell RNA-seq experiments”, en, in: *Genome Biology* 17.*1* (Dec. 2016), p. 222.

[79] Jonas M. Kübler et al., “A Witness Two-Sample Test”, en, in: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, ISSN: 2640-3498, PMLR, May 2022, pp. 1403–1419.

[80] Gioele La Manno et al., “RNA velocity of single cells”, in: *Nature* 560.*7719* (2018), Publisher: Nature Publishing Group, pp. 494–498.

[81] Quoc Le, Tamás Sarlós, and Alex Smola, "Fastfood-approximating kernel expansions in loglinear time", in: *Proceedings of the international conference on machine learning*, vol. 85, 2013, p. 8.

[82] Erich Leo Lehmann, Joseph P. Romano, and George Casella, *Testing statistical hypotheses*, vol. 3, Springer, 1986.

[83] Chengtao Li, Stefanie Jegelka, and Suvrit Sra, "Fast dpp sampling for nystrom with application to kernel methods", in: *International Conference on Machine Learning*, PMLR, 2016, pp. 2061–2070.

[84] Mingsheng Long et al., "Transfer Feature Learning with Joint Distribution Adaptation", in: *2013 IEEE International Conference on Computer Vision*, ISSN: 2380-7504, Dec. 2013, pp. 2200–2207.

[85] David Lopez-Paz and Maxime Oquab, *Revisiting Classifier Two-Sample Tests*, arXiv:1610.06545 [stat], Mar. 2018.

[86] Michael I Love, Wolfgang Huber, and Simon Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2", en, in: *Genome Biology* 15.*12* (Dec. 2014), p. 550.

[87] Malte D. Luecken and Fabian J. Theis, "Current best practices in single-cell RNA-seq analysis: a tutorial", in: *Molecular systems biology* 15.*6* (2019), e8746.

[88] Laurens van der Maaten and Geoffrey Hinton, "Visualizing Data using t-SNE", in: *Journal of Machine Learning Research* 9.*86* (2008), pp. 2579–2605.

[89] Evan Z. Macosko et al., "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets", en, in: *Cell* 161.*5* (May 2015), pp. 1202–1214.

[90] Henry B. Mann and Donald R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other", in: *The annals of mathematical statistics* (1947), Publisher: JSTOR, pp. 50–60.

[91] Raphael Margueron et al., "Role of the polycomb protein Eed in the propagation of repressive histone marks", in: *Nature* 461.*7265* (Oct. 2009), pp. 762–767.

[92] Justine Marsolier et al., "H3K27me3 conditions chemotolerance in triple-negative breast cancer", en, in: *Nature Genetics* 54.*4* (Apr. 2022), pp. 459–468.

[93] Frank J. Massey Jr, "The Kolmogorov-Smirnov test for goodness of fit", in: *Journal of the American statistical Association* 46.*253* (1951), Publisher: Taylor & Francis, pp. 68–78.

[94] Colin McDiarmid, "On the method of bounded differences", in: *Surveys in combinatorics* 141.*1* (1989), Publisher: Norwich, pp. 148–188.

[95] Christopher S. McGinnis, Lyndsay M. Murrow, and Zev J. Gartner, "DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors", in: *Cell systems* 8.*4* (2019), Publisher: Elsevier, pp. 329–337.

[96] Leland McInnes et al., "UMAP: Uniform Manifold Approximation and Projection", en, in: *Journal of Open Source Software* 3.*29* (Sept. 2018), p. 861.

[97] James Mercer, "Xvi. functions of positive and negative type, and their connection 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.*441-458* (1909), Publisher: The Royal Society London, pp. 415–446.

[98] Zhun Miao et al., "DEsingle for detecting three types of differential expression in single-cell RNA-seq data", en, in: *Bioinformatics* 34.*18* (Sept. 2018), ed. by Bonnie Berger, pp. 3223–3224.

[99] Sebastian Mika et al., "Fisher discriminant analysis with kernels", in: *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468)*, Ieee, 1999, pp. 41–48.

[100] Krikamol Muandet et al., "Kernel Mean Embedding of Distributions: A Review and Beyond", en, in: *Foundations and Trends® in Machine Learning* 10.*1-2* (2017), arXiv: 1605.09522, pp. 1–141.

[101] Somabha Mukherjee et al., "Distribution-Free Multisample Tests Based on Optimal Matchings With Applications to Single Cell Genomics", in: *Journal of the American Statistical Association* 117.*538* (Apr. 2022), pp. 627–638.

[102] Cameron Musco and Christopher Musco, "Recursive Sampling for the Nystrom Method", en, in: *arXiv:1605.07583 [cs, stat]* (Nov. 2017), arXiv: 1605.07583.

[103] Dino Oglic and Thomas Gärtner, "Nyström Method with Kernel K-means++ Samples as Landmarks", en, in: *Proceedings of the 34th International Conference on Machine Learning*, ISSN: 2640-3498, PMLR, July 2017, pp. 2652–2660.

[104] David J. Olive, David J. Olive, and Chernyk, *Robust multivariate analysis*, Springer, 2017.

[105] Marie Ouimet and Yoshua Bengio, "Greedy Spectral Embedding", en, in: *International Workshop on Artificial Intelligence and Statistics*, ISSN: 2640-3498, PMLR, Jan. 2005, pp. 253–260.

[106] Anthony Ozier-Lafontaine et al., "Kernel-Based Testing for Single-Cell Differential Analysis", in: *arXiv preprint arXiv:2307.08509* (2023).

[107] Sinno Jialin Pan et al., "Domain Adaptation via Transfer Component Analysis", in: *IEEE Transactions on Neural Networks* 22.*2* (Feb. 2011), Conference Name: IEEE Transactions on Neural Networks, pp. 199–210.

[108] Alessia Pini, Aymeric Stamm, and Simone Vantini, "Hotelling's T 2 in separable Hilbert spaces", en, in: *Journal of Multivariate Analysis* 167 (Sept. 2018), pp. 284–305.

[109] Ana M. Pires and João A. Branco, "Partial Influence Functions", en, in: *Journal of Multivariate Analysis* 83.*2* (Nov. 2002), pp. 451–468.

[110] John Platt, "FastMap, MetricMap, and Landmark MDS are all Nyström Algorithms", en, in: *International Workshop on Artificial Intelligence and Statistics*, ISSN: 2640-3498, PMLR, Jan. 2005, pp. 261–268.

[111] Sebastian Pott and Jason D. Lieb, "Single-cell ATAC-seq: strength in numbers", in: *Genome Biology* 16.*1* (Aug. 2015), p. 172.

[112] Luke A. Prendergast and Jodie A. Smith, "Influence functions for linear discriminant analysis: Sensitivity analysis and efficient influence diagnostics", in: *Journal of Multivariate Analysis* 190 (July 2022), p. 104993.

[113] Ali Rahimi and Benjamin Recht, "Random Features for Large-Scale Kernel Machines", in: *Advances in Neural Information Processing Systems*, vol. 20, Curran Associates, Inc., 2007.

[114] Ali Rahimi and Benjamin Recht, "Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning", in: *Advances in Neural Information Processing Systems 21*, ed. by D. Koller et al., Curran Associates, Inc., 2009, pp. 1313–1320.

[115] Aaditya Ramdas et al., "On the Decreasing Power of Kernel and Distance Based Nonparametric Hypothesis Tests in High Dimensions", en, in: *Proceedings of the AAAI Conference on Artificial Intelligence* 29.*1* (Mar. 2015), Number: 1.

[116] Paul A. Reyfman et al., "Single-Cell Transcriptomic Analysis of Human Lung Provides Insights into the Pathobiology of Pulmonary Fibrosis", eng, in: *American Journal of Respiratory and Critical Care Medicine* 199.*12* (June 2019), pp. 1517–1536.

[117] Angélique Richard et al., "Single-Cell-Based Analysis Highlights a Surge in Cell-to-Cell Molecular Variability Preceding Irreversible Commitment in a Differentiation Process", en, in: *PLOS Biology* 14.*12* (Dec. 2016), ed. by Sarah A. Teichmann, e1002585.

[118] Matthew E. Ritchie et al., "limma powers differential expression analyses for RNA-sequencing and microarray studies", eng, in: *Nucleic Acids Research* 43.*7* (Apr. 2015), e47.

[119] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data", en, in: *Bioinformatics* 26.*1* (Jan. 2010), pp. 139–140.

[120] Assaf Rotem et al., "Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state", en, in: *Nature Biotechnology* 33.*11* (Nov. 2015), pp. 1165–1172.

[121] meyer scetbon and Gael Varoquaux, "Comparing distributions: \textbackslash ell_1 geometry improves kernel two-sample testing", in: *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.

[122] Roman Schefzik, Julian Flesch, and Angela Goncalves, "Fast identification of differential distributions in single-cell RNA-sequencing data with waddR", en, in: *Bioinformatics* 37.*19* (Oct. 2021), ed. by Anthony Mathelier, pp. 3204–3211.

[123] Antonin Schrab et al., *MMD Aggregated Two-Sample Test*, arXiv:2110.15073 [cs, math, stat], June 2022.

[124] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller, "Kernel principal component analysis", in: *International conference on artificial neural networks*, Springer, 1997, pp. 583–588.

[125] Bernhard Schölkopf and Alexander J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT press, 2002.

[126] Piercesare Secchi, Aymeric Stamm, and Simone Vantini, "Inference for the mean of large p small n data: A finite-sample high-dimensional generalization of Hotelling's theorem", en, in: *Electronic Journal of Statistics* 7.*none* (Jan. 2013).

[127] Dino Sejdinovic et al., "Equivalence of Distance-Based and Rkhs-Based Statistics in Hypothesis Testing", in: *The Annals of Statistics* 41.*5* (2013), Publisher: Institute of Mathematical Statistics, pp. 2263–2291.

[128] John Shawe-Taylor, "On the Eigenspectrum of the Gram Matrix and the Generalisation Error of Kernel PCA", en, in: *IEEE Transactions on Information Theory* 51.*7* (2005), pp. 2510–2522.

[129] John Shawe-Taylor and Nello Cristianini, *Kernel Methods for Pattern Analysis*, en, Cambridge University Press, New York, NY, USA, June 2004.

[130] Efrat Shema, Bradley E. Bernstein, and Jason D. Buenrostro, "Single-cell and single-molecule epigenomics to uncover genome regulation at unprecedented resolution", eng, in: *Nature Genetics* 51.*1* (Jan. 2019), pp. 19–25.

[131] Carl-Johann Simon-Gabriel and Bernhard Schölkopf, "Kernel Distribution Embeddings: Universal Kernels, Characteristic Kernels and Kernel Metrics on Distributions", in: *Journal of Machine Learning Research* 19.*44* (2018), pp. 1–29.

[132] Nikolai V. Smirnov, "On the estimation of the discrepancy between empirical curves of distribution for two independent samples", in: *Bull. Math. Univ. Moscou* 2.*2* (1939), pp. 3–14.

[133] Alex J. Smola and Bernhard Schölkopf, "Sparse Greedy Matrix Approximation for Machine Learning", in: Morgan Kaufmann, 2000, pp. 911–918.

[134] Jordan W. Squair et al., "Confronting false discoveries in single-cell differential expression", en, in: *Nature Communications* 12.*1* (Sept. 2021), p. 5692.

[135] Student, "The probable error of a mean", in: *Biometrika* 6.*1* (1908), Publisher: Oxford University Press, pp. 1–25.

[136] Patrik L. Ståhl et al., "Visualization and analysis of gene expression in tissue sections by spatial transcriptomics", in: *Science* 353.*6294* (2016), Publisher: American Association for the Advancement of Science, pp. 78–82.

[137] Valentine Svensson, "Droplet scRNA-seq is not zero-inflated", en, in: *Nature Biotechnology* 38.*2* (Feb. 2020), Number: 2 Publisher: Nature Publishing Group, pp. 147–150.

[138] Gábor J. Székely and Maria L. Rizzo, "Testing for equal distributions in high dimension", in: *InterStat* 5.*16.10* (2004), Publisher: Citeseer, pp. 1249–1272.

[139] Simone Tiberi et al., *distinct: a novel approach to differential distribution analyses*, en, bioRxiv, Apr. 2022.

[140] Itay Tirosh et al., "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq", in: *Science* 352.*6282* (2016), Publisher: American Association for the Advancement of Science, pp. 189–196.

[141] Stephen Tu et al., "Large Scale Kernel Learning using Block Coordinate Descent", in: *arXiv:1602.05310 [cs, math, stat]* (Feb. 2016), arXiv: 1602.05310.

[142] Hugues Van Assel et al., "A Probabilistic Graph Coupling View of Dimension Reduction", en, in: *Advances in Neural Information Processing Systems* 35 (Dec. 2022), pp. 10696–10708.

[143] Isaac Virshup et al., "The scverse project provides a computational ecosystem for single-cell omics data analysis", in: *Nature biotechnology* (2023), Publisher: Nature Publishing Group US New York, pp. 1–3.

[144] Abraham Wald and Jacob Wolfowitz, "On a test whether two samples are from the same population", in: *The Annals of Mathematical Statistics* 11.*2* (1940), Publisher: JSTOR, pp. 147–162.

[145] Tianyu Wang and Sheida Nabavi, "SigEMD: A powerful method for differential gene expression analysis in single-cell RNA sequencing data", en, in: *Methods*, Data mining methods for analyzing biological data in terms of phenotypes 145 (Aug. 2018), pp. 25–32.

[146] Lionel Weiss, "Two-sample tests for multivariate distributions", in: *The Annals of Mathematical Statistics* 31.*1* (1960), Publisher: Institute of Mathematical Statistics, pp. 159–164.

[147] Christopher K. I. Williams and Matthias Seeger, "Using the Nystrom Method to Speed Up Kernel Machines", in: *Advances in Neural Information Processing Systems 13*, ed. by T. K. Leen, T. G. Dietterich, and V. Tresp, MIT Press, 2001, pp. 682–688.

[148] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis, "SCANPY: large-scale single-cell gene expression data analysis", in: *Genome biology* 19 (2018), Publisher: Springer, pp. 1–5.

[149] Hongliang Yan et al., “Mind the Class Weight Bias: Weighted Maximum Mean Discrepancy for Unsupervised Domain Adaptation”, en, in: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, July 2017, pp. 945–954.

[150] Matthew D. Young and Sam Behjati, “SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data”, in: *Gigascience* 9.*12* (2020), Publisher: Oxford University Press, giaa151.

[151] Kai Zhang, Ivor W. Tsang, and James T. Kwok, “Improved Nyström low-rank approximation and error analysis”, en, in: *Proceedings of the 25th international conference on Machine learning - ICML ’08*, Helsinki, Finland: ACM Press, 2008, pp. 1232–1239.

[152] Grace X. Y. Zheng et al., “Massively parallel digital transcriptional profiling of single cells”, eng, in: *Nature Communications* 8 (Jan. 2017), p. 14049.

[153] Souad Zreika et al., “Evidence for close molecular proximity between reverting and undifferentiated cells”, in: *BMC Biology* 20.*1* (July 2022), p. 155.

[154] Laurent Zwald and Gilles Blanchard, “On the Convergence of Eigenspaces in Kernel Principal Component Analysis”, in: *Advances in Neural Information Processing Systems*, ed. by Y. Weiss, B. Schölkopf, and J. Platt, vol. 18, MIT Press, 2005.

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

CENTRALE NANTES

Titre : Tests à noyaux, et leurs applications aux données de séquençage en cellule unique

Mot clés : tests à noyaux, séquençage en cellule unique, modèle linéaire à noyaux, Nyström, fonctions d'influence

Résumé : Les technologies de sequençage en cellule unique mesurent des informations à l'échelle de chaque cellule d'une population. Les données issues de ces technologies présentent de nombreux défis : beaucoup d'observations en grande dimension et souvent parcimonieuses. De nombreuses expériences de biologie consistent à comparer des conditions. L'objet de la thèse est de développer un ensemble d'outils qui compare des échantillons de données issues des technologies de séquençage en cellule unique afin de détecter et décrire les différences qui existent. Pour cela, nous proposons d'appliquer les tests de comparaison de deux échantillons basés sur les méthodes à noyaux existants. Nous proposons de généraliser ces tests à noyaux pour les designs expérimentaux quelconques, ce test s'inspire du test de la trace de Hotelling-Lawley. Nous implémentons pour la première fois ces tests à noyaux dans un package R et Python nommé *ktest*, et nos applications sur données simulées et issues d'expériences démontrent leurs performances. L'application de ces méthodes à des données expérimentales permet d'identifier les observations qui expliquent les différences détectées. Enfin, nous proposons une implémentation efficace de ces tests basée sur des factorisations matricielles de type Nyström, ainsi qu'un ensemble d'outils de diagnostic et d'interprétation des résultats pour rendre ces méthodes accessibles et compréhensibles par des non-spécialistes.

Title: Kernel-based testing and their applications to single-cell data

Keywords: kernel testing, single-cell, kernel linear model, Nyström, influence functions

Abstract: Single-cell technologies generate data at the single-cell level. They are coumposed of hundreds to thousands of observations (i.e. cells) and tens of thousands of variables (i.e. genes). New methodological challenges arose to fully exploit the potentialities of these complex data. A major statistical challenge is to distinguish biological information from technical noise in order to compare conditions or tissues. This thesis explores the application of kernel testing on single-cell datasets in order to detect and describe the potential differences between compared conditions. To overcome the limitations of existing kernel two-sample tests, we propose a kernel test inspired from the Hotelling-Lawley test that can apply to any experimental design. We implemented these tests in a R and Python package called *ktest* that is their first user-oriented implementation. We demonstrate the performances of kernel testing on simulated datasets and on various experimental single-cell datasets. The geometrical interpretations of these methods allows to identify the observations leading a detected difference. Finally, we propose a Nyström-based efficient implementation of these kernel tests as well as a range of diagnostic and interpretation tools.