

UNIVERSITÀ COMMERCIALE “LUIGI BOCCONI”
PHD SCHOOL

PHD PROGRAM: Statistics and Computer Science

CYCLE: XXXVII

DISCIPLINARY FIELD: SECS-S/01

HILBERT-BASED INDICES OF DEPENDENCE
BETWEEN RANDOM PROBABILITY MEASURES
FOR DISTRIBUTIONAL DATA AND BAYESIAN
NONPARAMETRICS

ADVISOR: *Hugo Lavenant*

CO-ADVISOR: *Marta Catalano*

PhD Thesis by *Francesco Mascari*

ID NUMBER: 3000721

Year: 2026

Acknowledgements

I would like to express my sincere gratitude to Hugo Lavenant and Marta Catalano for their guidance, insight, and continuous support throughout these years. Their intellectual curiosity and sharp perspectives have profoundly influenced my way of thinking and the development of this work. I am equally grateful for their empathy and encouragement, which helped me believe in myself and made this period particularly meaningful for my personal growth.

I would also like to thank the members of the Doctoral Committee for their encouraging reviews and valuable suggestions, and all Faculty members of the Decision Sciences and Computing Sciences Departments at Bocconi University for their stimulating lessons, constructive feedback, and thoughtful advice. I am particularly grateful to Omiros Papaspiliopoulos for the opportunity to teach as an instructor, as well as to all Faculty members with whom I worked as a teaching assistant.

I would like to extend my thanks to my colleagues and friends at Bocconi University. A special mention goes to Giovanni, Beatrice, Davide, Andrea, Brandon, Federico, Veronica, Patric, Claudio, Chiara, Alessandro, George, Alessandra, Beatrice, Francesco, Pieralberto, Filippo, and Raffaella. Thank you for the joyful moments together.

Finally, I wish to thank my loved ones for their unwavering support and understanding throughout this journey, and for the countless hugs that reminded me I was never alone in it.

Abstract

This thesis develops a general and practical framework for defining, estimating, and interpreting indices of dependence between random probability measures. The core idea is to embed each probability measure into a Hilbert space, enabling the application of classical multivariate dependence tools to the space of distributions. Building on this representation, we introduce a family of model-free indices of dependence that varies with the choice of the embedding and the choice of a summary of the induced cross-covariance operator. We focus on Wasserstein-based and kernel-based embeddings, characterizing the theoretical and numerical behavior of the indices at their extreme values. When the kernel mean embedding is composed with the trace of the cross-covariance operator, we obtain an index that generalizes the widely used set-wise correlation in Bayesian Nonparametrics, exhibiting several notable properties that do not hold for the standard set-wise correlation. Applications of our Hilbert-based indices include the hierarchical clustering of functional brain imaging, the study of the Hierarchical Dirichlet Process a posteriori, and the development of a principled model comparison between parametric and nonparametric Bayesian models.

Contents

Introduction	1
1 Measuring Stochastic Dependence	9
1.1 Dependence Measures for Random Variables	9
1.2 Hilbert-Valued Random Variables and Measures of Dependence	14
Discussion	25
Appendices	27
1.A Proofs of the Statements	27
2 Hilbert-Based Correlation Indices for Random Probability Measures	31
2.1 Introduction	32
2.2 Correlation Indices for Hilbert Embeddings of Measures	35
2.3 Hilbert Embeddings of Random Variables	37
2.4 Numerical Validation	46
2.5 Application: Hierarchical Clustering on Resting-State Functional Magnetic Resonance Imaging	54
Discussion	56
Acknowledgments	58
Appendices	59
2.A Proofs of the Statements	59
2.B Examples	62
2.C Numerical Implementation of the Indices	68

3	Measuring Exchangeability with Reproducing Kernel Hilbert Spaces	73
3.1	Introduction	75
3.2	Kernel Correlation	79
3.3	Detecting Full Exchangeability	83
3.4	Estimation from Samples	85
3.5	From Set-Wise to Kernel Correlation	87
3.6	Kernel Correlation for Mixture Models	90
3.7	Hierarchical Dirichlet Process	93
3.8	Numerical Simulations	96
3.9	Model Comparison	100
	Discussion	104
	Appendices	107
3.A	Examples	107
3.B	Proofs of the Statements	110
3.C	Distributional Properties of the Hierarchical Dirichlet Process for Proofs and Simulations	122
3.D	Algorithms for Numerical Simulations	125
3.E	Computations for Model Comparison	132
	References	135

Introduction

The study of dependence between random probability measures is a topic of growing interest in Statistics and Machine Learning. In particular, Bayesian Nonparametrics provides a natural motivation, since quantifying the borrowing of information is crucial in nonparametric Bayesian models for grouped data. Similarly, in the analysis of distributional data, where each observation is a probability distribution and can be modeled as a realization of a random probability measure, quantifying dependence between random probability measures is therefore essential for understanding relations across multiple samples. In this introduction, we outline these motivations and present the contributions of this thesis.

Bayesian modeling is widely embraced when data exhibit related groups, thanks to its flexibility and natural shrinkage effect ([Lindley and Smith, 1972](#); [Efron and Morris, 1973](#); [Gelman et al., 2003](#); [Gelman and Hill, 2007](#)). A natural hypothesis for grouped data is partial exchangeability ([de Finetti, 1937](#)), which assumes exchangeability of the observations within each group but not across different groups. This assumption is particularly relevant in hierarchical settings, where data are collected from multiple sources or populations that may share some common characteristics while exhibiting group-specific variations. Formally, we say that observations are partially exchangeable when their distribution is invariant under permutations within each group, i.e., for d groups of

observations, $(X_{i,j})_{j \in \mathbb{N}}$ for $i = 1, \dots, d$, we have

$$\left((X_{1,j})_{j \in \mathbb{N}}, \dots, (X_{d,j})_{j \in \mathbb{N}} \right) \stackrel{d}{=} \left((X_{1,\pi_1(j)})_{j \in \mathbb{N}}, \dots, (X_{d,\pi_d(j)})_{j \in \mathbb{N}} \right),$$

for $\pi_1, \dots, \pi_d$ finite permutations of $\mathbb{N}$, where $\stackrel{d}{=}$ denotes equality in distribution.

The relation between partially exchangeable sequences and multilevel models is established by de Finetti's theorem ([de Finetti, 1938](#)), which states that any partially exchangeable sequence can be represented as many conditionally independent and identically distributed sequences. This leads to the general class of multilevel models of the form

$$X_{i,j} | (\tilde{P}_1, \dots, \tilde{P}_d) \stackrel{\text{i.i.d.}}{\sim} \tilde{P}_i, \quad (\tilde{P}_1, \dots, \tilde{P}_d) \sim Q, \quad (1)$$

for $i = 1, \dots, d$ and $j \in \mathbb{N}$, where $\tilde{P}_1, \dots, \tilde{P}_d$ are d random probability measures defined on the same space having joint prior distribution Q .

Measuring dependence among the random probability measures $\tilde{P}_1, \dots, \tilde{P}_d$ is essential for understanding the behavior of the multilevel model in (1). Particularly, the dependence structure induced by Q across $\tilde{P}_1, \dots, \tilde{P}_d$ allows for a *borrowing of information*, a concept first introduced by Tukey as the need “to borrow strength from either other aspects of the same body of data or from other bodies of data” ([Tukey, 1972](#)). As a limiting case, when $\tilde{P}_1 = \dots = \tilde{P}_d$ almost surely (a.s.), both the dependence and the borrowing are maximal: since the observations are fully exchangeable, i.e., the observations in a group carry the same information as observations in the other groups. When $\tilde{P}_1, \dots, \tilde{P}_d$ are independent, there is no borrowing of information, i.e., the observations in the other groups will not affect the group-specific inference. Thus, quantifying the strength of dependence among $\tilde{P}_1, \dots, \tilde{P}_d$ induced by Q can be helpful in assessing model fit, guiding the choice of prior distributions, and comparing different models. For instance, by quantifying the dependence induced by different choices of Q , practitioners can make informed decisions about which prior is most appropriate for their specific application.

There exists a plethora of nonparametric models for partially exchangeable observations in the Bayesian literature. The interested reader may refer to [Quintana et al. \(2022\)](#) and [Wade and Inácio \(2025\)](#) for a complete review. However, there is comparatively less work on quantifying how these models induce dependence between random probability measures. So far, there are two primary methods for this scope. The first method is the set-wise correlation $\text{Corr}(\tilde{P}_1(A), \tilde{P}_2(A))$, for any measurable set A , which is the most used in practice due to its tractable and interpretable expression for the majority of Bayesian nonparametric models ([Franzolini et al., 2025](#)). The second method, recently proposed in [Catalano et al. \(2021, 2024\)](#), uses the Wasserstein distance to quantify the discrepancy between Q and the law in the same Fréchet class, i.e., the set of all joint distributions with the same given marginals ([Fréchet, 1951](#); [Sklar, 1959](#); [Dall’Aglio et al., 1991](#); [Nelsen, 2006](#)), inducing full exchangeability.

However, both methods have limitations, as they cannot be applied a posteriori. On the one hand, the set-wise correlation may depend on the choice of the measurable set A , which may lead to different quantification of the posterior dependence structure depending on this choice, as we show in [Chapter 3](#). On the other hand, the Wasserstein-based index can be employed only for Completely Random Measures with equal marginals, making it unsuitable for posterior analysis. Moreover, these methods are purely designed and validated for nonparametric models, hindering model comparison with parametric alternatives. Overall, these limitations highlight the need for more versatile and interpretable indices of dependence between random probability measures, which are applicable both a priori and a posteriori across parametric and nonparametric models.

Beyond its relevance to Bayesian Nonparametrics, the need to quantify dependence between random probability measures also emerges in Distributional Data Analysis (DDA), where the objects of interest are distributions. DDA has gained significant attention in recent years due to its wide range of applications. In Economics, one may study household income distribution across different ages ([Kneip and Utikal, 2001](#)) or analyze women’s

share in a heterosexual couple’s total labour income (Maier et al., 2025). In Finance, DDA can be employed to model the distribution of asset returns (Kokoszka et al., 2019). In Social Sciences, researchers may analyze population pyramids (Delicado, 2011; Hron et al., 2016; Bigot et al., 2017), or investigate mortality curves (Ghodrati and Panaretos, 2022). In Natural Sciences, DDA can be applied to study size spectra of zooplankton communities (Nerini and Ghattas, 2007) or particle-size curves of aquifer geomaterials (Menafoglio et al., 2014). In Medical Sciences, DDA can be utilized to study neural activity patterns in functional brain imaging (Petersen and Müller, 2019; Zhou et al., 2026) or glucose level profiles (Matabuena et al., 2021). In all these fields, understanding the dependence structure among distributional data is crucial for uncovering underlying patterns and relations. Since each observation in DDA can be modeled as a realization of a random probability measure, quantifying the dependence between random probability measures becomes essential for effective exploratory data analysis and statistical modeling.

Moreover, any finite dataset can be represented as a probability distribution via the empirical measure of the data points. This approach can be particularly useful in scenarios where the data are naturally grouped or when one wants to capture the overall distributional characteristics of a dataset. For instance, words’ appearances in a document (Levina and Bickel, 2001; Kusner et al., 2015), genes’ expressions in a single-cell RNA sequence (Nitzan et al., 2019), biomarkers’ signatures in a cell population (Orlova et al., 2016; Schiebinger et al., 2019; Bellazzi et al., 2021; Haviv et al., 2025), pixels’ intensities in an image (Bonet et al., 2025), and more general point clouds (Alvarez-Melis and Fusi, 2020; Kratsios et al., 2023; Geuter et al., 2025), can also be represented as probability distributions, making DDA relevant in Natural Language Processing, Genomics, Computer Vision, and Machine Learning.

However, analyzing distributional data presents unique challenges due to the peculiar nature of probability measures as data objects. Traditional statistical methods may not be directly applicable, as probability measures do not possess a natural vector space

structure: they must be non-negative and have total mass equal to 1. Consequently, several specialized approaches have been developed to address these challenges.

When the data consist of probability measures with a common discrete finite support, they can be represented as compositional data, i.e., distributions on a simplex. They can be analyzed with techniques from Compositional Data Analysis ([Aitchison, 1982](#); [Buccianti et al., 2006](#); [Pawlowsky-Glahn and Buccianti, 2011](#); [Pawlowsky-Glahn et al., 2015](#); [Greenacre, 2018, 2021](#); [Filzmoser et al., 2021](#)).

Conversely, if the random probability measures are absolutely continuous with respect to the Lebesgue measure almost surely, we can identify the distributions with their corresponding probability density functions. See [Petersen et al. \(2022\)](#) for a recent review on methods to analyze probability density functions as data objects. We can either treat them as elements of a functional space, as done in Functional Data Analysis ([Ramsay and Silverman, 2005](#); [Ferraty and Vieu, 2006](#); [Horváth and Kokoszka, 2012](#); [Hsing and Eubank, 2015](#); [Wang et al., 2016](#); [Kokoszka and Reimherr, 2017](#)), or consider specialized frameworks such as Bayes Spaces ([Egozcue et al., 2006](#); [van den Boogart et al., 2010, 2014](#)), Wasserstein Spaces ([Panaretos and Zemel, 2020](#)), or the Hilbert Sphere ([Srivastava et al., 2007](#)).

In order to overcome the assumptions of common discrete finite supports or almost sure absolute continuity, we define a principled framework for analyzing random probability measures by embedding them into a Hilbert space, as already proposed in [Petersen and Müller \(2016\)](#) for random densities. Once embedded into a Hilbert space, we can exploit the theory for random elements valued on a Hilbert space ([Hsing and Eubank, 2015](#)) to define various indices of dependence by summarizing the cross-covariance operator. This approach allows us to define indices of dependence that are applicable to a wide range of random probability measures, regardless of their support or continuity properties. Moreover, by choosing different embeddings, we can tailor the analysis to specific data characteristics, capturing different aspects of dependence and generalizing existing results

that exploit the geometry of the Wasserstein space ([Petersen and Müller, 2019](#); [Zhou et al., 2026](#)).

This thesis develops a general and practical framework for defining, estimating, and interpreting indices of dependence between random probability measures via Hilbert embeddings, with the ultimate goal of quantifying dependence for Bayesian models for partially exchangeable data and analyzing distributional data. The thesis is structured into three main chapters, which are organized as follows.

In [Chapter 1](#), we first review classical and modern indices of dependence for random variables, with particular focus on the ones taking values on Hilbert spaces. In this setting, we define different indices of dependence as summaries of the cross-covariance operator and study their theoretical properties, with particular focus on their behavior at the extremal values.

Then, [Chapter 2](#) introduces the principled framework of quantifying dependence between random probability measures via Hilbert embeddings. Specifically, we introduce and analyze different Hilbert embeddings for probability measures. For instance, one can exploit the Riemannian geometry of the Wasserstein space over the real line to define notions of correlation, as done in [Petersen and Müller \(2019\)](#) and [Zhou et al. \(2026\)](#). Alternatively, kernel mean embeddings ([Berlinet and Thomas-Agnan, 2004](#)) can be exploited, as in [Catalano et al. \(2025\)](#). Similarly, one may build a kernel directly on the space of probability measures, as in [Christmann and Steinwart \(2010\)](#) or [Kolouri et al. \(2016\)](#). This approach allows us to define a variety of indices of dependence for random probability measures, each with its own advantages and interpretability. We investigate, both analytically and numerically, the properties of these indices at the extremal values, with particular focus on whether the proposed indices can detect independence, almost sure equality, or almost sure equality under push-forward via a linear map. Finally, we demonstrate their practical utility in exploratory data analysis by applying hierarchical clustering to functional brain imaging data. The content of this chapter is part of [Mascari](#)

[et al. \(2026\)](#).

Lastly, in [Chapter 3](#), we focus on one of the proposed indices, the Trace Correlation for the kernel mean embedding, under the name of Kernel Correlation, and apply it to quantify the borrowing of information in partially exchangeable models. We show how this index can be used both a priori and a posteriori for parametric and nonparametric models, we derive its analytical expression for a broad class of nonparametric and mixture models, we compute its convergence rate to 0 a posteriori for a Hierarchical Dirichlet Process prior, and we illustrate its use for calibrating dependence and comparing Bayesian models through a numerical application. Moreover, we show that under some hypothesis on the kernel, the Kernel Correlation can identify full exchangeability both a priori and a posteriori for almost surely discrete random probability measures, which are the most common in Bayesian Nonparametrics. The content of this chapter can be found in [Catalano et al. \(2025\)](#).

Chapter 1

Measuring Stochastic Dependence

The definition of notions of correlation for random elements valued in vector spaces has been extensively explored in the literature on multivariate analysis. While the classical Pearson correlation coefficient naturally quantifies linear association between scalar variables, it cannot be directly generalized to vector-valued variables and cannot capture nonlinear dependencies. To overcome these limitations, a broad range of indices of dependence and independence criteria has been developed.

In [Section 1.1](#), we survey indices of dependence for random variables. In [Section 1.2](#), we focus on the case when the objects of interest are Hilbert-valued random variables, describing three indices of dependence based on the cross-covariance operator. We provide their definitions, empirical estimators, and discuss their properties at extremal values.

1.1 Dependence Measures for Random Variables

The study and the quantification of statistical dependence between two random variables defined on the same probability space have been extensively studied in the last century. To fix the notation, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $X : \Omega \rightarrow \mathbb{R}$ and $Y : \Omega \rightarrow \mathbb{R}$ be two real-valued random variables defined on it. The goal is to define an index of

dependence $\mathcal{I}(X, Y)$ that quantifies statistical dependence between X and Y .

In 1959, Rényi (1959) formalized a set of desired properties that any index of dependence should satisfy. These properties include the following list, which we will focus on throughout this chapter.

- (a) **Well-definedness:** $\mathcal{I}(X, Y)$ is defined for any pair of non-constant random variables X and Y .
- (b) **Symmetry:** $\mathcal{I}(X, Y) = \mathcal{I}(Y, X)$.
- (c) **Normalization:** $\mathcal{I}(X, Y) \in [0, 1]$.
- (d) **Independence:** $\mathcal{I}(X, Y) = 0$ if and only if X and Y are independent.
- (e) **Functional Dependence:** $\mathcal{I}(X, Y) = 1$ if and only if there exist measurable functions f and g from $\mathbb{R}$ to $\mathbb{R}$ such that $Y = f(X)$ or $X = g(Y)$ almost surely.
- (f) **Invariance to Isometries:** $\mathcal{I}(f(X), g(Y)) = \mathcal{I}(X, Y)$ for any isometries f and g on $\mathbb{R}$.

This list formalizes the desirable properties an index of dependence should satisfy to capture all forms of statistical dependence between two random variables. However, constructing indices of dependence that satisfy all of Rényi's properties has proven challenging, thus motivating a large body of research in statistics and information theory. Moreover, one should note that in many practical applications, it may be sufficient to consider indices that satisfy only a subset of these properties, depending on the specific context and analysis requirements. Particularly, in many applications, it may be interesting to relax the normalization property (c) by asking that $\mathcal{I}(X, Y) \in [-1, 1]$, distinguishing between positive and negative dependence. In other applications, it may be acceptable to use indices of dependence that do not satisfy property (e), as long as they can effectively capture the relevant forms of functional dependence for the problem at hand. Lastly, in

other applications, it may be better to use indices of dependence that are not invariant under isometries, as in property (f), but rather under other types of transformations, such as affine or monotone transformations. The interested reader may find a comprehensive review of indices of dependence in [Josse and Holmes \(2016\)](#), [Tjøstheim et al. \(2022\)](#), and [Catalano and Lavenant \(2025\)](#).

The first attempts to define such indices date back to the beginning of the 20th century with the introduction of Pearson’s correlation coefficient ([Pearson, 1920](#)). This index takes values in $[-1, 1]$ and captures only linear dependence, not all forms of functional dependence. As a result, independence implies that this index is 0, but a value of 0 does not, in general, imply independence. To overcome these limitations, [Rényi \(1959\)](#) proposed the maximal correlation coefficient, first introduced in [Hirschfeld \(1935\)](#) and [Gebelin \(1941\)](#). This index captures all dependence relationships between random variables since it is defined as the supremum of the Pearson correlation between measurable transformations of X and Y . It satisfies all Rényi’s properties but (e), as it might be 1 even without a relation of functional dependence between X and Y ([Rényi, 1959](#); [Chatterjee, 2021](#)). Nonetheless, property (e) can be achieved under minimal assumptions on the distributions of X and Y ([Rényi, 1959](#)). Given the difficulty of direct computation due to optimization over the infinite-dimensional space of measurable functions, the maximal correlation coefficient has led to the development of several algorithms that estimate its value and approximate the transformations of the variables that maximize Pearson’s correlation, such as the Alternating Conditional Expectations (ACE) algorithm in [Breiman and Friedman \(1985\)](#) and the Randomized Dependence Coefficient (RDC) in [Lopez-Paz et al. \(2013\)](#).

Historically, proposals based on cumulative distribution functions and ranks have tried to overcome the limitations of Pearson’s correlation. The most famous examples of this class are Spearman’s ρ ([Spearman, 1904](#)) and Kendall’s τ ([Kendall, 1938](#)). Other notable proposals include those in [Hoeffding \(1948\)](#) and [Yanagimoto \(1970\)](#). These indices are based on cumulative distribution functions or ranks; hence, they are more robust to outliers

and nonlinear relationships. However, they still do not satisfy all of Rényi's properties, as they may fail to capture certain types of dependence structures. The more recent proposal in [Chatterjee \(2021\)](#) is a prominent example in this category as it satisfies all of Rényi's properties. See also [Gamboa et al. \(2018\)](#).

Chatterjee's correlation coefficient is theoretically equal to the one proposed in [Dette et al. \(2013\)](#), which provides a copula-based estimation of it rather than a rank-based one. Other copula-based indices have been proposed in the literature to satisfy Rényi's properties. For instance, Schweizer-Wolff's σ ([Schweizer and Wolff, 1981](#)) fulfills all Rényi's properties but (e), as it is suitable to detect only functional dependence by strictly increasing functions. [Sklar \(1959\)](#) introduced copulas to capture the dependence structure between two random variables with fixed marginal distributions. See [Nelsen \(2006\)](#) for a comprehensive treatment of copulas and their applications in statistics.

Copula methods can be reread in the Optimal Transport framework ([Marti et al., 2017](#)) since copulas can be interpreted as optimal transport plans that minimize the cost of transporting mass from the joint distribution of two random variables to the product of their marginals. Relevant indices of dependence stemming from the use of the Wasserstein distance and transport maps are the α -transport correlation, the marginal transport correlation, and the isometric transport correlation, all introduced in [Nies et al. \(2025\)](#), which generalizes the Earth mover's correlation in [Móri and Székely \(2020\)](#). These indices can be shown to satisfy most of Rényi's properties and to equal 1 for different functional dependencies between the variables. Similarly, they are invariant under different classes of transformations, depending on the specific index considered. For a comprehensive review of dependence measures based on Wasserstein distance, we refer to [Catalano and Lavenant \(2025\)](#).

[Catalano and Lavenant \(2025\)](#) describe a general framework for constructing indices of dependence by normalizing the distributional distances between the joint law of X and Y ($\mathcal{L}_{X,Y}$) and the product of their marginals ($\mathcal{L}_X \otimes \mathcal{L}_Y$). The first attempt in this

direction can be traced back to the Mutual Information (MI) ([Shannon, 1948](#)), which can be expressed in terms of the Kullback-Leibler divergence between $\mathcal{L}_{X,Y}$ and $\mathcal{L}_X \otimes \mathcal{L}_Y$. MI equals 0 if and only if the two variables are independent, but it does not satisfy both the normalization property, as it can take values in $[0, \infty]$, and the functional dependence property, as it can be infinite even when there is no functional dependence between the variables. To normalize MI, several variants have been proposed in the literature ([Kvålseth, 1987](#)), with the Normalized Mutual Information (NMI) ([Strehl and Ghosh, 2002](#)) being the most popular. See [Cover and Thomas \(2006\)](#) for a comprehensive treatment of Mutual Information and its properties, and [Kvålseth \(2017\)](#) for a survey of its different normalization proposals. Notable procedures for estimating MI from data have been proposed over the years ([Moon et al., 1995](#); [Darbellay and Vajda, 1999](#); [Paninski, 2003](#); [Kraskov et al., 2004](#)). Another information-based index is the Maximal Information Coefficient (MIC) in [Reshef et al. \(2011\)](#), which aims to capture a wide range of dependence relationships between random variables by considering all possible grid discretizations of the joint distribution of X and Y and computing the maximal NMI over these grids. However, MIC does not satisfy all of Rényi’s properties, as it can be 1 even when there is a lack of functional dependence between the variables, as shown in [Chatterjee \(2021\)](#).

Other distance-based indices of dependence include the distance correlation ([Székely et al., 2007](#); [Székely and Rizzo, 2009](#)), which is a measure of association, as it is essentially equal to Pearson’s correlation between the energy distances between X and X' and between Y and Y' , where (X', Y') is an independent copy of (X, Y) . This index takes values in $[0, 1]$ and equals 0 if and only if X and Y are independent. However, it does not satisfy the functional dependence property (e), as it equals 1 only for an affine map whose linear part is an orthogonal transformation composed with isotropic scaling. This quantity can also be generalized to random variables taking values in general metric spaces ([Lyons, 2013](#)). Notably, the measure of association proposed in [Friedman and Rafsky \(1983\)](#) can be reread as a finite-sample version of distance correlation.

Other measures of association have been proposed in the literature as indices of dependence, such as the RV-coefficient ([Escoufier, 1973](#)), which is essentially equal to the Pearson's correlation between the inner product of X and X' and the inner product of Y and Y' , where (X', Y') is an independent copy of (X, Y) . A direct extension of this index is to replace the inner product with a symmetric, positive-definite kernel k . This yields the definition of Centered Kernel Alignment (CKA), first proposed under the name of Kernel Target Alignment in [Cristianini et al. \(2001\)](#) and [Cortes et al. \(2012\)](#) and later popularized in [Kornblith et al. \(2019\)](#) as a normalization of the Hilbert-Schmidt Independence Criterion (HSIC) ([Gretton et al., 2005a, 2007](#)). Other kernel-based dependence measures include the Kernelized Canonical Correlation (KCC) ([Bach and Jordan, 2002](#)), which defines a kernelized notion of Canonical Correlation, firstly introduced in [Hotelling \(1936\)](#), as a way to approximate all functional relations in the definition of the maximal correlation coefficient. Similarly, [Gretton et al. \(2005b\)](#) propose the Constrained Covariance (COCO), which can be seen as a different normalization of KCC.

1.2 Hilbert-Valued Random Variables and Measures of Dependence

The kernel-based indices defined above, such as CKA, KCC, and COCO, are based on the idea of embedding each random variable into a Reproducing Kernel Hilbert Space (RKHS) and exploiting its linear structure and the notion of the cross-covariance operator to define indices that capture complex relationships between random variables. This approach can be directly applied to random probability measures by considering suitable Hilbert embeddings of probability measures, as we describe in [Chapter 2](#).

Below, we recall the main definitions related to Hilbert spaces and bounded linear operators. Then, we define the notions of mean, covariance, and cross-covariance operators for Hilbert-valued random variables, and we describe three popular indices of dependence

based on these operators. We refer to [Hsing and Eubank \(2015\)](#) for a comprehensive treatment of Hilbert-valued random variables.

1.2.1 Hilbert Spaces and Bounded Linear Operators

Let $\mathbb{H}$ be a separable Hilbert space endowed with the inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}} : \mathbb{H} \times \mathbb{H} \rightarrow \mathbb{R}$ and the norm $\| \cdot \|_{\mathbb{H}} : \mathbb{H} \rightarrow \mathbb{R}$ defined as $\|f\|_{\mathbb{H}} := \sqrt{\langle f, f \rangle_{\mathbb{H}}}$ for any $f \in \mathbb{H}$. Since $\mathbb{H}$ is separable, it contains a countable dense subset, implying that $\mathbb{H}$ has an orthonormal basis $\{e_j\}_{j=1}^{\infty}$, i.e., a set of elements in $\mathbb{H}$ such that $\langle e_i, e_j \rangle_{\mathbb{H}} = \delta_{i,j}$ for any $i, j \in \mathbb{N}$, where $\delta_{i,j}$ is the Kronecker delta, and such that any element $f \in \mathbb{H}$ can be expressed as $f = \sum_{j=1}^{\infty} \langle f, e_j \rangle_{\mathbb{H}} e_j$, where the convergence of the series has to be intended in the norm of $\mathbb{H}$.

A bounded linear operator $\mathcal{T} : \mathbb{H} \rightarrow \mathbb{H}$ is a linear map such that there exists a constant $B \geq 0$ such that $\|\mathcal{T}f\|_{\mathbb{H}} \leq B\|f\|_{\mathbb{H}}$ for any $f \in \mathbb{H}$. The composition of two bounded linear operators $\mathcal{T}_1, \mathcal{T}_2 : \mathbb{H} \rightarrow \mathbb{H}$ is defined as $(\mathcal{T}_1\mathcal{T}_2)f := \mathcal{T}_1(\mathcal{T}_2f)$ for any $f \in \mathbb{H}$.

For any bounded linear operator $\mathcal{T} : \mathbb{H} \rightarrow \mathbb{H}$, the adjoint operator of $\mathcal{T}$ is defined as the unique bounded linear operator $\mathcal{T}^* : \mathbb{H} \rightarrow \mathbb{H}$ such that $\langle f, \mathcal{T}g \rangle_{\mathbb{H}} = \langle \mathcal{T}^*f, g \rangle_{\mathbb{H}}$ for any $f, g \in \mathbb{H}$. If $\mathcal{T} = \mathcal{T}^*$, then $\mathcal{T}$ is called a self-adjoint operator. A self-adjoint operator $\mathcal{T} : \mathbb{H} \rightarrow \mathbb{H}$ is said to be positive definite if $\langle f, \mathcal{T}f \rangle_{\mathbb{H}} \geq 0$ for any $f \in \mathbb{H}$. For any self-adjoint positive operator $\mathcal{T} : \mathbb{H} \rightarrow \mathbb{H}$, its square root operator is defined as the unique self-adjoint positive operator $\mathcal{T}^{1/2} : \mathbb{H} \rightarrow \mathbb{H}$ such that $\mathcal{T} = \mathcal{T}^{1/2}\mathcal{T}^{1/2}$. Similarly, a bounded linear operator $\mathcal{T} : \mathbb{H} \rightarrow \mathbb{H}$ is invertible if there exists an inverse bounded linear operator $\mathcal{T}^{-1} : \mathbb{H} \rightarrow \mathbb{H}$ such that $\mathcal{T}\mathcal{T}^{-1} = \mathcal{T}^{-1}\mathcal{T} = \text{Id}$, where $\text{Id} : \mathbb{H} \rightarrow \mathbb{H}$ is the identity operator on $\mathbb{H}$, defined as $\text{Id}f := f$ for any $f \in \mathbb{H}$. We write $\mathcal{T}^{-1/2} := (\mathcal{T}^{1/2})^{-1}$ for any invertible self-adjoint positive operator $\mathcal{T} : \mathbb{H} \rightarrow \mathbb{H}$.

A bounded linear operator $\mathcal{T} : \mathbb{H} \rightarrow \mathbb{H}$ is said to be compact if for any bounded sequence $\{f_j\}_{j=1}^{\infty}$ in $\mathbb{H}$, the sequence $\{\mathcal{T}f_j\}_{j=1}^{\infty}$ has a convergent subsequence in $\mathbb{H}$. A compact bounded linear operator $\mathcal{T} : \mathbb{H} \rightarrow \mathbb{H}$ admits a singular value decomposition, i.e., there exist two orthonormal bases $\{u_j\}_{j=1}^{\infty}$ and $\{v_j\}_{j=1}^{\infty}$ of $\mathbb{H}$ and a non-increasing sequence

of non-negative real numbers $\{\sigma_j\}_{j=1}^\infty$ converging monotonically to 0 such that

$$\mathcal{T} = \sum_{j=1}^{\infty} \sigma_j u_j \otimes v_j, \quad (1.1)$$

where $\otimes$ is the tensor product operator, defined as $(f \otimes g)h := \langle g, h \rangle_{\mathbb{H}} f$ for any $f, g, h \in \mathbb{H}$. If the representation in (1.1) holds, then the elements of the sequence $\{\sigma_j\}_{j=1}^\infty$ are called the singular values of $\mathcal{T}$, while the elements of the sequences $\{u_j\}_{j=1}^\infty$ and $\{v_j\}_{j=1}^\infty$ are called the left and right singular vectors of $\mathcal{T}$, respectively. Particularly, if there exists $m \in \mathbb{N}$ such that $\sigma_j > 0$ for any $j \leq m$ and $\sigma_j = 0$ for any $j > m$, then $\mathcal{T}$ is said to be a finite-rank operator of rank m and the representation in (1.1) reduces to a finite sum.

A compact linear operator $\mathcal{T}$ is said to be Hilbert-Schmidt if its Hilbert-Schmidt norm is finite, while it is said to be trace-class if its trace norm is finite. These norms are defined as

$$\|\mathcal{T}\|_{\text{HS}}^2 := \sum_{j=1}^{\infty} \|\mathcal{T}e_j\|_{\mathbb{H}}^2 \quad \text{and} \quad \|\mathcal{T}\|_{\text{TR}} := \sum_{j=1}^{\infty} \langle (\mathcal{T}^* \mathcal{T})^{1/2} e_j, e_j \rangle_{\mathbb{H}},$$

respectively, for some orthonormal basis $\{e_j\}_{j=1}^\infty$ of $\mathbb{H}$. The values of these norms do not depend on the choice of the orthonormal basis, since for $\{v_j\}_{j=1}^\infty$, the orthonormal basis of $\mathbb{H}$ made of the right singular vectors of $\mathcal{T}$, the above sums reduce to

$$\|\mathcal{T}\|_{\text{HS}}^2 := \sum_{j=1}^{\infty} \sigma_j^2 \quad \text{and} \quad \|\mathcal{T}\|_{\text{TR}} := \sum_{j=1}^{\infty} \sigma_j.$$

The space of Hilbert-Schmidt operators from $\mathbb{H}$ to $\mathbb{H}$ can be endowed with the structure of a Hilbert space itself. In particular, the Hilbert-Schmidt inner product between two Hilbert-Schmidt operators $\mathcal{T}_1, \mathcal{T}_2$ is defined as $\langle \mathcal{T}_1, \mathcal{T}_2 \rangle_{\text{HS}} := \sum_{j=1}^{\infty} \langle \mathcal{T}_1 e_j, \mathcal{T}_2 e_j \rangle_{\mathbb{H}}$ for any orthonormal basis $\{e_j\}_{j=1}^\infty$ of $\mathbb{H}$. Again, the value of this inner product does not depend on the choice of the orthonormal basis. The Hilbert-Schmidt inner product induces the Hilbert-Schmidt norm, as $\|\mathcal{T}\|_{\text{HS}}^2 = \langle \mathcal{T}, \mathcal{T} \rangle_{\text{HS}}$ for any Hilbert-Schmidt operator $\mathcal{T}$.

For a trace-class operator $\mathcal{T}$, we define its trace as $\text{tr}(\mathcal{T}) := \sum_{j=1}^{\infty} \langle \mathcal{T}e_j, e_j \rangle_{\mathbb{H}}$ and

$\|\mathcal{T}\|_{\text{TR}} = \text{tr}((\mathcal{T}^*\mathcal{T})^{1/2})$. A trace-class operator is also a Hilbert-Schmidt operator since $\|\mathcal{T}\|_{\text{HS}} \leq \|\mathcal{T}\|_{\text{TR}}$.

A particular class of bounded linear operators is the one of rank-1 operators, defined as operators of the form $f \otimes g$ for some $f, g \in \mathbb{H}$. Rank-1 operators are trace-class since $\|f \otimes g\|_{\text{TR}} = \|f\|_{\mathbb{H}}\|g\|_{\mathbb{H}}$. This can be easily verified by direct computation considering an orthonormal basis of $\mathbb{H}$, noticing that $(f \otimes g)^*(f \otimes g) = \|g\|_{\mathbb{H}}^2(f \otimes f)$ since $(f \otimes g)^* = g \otimes f$. A fortiori, rank-1 operators are also Hilbert-Schmidt with $\|f \otimes g\|_{\text{HS}} = \|f\|_{\mathbb{H}}\|g\|_{\mathbb{H}}$ since $\langle f_1 \otimes g_1, f_2 \otimes g_2 \rangle_{\text{HS}} = \langle f_1, f_2 \rangle_{\mathbb{H}}\langle g_1, g_2 \rangle_{\mathbb{H}}$ for any $f_1, f_2, g_1, g_2 \in \mathbb{H}$, which again can be easily verified by direct computation considering an orthonormal basis of $\mathbb{H}$.

1.2.2 Hilbert-Valued Random Variables

Let $(\mathbb{H}, \langle \cdot, \cdot \rangle_{\mathbb{H}})$ be a separable Hilbert space endowed with the Borel σ -algebra $\mathcal{H}$. Let X be a random variable taking values on $\mathbb{H}$. Provided that $\mathbb{E}[\|X\|_{\mathbb{H}}] < \infty$, the expectation of X is defined as the unique element $m_X \in \mathbb{H}$ such that $\langle f, m_X \rangle_{\mathbb{H}} = \mathbb{E}[\langle f, X \rangle_{\mathbb{H}}]$ for any $f \in \mathbb{H}$. Such an element exists and is unique due to the Riesz representation theorem, since the map $f \mapsto \mathbb{E}[\langle f, X \rangle_{\mathbb{H}}]$ for $f \in \mathbb{H}$ is a bounded linear functional on $\mathbb{H}$. For the sake of simplicity, we define $\mathbb{E}[X] := m_X$ and use this notation throughout this thesis.

Similarly, provided that $\mathbb{E}[\|X\|_{\mathbb{H}}^2] < \infty$, we can define the covariance operator of X as the unique bounded linear operator $C_{X,X} : \mathbb{H} \rightarrow \mathbb{H}$ such that $\langle f, C_{X,X}g \rangle_{\mathbb{H}} = \mathbb{E}[\langle f, X - \mathbb{E}[X] \rangle_{\mathbb{H}}\langle g, X - \mathbb{E}[X] \rangle_{\mathbb{H}}]$ for any $f, g \in \mathbb{H}$, or equivalently, as

$$C_{X,X} = \mathbb{E}[(X - \mathbb{E}[X]) \otimes (X - \mathbb{E}[X])]. \quad (1.2)$$

Again, such an operator exists and is unique due to the Riesz representation theorem, since the map $g \in \mathbb{H} \mapsto \mathbb{E}[\langle f, X - \mathbb{E}[X] \rangle_{\mathbb{H}}\langle g, X - \mathbb{E}[X] \rangle_{\mathbb{H}}]$ is a bounded linear functional on $\mathbb{H}$ for any fixed $f \in \mathbb{H}$. The covariance operator summarizes the second-order structure of the Hilbert-valued random variable X and generalizes the variance-covariance matrix

for random vectors to the infinite-dimensional setting. It is self-adjoint, positive, and trace class with $\|C_{X,X}\|_{\text{HS}} \leq \|C_{X,X}\|_{\text{TR}} \leq \mathbb{E}[\|X - \mathbb{E}[X]\|_{\mathbb{H}}^2]$, which is finite since $\mathbb{E}[\|X\|_{\mathbb{H}}^2] < \infty$. This implies that $C_{X,X}$ is null as an operator if and only if $X = \mathbb{E}[X]$ almost surely, i.e., X is almost surely constant.

Now, let X and Y be random variables both taking values on $\mathbb{H}$. Provided that $\mathbb{E}[\|X\|_{\mathbb{H}}^2] < \infty$ and $\mathbb{E}[\|Y\|_{\mathbb{H}}^2] < \infty$, the cross-covariance operator between X and Y is defined as the unique bounded linear operator $C_{X,Y} : \mathbb{H} \rightarrow \mathbb{H}$ such that $\langle f, C_{X,Y}g \rangle_{\mathbb{H}} = \mathbb{E}[\langle f, X - \mathbb{E}[X] \rangle_{\mathbb{H}} \langle g, Y - \mathbb{E}[Y] \rangle_{\mathbb{H}}]$ for any $f, g \in \mathbb{H}$, or equivalently, as

$$C_{X,Y} := \mathbb{E}[(X - \mathbb{E}[X]) \otimes (Y - \mathbb{E}[Y])]. \quad (1.3)$$

Such an operator is well-defined due to the Riesz representation theorem, since the map $g \in \mathbb{H} \mapsto \mathbb{E}[\langle f, X - \mathbb{E}[X] \rangle_{\mathbb{H}} \langle g, Y - \mathbb{E}[Y] \rangle_{\mathbb{H}}]$ is a bounded linear functional on $\mathbb{H}$ for any fixed $f \in \mathbb{H}$. The cross-covariance operator is a bounded linear operator. The cross-covariance operator generalizes the covariance matrix between two random vectors to the infinite-dimensional setting and summarizes the joint second-order structure between X and Y . Note that $C_{Y,X} = C_{X,Y}^*$ and $C_{X,Y}$ is a trace-class operator with $\|C_{X,Y}\|_{\text{HS}} \leq \|C_{X,Y}\|_{\text{TR}} \leq \sqrt{\mathbb{E}[\|X - \mathbb{E}[X]\|_{\mathbb{H}}^2]} \sqrt{\mathbb{E}[\|Y - \mathbb{E}[Y]\|_{\mathbb{H}}^2]}$, which is finite since $\mathbb{E}[\|X\|_{\mathbb{H}}^2] < \infty$ and $\mathbb{E}[\|Y\|_{\mathbb{H}}^2] < \infty$.

1.2.3 Indices of Dependence for Hilbert-Valued Random Variables

We now describe some popular indices of dependence for two Hilbert-valued random variables X and Y . Below, we present three indices based on summaries of the cross-covariance operator $C_{X,Y}$ in (1.3) and the covariance operators $C_{X,X}$ and $C_{Y,Y}$ in (1.2): Canonical Correlation, Centered Alignment, and Trace Correlation. These indices take values in $[0, 1]$ or $[-1, 1]$ and generalize well-known indices for multivariate random variables.

We assume $\mathbb{E}[\|X\|_{\mathbb{H}}^2] < \infty$ and $\mathbb{E}[\|Y\|_{\mathbb{H}}^2] < \infty$, so that all the operators above are well-defined. Similarly, we require that X and Y are non-constant almost surely, so that $C_{X,X}$ and $C_{Y,Y}$ are not null operators and the indices are well-defined. These assumptions hold throughout this thesis, unless otherwise specified.

Canonical Correlation The first index under investigation is the *Canonical Correlation* (CC). This index, first introduced in [Hotelling \(1936\)](#), has been applied to Reproducing Kernel Hilbert Spaces to define the Kernelized Canonical Correlation (KCC) in [Bach and Jordan \(2002\)](#) and has been further generalized to Hilbert-valued random variables in [Hsing and Eubank \(2015\)](#)[Chapter 10]. Canonical Correlation is defined as the supremum of the Pearson correlation between all scalar projections of X and Y , i.e.,

$$\text{CC}(X, Y) := \sup_{f, g \in \mathbb{H}} \frac{\langle f, C_{X,Y} g \rangle_{\mathbb{H}}}{\sqrt{\langle f, C_{X,X} f \rangle_{\mathbb{H}}} \sqrt{\langle g, C_{Y,Y} g \rangle_{\mathbb{H}}}}, \quad (1.4)$$

provided that the denominators are non-zero. This index takes values in $[0, 1]$ and can be proved to be equal to the first singular value of the operator $C_{X,X}^{-1/2} C_{X,Y} C_{Y,Y}^{-1/2}$ ([Hsing and Eubank, 2015](#))[Chapter 10].

However, the quantity in (1.4) might be ill-defined if the covariance operators $C_{X,X}$ and $C_{Y,Y}$ are not invertible, which is often the case for finite-rank operators in infinite-dimensional Hilbert spaces. To overcome this issue, we consider a regularized version of Canonical Correlation, where we add a Tikhonov regularization term to the denominators ([Tikhonov, 1963](#)). The resulting regularized Canonical Correlation with regularization parameter $\varepsilon \geq 0$ is defined as

$$\text{CC}(X, Y; \varepsilon) := \sup_{f, g \in \mathbb{H}} \frac{\langle f, C_{X,Y} g \rangle_{\mathbb{H}}}{\sqrt{\langle f, C_{X,X} f \rangle_{\mathbb{H}} + \varepsilon \|f\|_{\mathbb{H}}^2} \sqrt{\langle g, C_{Y,Y} g \rangle_{\mathbb{H}} + \varepsilon \|g\|_{\mathbb{H}}^2}}. \quad (1.5)$$

As for the unregularized version, this index still takes values in $[0, 1]$ and coincides with the first singular value of the operator $(C_{X,X} + \varepsilon \text{Id})^{-1/2} C_{X,Y} (C_{Y,Y} + \varepsilon \text{Id})^{-1/2}$. Particularly,

$\text{CC}(X, Y; 0) = \text{CC}(X, Y)$ provided that the operators $C_{X,X}$ and $C_{Y,Y}$ are invertible. Conversely, for any $\varepsilon > 0$, $\text{CC}(X, Y; \varepsilon)$ is always well-defined, regardless of the invertibility of $C_{X,X}$ and $C_{Y,Y}$. However, it is worth noticing that the value of ε influences the behavior of the index, as already noticed in [Leurgans et al. \(1993\)](#) for functional data.

Centered Alignment The second index that we consider is the *Centered Alignment* (CA). This index is based on the normalization of the Hilbert-Schmidt norm of the cross-covariance operator $C_{X,Y}$, defined as

$$\text{CA}(X, Y) := \frac{\|C_{X,Y}\|_{\text{HS}}^2}{\|C_{X,X}\|_{\text{HS}}\|C_{Y,Y}\|_{\text{HS}}}, \quad (1.6)$$

This index takes values in $[0, 1]$ and can be viewed as a cosine-type similarity, since it normalizes the squared Hilbert-Schmidt norm of $C_{X,Y}$ by the product of the Hilbert-Schmidt norms of the marginal covariance operators $C_{X,X}$ and $C_{Y,Y}$.

This index is a generalization of the RV-coefficient ([Escoufier, 1973](#)) for multivariate random variables. The normalization of the Hilbert-Schmidt norm of the cross-covariance operator was initially proposed in the context of Reproducing Kernel Hilbert Spaces. [Cristianini et al. \(2001\)](#) and [Cortes et al. \(2012\)](#) defined it under the name of Kernel Target Alignment to compare kernel functions. [Kornblith et al. \(2019\)](#) applied the same kernel-based notion under the name of Centered Kernel Alignment as a similarity index for random variables, since it can be interpreted as a normalization of the Hilbert-Schmidt Independence Criterion (HSIC), introduced in [Gretton et al. \(2005a, 2007\)](#).

Trace Correlation The third and last index is the *Trace Correlation*. This index is defined as a normalization of the trace of the cross-covariance operator $C_{X,Y}$, i.e.

$$\text{Corr}_{\mathbb{H}}(X, Y) := \frac{\text{tr}(C_{X,Y})}{\sqrt{\text{tr}(C_{X,X})}\sqrt{\text{tr}(C_{Y,Y})}}. \quad (1.7)$$

This index, differently from the previous two, takes values in $[-1, 1]$. Thus, it can capture both positive and negative dependence between X and Y .

1.2.4 Properties of the Indices at Extremal Values

The three indices discussed above have been designed to measure dependence between Hilbert-valued random variables. Thus, they are expected to capture some notions of independence and functional dependence between X and Y . To study the types of dependence these indices capture, we analyze their properties at the extremes of their ranges. Consequently, we investigate the behavior of CC and CA at 0 and 1, and the behavior of $\text{Corr}_{\mathbb{H}}$ at -1 , 0, and 1, highlighting the relations among them.

A first important remark is that $X \perp Y$ implies that $\text{CC}(X, Y; \varepsilon) = \text{CA}(X, Y) = \text{Corr}_{\mathbb{H}}(X, Y) = 0$, while the converse implication is not granted without further knowledge. Nonetheless, we can characterize the case of null Canonical Correlation and Centered Alignment as the lack of correlation between any pair of scalar projections of X and Y , as shown in the following result.

Proposition 1.2.1. *The following conditions are equivalent.*

- (i) $C_{X,Y} = 0$ as a bounded linear operator;
- (ii) $\langle f, X - \mathbb{E}[X] \rangle_{\mathbb{H}}$ and $\langle g, Y - \mathbb{E}[Y] \rangle_{\mathbb{H}}$ are uncorrelated for any $f, g \in \mathbb{H}$;
- (iii) $\text{CC}(X, Y; \varepsilon) = 0$ for any $\varepsilon \geq 0$;
- (iv) $\text{CA}(X, Y) = 0$.

Moreover, any of these conditions implies that $\text{Corr}_{\mathbb{H}}(X, Y) = 0$.

Another important aspect to analyze is the behavior of the indices at their maximum values. First, it is straightforward to see that if $X = Y$ almost surely, then $\text{CC}(X, Y; 0) = \text{CA}(X, Y) = \text{Corr}_{\mathbb{H}}(X, Y) = 1$. Once again, the converse implication is not generally

true. Nonetheless, we can characterize Trace Correlation as an index of linear dependence between X and Y at ± 1 as shown in the following proposition, which is a direct consequence of Cauchy-Schwarz.

Proposition 1.2.2. *The following conditions are equivalent*

- (i) $\mathbb{C}\text{orr}_{\mathbb{H}}(X, Y) = \pm 1$;
- (ii) *there exists $\alpha \in \mathbb{R} \setminus \{0\}$ with $\text{sgn}(\alpha) = \mathbb{C}\text{orr}_{\mathbb{H}}(X, Y)$ such that $X - \mathbb{E}[X] = \alpha(Y - \mathbb{E}[Y])$ almost surely;*
- (iii) *there exists $\alpha \in \mathbb{R} \setminus \{0\}$ with $\text{sgn}(\alpha) = \mathbb{C}\text{orr}_{\mathbb{H}}(X, Y)$ such that $C_{X,Y} = \alpha C_{Y,Y}$ and $C_{X,X} = \alpha^2 C_{Y,Y}$.*

Similarly, Centered Alignment at 1 can be characterized as an affine relation between the second-order structures of the random probability measures, as shown in the following result.

Proposition 1.2.3. *It holds $\text{CA}(X, Y) = 1$ if and only if there exists $\alpha > 0$ such that $\langle X - \mathbb{E}[X], X' - \mathbb{E}[X'] \rangle_{\mathbb{H}} = \alpha \langle Y - \mathbb{E}[Y], Y' - \mathbb{E}[Y'] \rangle_{\mathbb{H}}$ almost surely, where (X', Y') is an independent copy of (X, Y) .*

Such a strong characterization for Canonical Correlation at 1 is hard to find. Nonetheless, we have the following result, which shows the relations among the indices when the Trace Correlation takes values ± 1 .

Corollary 1.2.4. *If $\mathbb{C}\text{orr}_{\mathbb{H}}(X, Y) = \pm 1$, then $\text{CA}(X, Y) = 1$ and $\text{CC}(X, Y; 0) = 1$.*

1.2.5 Estimation of the Indices

We now want to show how to estimate the indices of dependence introduced above from data. Since all the indices are expressed in terms of the cross-covariance operator $C_{X,Y}$ and the covariance operators $C_{X,X}$ and $C_{Y,Y}$, we first estimate these operators from data.

For $(X_1, Y_1), \dots, (X_n, Y_n)$, n independent realizations of (X, Y) , an unbiased estimator of $C_{X,Y}$ is given by the empirical cross-covariance operator defined as

$$\hat{C}_{X,Y} = \frac{1}{n-1} \sum_{i=1}^n X_i^{(c)} \otimes Y_i^{(c)},$$

where

$$X_i^{(c)} := X_i - \frac{1}{n} \sum_{j=1}^n X_j, \quad Y_i^{(c)} := Y_i - \frac{1}{n} \sum_{j=1}^n Y_j$$

for $i = 1, \dots, n$, while unbiased estimators of $C_{X,X}$ and $C_{Y,Y}$ are given, respectively, by

$$\hat{C}_{X,X} = \frac{1}{n-1} \sum_{i=1}^n X_i^{(c)} \otimes X_i^{(c)}, \quad \hat{C}_{Y,Y} = \frac{1}{n-1} \sum_{i=1}^n Y_i^{(c)} \otimes Y_i^{(c)}.$$

Given these estimators, we can define the empirical versions of the indices by plugging in the operators $\hat{C}_{X,Y}$, $\hat{C}_{X,X}$, and $\hat{C}_{Y,Y}$ in place of $C_{X,Y}$, $C_{X,X}$, and $C_{Y,Y}$ in (1.5), (1.6), and (1.7), respectively. Accordingly, the empirical estimators of the indices are expressed in terms of the centered Gram matrix $\hat{\Sigma}_{X,Y}$ having entries

$$(\hat{\Sigma}_{X,Y})_{i,j} = \langle X_i^{(c)}, Y_j^{(c)} \rangle_{\mathbb{H}}, \quad \text{for } i, j = 1, \dots, n.$$

Similarly, we can define the centered Gram matrices $\hat{\Sigma}_{X,X}$ and $\hat{\Sigma}_{Y,Y}$ as $(\hat{\Sigma}_{X,X})_{i,j} = \langle X_i^{(c)}, X_j^{(c)} \rangle_{\mathbb{H}}$ and $(\hat{\Sigma}_{Y,Y})_{i,j} = \langle Y_i^{(c)}, Y_j^{(c)} \rangle_{\mathbb{H}}$ for $i, j = 1, \dots, n$, respectively.

The empirical estimators of Centered Alignment and Trace Correlation can be directly expressed in closed form by plugging in the empirical cross-covariance operator $\hat{C}_{X,Y}$ and the empirical covariance operators $\hat{C}_{X,X}$ and $\hat{C}_{Y,Y}$ in the expressions of (1.6) and (1.7), respectively. Since by direct computation

$$\|\hat{C}_{X,Y}\|_{\text{HS}}^2 = \frac{1}{(n-1)^2} \text{tr}(\hat{\Sigma}_{X,X} \hat{\Sigma}_{Y,Y}) \quad \text{and} \quad \text{tr}(\hat{C}_{X,Y}) = \frac{1}{n-1} \text{tr}(\hat{\Sigma}_{X,Y}),$$

we define the empirical Centered Alignment as

$$\widehat{\text{CA}}(X, Y) := \frac{\text{tr}(\widehat{\Sigma}_{X,X} \widehat{\Sigma}_{Y,Y})}{\sqrt{\text{tr}(\widehat{\Sigma}_{X,X}^2)} \sqrt{\text{tr}(\widehat{\Sigma}_{Y,Y}^2)}} \quad (1.8)$$

and the empirical Trace Correlation as

$$\widehat{\text{Corr}}_{\mathbb{H}}(X, Y) := \frac{\text{tr}(\widehat{\Sigma}_{X,Y})}{\sqrt{\text{tr}(\widehat{\Sigma}_{X,X})} \sqrt{\text{tr}(\widehat{\Sigma}_{Y,Y})}}. \quad (1.9)$$

For the empirical regularized Canonical Correlation, the situation is more complex since we need to solve an optimization problem over an infinite-dimensional space in (1.5) after plugging in the empirical operators $\widehat{C}_{X,Y}$, $\widehat{C}_{X,X}$, and $\widehat{C}_{Y,Y}$. Consequently, we need to decompose the functions $f, g \in \mathbb{H}$ into a finite-dimensional subspace of $\mathbb{H}$ to reduce the problem to a finite-dimensional optimization problem. Two main approaches have been proposed in the literature for performing this decomposition. [Bach and Jordan \(2002\)](#) propose a data-driven approach by decomposing f and g with respect to the data points $\{X_i^{(c)}\}_{i=1}^n$ and $\{Y_i^{(c)}\}_{i=1}^n$, respectively. Instead, [Zhou et al. \(2026\)](#) propose to approximate f and g by truncating their decompositions with respect to a fixed orthonormal basis of $\mathbb{H}$.

We favor the data-driven approach of [Bach and Jordan \(2002\)](#) since it leads to more interpretable estimators, as the choice of the finite-dimensional subspaces is driven by the data themselves. Consequently, we decompose f and g as

$$f = \sum_{i=1}^n \alpha_i X_i^{(c)} + f^\perp, \quad g = \sum_{i=1}^n \beta_i Y_i^{(c)} + g^\perp$$

for $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)^t, \boldsymbol{\beta} = (\beta_1, \dots, \beta_n)^t \in \mathbb{R}^n$ and $f^\perp \in \text{Span}(\{X_i^{(c)} : i = 1, \dots, n\})^\perp$, $g^\perp \in \text{Span}(\{Y_i^{(c)} : i = 1, \dots, n\})^\perp$. By direct computation, these decompositions lead to

$$\langle f, C_{X,Y} g \rangle_{\mathbb{H}} = \frac{1}{n-1} \boldsymbol{\alpha}^t \widehat{\Sigma}_{X,X} \widehat{\Sigma}_{Y,Y} \boldsymbol{\beta}$$

and

$$\langle f, C_{X,X}f \rangle_{\mathbb{H}} = \frac{1}{n-1} \boldsymbol{\alpha}^t \widehat{\boldsymbol{\Sigma}}_{X,X}^2 \boldsymbol{\alpha}, \quad \langle g, C_{Y,Y}g \rangle_{\mathbb{H}} = \frac{1}{n-1} \boldsymbol{\beta}^t \widehat{\boldsymbol{\Sigma}}_{Y,Y}^2 \boldsymbol{\beta},$$

with $f^\perp$ and $g^\perp$ vanishing from the numerators due to orthogonality. However, the regularization terms $\|f\|_{\mathbb{H}}^2$ and $\|g\|_{\mathbb{H}}^2$ still depend on $f^\perp$ and $g^\perp$ since, by direct computation, $\|f\|_{\mathbb{H}}^2 = \boldsymbol{\alpha}^t \widehat{\boldsymbol{\Sigma}}_{X,X} \boldsymbol{\alpha} + \|f^\perp\|_{\mathbb{H}}^2$ and $\|g\|_{\mathbb{H}}^2 = \boldsymbol{\beta}^t \widehat{\boldsymbol{\Sigma}}_{Y,Y} \boldsymbol{\beta} + \|g^\perp\|_{\mathbb{H}}^2$, again due to orthogonality. Hence, for the goal of finding the supremum in the expression of the regularized Canonical Correlation, we must take f and g to be the function with minimal norm in $\mathbb{H}$, leading to $\|f\|_{\mathbb{H}}^2 = \boldsymbol{\alpha}^t \widehat{\boldsymbol{\Sigma}}_{X,X} \boldsymbol{\alpha}$ and $\|g\|_{\mathbb{H}}^2 = \boldsymbol{\beta}^t \widehat{\boldsymbol{\Sigma}}_{Y,Y} \boldsymbol{\beta}$ by taking $f^\perp = g^\perp = 0$.

Consequently, the empirical regularized Canonical Correlation can be approximated as

$$\sup_{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n} \frac{\boldsymbol{\alpha}^t \widehat{\boldsymbol{\Sigma}}_{X,X} \widehat{\boldsymbol{\Sigma}}_{Y,Y} \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^t (\widehat{\boldsymbol{\Sigma}}_{X,X}^2 + (n-1)\varepsilon \widehat{\boldsymbol{\Sigma}}_{X,X}) \boldsymbol{\alpha}} \sqrt{\boldsymbol{\beta}^t (\widehat{\boldsymbol{\Sigma}}_{Y,Y}^2 + (n-1)\varepsilon \widehat{\boldsymbol{\Sigma}}_{Y,Y}) \boldsymbol{\beta}}},$$

However, we can notice that $\widehat{\boldsymbol{\Sigma}}_{X,X}^2 + (n-1)\varepsilon \widehat{\boldsymbol{\Sigma}}_{X,X}$ can be read as a first-order approximation with respect to ε of $(\widehat{\boldsymbol{\Sigma}}_{X,X} + (n-1)\varepsilon/2 \mathbf{I}_n)^2$ and, similarly, $\widehat{\boldsymbol{\Sigma}}_{Y,Y}^2 + (n-1)\varepsilon \widehat{\boldsymbol{\Sigma}}_{Y,Y}$ can be read as a first-order approximation with respect to ε of $(\widehat{\boldsymbol{\Sigma}}_{Y,Y} + (n-1)\varepsilon/2 \mathbf{I}_n)^2$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix. Hence, we can define the empirical regularized Canonical Correlation as

$$\widehat{\text{CC}}(X, Y; \varepsilon) := \sup_{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n} \frac{\boldsymbol{\alpha}^t \widehat{\boldsymbol{\Sigma}}_{X,X} \widehat{\boldsymbol{\Sigma}}_{Y,Y} \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^t (\widehat{\boldsymbol{\Sigma}}_{X,X} + \varepsilon_n \mathbf{I}_n)^2 \boldsymbol{\alpha}} \sqrt{\boldsymbol{\beta}^t (\widehat{\boldsymbol{\Sigma}}_{Y,Y} + \varepsilon_n \mathbf{I}_n)^2 \boldsymbol{\beta}}}, \quad (1.10)$$

for $\varepsilon_n = (n-1)\varepsilon/2$. The last approximation is fundamental since it allows us to express $\widehat{\text{CC}}(X, Y; \varepsilon)$ as the maximal singular value of $(\widehat{\boldsymbol{\Sigma}}_{X,X} + \varepsilon_n \mathbf{I}_n)^{-1} \widehat{\boldsymbol{\Sigma}}_{X,X} \widehat{\boldsymbol{\Sigma}}_{Y,Y} (\widehat{\boldsymbol{\Sigma}}_{Y,Y} + \varepsilon_n \mathbf{I}_n)^{-1}$, as shown in [Anderson \(2003, Chapter 12\)](#).

Discussion

In this chapter, we have presented general indices of dependence for Hilbert-valued random variables and their estimators. In the following chapter, we apply them to define a wide range of indices of dependence for random probability measures by choosing different combinations of indices and embeddings, and we analyze their properties in detail.

We have restricted our attention to indices of dependence based on the cross-covariance operator. However, other approaches can be pursued. For instance, we may consider distance-based indices, such as the distance correlation ([Székely et al., 2007](#); [Székely and Rizzo, 2009](#)), which can be generalized to any metric space ([Lyons, 2013](#)) and, in particular, to Hilbert spaces. Exploring these alternatives in the context of random probability measures is an interesting direction to extend our framework to other types of indices of dependence.

Appendices

In the following [Appendix 1.A](#), we provide the proofs of the statements that were omitted or summarized in [Chapter 1](#).

1.A Proofs of the Statements

1.A.1 Proof of [Proposition 1.2.1](#)

(i) $\iff$ (ii)

The result holds by the definition of the cross-covariance operator. Since the covariance operator $C_{X,Y}$ is defined so that for any $f, g \in \mathbb{H}$,

$$\langle f, C_{X,Y}g \rangle_{\mathbb{H}} = \mathbb{E} [\langle f, X - \mathbb{E}[X] \rangle_{\mathbb{H}} \langle g, Y - \mathbb{E}[Y] \rangle_{\mathbb{H}}],$$

if $C_{X,Y}$ is the null operator, then $\langle f, C_{X,Y}g \rangle_{\mathbb{H}} = 0$ for any $f, g \in \mathbb{H}$, which is equivalent to $\langle f, X - \mathbb{E}[X] \rangle_{\mathbb{H}}$ and $\langle g, Y - \mathbb{E}[Y] \rangle_{\mathbb{H}}$ being uncorrelated for any $f, g \in \mathbb{H}$.

(i) $\iff$ (iii)

Again, the result holds by the definition of the cross-covariance operator. $C_{X,Y}$ is the null operator if and only if for any $f, g \in \mathbb{H}$, $\langle f, C_{X,Y}g \rangle_{\mathbb{H}} = 0$, which is equivalent to $\text{CC}(X, Y; \varepsilon) = 0$ for any $\varepsilon \geq 0$ by the definition of Canonical Correlation.

(i) $\iff$ (iv)

The result holds by noticing that the Hilbert-Schmidt norm $\|\cdot\|_{\text{HS}}$ is a norm on the

space of Hilbert-Schmidt operators on $\mathbb{H}$. Hence, $C_{X,Y}$ is the null operator if and only if $\|C_{X,Y}\|_{\text{HS}} = 0$, which is equivalent to $\text{CA}(X,Y) = 0$ by the definition of Centered Alignment.

Finally, any of the conditions (i) to (iv) implies that $\text{Corr}_{\mathbb{H}}(X,Y) = 0$ by (1.7) since $\text{tr}(C_{X,Y}) = 0$ when $C_{X,Y}$ is the null operator.

1.A.2 Proof of Proposition 1.2.2

(i) $\iff$ (ii)

This result is based on the following corollary of the Cauchy-Schwarz inequality, which we report here as a lemma.

Lemma 1.A.1. *Let $(\mathbb{H}, \langle \cdot, \cdot \rangle_{\mathbb{H}})$ be a Hilbert space and let X, Y be $\mathbb{H}$ -valued non-constant random variables with finite second moments. Then $\text{Corr}_{\mathbb{H}}(X,Y) = \pm 1$ if and only if $X - \mathbb{E}[X] = \alpha(Y - \mathbb{E}[Y])$ a.s. for $\alpha \neq 0$ such that $\text{sgn}(\alpha) = \text{Corr}_{\mathbb{H}}(X,Y)$.*

Proof. Without loss of generality, we assume that X and Y are centered, i.e. $\mathbb{E}[X] = \mathbb{E}[Y] = 0$. Then, $\text{Corr}_{\mathbb{H}}(X,Y) = \pm 1$ if and only if $\mathbb{E}[\langle X, Y \rangle_{\mathbb{H}}]^2 = \mathbb{E}[\|X\|_{\mathbb{H}}^2] \mathbb{E}[\|Y\|_{\mathbb{H}}^2]$, which means that there is equality in Cauchy-Schwarz. Thus, X and Y are collinear, in the sense that $X = \alpha Y$ almost surely for $\alpha = \mathbb{E}[\langle X, Y \rangle_{\mathbb{H}}] / \mathbb{E}[\|Y\|_{\mathbb{H}}^2]$, which is well-defined since Y has finite second moment and is non-zero, as it is non-constant and centered. In particular, α has the same sign as $\text{Corr}_{\mathbb{H}}(X,Y)$. $\square$

(ii) $\implies$ (iii)

By the definition of the cross-covariance operator,

$$C_{X,Y} = \mathbb{E}[(X - \mathbb{E}[X]) \otimes (Y - \mathbb{E}[Y])] = \alpha \mathbb{E}[(Y - \mathbb{E}[Y]) \otimes (Y - \mathbb{E}[Y])] = \alpha C_{Y,Y}.$$

Similarly, $C_{X,X} = \alpha^2 C_{Y,Y}$.

(iii) $\implies$ (i)

If we apply the trace operator on both sides of $C_{X,Y} = \alpha C_{Y,Y}$, we obtain $\text{tr}(C_{X,Y}) = \alpha \text{tr}(C_{Y,Y})$ by linearity. Similarly, $\text{tr}(C_{X,X}) = \alpha^2 \text{tr}(C_{Y,Y})$. Thus, $\text{Corr}_{\mathbb{H}}(X, Y) = \alpha/|\alpha| = \text{sgn}(\alpha)$.

1.A.3 Proof of Proposition 1.2.3

The result follows from noticing that by Fubini and the definition of Hilbert-Schmidt inner product,

$$\begin{aligned} \|C_{X,Y}\|_{\text{HS}}^2 &= \left\langle \mathbb{E}[(X - \mathbb{E}[X]) \otimes (Y - \mathbb{E}[Y])], \mathbb{E}[(X' - \mathbb{E}[X']) \otimes (Y' - \mathbb{E}[Y'])] \right\rangle_{\text{HS}} \\ &= \mathbb{E} \left[\left\langle X - \mathbb{E}[X], X' - \mathbb{E}[X'] \right\rangle_{\mathbb{H}} \left\langle Y - \mathbb{E}[Y], Y' - \mathbb{E}[Y'] \right\rangle_{\mathbb{H}} \right] \end{aligned}$$

since (X', Y') is an independent copy of (X, Y) .

Similar computations hold for $\|C_{X,X}\|_{\text{HS}}^2$ and $\|C_{Y,Y}\|_{\text{HS}}^2$, leading to

$$\text{CA}(X, Y) = \frac{\mathbb{E} \left[\left\langle X - \mathbb{E}[X], X' - \mathbb{E}[X'] \right\rangle_{\mathbb{H}} \left\langle Y - \mathbb{E}[Y], Y' - \mathbb{E}[Y'] \right\rangle_{\mathbb{H}} \right]}{\sqrt{\mathbb{E} \left[\left\langle X - \mathbb{E}[X], X' - \mathbb{E}[X'] \right\rangle_{\mathbb{H}}^2 \right]} \sqrt{\mathbb{E} \left[\left\langle Y - \mathbb{E}[Y], Y' - \mathbb{E}[Y'] \right\rangle_{\mathbb{H}}^2 \right]}},$$

which equates $\text{Corr}(\left\langle X - \mathbb{E}[X], X' - \mathbb{E}[X'] \right\rangle_{\mathbb{H}}, \left\langle Y - \mathbb{E}[Y], Y' - \mathbb{E}[Y'] \right\rangle_{\mathbb{H}})$. Hence, the result follows by applying Cauchy-Schwarz to $\left\langle X - \mathbb{E}[X], X' - \mathbb{E}[X'] \right\rangle_{\mathbb{H}}$ and $\left\langle Y - \mathbb{E}[Y], Y' - \mathbb{E}[Y'] \right\rangle_{\mathbb{H}}$.

1.A.4 Proof of Corollary 1.2.4

Let $\text{Corr}_{\mathbb{H}}(X, Y) = \pm 1$. Then, by Proposition 1.2.2, there exists $\alpha \in \mathbb{R} \setminus \{0\}$ such that $X - \mathbb{E}[X] = \alpha(Y - \mathbb{E}[Y])$ almost surely.

Let (X', Y') be an independent copy of (X, Y) . Then,

$$\left\langle X - \mathbb{E}[X], X' - \mathbb{E}[X'] \right\rangle_{\mathbb{H}} = \alpha^2 \left(\left\langle Y - \mathbb{E}[Y], Y' - \mathbb{E}[Y'] \right\rangle_{\mathbb{H}} \right),$$

leading to $\text{CA}(X, Y) = 1$ due to [Proposition 1.2.3](#).

Similarly, under the hypothesis above, if we pick $g = f$ in the definition of Canonical Correlation for $\varepsilon = 0$ in [\(1.5\)](#), we obtain $\alpha/|\alpha| = \text{sgn}(\alpha)$. Conversely, if we pick $g = -f$, we obtain $-\alpha/|\alpha| = -\text{sgn}(\alpha)$. Hence, the supremum is attained at 1, i.e., $\text{CC}(X, Y; 0) = 1$.

Chapter 2

Hilbert-Based Correlation Indices for Random Probability Measures¹

We extend the Hilbert-based indices introduced in [Chapter 1](#) to random probability measures by embedding them into suitable Hilbert spaces. We analyze these indices at their extremal values, demonstrate their practical utility, and examine their behavior on synthetic data and in hierarchical clustering of cortical regions for functional brain imaging.

In [Section 2.1](#), we motivate the need for analyzing random probability measures as data objects and review the existing literature on this topic, focusing on Hilbert-based approaches.

In [Section 2.2](#), we formalize a general mathematical framework for defining correlation indices between random probability measures by embedding them into Hilbert spaces, and we recall the definitions of the correlation indices introduced in [Chapter 1](#) to fix the notations.

In [Section 2.3](#), we specialize these indices for different Hilbert embeddings: the Wasserstein embedding, the kernel mean embedding, and the feature map embedding. Particularly,

¹The content of this chapter is part of [Mascari et al. \(2026\)](#).

we reinterpret both the Wasserstein-based indices in Petersen and Müller (2019); Zhou et al. (2026) and the kernel-based index in Catalano et al. (2025) within this unified theoretical framework. Similarly, we define notions of Kernelized Canonical Correlation (Bach and Jordan, 2002) and Centered Kernel Alignment (Kornblith et al., 2019) for the setting of probability measures by using a kernel on the space of probability measures, as defined in Christmann and Steinwart (2010) and Kolouri et al. (2016)

Lastly, in Section 2.4, we study the analytical behavior of these indices under three assumptions on the relation between the two random probability measures (complete independence, almost sure equality, and almost sure equality under push-forward by a linear map), validating these findings with numerical simulation on synthetic datasets. In Section 2.5, we apply these indices to perform hierarchical clustering of cortical areas in functional brain imaging.

The code to reproduce all the numerical experiments in this chapter is written in **Python** and is available, upon request, at <https://github.com/francescomascari/nparcorr>.

2.1 Introduction

The idea of embedding random probability measures into Hilbert spaces for statistical purposes arises naturally whenever we want to analyze probability measures as data objects. The most notable examples come from the fields of Compositional Data Analysis (CoDA) and Functional Data Analysis (FDA).

For instance, when the data consist of probability measures with a common discrete finite support, they can be represented as compositional data, i.e., distributions on the simplex. This representation allows for the application of Compositional Data Analysis (Aitchison, 1982; Buccianti et al., 2006; Pawlowsky-Glahn and Buccianti, 2011; Pawlowsky-Glahn et al., 2015; Greenacre, 2018, 2021; Filzmoser et al., 2021) to analyze such data.

The seminal work by [Aitchison \(1982\)](#) introduces the so-called Aitchison geometry on the simplex, which provides a vector space structure for compositional data by considering equivalence classes under scaling. This approach allows the adaptation of classical statistical techniques to the compositional setting. However, in many applications, the distributions under study may be continuous or have different supports, making it impossible to represent them as compositional data.

Particularly, if the random probability measures are absolutely continuous with respect to the Lebesgue measure almost surely, we can identify the distributions with their corresponding probability density functions. See [Petersen et al. \(2022\)](#) for a recent review on methods to analyze probability density functions as data objects. The most direct way to analyze random densities is to treat them as elements of the Hilbert space of square-integrable functions with respect to some reference measure, as done in [Kneip and Utikal \(2001\)](#) and [Park and Qian \(2012\)](#). This approach allows us to leverage the rich theory of Functional Data Analysis ([Ramsay and Silverman, 2005](#); [Ferraty and Vieu, 2006](#); [Horváth and Kokoszka, 2012](#); [Hsing and Eubank, 2015](#); [Wang et al., 2016](#); [Kokoszka and Reimherr, 2017](#)) to define various statistical methods for random densities.

However, traditional FDA techniques may overlook the peculiarity of the distributional setting, as probability density functions do not form a vector space due to non-negativity and unit-integral constraints. Consequently, in the last few years, Bayes Spaces have been introduced as a suitable framework for analyzing random densities from the intersection of CoDA and FDA ([Egozcue et al., 2006](#); [van den Boogart et al., 2010](#)). These spaces extend the Aitchison geometry to the infinite-dimensional distributional setting by embedding probability density functions via an isometric log-ratio transformation. The resulting space can be proved to be a Hilbert space, known as the Bayes Hilbert space, provided that the common reference measure is finite ([van den Boogart et al., 2014](#)). This approach has been widely explored in the literature ([Delicado, 2011](#); [Menafooglio et al., 2014](#); [Hron et al., 2016](#); [Talská et al., 2021](#); [Maier et al., 2025](#)) because it allows for the adaptation of FDA

techniques to the distributional setting without violating the constraints of probability density functions.

Similarly, [Petersen and Müller \(2016\)](#) propose two alternative approaches to map random probability density functions into Hilbert spaces for data analysis. They consider the case where the densities are defined on a compact interval and propose either a log-hazard or a log-quantile density transformation to embed the densities into L^2 spaces. Alternatively, [Zhang and Müller \(2011\)](#) suggest using the corresponding quantile functions for the random densities under investigation. The study of random quantiles has also been exploited in [Bigot et al. \(2017\)](#), [Matabuena et al. \(2021\)](#), and [Ghosal et al. \(2023\)](#). Quantile functions can be interpreted as embeddings into the space of square-integrable functions on the unit interval, or equivalently, as embeddings into the Wasserstein space of order 2 ([Panaretos and Zemel, 2020](#)).

The last embedding can be extended to the case when the random probability measures are one-dimensional, even when they are not absolutely continuous with respect to a common dominating measure. The recent works by [Petersen and Müller \(2019\)](#) and [Zhou et al. \(2026\)](#) propose to quantify dependence between one-dimensional random probability measures by leveraging the geometry of the Wasserstein space. While these methods provide valuable insights into the dependence structure of distributional data, there is still a need for more general and flexible approaches that go beyond the one-dimensional setting, enabling the analysis of multivariate or high-dimensional distributional data.

To this end, a promising approach for analyzing random probability measures in multivariate settings is to embed them into Reproducing Kernel Hilbert Spaces (RKHSs) ([Berlinet and Thomas-Agnan, 2004](#)) and leverage their properties for statistical analysis. The kernel mean embedding ([Muandet et al., 2017](#)) is a well-established technique for embedding probability measures into RKHSs. Notably, [Catalano et al. \(2025\)](#) exploit this embedding to introduce a kernel-based correlation index for random probability measures motivated by applications in Bayesian Nonparametrics. An alternative kernel-based

embedding is the feature map embedding (Christmann and Steinwart, 2010; Kolouri et al., 2016), which maps probability measures into an RKHS defined on the space of probability measures itself.

2.2 Correlation Indices for Hilbert Embeddings of Measures

Let $\mathbb{X}$ be a locally compact Polish Space, endowed with its Borel σ -algebra $\mathcal{X}$. Let us denote the space of bounded signed measures on $\mathbb{X}$ as $\mathcal{M}_b(\mathbb{X})$, and the subspace of probability measures as $\mathcal{P}(\mathbb{X})$, endowing both these spaces with the σ -algebra induced by evaluation maps $\varphi_A(\xi) = \xi(A)$, for any $A \in \mathcal{X}, \xi \in \mathcal{M}_b(\mathbb{X})$ (Kallenberg, 2017). For $\tilde{P}$ random element on $\mathcal{P}(\mathbb{X})$, the mean measure $\mathbb{E}[\tilde{P}]$ is the element in $\mathcal{P}(\mathbb{X})$ satisfying $\int f(x) d\mathbb{E}[\tilde{P}](x) = \mathbb{E}[\int f(x) d\tilde{P}(x)]$ for any bounded and measurable function $f : \mathbb{X} \rightarrow \mathbb{R}$.

Let $\tilde{P}$ and $\tilde{Q}$ be two random probability measures defined on the same probability space taking values on $\mathcal{S} \subset \mathcal{P}(\mathbb{X})$. The goal is to quantify the dependence structure between these two random measures through an index $\mathcal{I}(\tilde{P}, \tilde{Q})$ that captures the strength and nature of their association. For this goal, we consider a separable Hilbert space $(\mathbb{H}, \langle \cdot, \cdot \rangle_{\mathbb{H}})$ and a map ι from $\mathcal{S}$ to $\mathbb{H}$ satisfying the following assumption, which ensures that no information is lost when embedding probability measures into $\mathbb{H}$.

Assumption 1. *The embedding $\iota : \mathcal{S} \rightarrow \mathbb{H}$ is injective.*

This approach allows us to apply the indices of dependence for Hilbert-valued random variables considered in Section 1.2 to $\iota(\tilde{P})$ and $\iota(\tilde{Q})$ for different ι 's. Note that there are two sources of variability: the type of index, which we have described in Section 1.2, and the choice of Hilbert embedding ι , which will be the topic of Section 2.3. In Table 2.1 you can find a summary of the resulting indices, together with the original references that applied it to probability measures, if available.

As already mentioned in [Section 1.2](#), we consider indices relying on the cross-covariance operator between two Hilbert-valued random variables. We fix the notation of the cross-covariance operator between $\iota(\tilde{P})$ and $\iota(\tilde{Q})$ and denote it as

$$C_{\tilde{P}, \tilde{Q}} := \mathbb{E} \left[\left(\iota(\tilde{P}) - \mathbb{E}[\iota(\tilde{P})] \right) \otimes \left(\iota(\tilde{Q}) - \mathbb{E}[\iota(\tilde{Q})] \right) \right].$$

Again, as in [Section 1.2](#), the indices are estimated via a plug-in approach, replacing the cross-covariance operator with the empirical unbiased estimator

$$\hat{C}_{\tilde{P}, \tilde{Q}} = \frac{1}{n-1} \sum_{i=1}^n \hat{\iota}^{(c)}(P_i) \otimes \hat{\iota}^{(c)}(Q_i),$$

where the sample-centered Hilbert embeddings are defined as

$$\hat{\iota}^{(c)}(P_i) := \iota(P_i) - \frac{1}{n} \sum_{j=1}^n \iota(P_j), \quad \hat{\iota}^{(c)}(Q_i) := \iota(Q_i) - \frac{1}{n} \sum_{j=1}^n \iota(Q_j),$$

for $(P_1, Q_1), \dots, (P_n, Q_n)$, n independent realizations of $(\tilde{P}, \tilde{Q})$. Accordingly, the empirical estimators of the indices are expressed in terms of the centered Gram matrices $\hat{\Sigma}_{\tilde{P}, \tilde{Q}}$, $\hat{\Sigma}_{\tilde{P}, \tilde{P}}$, and $\hat{\Sigma}_{\tilde{Q}, \tilde{Q}}$. Recall that the entries of $\hat{\Sigma}_{\tilde{P}, \tilde{Q}}$ are given by $(\hat{\Sigma}_{\tilde{P}, \tilde{Q}})_{i,j} = \langle \hat{\iota}^{(c)}(P_i), \hat{\iota}^{(c)}(Q_j) \rangle_{\mathbb{H}}$ for $i, j = 1, \dots, n$, and the entries of $\hat{\Sigma}_{\tilde{P}, \tilde{P}}$ and $\hat{\Sigma}_{\tilde{Q}, \tilde{Q}}$ are similarly defined.

Now that we have fixed the notation, we can specialize the definitions of Canonical Correlation, Centered Alignment, and Trace Correlation for the Hilbert embeddings of random probability measures by taking $X = \iota(\tilde{P})$ and $Y = \iota(\tilde{Q})$ in the formulae in [Section 1.2](#). We denote the regularized Canonical Correlation in [\(1.5\)](#) and its empirical estimator in [\(1.10\)](#) for $\varepsilon \geq 0$ and $\varepsilon_n = (n-1)\varepsilon/2$ as

$$\begin{aligned} \text{CC}_{\iota}(\tilde{P}, \tilde{Q}; \varepsilon) &:= \sup_{f, g \in \mathbb{H}} \frac{\langle f, C_{\tilde{P}, \tilde{Q}} g \rangle_{\mathbb{H}}}{\sqrt{\langle f, C_{\tilde{P}, \tilde{P}} f \rangle_{\mathbb{H}} + \varepsilon \|f\|_{\mathbb{H}}^2} \sqrt{\langle g, C_{\tilde{Q}, \tilde{Q}} g \rangle_{\mathbb{H}} + \varepsilon \|g\|_{\mathbb{H}}^2}}, \\ \widehat{\text{CC}}_{\iota}(\tilde{P}, \tilde{Q}; \varepsilon) &:= \sup_{\alpha, \beta \in \mathbb{R}^n} \frac{\alpha^t \hat{\Sigma}_{\tilde{P}, \tilde{P}} \hat{\Sigma}_{\tilde{Q}, \tilde{Q}} \beta}{\sqrt{\alpha^t (\hat{\Sigma}_{\tilde{P}, \tilde{P}} + \varepsilon_n \mathbf{I}_n)^2 \alpha} \sqrt{\beta^t (\hat{\Sigma}_{\tilde{Q}, \tilde{Q}} + \varepsilon_n \mathbf{I}_n)^2 \beta}} \end{aligned}$$

Similarly, we specialize the Centered Alignment in (1.6) and its sample version in (1.8) as

$$\text{CA}_\iota(\tilde{P}, \tilde{Q}) := \frac{\|C_{\tilde{P}, \tilde{Q}}\|_{\text{HS}}^2}{\|C_{\tilde{P}, \tilde{P}}\|_{\text{HS}} \|C_{\tilde{Q}, \tilde{Q}}\|_{\text{HS}}}, \quad \widehat{\text{CA}}_\iota(\tilde{P}, \tilde{Q}) := \frac{\text{tr}(\hat{\Sigma}_{\tilde{P}, \tilde{P}} \hat{\Sigma}_{\tilde{Q}, \tilde{Q}})}{\sqrt{\text{tr}(\hat{\Sigma}_{\tilde{P}, \tilde{P}}^2)} \sqrt{\text{tr}(\hat{\Sigma}_{\tilde{Q}, \tilde{Q}}^2)}}.$$

Lastly, we can rewrite the Trace Correlation in (1.7) and its empirical estimator in (1.9) as

$$\text{Corr}_\iota(\tilde{P}, \tilde{Q}) := \frac{\text{Cov}_\iota(\tilde{P}, \tilde{Q})}{\sqrt{\text{Var}_\iota(\tilde{P})} \sqrt{\text{Var}_\iota(\tilde{Q})}}, \quad \widehat{\text{Corr}}_\iota(\tilde{P}, \tilde{Q}) := \frac{\text{tr}(\hat{\Sigma}_{\tilde{P}, \tilde{Q}})}{\sqrt{\text{tr}(\hat{\Sigma}_{\tilde{P}, \tilde{P}})} \sqrt{\text{tr}(\hat{\Sigma}_{\tilde{Q}, \tilde{Q}})}}.$$

2.3 Hilbert Embeddings of Random Variables

We describe the use of Canonical Correlation, Centered Alignment, and Trace Correlation for different choices of a space of measures $\mathcal{S} \subset \mathcal{P}(\mathbb{X})$, a separable Hilbert space $\mathbb{H}$, and an embedding $\iota : \mathcal{S} \rightarrow \mathbb{H}$ satisfying [Assumption 1](#). For the sake of simplicity, we consider $\mathbb{X} \subset \mathbb{R}^d$ with $d \geq 1$ throughout the rest of the chapter, unless otherwise specified.

We consider the Wasserstein embedding for one-dimensional distributions and two kernel-based embeddings in the more general multivariate setting: the kernel mean embedding and the feature map embedding. [Table 2.1](#) summarizes these choices and their notations, while [Figure 2.1](#) reports the behaviors of the indices graphically for different stochastic relations between $\tilde{P}$ and $\tilde{Q}$. Overall, the indices based on the Wasserstein embedding can suggest the presence of a linear map that transforms one random probability measure into another via push-forward. Conversely, the indices based on kernel mean embedding and feature map embedding are helpful for detecting almost sure equality between two random probability measures. Notably, the feature map embedding can

Table 2.1: List of correlation indices for different choices of the embedding

Hilbert embedding			Correlation indices		
Space $\mathcal{S}$	Hilbert space $\mathbb{H}$	$\iota(\nu)$	Canonical Corr.	Centered Align.	Trace Corr.
$\mathcal{P}_2(\mathbb{X})$	$L^2(\text{Leb}; [0, 1])$	F_ν^{-1}	CC_W Zhou et al. (2026)	CA_W	Corr_W Petersen and Müller (2019)
$\mathcal{P}(\mathbb{X})$	$\mathbb{H}_k \subset \mathbb{R}^{\mathbb{X}}$	$\mu_k(\nu)$	CC_k	CA_k	Corr_k Catalano et al. (2025)
$\mathcal{P}(\mathbb{X})$, $\mathbb{X}$ compact	$\mathbb{H}_K \subset \mathbb{R}^{P(\mathbb{X})}$	$\phi_K(\nu)$	CC_K	CA_K	Corr_K

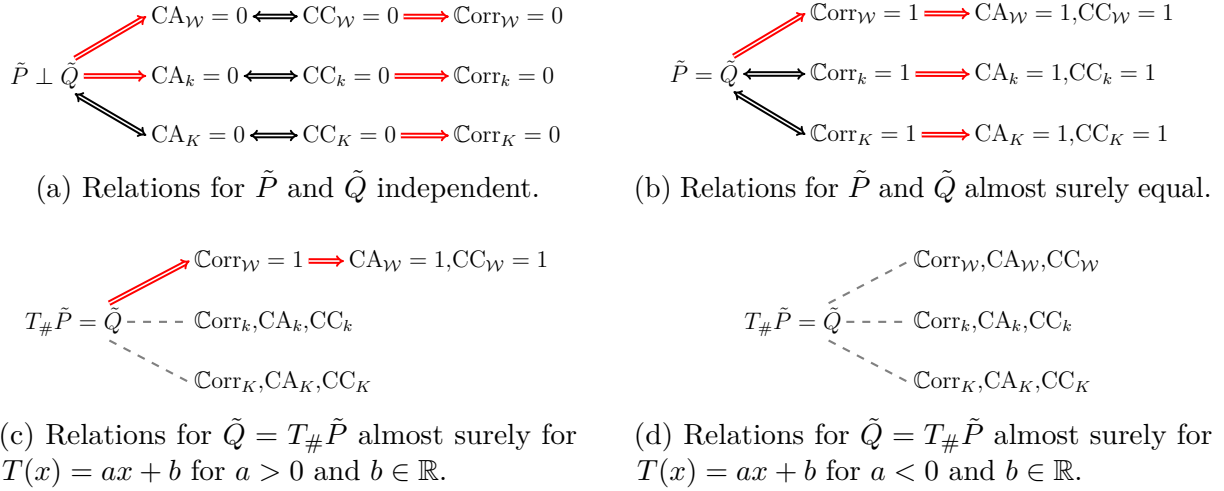


Figure 2.1: Values of the correlation indices for different stochastic relations between $\tilde{P}$ and $\tilde{Q}$. The implication arrows are red whenever there exists an example showing that the inverse implication does not hold. Similarly, the dotted gray lines indicate that there is no relation between the indices' values and the stochastic relation. In order for all relations to hold, we consider $\tilde{P}, \tilde{Q} \in \mathcal{P}_2(\mathbb{X})$ for $\mathbb{X} \subset \mathbb{R}$ compact with $\tilde{P}$ and $\tilde{Q}$ almost surely discrete with atomless laws and $\mathbb{E}[\tilde{P}]$ and $\mathbb{E}[\tilde{Q}]$ atomless.

also be used to detect independence between the two when combined with Canonical Correlation and Centered Alignment.

2.3.1 Wasserstein Embedding

The first proposal is the Wasserstein embedding, which can be defined in the one-dimensional setting.

A Primer on Optimal Transport

Let $(\mathbb{X}, d_{\mathbb{X}})$ be a complete and separable metric space, and let $\nu, \xi \in \mathcal{P}(\mathbb{X})$ be two probability measures on it. Let $\Pi(\nu, \xi)$ denote the set of couplings between ν and ξ , i.e., the set of probability measures on $\mathbb{X} \times \mathbb{X}$ with marginals ν and ξ . The Wasserstein distance of order $p \geq 1$ between ν and ξ is defined as

$$\mathcal{W}_p(\nu, \xi) := \left(\inf_{\gamma \in \Pi(\nu, \xi)} \int d_{\mathbb{X}}(x, y)^p d\gamma(x, y) \right)^{1/p}. \quad (2.1)$$

This distance is equal to the p -th root of the optimal transport cost between ν and ξ with cost function given by the p -th power of the ground distance $d_{\mathbb{X}}$ on $\mathbb{X}$. The optimal transport problem was first formulated in [Monge \(1781\)](#) and later relaxed in [Kantorovich \(1941\)](#), leading to the modern theory of optimal transport.

Let $\mathcal{P}_p(\mathbb{X})$ denote the subspace of $\mathcal{P}(\mathbb{X})$ of probability measures on $\mathbb{X}$ with finite p -th moment. The Wasserstein distance is a true metric on $\mathcal{P}_p(\mathbb{X})$ that metrizes weak convergence plus convergence of the p -th moments when $d_{\mathbb{X}}$ is a lower semi-continuous metric ([Villani, 2008](#)). Moreover, Wasserstein distances have a rich geometric structure, as they induce a Riemannian-like geometry on $\mathcal{P}_p(\mathbb{X})$, as described in [Ambrosio et al. \(2008\)](#). We refer to [Santambrogio \(2015\)](#) and [Peyré and Cuturi \(2019\)](#) for comprehensive treatments of optimal transport theory and its applications, while we consider [Panaretos and Zemel \(2020\)](#) and [Chewi et al. \(2025\)](#) for introductions to statistical aspects of optimal transport.

Quantile Embedding

Let $\mathbb{X} \subset \mathbb{R}$ closed with $d_{\mathbb{X}}(x, y) = |x - y|$. Then, the Wasserstein distance between ν and ξ in (2.1) can be written explicitly in terms of the quantile functions. Let $L^p(\text{Leb}; [0, 1])$ be the space of p -integrable real-valued functions on $[0, 1]$ with respect to the Lebesgue measure. Then, it holds ([Gangbo and McCann, 1996](#)) that for any $\nu, \xi \in \mathcal{P}_p(\mathbb{X})$, the Wasserstein distance of order p between ν and ξ reads

$$\mathcal{W}_p^p(\nu, \xi) = \|F_{\nu}^{-1} - F_{\xi}^{-1}\|_{L^p(\text{Leb}; [0, 1])}^p = \int_0^1 |F_{\nu}^{-1}(t) - F_{\xi}^{-1}(t)|^p dt,$$

where $F_{\nu}^{-1}(t) = \inf\{x \in \mathbb{X} \text{ s.t. } \nu((-\infty, x]) \geq t\}$ for $t \in [0, 1]$ is the quantile function of ν . Since, for $p = 2$, $L^2(\text{Leb}; [0, 1])$ is a Hilbert space with inner product $\langle f, g \rangle_{L^2(\text{Leb}; [0, 1])} := \int_0^1 f(t)g(t)dt$, we can take $\mathcal{S} = \mathcal{P}_2(\mathbb{X})$, $\mathbb{H} = L^2(\text{Leb}; [0, 1])$ and ι to be the quantile embedding $q : \nu \in \mathcal{P}_2(\mathbb{X}) \mapsto F_{\nu}^{-1} \in L^2(\text{Leb}; [0, 1])$ mapping any probability measure to

its quantile function. This choice of ι satisfies [Assumption 1](#) since the quantile function uniquely determines the distribution.

For this embedding, we define all indices in [Section 2.2](#): the Wasserstein Canonical Correlation ($\text{CC}_{\mathcal{W}}$), the Wasserstein Centered Alignment ($\text{CA}_{\mathcal{W}}$), and the Wasserstein Trace Correlation ($\text{Corr}_{\mathcal{W}}$). We note that at least two popular measures of dependence for random probabilities fit into this framework: $\text{CC}_{\mathcal{W}}$ naturally arises from centering the cross-covariance operator of the *Intrinsic Wasserstein Correlation* in [Zhou et al. \(2026\)](#), whereas $\text{Corr}_{\mathcal{W}}$ is the normalization of the *Wasserstein Covariance* in [Petersen and Müller \(2019\)](#). In contrast, the fusion of Wasserstein embedding and Centered Alignment is not already present in the literature.

Behavior at Extremal Values

For this embedding, all indices are zero at independence, as shown in [Proposition 1.2.1](#). However, these values fail to be sufficient for independence, as shown in [Example 2.B.1](#). Thus, we cannot conclude independence between random probability measures when any of the indices are zero.

Nonetheless, for this embedding, the Trace Correlation is 1 if one random probability measure is the push-forward measure of the other through an increasing linear map, as stated in the following corollary to [Proposition 1.2.2](#).

Corollary 2.3.1. *Let $\tilde{Q} = T_{\#}\tilde{P}$ a.s. for some $T : x \mapsto ax + b$ with $a \in \mathbb{R} \setminus \{0\}$ and $b \in \mathbb{R}$. Then, $\text{CC}_{\mathcal{W}}(\tilde{P}, \tilde{Q}; 0) = 1$ and $\text{CA}_{\mathcal{W}}(\tilde{P}, \tilde{Q}) = 1$. Moreover, $\text{Corr}_{\mathcal{W}}(\tilde{P}, \tilde{Q}) = 1$ if $a > 0$.*

Remark 1. Recall that, for a measurable space $\mathbb{X}$ endowed with the σ -algebra $\mathcal{X}$, the push-forward of a measure ν by a measurable map $T : \mathbb{X} \rightarrow \mathbb{X}$ is defined as $T_{\#}\nu(A) := \nu(T^{-1}(A))$ for $A \in \mathcal{X}$. This implies that for any bounded measurable function $f : \mathbb{X} \rightarrow \mathbb{R}$, it holds that $\int_{\mathbb{X}} f(x) dT_{\#}\nu(x) = \int_{\mathbb{X}} f(T(x)) d\nu(x)$.

However, the converse implication to [Corollary 2.3.1](#) does not hold, i.e., we cannot conclude that a random probability measure is the push-forward of the other by a linear

map if any of the indices equals 1 as shown in [Example 2.B.2](#). We also cannot draw conclusions similar to the ones in [Corollary 2.3.1](#) on the value of $\text{Corr}_{\mathcal{W}}$ for $T(x) = ax + b$ with $a < 0$ and $b \in \mathbb{R}$. Particularly, we show that $\text{Corr}_{\mathcal{W}}(\tilde{P}, T_{\#}\tilde{P})$ can take values -1 , $-1/2$ or 1 for different choices of $\tilde{P}$ in [Example 2.B.3](#).

2.3.2 Kernel-Based Embeddings

The Wasserstein embedding is only defined for one-dimensional distributions. For higher-dimensional cases, we can instead rely on the theory of Reproducing Kernel Hilbert Spaces (RKHSs), first introduced in [Aronszajn \(1950\)](#).

A Primer on Reproducing Kernel Hilbert Spaces

Let $\mathbb{Y}$ be a Polish space endowed with the Borel σ -algebra $\mathcal{Y}$. A kernel $k : \mathbb{Y} \times \mathbb{Y} \rightarrow \mathbb{R}$ is a measurable function that, in the sequel, will always be assumed to be bounded, symmetric, and positive-definite. This means that for some $B > 0$, we have $k(y_1, y_2) = k(y_2, y_1)$ and $|k(y_1, y_2)| \leq B$ for any $y_1, y_2 \in \mathbb{Y}$, and, moreover,

$$\sum_{i=1}^m \sum_{j=1}^m a_i a_j k(y_i, y_j) \geq 0,$$

for any $m \in \mathbb{N}$, $y_1, \dots, y_m \in \mathbb{Y}$ and $a_1, \dots, a_m \in \mathbb{R}$.

Any kernel defines a natural mapping of $\mathbb{Y}$ into the space $\mathbb{R}^{\mathbb{Y}}$ of functions from $\mathbb{Y}$ to $\mathbb{R}$ through the feature map $\phi_k : y \mapsto k(y, \cdot)$. The natural notion of inner product between feature maps reads as $\langle \phi_k(y), \phi_k(y') \rangle = \langle k(y, \cdot), k(y', \cdot) \rangle := k(y, y')$ for any $y, y' \in \mathbb{Y}$, and it can be naturally extended to all linear combinations of feature maps. The closure of this space in $\mathbb{R}^{\mathbb{Y}}$ defines the unique Reproducing Kernel Hilbert Space (RKHS) $\mathbb{H}_k$ induced by the kernel k ([Aronszajn, 1950](#)). We denote its scalar product as $\langle \cdot, \cdot \rangle_{\mathbb{H}_k}$. The elements of $\mathbb{H}_k$ satisfy the so-called reproducing property, i.e., for any $f \in \mathbb{H}_k$ and $y \in \mathbb{Y}$, it holds that $f(y) = \langle f, k(y, \cdot) \rangle_{\mathbb{H}_k}$.

For a kernel k , we can define the kernel mean embedding μ_k , which associates every bounded signed measure $\xi \in \mathcal{M}_b(\mathbb{Y})$ to a unique element $\mu_k(\xi) \in \mathbb{H}_k$ as

$$\mu_k(\xi)(y) := \int_{\mathbb{Y}} k(y, y') d\xi(y'). \quad (2.2)$$

The reproducing property ensures the following identities for $\xi, \xi_1, \xi_2 \in \mathcal{M}_b(\mathbb{Y})$ and $f \in \mathbb{H}_k$,

$$\langle f, \mu_k(\xi) \rangle_{\mathbb{H}_k} = \int_{\mathbb{Y}} f(y) d\xi(y), \quad \langle \mu_k(\xi_1), \mu_k(\xi_2) \rangle_{\mathbb{H}_k} = \iint_{\mathbb{Y} \times \mathbb{Y}} k(y, y') d\xi_1(y) d\xi_2(y'). \quad (2.3)$$

A kernel k is c_0 if it vanishes at infinity, that is, if the set $\{y : |k(y, y)| \geq \varepsilon\}$ is compact for every $\varepsilon > 0, y \in \mathbb{Y}$; it is injective if the feature map $y \mapsto k(y, \cdot)$ is injective; it is characteristic if the kernel mean embedding is injective on $\mathcal{P}(\mathbb{Y})$. Since $k(y, \cdot) = \mu_k(\delta_y)$, a characteristic kernel is always injective. For a c_0 -kernel, the kernel mean embedding is injective on $\mathcal{M}_b(\mathbb{Y})$ if and only if $\mathbb{H}_k$ is dense in $C_0(\mathbb{Y})$, the space of continuous functions over $\mathbb{Y}$ vanishing at infinity endowed with the uniform norm (Sriperumbudur et al., 2011). In this case, we call k c_0 -universal, which implies that k is characteristic since $\mathcal{P}(\mathbb{Y}) \subset \mathcal{M}_b(\mathbb{Y})$. We refer to Schölkopf and Smola (2001); Rasmussen and Williams (2006); Steinwart and Christmann (2008) for general references on RKHSs and kernels, to Muandet et al. (2017) for a comprehensive treatment of kernel mean embeddings and characteristic kernels, to Sriperumbudur et al. (2011) for a review of different notions of universality, and to Berlinet and Thomas-Agnan (2004) for an overview on the use of RKHSs in Statistics and Probability.

Choice of the Embedding

We consider two different kernel-based embeddings for random probability measures. In both cases, we take $\mathcal{S} = \mathcal{P}(\mathbb{X})$ for $\mathbb{X} \subset \mathbb{R}^d$ with $d \geq 1$, while the choice of the Hilbert space $\mathbb{H}$ and the embedding ι differ. In the first case, we consider $\mathbb{Y} = \mathbb{X}$ and $\mathbb{H} = \mathbb{H}_k$, the RKHS associated with a kernel $k : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$; the embedding is $\iota = \mu_k$, i.e., the kernel

mean embedding for k , as defined in (2.2). In the second case, we consider $\mathbb{Y} = \mathcal{P}(\mathbb{X})$ and $\mathbb{H} = \mathbb{H}_K$, the RKHS associated with a kernel $K : \mathcal{P}(\mathbb{X}) \times \mathcal{P}(\mathbb{X}) \rightarrow \mathbb{R}$; the embedding is $\iota = \phi_K$, i.e., the feature map embedding for K , as proposed in Christmann and Steinwart (2010) and Kolouri et al. (2016).

Kernel Mean Embedding Let ι be the kernel mean embedding μ_k for a kernel k on $\mathbb{X}$. This embedding has been widely used in Statistics and Machine Learning to define distances and test statistics between probability measures (Gretton et al., 2012). Notably, it has been applied in Muandet et al. (2012) for classification. See Muandet et al. (2017) for a complete reference. Asking that Assumption 1 is fulfilled is equivalent to asking that k is characteristic. Consequently, for k characteristic we can consider the three indices in Section 2.2 stemming from the kernel mean embedding: the kernel mean Canonical Correlation (CC_k), the kernel mean Centered Alignment (CA_k), and the kernel mean Trace Correlation (Corr_k). Among these indices, CC_k and CA_k are proposed here for the first time. Conversely, Corr_k has been proposed in Catalano et al. (2025) to detect almost sure equality between random probability measures, as shown in Chapter 3.

Notably, if k is also c_0 -universal, thus characteristic a fortiori, the following corollary to Proposition 1.2.2 holds under some assumptions on the support of $\tilde{P}$ and $\tilde{Q}$ and their expectations.

Corollary 2.3.2. *Let $\tilde{P}$ and $\tilde{Q}$ be almost-surely discrete such that $\mathbb{E}[\tilde{P}]$ and $\mathbb{E}[\tilde{Q}]$ are atomless, and let k be c_0 -universal. Then, $\text{Corr}_k(\tilde{P}, \tilde{Q}) = 1$ if and only if $\tilde{P} = \tilde{Q}$ almost surely, implying $\text{CA}_k(\tilde{P}, \tilde{Q}) = \text{CC}_k(\tilde{P}, \tilde{Q}; 0) = 1$.*

Speaking of the behavior at 0, again, all the kernel mean indices are null at independence as in Proposition 1.2.1. However, none of these values is sufficient for independence, as shown in Example 2.B.4, where we consider two random probability measures that are dependent but have zero values for all the indices with kernel mean embedding.

A notable example of a kernel satisfying both Assumption 1 and the hypotheses of

[Corollary 2.3.2](#) is the Gaussian kernel, as it is both characteristic and c_0 -universal. The Gaussian kernel is defined for any $x, y \in \mathbb{X}$ as $k(x, y) := \exp(-\|x - y\|^2/(2\sigma^2))$ for $\sigma > 0$, where $\|\cdot\|$ is the standard Euclidean norm in $\mathbb{R}^d$. Therefore, all the kernel mean indices for this kernel can be used to detect almost sure equality between random probability measures. We will assume k to be Gaussian in the applications that follow.

Remark 2. Another typical example in the literature is the linear kernel $k(x, y) = \langle x, y \rangle$ for $\mathbb{X}$ bounded, where $\langle \cdot, \cdot \rangle$ is the canonical inner product in $\mathbb{R}^d$. This kernel does not satisfy [Assumption 1](#) because it is not characteristic ([Muandet et al., 2017](#)). Nonetheless, it is worth mentioning that in the one-dimensional case, i.e., $\mathbb{X} \subset \mathbb{R}$, the indices based on the linear kernel mean embedding are strongly related to those based on the Wasserstein embedding, since, due to the identity $\int_0^1 F_\nu^{-1}(t)dt = \int_{\mathbb{X}} x d\nu(x)$ for any $\nu \in \mathcal{P}(\mathbb{X})$, it can be proved by direct computation that for k linear, $\text{Corr}_{\mathcal{W}}(\tilde{P}, \tilde{Q}) = 1$ implies $\text{Corr}_k(\tilde{P}, \tilde{Q}) = 1$. Similarly, if either $\text{CC}_{\mathcal{W}}(\tilde{P}, \tilde{Q}; 0) = 0$ or $\text{CA}_{\mathcal{W}}(\tilde{P}, \tilde{Q}) = 0$, then both $\text{CC}_k(\tilde{P}, \tilde{Q}; 0) = 0$ and $\text{CA}_k(\tilde{P}, \tilde{Q}) = 0$. Hence, the linear kernel mean embedding is a looser embedding than the Wasserstein one since it is based only on the first moments rather than the whole quantile function.

Feature Map Embedding The second idea is to define a kernel directly on the space of distributions and consider its feature map as an embedding. Several attempts have been made to define a kernel between distributions. Most of them are based on the common assumption of absolute continuity of the measures with respect to a common reference measure ([Hein and Bousquet, 2005](#)).

We want to avoid such a strong assumption. Consequently, we consider a radial kernel on the space of probability measures defined for a distance d on $\mathcal{P}(\mathbb{X})$. Then, we can define a radial kernel K for any $\nu, \xi \in \mathcal{P}(\mathbb{X})$ as

$$K(\nu, \xi) := \exp\left(-\frac{1}{2\gamma^2}d(\nu, \xi)^2\right), \quad (2.4)$$

for $\gamma > 0$.

Two notable examples of distances d that yield positive-definite kernels K in (2.4) are the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) and the 2-Sliced Wasserstein distance (SW_2) (Rabin et al., 2012). The former is the distance induced by the kernel mean embedding for a characteristic kernel k on $\mathbb{H}_k$ and is defined as

$$\text{MMD}_k(\nu, \xi) := \|\mu_k(\nu) - \mu_k(\xi)\|_{\mathbb{H}_k} \quad \text{for } \nu, \xi \in \mathcal{P}(\mathbb{X}). \quad (2.5)$$

The latter is based on averaging the 2-Wasserstein distance across one-dimensional projections of the measures and is equal to

$$\mathcal{SW}_2^2(\nu, \xi) := \int_{\mathbb{S}^{d-1}} \mathcal{W}_2^2(\pi_{\theta\#}\nu, \pi_{\theta\#}\xi) d\theta \quad \text{for } \nu, \xi \in \mathcal{P}_2(\mathbb{X}), \quad (2.6)$$

where $\mathbb{S}^{d-1} := \{\theta \in \mathbb{R}^d : \|\theta\| = 1\}$ is the unit sphere in $\mathbb{R}^d$ and $\pi_\theta : x \mapsto \langle x, \theta \rangle$ is the projection along the direction of $\theta \in \mathbb{S}^{d-1}$.

The kernel K induced by the MMD in (2.5) has been introduced in Christmann and Steinwart (2010), and further explored in Ziegel et al. (2024). Moreover, it has been applied to define functionals on $\mathcal{P}(\mathbb{X})$ for classification (Blanchard et al., 2011) and distribution regression (Szabó et al., 2016), as well as in both single-agent and multi-agent multi-armed bandit problems (Deshmukh et al., 2017; Dubey and Pentland, 2020). Conversely, the kernel K induced by the $\mathcal{SW}_2$ in (2.6) has been introduced as *Sliced Wasserstein Kernel* in Kolouri et al. (2016) and further studied in Meunier et al. (2022). It has been applied to Wasserstein-over-Wasserstein gradient flows (Bonet et al., 2025) and persistence diagrams (Carrière et al., 2017).

Both kernels have been proved to be universal on a compact space (Christmann and Steinwart, 2010; Meunier et al., 2022) provided that $\mathbb{X}$ is compact, since both $\mathcal{P}(\mathbb{X})$ and $\mathcal{P}_2(\mathbb{X})$ are compact with respect to convergence in distribution. Consequently, both kernels can be used to define embeddings satisfying Assumption 1. However, we are going to

consider only the kernel K induced by MMD in (2.5) since it is defined on the whole $\mathcal{P}(\mathbb{X})$ rather than only on $\mathcal{P}_2(\mathbb{X})$ as for $\mathcal{SW}_2$. Nonetheless, all the upcoming results can be straightforwardly extended to the Sliced Wasserstein kernel.

Consequently, let $\mathbb{X}$ be compact, $\mathcal{S} = \mathcal{P}(\mathbb{X})$, $\mathbb{H} = \mathbb{H}_K$ and ι as the feature map ϕ_K defined as $\phi_K(\nu) = K(\nu, \cdot) \in \mathbb{H}_K$ for any $\nu \in \mathcal{P}(\mathbb{X})$. Hence, we consider the three indices in Section 2.2 stemming from the feature map embedding: the feature map Canonical Correlation (CC_K), the feature map Centered Alignment (CA_K), and the feature map Trace Correlation (Corr_K).

Let us consider Corr_K . This index detects almost sure equality between random probability measures under some mild assumptions on the joint law of $\tilde{P}$ and $\tilde{Q}$, as shown in the following corollary to Proposition 1.2.2.

Corollary 2.3.3. *Let $\tilde{P}$ and $\tilde{Q}$ have atomless laws on $\mathcal{P}(\mathbb{X})$, i.e., $\mathbb{P}(\tilde{P} = \nu) = \mathbb{P}(\tilde{Q} = \nu) = 0$ for any $\nu \in \mathcal{P}(\mathbb{X})$. Then, $\text{Corr}_K(\tilde{P}, \tilde{Q}) = 1$ if and only if $\tilde{P} = \tilde{Q}$ almost surely, implying $\text{CA}_K(\tilde{P}, \tilde{Q}) = \text{CC}_K(\tilde{P}, \tilde{Q}; 0) = 1$.*

Notably, CC_K and CA_K extend the definitions of *Kernelized Canonical Correlation* (Bach and Jordan, 2002; Gretton et al., 2005b) and *Centered Kernel Alignment* (Cristianini et al., 2001; Cortes et al., 2012), respectively, beyond standard input spaces to random probability measures. Consequently, we can apply the corresponding independence criteria (Gretton et al., 2005a,b) to detect independence between random probability measures as shown in the following result.

Theorem 2.3.4. *For any $\varepsilon \geq 0$, $\text{CC}_K(\tilde{P}, \tilde{Q}; \varepsilon) = 0$ if and only if $\text{CA}_K(\tilde{P}, \tilde{Q}) = 0$ if and only if $\tilde{P} \perp \tilde{Q}$, implying $\text{Corr}_K(\tilde{P}, \tilde{Q}) = 0$.*

2.4 Numerical Validation

We now compare the estimates of the correlation indices defined above across different datasets by considering (P_i, Q_i) for $i = 1, \dots, n$, n i.i.d. realizations of $(\tilde{P}, \tilde{Q})$. In

the experiments, we use different strategies to approximate each realization, i.e., each probability measure, as a mixture of R Dirac masses at some locations in $\mathbb{X}$ with non-negative weights summing to one. More explicitly, for every $i = 1, \dots, n$, we find some $\theta_{i,r}, \eta_{i,r} \in \mathbb{X}$ and $\pi_{i,r}, \omega_{i,r} \geq 0$ for $r = 1, \dots, R$ with $\sum_{r=1}^R \pi_{i,r} = \sum_{r=1}^R \omega_{i,r} = 1$, such that

$$P_i \approx \sum_{r=1}^R \pi_{i,r} \delta_{\theta_{i,r}}, \quad Q_i \approx \sum_{r=1}^R \omega_{i,r} \delta_{\eta_{i,r}}. \quad (2.7)$$

We refer to [Appendix 2.C](#) for details on how we obtain such approximations in each experiment.

For Canonical Correlation indices, we follow the choice of regularization parameter ε suggested in [Bach and Jordan \(2002\)](#). Hence, we take $\varepsilon = 2 \cdot 10^{-2}$ for $n \leq 1000$ and $\varepsilon = 2 \cdot 10^{-3}$ for $n > 1000$.

2.4.1 Monotone Push-Forward

In light of [Corollary 2.3.1](#), we consider the case of $\tilde{Q} = T_{\#}\tilde{P}$ almost surely for different monotone linear maps $T(x) = ax + b$ on $\mathbb{R}$ for $a \in \{-10, -1, -0.1, 0.1, 1, 10\}$ and $b \in \{-7, -3, -1, 0, 1, 3, 7\}$. We consider both a nonparametric and a parametric case for the choice of $\tilde{P}$.

Firstly, in the nonparametric case, we take $\tilde{P} \sim \text{DP}(c, \nu_0)$, where we write $\text{DP}(c, \nu_0)$ for the Dirichlet Process with concentration parameter c and atomless baseline measure $\nu_0 \in \mathcal{P}(\mathbb{X})$ ([Ferguson, 1973](#)). In other words, $\tilde{P}$ is a random probability measure such that

$$(\tilde{P}(A_1), \tilde{P}(A_2), \dots, \tilde{P}(A_m)) \sim \text{Dir}(c\nu_0(A_1), c\nu_0(A_2), \dots, c\nu_0(A_m))$$

for any partition $\{A_1, A_2, \dots, A_m\}$ of $\mathbb{X}$, where Dir indicates the finite-dimensional Dirichlet distribution. We set $c = 10$ and $\nu_0 = \mathcal{N}(0, 1)$; this setting goes beyond the constraint of $\mathbb{X}$ being compact required for the kernel K in [\(2.4\)](#) to be universal since ν_0 is supported on

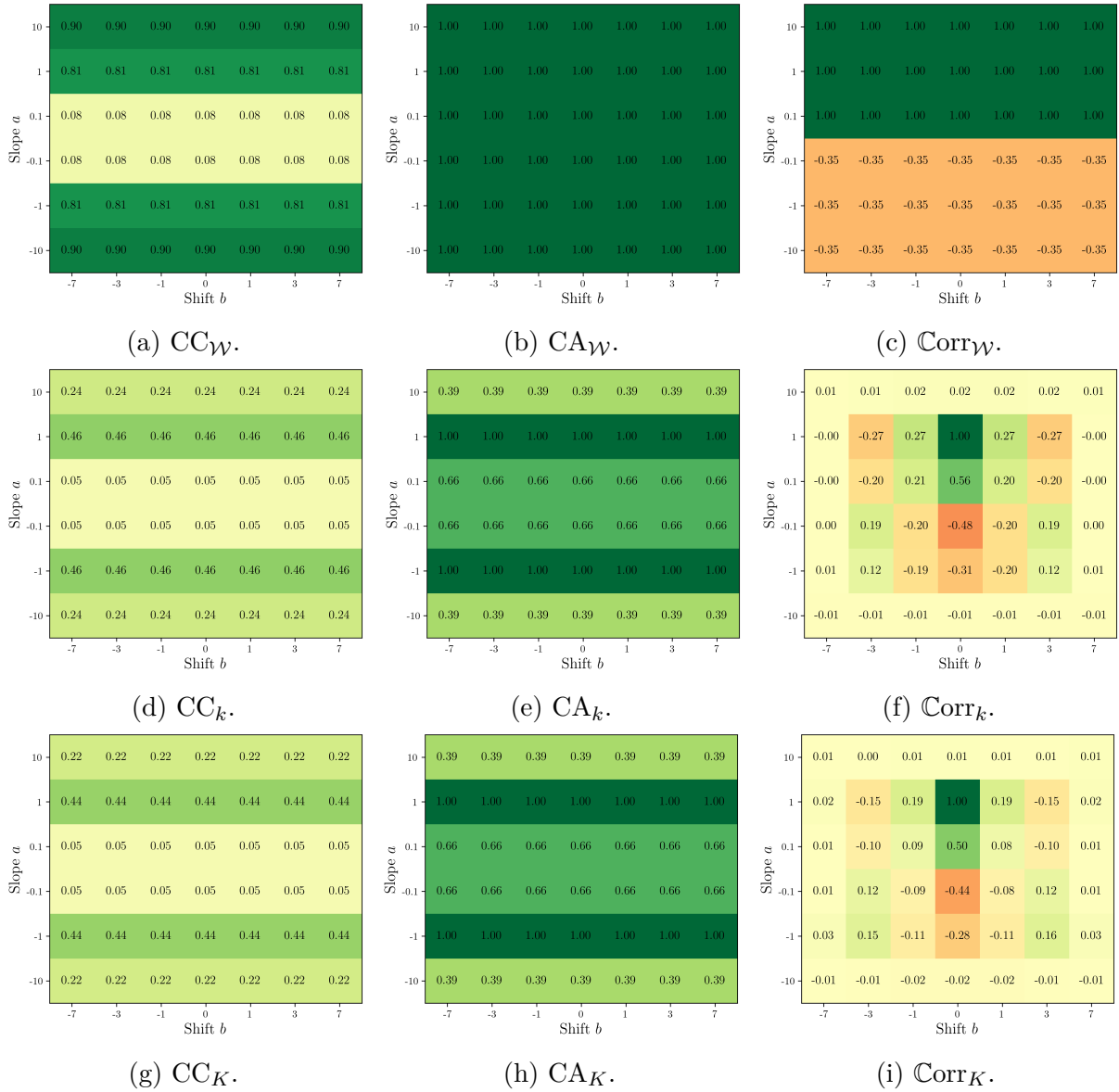


Figure 2.2: Values of the correlation indices between $\tilde{P}$ and $\tilde{Q} = T_{\#}\tilde{P}$ for $T(x) = ax + b$ for $a \in \{-10, -1, -0.1, 0.1, 1, 10\}$ and $b \in \{-7, -3, -1, 0, 1, 3, 7\}$ for the nonparametric case with $\tilde{P} \sim \text{DP}(10, \mathcal{N}(0, 1))$.

the whole $\mathbb{R}$. Nonetheless, in practice, we approximate $\tilde{P}$ and $\tilde{Q}$ through (2.7) as discrete measures with compact support. We obtain each realization (P_i, Q_i) for $i = 1, \dots, n$ using the approximate sampling scheme proposed in Ishwaran and James (2001). Specifically, in (2.7), we take the atoms to be $\theta_{i,1}, \dots, \theta_{i,R} \stackrel{\text{i.i.d.}}{\sim} \nu_0$ and $\eta_{i,r} = T(\theta_{i,r})$ for $r = 1, \dots, R$ and $i = 1, \dots, n$, and the weights to be $\pi_{i,r} = \omega_{i,r} = Y_r / \sum_{s=1}^R Y_s$ for $r = 1, \dots, R$ and

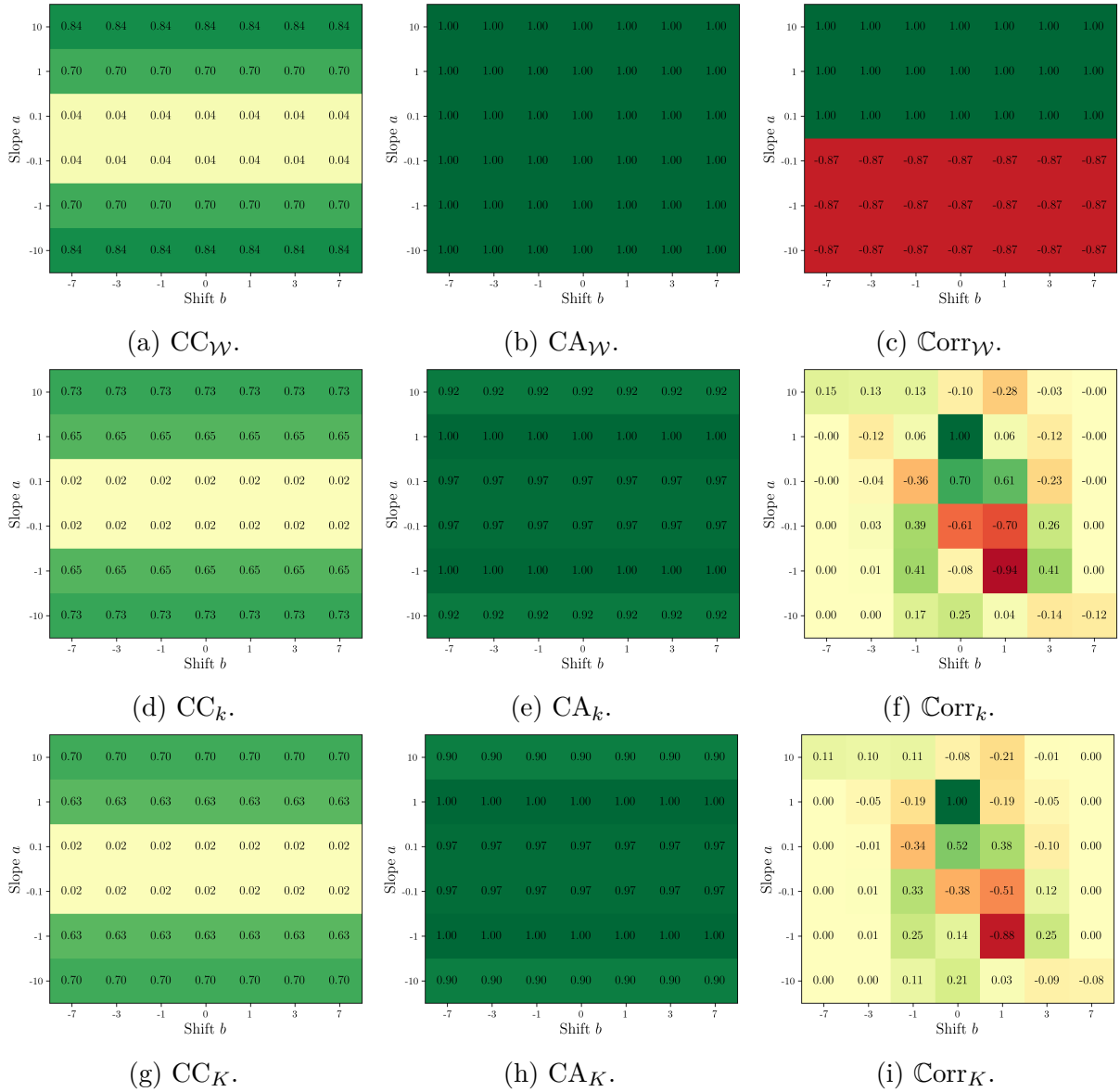


Figure 2.3: Values of the correlation indices between $\tilde{P}$ and $\tilde{Q} = T_{\#}\tilde{P}$ for $T(x) = ax + b$ for $a \in \{-10, -1, -0.1, 0.1, 1, 10\}$ and $b \in \{-7, -3, -1, 0, 1, 3, 7\}$ for the parametric case with $\tilde{P}|\alpha, \beta = \text{Beta}(\alpha, \beta)$ and $\alpha, \beta \stackrel{\text{i.i.d.}}{\sim} \text{Gamma}(2, 1)$.

$i = 1, \dots, n$ with $Y_1, \dots, Y_R \stackrel{\text{i.i.d.}}{\sim} \text{Gamma}(c/R, 1)$. We set $n = 1000$ and $R = 10,000$.

Secondly, in the parametric case, we take $\tilde{P}$ to be a Beta distribution with random parameters $\alpha, \beta > 0$ sampled from a Gamma distribution. More explicitly, we set $\tilde{P} = \text{Beta}(\alpha, \beta)$ almost surely with $\alpha, \beta \stackrel{\text{i.i.d.}}{\sim} \text{Gamma}(2, 1)$. Let us notice that $\tilde{P}$ is almost surely absolutely continuous with respect to the Lebesgue measure and has compact

support on $[0, 1]$. Specifically, we approximate each realization (P_i, Q_i) for $i = 1, \dots, n$ as in (2.7). First, we sample $\alpha_i, \beta_i \stackrel{\text{i.i.d.}}{\sim} \text{Gamma}(2, 1)$ for $i = 1, \dots, n$. Then, we take the atoms to be $\theta_{i,1}, \dots, \theta_{i,R} \stackrel{\text{i.i.d.}}{\sim} \text{Beta}(\alpha_i, \beta_i)$ and $\eta_{i,r} = T(\theta_{i,r})$ for $r = 1, \dots, R$ and $i = 1, \dots, n$, and the weights to be uniform, i.e., $\pi_{i,r} = \omega_{i,r} = 1/R$ for $r = 1, \dots, R$ and $i = 1, \dots, n$. Again, we set $n = 1000$ and $R = 10,000$.

In Figure 2.2, we summarize the mean values of the indices for different linear maps T across 10 runs for the nonparametric case. We can see that all values of Canonical Correlation and Centered Alignment are invariant to shifts and symmetric with respect to the sign of the slope. The values of the Wasserstein-based indices are in accordance with Corollary 2.3.1 since $\text{CA}_{\mathcal{W}} = 1$ for any non-constant linear map, while $\text{Corr}_{\mathcal{W}} = 1$ only for increasing ones. The same theoretical guarantees of Corollary 2.3.1 do not hold for $\text{CC}_{\mathcal{W}}$ since the regularization has the effect of changing the value of the correlation dramatically as the slope varies. A similar result holds for the Canonical Correlation indices for kernel mean embedding and feature map embedding. Firstly, we can observe that the same index yields similar values for both embeddings. Secondly, we can notice that Corr_k and Corr_K attain the value 1 only at the identity, i.e., $a = 1, b = 0$, as shown in Corollaries 2.3.2 and 2.3.3. Nonetheless, these two indices exhibit a peculiar behavior as the shift value varies. In particular, when a is positive, the indices are positive only for smaller values of $|b|$, while when a is negative, the indices are positive only for larger values of $|b|$. This behavior for Corr_k is described analytically in Example 2.B.5. The same phenomenon holds for Corr_K , even if analytical computations are harder to carry out. Lastly, CA_k and CA_K are equal to 1 for $T(x) = \pm x + b$ for $b \in \mathbb{R}$. However, these are not the only cases when these indices are 1. Example 2.B.6 shows that CA_k can be 1 even without the existence of such a relation between $\tilde{P}$ and $\tilde{Q}$. Again, the computations for CA_K are hard to carry on, but we can expect a similar behavior.

The parametric case is shown in Figure 2.3. Specifically, the Wasserstein-based indices and the Canonical Correlation and Centered Alignment indices behave similarly to the

nonparametric case. Conversely, Corr_k and Corr_K still exhibit a peculiar behavior with respect to the choice of slope and shift, but it is different to the one observed in the previous case due to the fact that the Beta distribution is supported on $[0, 1]$, and hence it is not symmetric with respect to reflections around 0. Nonetheless, both CA_k and CA_K attain the value 1 for $T(x) = \pm x + b$ for any $b \in \mathbb{R}$, while Corr_k and Corr_K are equal to 1 only for the identity map. Thus, these results confirm both the theoretical findings in [Corollaries 2.3.2](#) and [2.3.3](#) and the observations made for the nonparametric case.

Overall, these results suggest that the Canonical Correlation is the least favorable index, as the regularization parameter has a significant influence on it. Hence, we favor the use of Centered Alignment and Trace Correlation. When it comes to embedding choices, Wasserstein-based indices can be used to detect a linear relationship between two random probability measures. In contrast, the indices for kernel mean embedding and feature map embedding are suited for detecting almost sure equality.

2.4.2 Hierarchical Dirichlet Process

In [Catalano et al. \(2025\)](#), the kernel mean Trace Correlation was used to study, both a priori and a posteriori, the dependence structure of the *hierarchical Dirichlet Process* (hDP). The hDP, first introduced in [Teh et al. \(2006\)](#) and further investigated in [Camerlenghi et al. \(2019\)](#), is a multi-level nonparametric Bayesian model for random probability measures. Specifically, we say that $(\tilde{P}, \tilde{Q}) \sim \text{hDP}(c, c_0, \nu_0)$ for $c, c_0 > 0$ and ν_0 an atomless probability measure on $\mathbb{X}$ if

$$\tilde{P}, \tilde{Q} \mid \tilde{P}_0 \stackrel{\text{i.i.d.}}{\sim} \text{DP}(c, \tilde{P}_0), \quad \tilde{P}_0 \sim \text{DP}(c_0, \nu_0). \quad (2.8)$$

We obtain each realization (P_i, Q_i) for $i = 1, \dots, n$ adapting the approximate sampling scheme for the DP in [Ishwaran and James \(2001\)](#) to the hierarchical setting, as described in [Teh et al. \(2006\)](#). Specifically, in [\(2.7\)](#), we take the atoms to be $\theta_{i,1}, \dots, \theta_{i,R} \stackrel{\text{i.i.d.}}{\sim} \nu_0$ and $\eta_{i,r} = \theta_{i,r}$ for $r = 1, \dots, R$. Similarly, we take the weights to be $\pi_{i,r} = Y_{1,i,r} / \sum_{s=1}^R Y_{1,i,s}$,

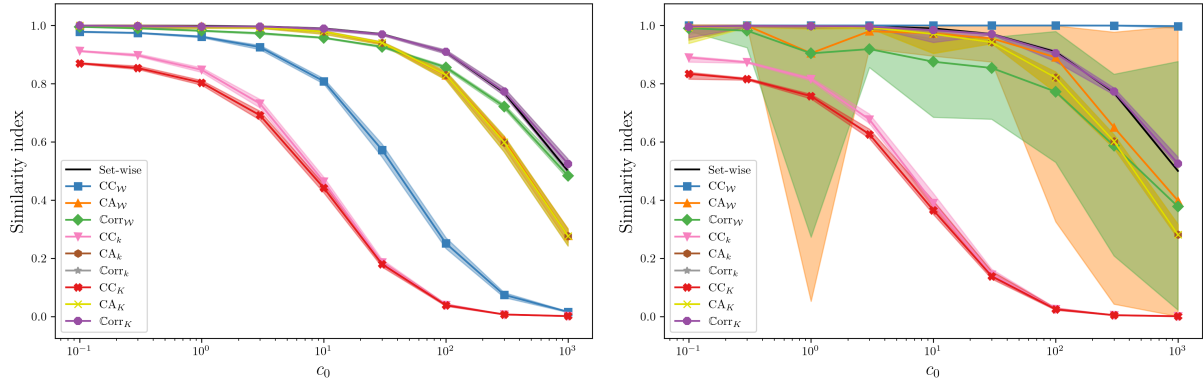


Figure 2.4: Correlation indices between $\tilde{P}$ and $\tilde{Q}$ being jointly distributed as $\text{hDP}(c, c_0, \nu_0)$ for $c = 1000$ and $c_0 \in \{0.1, 0.3, 1, 3, 10, 30, 100, 300, 1000\}$ for $\nu_0 = \mathcal{N}(0, 1)$ (**left**) and $\nu_0 = \text{Cauchy}(0, 1)$ (**right**).

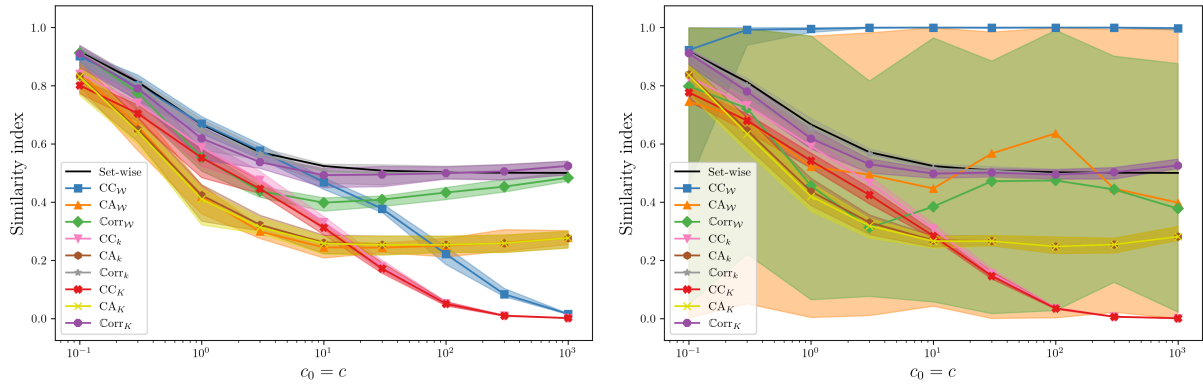


Figure 2.5: Correlation indices between $\tilde{P}$ and $\tilde{Q}$ being jointly distributed as $\text{hDP}(c, c_0, \nu_0)$ for $c = c_0 \in \{0.1, 0.3, 1, 3, 10, 30, 100, 300, 1000\}$ for $\nu_0 = \mathcal{N}(0, 1)$ (**left**) and $\nu_0 = \text{Cauchy}(0, 1)$ (**right**).

$\omega_{i,r} = Y_{2,i,r} / \sum_{s=1}^R Y_{2,i,s}$, with $Y_{1,i,r}, Y_{2,i,r} | \beta_r \stackrel{\text{i.i.d.}}{\sim} \text{Gamma}(c\beta_r, 1)$, where $\beta_r = X_r / \sum_{s=1}^R X_s$ with $X_r \stackrel{\text{i.i.d.}}{\sim} \text{Gamma}(c_0/R, 1)$ for $r = 1, \dots, R$. Again, we set $n = 1000$ and $R = 10,000$.

In [Catalano et al. \(2025\)](#), it is proved that $\text{Corr}_k(\tilde{P}, \tilde{Q})$ does not depend on the choice of the kernel k and is always equal to the corresponding set-wise correlation $\text{Corr}(\tilde{P}(A), \tilde{Q}(A)) = (c + 1)/(c + c_0 + 1)$ for any $A \in \mathcal{X}$ such that $\nu_0(A) \notin \{0, 1\}$. This correlation index is the most popular in the Bayesian Nonparametrics literature for assessing the level of dependence between two random probability measures a priori, as it has a closed-form expression for most models, as proved in [Franzolini et al. \(2025\)](#). We check whether other correlation indices attain such value. To this end, we consider two different

scenarios: in the first one, we fix $c = 1000$ and vary $c_0 \in \{0.1, 0.3, 1, 3, 10, 30, 100, 300, 1000\}$; in the second one, we set $c = c_0$ and vary their common value in the same set. We consider two different choices for ν_0 : the standard Gaussian distribution $\mathcal{N}(0, 1)$ and the standard Cauchy distribution $\text{Cauchy}(0, 1)$. The main difference between these two choices of ν_0 is that the former has finite variance, while the latter does not. This aspect is crucial for the Wasserstein-based indices since they are well-defined only for probability measures with finite second moment. Figures 2.4 and 2.5 show the mean values as well as the min-max bands of the indices over 10 runs for $\nu_0 = \mathcal{N}(0, 1)$ (on the left) and $\nu_0 = \text{Cauchy}(0, 1)$ (on the right) for the two scenarios described above, respectively. In every plot, the set-wise correlation is reported as a solid black line.

Figure 2.4 illustrates the behaviors of the values for $c = 1000$ as c_0 varies in the set $\{0.1, 0.3, 1, 3, 10, 30, 100, 300, 1000\}$. For $\nu_0 = \mathcal{N}(0, 1)$, all the indices tend to 0, as expected, since $\tilde{\nu}_0$ converges to ν_0 as c_0 increases, leading to $\tilde{P} \perp \tilde{Q}$. For $\nu_0 = \text{Cauchy}(0, 1)$, all the indices tend to 0 as well, with the exception of $\text{CC}_{\mathcal{W}}$, which fails to converge to 0 due to the joint effect of the infinite variance of ν_0 and the regularization parameter. The effect of the infinite variance of the baseline measure on the Wasserstein embedding is further confirmed by the high variability across the runs for $\text{Corr}_{\mathcal{W}}$ and $\text{CC}_{\mathcal{W}}$, shown by the large min-max bands in the plot on the right. The mean values of the indices for kernel mean embedding and feature map embedding exhibit similar behaviors across the two plots. Particularly, Corr_k follows the set-wise correlation, as expected, and Corr_K follows a similar trend too, which is not guaranteed by theory. Overall, the Canonical Correlation indices are the fastest to converge to 0, while the Centered Alignment indices exhibit similar behavior and attain a mid-value for higher values of c_0 .

Figure 2.5 shows the behaviors of the values for $c = c_0$ as they vary in the set $\{0.1, 0.3, 1, 3, 10, 30, 100, 300, 1000\}$. We notice that for $\nu_0 = \mathcal{N}(0, 1)$ the mean values of the Canonical Correlation indices converge to 0 as c_0 and c increase being equal. In contrast, the Centered Alignment converges to $1/4$, and the Trace Correlation indices

converge to $1/2$, as well as the set-wise correlation. This setting is peculiar since, as said above, $\tilde{P}$ and $\tilde{Q}$ are closer to independence as c_0 increase; nonetheless, both $\tilde{P}$ and $\tilde{Q}$ tend to be equal to $\tilde{\nu}_0$ as c increases. Again, all the kernel-based indices retain a similar behavior for both choices of ν_0 . In contrast, the Wasserstein indices show the same issues as in [Figure 2.4](#) due to the infinite variance of the baseline measure for $\nu_0 = \text{Cauchy}(0, 1)$.

2.5 Application: Hierarchical Clustering on Resting-State Functional Magnetic Resonance Imaging

We apply the indices above to study resting-state functional magnetic resonance imaging (rs-fMRI) data from the Human Connectome Project (HCP) Young Adult cohort ([Van Essen et al., 2013](#)) for the brain cortical areas of 1080 individuals. Specifically, we use data from the minimally preprocessed `REST1_LR` run, corresponding to the first resting-state scan with a left-to-right phase-encoding direction. As described in [Smith et al. \(2013\)](#), a full-volume brain scan, composed of isotropic voxels with 2 mm sides, is acquired every 720 ms, resulting in 1,200 frames per run and a total duration of approximately 15 minutes. The minimal preprocessing pipeline, detailed in [Glasser et al. \(2013\)](#), was applied to this data. This pipeline was designed to remove common imaging artifacts, such as spatial distortions and modality- and subject-level misalignments, while preserving the signal characteristics necessary for functional connectivity analyses.

At each time point, the scanner records a brain map of blood oxygenation levels, producing a Blood Oxygenation Level-Dependent (BOLD) signal. This yields a BOLD time series for each voxel: a sequence of signal intensities reflecting local fluctuations in neural activity. To facilitate our analysis, we employed the Schaefer parcellation, which comprises 100 regions of interest (ROIs) ([Schaefer et al., 2017](#)), to obtain a computationally manageable representation of cortical organization. Each ROI in the Schaefer parcellation can be associated with one of the Yeo 7-networks identified in [Yeo et al. \(2011\)](#). This

functional network framework represents a canonical clustering of the brain based on seven functional networks (visual, somatomotor, dorsal attention, ventral attention, limbic, frontoparietal, and default mode). For each subject, we extracted the mean BOLD time series from each of the 100 ROIs defined by the Schaefer parcellation.

For each of the 100 ROIs in the Schaefer parcellation, we model the distribution of BOLD signals across the 1,200 time frames as a random probability measure. Each subject’s observed BOLD time series at a given ROI is treated as one realization from this random probability measure, yielding $n = 1,080$ realizations per ROI. Then, we construct functional connectivity matrices by computing the correlation coefficients defined above between pairs of ROIs, based on their corresponding random probability measure. This results in a symmetrical 100×100 connectivity matrix $\mathbf{R}$. We explicitly avoid Canonical Correlation as it shows high dependence on the regularization parameter, as shown in [Section 2.4](#).

Then, we compute the dissimilarity matrix $\sqrt{2(1 - \mathbf{R}_{i,j})}$ for $i, j = 1, \dots, 100$ and apply hierarchical clustering with complete linkage to partition the ROIs into 7 groups. We assess the clustering performance using the Yeo 7-network as a reference. The Normalized Mutual Information (NMI), Adjusted Mutual Information (AMI), and Adjusted Rand Score (ARS) between the obtained clusters and the Yeo network are reported in [Table 2.2](#). The results suggest that both kernel-based Centered Alignment indices achieve superior performance in this task.

[Figure 2.6](#) displays the original Yeo 7-network classification for the Schaefer parcellation

Table 2.2: Normalized Mutual Information (NMI), Adjusted Mutual Information (AMI), Adjusted Rand Score (ARS) for correlation-based hierarchical clustering for different indices with respect to the Yeo 7-network.

Index	$\text{CA}_{\mathcal{W}}$	$\text{Corr}_{\mathcal{W}}$	CA_k	Corr_k	CA_K	Corr_K
NMI	0.6484	0.6830	0.7445	0.5646	0.7445	0.5646
AMI	0.6040	0.6427	0.7120	0.5081	0.7120	0.5081
ARS	0.4464	0.5012	0.5990	0.3029	0.5990	0.3029

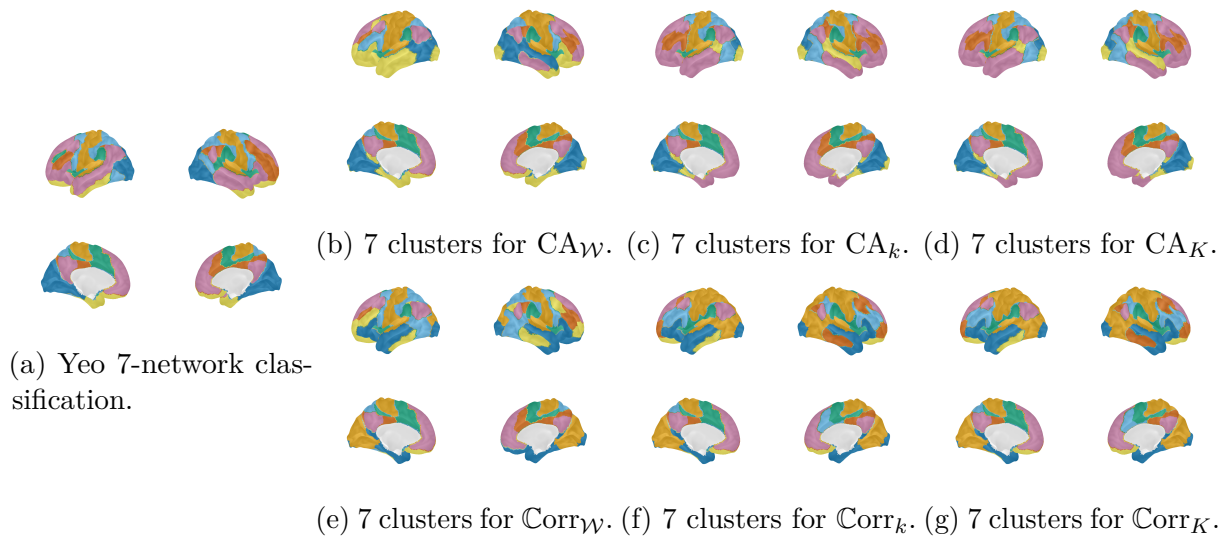


Figure 2.6: The Yeo 7-network clusters and the results of correlation-based hierarchical clustering for Centered Alignment and Trace Correlation indices for different choices of the embedding. Each image consists of 4 views of the brain: the left and right lateral views (the outer sides of the brain) are displayed above, while the left and right medial views (the inner parts of the brain) are displayed below.

along with the 7 groups obtained via hierarchical clustering for different correlation indices, as described above. For clarity, we color each cluster to match the closest corresponding cluster in the Yeo 7-network. Overall, the clusters obtained using the Centered Alignment indices for both kernel-based embeddings most closely resemble the Yeo 7-network classification. These results suggest their potential utility for developing new functional brain parcellations.

Discussion

In this work, we have introduced several correlation measures for distributional data by projecting random probability measures into Hilbert spaces via a Wasserstein embedding, a kernel mean embedding, and a feature map embedding. For each embedding, we have defined three correlation indices, Canonical Correlation, Centered Alignment, and Trace Correlation, studying their theoretical properties and conducting numerical experiments

to compare their values in various scenarios.

Overall, no index can capture both independence and some notion of deterministic dependence, such as almost sure equality or the existence of a linear map pushing forward one random probability measure onto the other. The definition of such an index remains an open problem for future research. Specifically, under mild assumptions, $\mathbb{C}A_K$ and $\mathbb{C}C_K$ detect independence at 0, while, under different sets of assumptions, both $\mathbb{C}\text{Corr}_K$ and $\mathbb{C}\text{Corr}_k$ detect almost sure equality at 1. Conversely, the Wasserstein-based index $\mathbb{C}\text{Corr}_W$ is the only one that is equal to 1 whenever there exists a positive linear map pushing forward one random probability measure. However, the converse implication does not generally hold. It would be of some interest to investigate if there exists a set of assumptions that guarantees the feasibility of the converse implication.

Moreover, we have observed that, numerically, the kernel mean embedding and the feature map embedding yield similar values for a given index, suggesting further investigation to determine whether there is a theoretical explanation for this phenomenon. This similarity is particularly noteworthy since $\mathbb{C}\text{Corr}_k$ has the advantage of having a closed-form expression, which is equal to the set-wise correlation, for injective kernels when applied to a Multivariate Species Sampling Process (mSSP), as shown in [Catalano et al. \(2025\)](#). mSSPs, first introduced in [Franzolini et al. \(2025\)](#), constitute a broad class of models for dependent random probability measures widely employed in Bayesian Nonparametrics. The equality between $\mathbb{C}\text{Corr}_k$ and the set-wise correlation for injective kernels for this broad class of models is noteworthy since [Franzolini et al. \(2025\)](#) also show that the set-wise correlation can detect both almost sure equality and independence under some assumptions for mSSPs. Thus, $\mathbb{C}\text{Corr}_k$ for k injective inherits these properties. It would be interesting to explore whether similar results hold for $\mathbb{C}\text{Corr}_K$ when applied to mSSPs.

Lastly, another interesting direction for future research is to explore alternative Wasserstein-based embeddings into Hilbert spaces, going beyond the one-dimensional setting by considering the geometry of the Wasserstein space in higher dimensions or the

sliced-Wasserstein metric ([Rabin et al., 2012](#)).

Acknowledgments

In [Section 2.5](#), data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657), funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

Appendices

Below, we provide additional materials that were omitted or summarized in [Chapter 2](#). The content is organized as follows.

- In [Appendix 2.A](#), we prove the statements in [Chapter 2](#).
- In [Appendix 2.B](#), we provide some examples related to the statements and the results in [Chapter 2](#).
- In [Appendix 2.C](#), we describe the numerical methods used to compute the entries of the centered Gram matrix $\hat{\Sigma}_{\tilde{P}, \tilde{Q}}$.

2.A Proofs of the Statements

2.A.1 Proof of [Corollary 2.3.1](#)

The result follows directly from the following lemma.

Lemma 2.A.1. *Let ν be a probability measure on $\mathbb{R}$ and T be a strictly monotone continuous map. Then, for any $t \in [0, 1]$,*

$$F_{T\#\nu}^{-1}(t) = \begin{cases} T \circ F_{\nu}^{-1}(t), & \text{for } T \text{ strictly increasing,} \\ T \circ F_{\nu}^{-1,+}(1-t), & \text{for } T \text{ strictly decreasing,} \end{cases}$$

where $F_\nu^{-1,+}(u) = \inf\{x \in \mathbb{R} \text{ s.t. } \nu((-\infty, x]) > u\}$ for $u \in [0, 1]$. If ν is atomless, $F_\nu^{-1,+}(u) = F_\nu^{-1}(u)$ for any $u \in [0, 1]$.

Proof. For T strictly increasing, we have

$$F_{T_\# \nu}^{-1}(t) = \inf\{x \text{ s.t. } T_\# \nu((-\infty, x]) \geq t\} = \inf\{x \text{ s.t. } \nu((-\infty, T^{-1}(x)]) \geq t\}.$$

Now, let $y = T^{-1}(x)$; thus, $x = T(y)$. Then, $F_{T_\# \nu}^{-1}(t) = \inf\{T(y) \text{ s.t. } \nu((-\infty, y]) \geq t\}$. Hence, the first equality in the statement follows by noticing that

$$\inf\{T(y) \text{ s.t. } \nu((-\infty, y]) \geq t\} = T\left(\inf\{y \text{ s.t. } \nu((-\infty, y]) \geq t\}\right),$$

since T is increasing and continuous.

For T strictly decreasing, first, we notice that any strictly decreasing function can be expressed as the composition of a strictly increasing function and the map $x \mapsto -x$. Hence, without loss of generality, we can assume that T is the map $x \mapsto -x$. In this case, $F_{T_\# \nu}^{-1}(t) = \inf\{x \text{ s.t. } \nu([-x, +\infty)) \geq t\}$. Now, for $x = -y$, the second equality in the statement follows by noticing that $F_{T_\# \nu}^{-1}(t)$ can be rewritten as $F_{T_\# \nu}^{-1}(t) = -\sup\{y \text{ s.t. } \nu((-\infty, y)) \leq 1 - t\}$, while $F_\nu^{-1,+}(1 - t)$ can be rewritten as $F_\nu^{-1,+}(1 - t) = \inf\{y \text{ s.t. } \nu((-\infty, y)) > 1 - t\}$. Hence, we only need to prove that

$$\sup\{y \text{ s.t. } \nu((-\infty, y)) \leq 1 - t\} = \inf\{y \text{ s.t. } \nu((-\infty, y)) > 1 - t\}.$$

In order to prove this equality, let the left-hand side be denoted by η and the right-hand side by λ . We want to prove $\eta = \lambda$. For any z such that $\nu((-\infty, z)) \leq 1 - t$ and any y such that $\nu((-\infty, y)) > 1 - t$ it holds $\nu((-\infty, z)) < \nu((-\infty, y))$. Since the function $x \mapsto \nu((-\infty, x))$ is non-decreasing, it holds $z \leq y$. If we take the supremum on the left-hand side and the infimum on the right-hand side, we obtain $\eta \leq \lambda$. Conversely, for any $z < \lambda$, by definition it holds $\nu((-\infty, z)) < 1 - t$. Hence, $\nu((-\infty, z)) \leq 1 - t$, leading

to $\{z < \lambda\} \subset \{z \text{ s.t. } \nu((-\infty, z)) \leq 1 - t\}$. By taking the supremum on both sides, we obtain $\eta \geq \lambda$.

Lastly, the equality between F_ν^{-1} and $F_\nu^{-1,+}$ for ν atomless follows directly by definition. $\square$

2.A.2 Proof of Corollary 2.3.2

It is trivial to prove that $\text{Corr}_k(\tilde{P}, \tilde{Q}) = 1$ whenever $\tilde{P} = \tilde{Q}$ almost surely.

Let us prove the converse implication. By Proposition 1.2.2 and by linearity of μ_k , there exists $\alpha > 0$ such that $\mu_k(\tilde{P} - \mathbb{E}[\tilde{P}]) = \mu_k(\alpha(\tilde{Q} - \mathbb{E}[\tilde{Q}]))$ almost surely. Now, since k is c_0 -universal, μ_k is injective for finite signed measures as proven in Sriperumbudur et al. (2011). Hence, it holds that $\tilde{P} - \mathbb{E}[\tilde{P}] = \alpha(\tilde{Q} - \mathbb{E}[\tilde{Q}])$ almost surely, which we rewrite as

$$\tilde{P} - \alpha\tilde{Q} = \mathbb{E}[\tilde{P}] - \alpha\mathbb{E}[\tilde{Q}] \quad \text{almost surely.}$$

Since the random measure on the left is almost surely discrete, while the measure on the right is atomless, both sides must be null almost surely. In other words, $\tilde{P} = \alpha\tilde{Q}$ almost surely. Evaluating both sides of this equality at $\mathbb{X}$, we obtain that $\alpha = 1$ as $\tilde{P}(\mathbb{X}) = \tilde{Q}(\mathbb{X}) = 1$ almost surely.

2.A.3 Proof of Corollary 2.3.3

This proof is analogous to that of Corollary 2.3.2.

It is trivial to prove that $\text{Corr}_K(\tilde{P}, \tilde{Q}) = 1$ whenever $\tilde{P} = \tilde{Q}$ almost surely.

Let us prove the converse implication. Firstly, let us notice that

$$\phi_K(\tilde{P}) = K(\tilde{P}, \cdot) = \mu_K(\delta_{\tilde{P}}) \quad \text{and} \quad \phi_K(\tilde{Q}) = K(\tilde{Q}, \cdot) = \mu_K(\delta_{\tilde{Q}}).$$

By Proposition 1.2.2 and by linearity of μ_K , there exists $\alpha > 0$ such that $\mu_K(\delta_{\tilde{P}} - \mathbb{E}[\delta_{\tilde{P}}]) =$

$\mu_K(\alpha(\delta_{\tilde{Q}} - \mathbb{E}[\delta_{\tilde{Q}}]))$ almost surely. Now, since K is universal on a compact set $\mathcal{P}(\mathbb{X})$ as in Christmann and Steinwart (2010)[Example 1], μ_K is injective for finite signed measures as proven in Sriperumbudur et al. (2011). Hence, $\delta_{\tilde{P}} - \mathbb{E}[\delta_{\tilde{P}}] = \alpha(\delta_{\tilde{Q}} - \mathbb{E}[\delta_{\tilde{Q}}])$ almost surely, which we rewrite as

$$\delta_{\tilde{P}} - \alpha\delta_{\tilde{Q}} = \mathbb{E}[\delta_{\tilde{P}}] - \alpha\mathbb{E}[\delta_{\tilde{Q}}] \quad \text{almost surely.}$$

Noticing that $\mathbb{E}[\delta_{\tilde{P}}] = \mathcal{L}_{\tilde{P}}$ and $\mathbb{E}[\delta_{\tilde{Q}}] = \mathcal{L}_{\tilde{Q}}$, where $\mathcal{L}_{\tilde{P}}$ and $\mathcal{L}_{\tilde{Q}}$ are the laws of $\tilde{P}$ and $\tilde{Q}$ respectively, we can conclude that the measure on the right-hand side is atomless. Consequently, since the random measure on the left-hand side is almost surely discrete, both sides must be null almost surely. In other words, $\delta_{\tilde{P}} = \alpha\delta_{\tilde{Q}}$ almost surely. Evaluating both sides of this equality at $\mathcal{P}(\mathbb{X})$, we obtain that $\alpha = 1$ since $\delta_{\tilde{P}}(\mathcal{P}(\mathbb{X})) = \delta_{\tilde{Q}}(\mathcal{P}(\mathbb{X})) = 1$ almost surely. This leads to $\delta_{\tilde{P}} = \delta_{\tilde{Q}}$, i.e., $\tilde{P} = \tilde{Q}$, almost surely.

2.A.4 Proof of Theorem 2.3.4

Firstly, let us recall that $\|C_{\tilde{P}, \tilde{Q}}\|_{\mathbb{H}}^2 = \text{HSIC}(\tilde{P}, \tilde{Q})$ as in Gretton et al. (2005a) for the RKHS $\mathbb{H}_K$ associated to the kernel K . This implies that $\|C_{\tilde{P}, \tilde{Q}}\|_{\mathbb{H}}^2 = 0$ is equivalent to $\text{HSIC}(\tilde{P}, \tilde{Q}) = 0$.

Now, since K is universal on a compact set $\mathcal{P}(\mathbb{X})$ as in Christmann and Steinwart (2010)[Example 1], the independence criterion in Gretton et al. (2005b) implies that $\text{CC}_K(\tilde{P}, \tilde{Q}; \varepsilon) = 0$ is equivalent to $\tilde{P} \perp \tilde{Q}$, while the independence criterion for HSIC in Gretton et al. (2005a) implies that $\text{HSIC}(\tilde{P}, \tilde{Q}) = 0$ is equivalent to $\tilde{P} \perp \tilde{Q}$.

2.B Examples

Example 2.B.1. Let $\tilde{P}$ and $\tilde{Q}$ be random probability measures such that $\tilde{P} = \text{Unif}_{[X, X+1]}$ and $\tilde{Q} = \text{Unif}_{[X^2, X^2+1]}$ almost surely for a symmetric real-valued random variable X with

0 mean and finite positive second moment. Then, $F_{\tilde{P}}^{-1}(t) = t + X$ and $F_{\tilde{Q}}^{-1}(t) = t + X^2$ for any $t \in [0, 1]$ almost surely. Clearly, $\tilde{P}$ and $\tilde{Q}$ are not independent. However, we show that $\langle \iota(\tilde{P}), f \rangle_{\mathbb{H}}$ and $\langle \iota(\tilde{Q}), g \rangle_{\mathbb{H}}$ are uncorrelated for any $f, g \in \mathbb{H} = L^2(\text{Leb}; [0, 1])$. Firstly, let us notice that

$$\langle F_{\tilde{P}}^{-1}(t), f \rangle_{\mathbb{H}} - \mathbb{E}[\langle F_{\tilde{P}}^{-1}(t), f \rangle_{\mathbb{H}}] = \int_0^1 (F_{\tilde{P}}^{-1}(t) - \mathbb{E}[F_{\tilde{P}}^{-1}(t)])f(t)dt = X \int_0^1 f(t)dt$$

and similarly

$$\langle F_{\tilde{Q}}^{-1}(t), g \rangle_{\mathbb{H}} - \mathbb{E}[\langle F_{\tilde{Q}}^{-1}(t), g \rangle_{\mathbb{H}}] = \int_0^1 (F_{\tilde{Q}}^{-1}(t) - \mathbb{E}[F_{\tilde{Q}}^{-1}(t)])g(t)dt = (X^2 - \mathbb{E}[X^2]) \int_0^1 g(t)dt.$$

Hence, we obtain

$$\mathbb{Cov}(\langle F_{\tilde{P}}^{-1}, f \rangle_{\mathbb{H}}, \langle F_{\tilde{Q}}^{-1}, g \rangle_{\mathbb{H}}) = (\mathbb{E}[X^3] - \mathbb{E}[X^2]\mathbb{E}[X]) \int_0^1 f(t)dt \int_0^1 g(t)dt = 0$$

since $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^3] = 0$ as X is symmetric with mean 0.

Hence, due to [Proposition 1.2.1](#), $\text{CC}_{\mathcal{W}}(\tilde{P}, \tilde{Q}; \varepsilon) = 0$ for any $\varepsilon \geq 0$, $\text{CA}_{\mathcal{W}}(\tilde{P}, \tilde{Q}) = 0$, and $\text{Corr}_{\mathcal{W}}(\tilde{P}, \tilde{Q}) = 0$.

Example 2.B.2. Let $\tilde{P}$ and $\tilde{Q}$ be random probability measures such that $\tilde{P} = \text{Unif}_{[X, X+1]}$ and $\tilde{Q} = \text{Unif}_{[X, X+2]}$ almost surely for a real-valued random variable X with finite positive variance. Then, $F_{\tilde{P}}^{-1}(t) = t + X$ and $F_{\tilde{Q}}^{-1}(t) = 2t + X$ for any $t \in [0, 1]$ almost surely. Consequently, we can write $F_{\tilde{P}}^{-1}(t) = F_{\tilde{Q}}^{-1}(t) - t$ almost surely. Now, since

$$F_{\tilde{P}}^{-1}(t) - \mathbb{E}[F_{\tilde{P}}^{-1}(t)] = F_{\tilde{Q}}^{-1}(t) - \mathbb{E}[F_{\tilde{Q}}^{-1}(t)] = X - \mathbb{E}[X].$$

we have $\mathbb{Cov}_{\mathcal{W}}(\tilde{P}, \tilde{Q}) = \mathbb{Var}(X)$, and $\mathbb{Var}_{\mathcal{W}}(\tilde{P}) = \mathbb{Var}_{\mathcal{W}}(\tilde{Q}) = \mathbb{Var}(X)$. Thus, we obtain $\text{Corr}_{\mathcal{W}}(\tilde{P}, \tilde{Q}) = 1$.

However, we cannot write $\tilde{P} = T_{\#}\tilde{Q}$ for $T : x \mapsto ax + b$ for any $a \in \mathbb{R} \setminus \{0\}$ and $b \in \mathbb{R}$

almost surely, since $\tilde{P} = T_{\#}\tilde{Q} = \text{Unif}_{[aX+b, aX+2a+b]}$ almost surely would lead to the system

$$\begin{cases} aX + b = X & \text{almost surely,} \\ aX + 2a + b = X + 1 & \text{almost surely,} \end{cases}$$

which is unsolvable for deterministic a and b .

Example 2.B.3. Let X be a non-constant positive-valued random variable with finite positive variance. We are going to compute $\mathbb{C}\text{orr}_{\mathcal{W}}(\tilde{P}, \tilde{Q})$ for three different choices of $\tilde{P}$: $\tilde{P} = \text{Unif}_{[X, X+1]}$ almost surely, $\tilde{P} = \text{Unif}_{[-X, X]}$ almost surely, or $\tilde{P} = \text{Unif}_{[0, X]}$ almost surely, with $\tilde{Q} = T_{\#}\tilde{P}$ for $T(x) = ax + b$ for $a < 0$ and $b \in \mathbb{R}$.

First, let us notice that if $\tilde{P}$ is atomless almost surely and $\tilde{Q} = T_{\#}\tilde{P}$ for $T(x) = ax + b$ with $a < 0$, by [Lemma 2.A.1](#), $F_{\tilde{Q}}^{-1}(t) = aF_{\tilde{P}}^{-1}(1-t) + b$ for any $t \in [0, 1]$. Hence,

$$\mathbb{C}\text{orr}_{\mathcal{W}}(\tilde{P}, \tilde{Q}) = -\frac{\int_0^1 \mathbb{C}\text{ov}(F_{\tilde{P}}^{-1}(t), F_{\tilde{P}}^{-1}(1-t))dt}{\int_0^1 \mathbb{V}\text{ar}(F_{\tilde{P}}^{-1}(t))dt}.$$

For $\tilde{P} = \text{Unif}_{[X, X+1]}$ almost surely, $F_{\tilde{P}}^{-1}(t) = X + t$. Thus, $\mathbb{C}\text{orr}_{\mathcal{W}}(\tilde{P}, \tilde{Q}) = -1$ since

$$\begin{aligned} \int_0^1 \mathbb{C}\text{ov}(F_{\tilde{P}}^{-1}(t), F_{\tilde{P}}^{-1}(1-t))dt &= \mathbb{V}\text{ar}(X) \int_0^1 dt = \mathbb{V}\text{ar}(X), \\ \int_0^1 \mathbb{V}\text{ar}(F_{\tilde{P}}^{-1}(t))dt &= \mathbb{V}\text{ar}(X) \int_0^1 dt = \mathbb{V}\text{ar}(X). \end{aligned}$$

For $\tilde{P} = \text{Unif}_{[-X, X]}$ almost surely, $F_{\tilde{P}}^{-1}(t) = X(2t - 1)$. Thus, $\mathbb{C}\text{orr}_{\mathcal{W}}(\tilde{P}, \tilde{Q}) = 1$ since

$$\begin{aligned} \int_0^1 \mathbb{C}\text{ov}(F_{\tilde{P}}^{-1}(t), F_{\tilde{P}}^{-1}(1-t))dt &= -\mathbb{V}\text{ar}(X) \int_0^1 (2t - 1)^2 dt = -\frac{1}{3}\mathbb{V}\text{ar}(X), \\ \int_0^1 \mathbb{V}\text{ar}(F_{\tilde{P}}^{-1}(t))dt &= \mathbb{V}\text{ar}(X) \int_0^1 (2t - 1)^2 dt = \frac{1}{3}\mathbb{V}\text{ar}(X). \end{aligned}$$

For $\tilde{P} = \text{Unif}_{[0,X]}$ almost surely, $F_{\tilde{P}}^{-1}(t) = Xt$. Thus, $\text{Corr}_{\mathcal{W}}(\tilde{P}, \tilde{Q}) = -1/2$ since

$$\begin{aligned} \int_0^1 \text{Cov}(F_{\tilde{P}}^{-1}(t), F_{\tilde{P}}^{-1}(1-t)) dt &= \text{Var}(X) \int_0^1 t(1-t) dt = \frac{\text{Var}(X)}{6}, \\ \int_0^1 \text{Var}(F_{\tilde{P}}^{-1}(t)) dt &= \text{Var}(X) \int_0^1 t^2 dt = \frac{\text{Var}(X)}{3}. \end{aligned}$$

Example 2.B.4. Let $\tilde{P}$ and $\tilde{Q}$ be random probability measures such that $\tilde{P} = \pi\delta_{\theta} + (1-\pi)\nu_0$ and $\tilde{Q} = \pi\delta_{\eta} + (1-\pi)\nu_0$ almost surely with $\pi \sim \text{Beta}(\alpha, \beta)$ and $\theta, \eta \stackrel{\text{i.i.d.}}{\sim} \nu_0$ for ν_0 atomless probability distribution on $\mathbb{X}$ and $\pi \perp \theta, \eta$. Thus, $\mathbb{E}[\tilde{P}] = \mathbb{E}[\tilde{Q}] = \nu_0$.

Clearly, $\tilde{P}$ and $\tilde{Q}$ are not independent. However, we show that $\langle \mu_k(\tilde{P}), f \rangle_{\mathbb{H}}$ and $\langle \mu_k(\tilde{Q}), g \rangle_{\mathbb{H}}$ are uncorrelated for any $f, g \in \mathbb{H}_k$. Firstly, let us notice that

$$\langle \mu_k(\tilde{P}), f \rangle_{\mathbb{H}} - \mathbb{E}[\langle \mu_k(\tilde{P}), f \rangle_{\mathbb{H}}] = \pi \left(f(\theta) - \int f(x) d\nu_0(x) \right)$$

and similarly

$$\langle \mu_k(\tilde{Q}), g \rangle_{\mathbb{H}} - \mathbb{E}[\langle \mu_k(\tilde{Q}), g \rangle_{\mathbb{H}}] = \pi \left(g(\eta) - \int g(x) d\nu_0(x) \right)$$

Hence, we obtain

$$\text{Cov}\left(\langle \mu_k(\tilde{P}), f \rangle_{\mathbb{H}}, \langle \mu_k(\tilde{Q}), g \rangle_{\mathbb{H}}\right) = \mathbb{E}[\pi^2] \mathbb{E}\left[f(\theta) - \int f(x) d\nu_0(x)\right] \mathbb{E}\left[g(\eta) - \int g(x) d\nu_0(x)\right],$$

which is null. Hence, due to [Proposition 1.2.1](#), $\text{CC}_k(\tilde{P}, \tilde{Q}; \varepsilon) = 0$ for any $\varepsilon \geq 0$, $\text{CA}_k(\tilde{P}, \tilde{Q}) = 0$, and $\text{Corr}_k(\tilde{P}, \tilde{Q}) = 0$.

Example 2.B.5. If $\tilde{P}$ is a Dirichlet Process, it is an instance of a Species Sampling Process (SSP) as introduced in [Pitman \(1996\)](#). A Species Sampling Process (SSP) with baseline measure ν_0 can be represented as

$$\tilde{P} = \sum_{i=1}^{\infty} \pi_i \delta_{\theta_i} + \left(1 - \sum_{i=1}^{\infty} \pi_i\right) \nu_0 \quad \text{almost surely,} \quad (2.B.1)$$

for $\{\pi_i\}_{i \geq 1}$ random weights such that $\pi_i \geq 0$ for $i \in \mathbb{N}$ and $0 < \sum_{i=1}^{\infty} \pi_i \leq 1$ almost surely and $\{\theta_i\}_{i \geq 1}$ random atoms such that $\theta_i \stackrel{\text{i.i.d.}}{\sim} \nu_0$ for $i \in \mathbb{N}$, with $\{\pi_i\}_{i \geq 1} \perp \{\theta_i\}_{i \geq 1}$.

Let $\tilde{Q} = T_{\#}\tilde{P}$ for $T(x) = ax + b$ for $a, b \in \mathbb{R}$ with $a \neq 0$. Then, $\mathbb{E}[\tilde{P}] = \nu_0$ and $\mathbb{E}[\tilde{Q}] = T_{\#}\nu_0$. For this choice of $\tilde{P}$ and $\tilde{Q}$, the value of $\mathbb{C}\text{orr}_k(\tilde{P}, \tilde{Q})$ is positively proportional to

$$\mathbb{E} \left[\langle \mu_k(\tilde{P}) - \mathbb{E}[\mu_k(\tilde{P})], \mu_k(\tilde{Q}) - \mathbb{E}[\mu_k(\tilde{Q})] \rangle_{\mathbb{H}_k} \right],$$

which can be rewritten as

$$\mathbb{E} \left[\int k(x, T(y)) d\tilde{P}(x) d\tilde{P}(y) \right] - \iint k(x, T(y)) d\nu_0(x) d\nu_0(y)$$

due to the reproducing property of $\mathbb{H}_k$ and the linearity of $\mathbb{E}$ and μ_k . Now, due to the representation of $\tilde{P}$ in (2.B.1) and the independence between $\{\pi_i\}_{i \geq 1}$ and $\{\theta_i\}_{i \geq 1}$, the first term can be rewritten as

$$\begin{aligned} \mathbb{E} \left[\int k(x, T(y)) d\tilde{P}(x) d\tilde{P}(y) \right] &= \mathbb{E} \left[\sum_{i=1}^{\infty} \pi_i^2 \int k(x, T(x)) d\nu_0(x) \right. \\ &\quad \left. + \left(1 - \mathbb{E} \left[\sum_{i=1}^{\infty} \pi_i^2 \right] \right) \iint k(x, T(y)) d\nu_0(x) d\nu_0(y) \right]. \end{aligned}$$

Hence, $\mathbb{C}\text{orr}_k(\tilde{P}, \tilde{Q})$ is positively proportional to

$$\int k(x, T(x)) d\nu_0(x) - \iint k(x, T(y)) d\nu_0(x) d\nu_0(y),$$

since $\mathbb{E} \left[\sum_{i=1}^{\infty} \pi_i^2 \right] > 0$ as all $\pi_i \geq 0$ and $0 < \sum_{i=1}^{\infty} \pi_i \leq 1$ almost surely.

Now, let us study this expression for the Gaussian kernel $k(x, y) = \exp(-\|x - y\|^2 / (2\sigma^2))$ for $\sigma > 0$ and $\nu_0 = \mathcal{N}(0, \gamma^2)$ for $\gamma > 0$ on $\mathbb{X} = \mathbb{R}$. Then, $T(x) = ax + b$ for $a, b \in \mathbb{R}$ with $a \neq 0$. Let

$$I_1 := \int k(x, T(x)) d\nu_0(x), \quad \text{and} \quad I_2 := \iint k(x, T(y)) d\nu_0(x) d\nu_0(y).$$

Then, $\mathbb{C}\text{orr}_k(\tilde{P}, \tilde{Q}) > 0$ if and only if $I_1 > I_2$. By direct computation,

$$I_1 = \frac{\sigma}{\sqrt{D_1}} \exp\left(-\frac{b^2}{2D_1}\right) \quad \text{and} \quad I_2 = \frac{\sigma}{\sqrt{D_2}} \exp\left(-\frac{b^2}{2D_2}\right)$$

with $D_1 := \sigma^2 + (1-a)^2\gamma^2$ and $D_2 = \sigma^2 + (1+a^2)\gamma^2$. Since, $D_1, D_2 > 0$ with $D_2 = D_1 + 2a\gamma^2$, if $a > 0$, then $D_2 > D_1$ and $\ln(D_2/D_1) > 0$, which implies that $I_1 > I_2$ if and only if

$$|b| < \sqrt{\frac{D_1 D_2}{2a\gamma^2} \ln\left(\frac{D_2}{D_1}\right)}.$$

Conversely, if $a < 0$, then $D_2 < D_1$ and $\ln(D_2/D_1) < 0$, which implies that $I_1 > I_2$ if and only if

$$|b| > \sqrt{\frac{D_1 D_2}{2a\gamma^2} \ln\left(\frac{D_2}{D_1}\right)}.$$

Example 2.B.6. Let $\tilde{P}$ and $\tilde{Q}$ be random probability measures such that $\tilde{P} = \pi\delta_0 + (1-\pi)\nu$ and $\tilde{Q} = (1-\pi)\delta_0 + \pi\nu$ for $\nu = \text{Unif}_{[0,1]}$ and π random variable valued on $[0, 1]$ with $\mathbb{E}[\pi] = 1/2$ almost surely. Thus, $\mathbb{E}[\tilde{P}] = \mathbb{E}[\tilde{Q}] = \frac{1}{2}\delta_0 + \frac{1}{2}\nu$.

Let $(\tilde{P}', \tilde{Q}')$ be an independent copy of $(\tilde{P}, \tilde{Q})$. Then, since

$$\mu_k(\tilde{P}) - \mathbb{E}[\mu_k(\tilde{P})] = \mathbb{E}[\mu_k(\tilde{Q})] - \mu_k(\tilde{Q}) = \left(\pi - \frac{1}{2}\right) (\mu_k(\delta_0) - \mu_k(\nu)),$$

both $\langle \mu_k(\tilde{P}) - \mathbb{E}[\mu_k(\tilde{P})], \mu_k(\tilde{P}') - \mathbb{E}[\mu_k(\tilde{P}')] \rangle_{\mathbb{H}}$ and $\langle \mu_k(\tilde{Q}) - \mathbb{E}[\mu_k(\tilde{Q})], \mu_k(\tilde{Q}') - \mathbb{E}[\mu_k(\tilde{Q}')] \rangle_{\mathbb{H}}$ are equal to

$$\left(\pi - \frac{1}{2}\right) \left(\pi' - \frac{1}{2}\right) \left(1 - 2 \int_0^1 k(0, y) dy + \int_0^1 \int_0^1 k(x, y) dx dy\right).$$

Consequently, we can conclude by [Proposition 1.2.3](#) that $\text{CA}_k(\tilde{P}, \tilde{Q}) = 1$. However, we cannot write $\tilde{Q} = T_{\#}\tilde{P}$ almost surely for some $T(x) = \pm x + b$ for $b \in \mathbb{R}$.

2.C Numerical Implementation of the Indices

In this section, we summarize the numerical implementation of the correlation indices defined in [Chapter 2](#) for the different choices of space of measures $\mathcal{S}$, Hilbert space $\mathbb{H}$, and embedding $\iota : \mathcal{S} \rightarrow \mathbb{H}$, as summarized in [Table 2.1](#), for n realizations $(P_1, Q_1), \dots, (P_n, Q_n)$ of independent copies of $(\tilde{P}, \tilde{Q})$.

The sample versions of our correlation indices are based on the evaluation of the centered Gram matrices $\hat{\Sigma}_{\tilde{P}, \tilde{Q}}$, $\hat{\Sigma}_{\tilde{P}, \tilde{P}}$, $\hat{\Sigma}_{\tilde{Q}, \tilde{Q}}$. It can be easily proved that

$$\hat{\Sigma}_{\tilde{P}, \tilde{Q}} = \mathbf{H}^t \hat{\Sigma}_{\tilde{P}, \tilde{Q}}^\dagger \mathbf{H} \quad \text{for } \mathbf{H} = \mathbf{I}_n - \frac{1}{n} \mathbf{1}_n^t \mathbf{1}_n,$$

where $(\hat{\Sigma}_{\tilde{P}, \tilde{Q}}^\dagger)_{i,j} = \langle \iota(P_i), \iota(Q_j) \rangle_{\mathbb{H}}$ for $i, j = 1, \dots, n$, $\mathbf{I}_n$ is the $n \times n$ identity matrix and $\mathbf{1}_n$ is the column vector of length n having all entries 1. Similarly, we can rewrite $\hat{\Sigma}_{\tilde{P}, \tilde{P}} = \mathbf{H}^t \hat{\Sigma}_{\tilde{P}, \tilde{P}}^\dagger \mathbf{H}$ and $\hat{\Sigma}_{\tilde{Q}, \tilde{Q}} = \mathbf{H}^t \hat{\Sigma}_{\tilde{Q}, \tilde{Q}}^\dagger \mathbf{H}$, where $(\hat{\Sigma}_{\tilde{P}, \tilde{P}}^\dagger)_{i,j} = \langle \iota(P_i), \iota(P_j) \rangle_{\mathbb{H}}$ and $(\hat{\Sigma}_{\tilde{Q}, \tilde{Q}}^\dagger)_{i,j} = \langle \iota(Q_i), \iota(Q_j) \rangle_{\mathbb{H}}$ for $i, j = 1, \dots, n$.

Firstly, recall that we approximate all probability measures with finite sums of atoms and weights of size R , i.e., for every $i = 1, \dots, n$, we find some $\theta_{i,r}, \eta_{i,r} \in \mathbb{X}$ and $\pi_{i,r}, \omega_{i,r} \geq 0$ for $r = 1, \dots, R$ with $\sum_{r=1}^R \pi_{i,r} = \sum_{r=1}^R \omega_{i,r} = 1$, such that

$$P_i \approx \sum_{r=1}^R \pi_{i,r} \delta_{\theta_{i,r}}, \quad Q_i \approx \sum_{r=1}^R \omega_{i,r} \delta_{\eta_{i,r}}. \quad (2.C.1)$$

Hence, the numerical implementation of the different indices boils down to the approximation of the inner products $\langle \iota(P_i), \iota(Q_j) \rangle_{\mathbb{H}}$ for $i, j = 1, \dots, n$.

Wasserstein embedding For the Wasserstein indices, we take $\mathcal{S} = \mathcal{P}_2(\mathbb{R})$, $\mathbb{H} = L^2(\text{Leb}; [0, 1])$ and ι as the function q mapping each probability measure to its quantile

function. Consequently, for any $P, Q \in \mathcal{P}_2(\mathbb{R})$, it holds

$$\langle q(P_i), q(Q_j) \rangle_{\mathbb{H}} = \int_0^1 F_{P_i}^{-1}(t) F_{Q_j}^{-1}(t) dt$$

We approximate this quantity using a Monte Carlo scheme as

$$\langle q(P_i), q(Q_j) \rangle_{\mathbb{H}} \approx \frac{1}{M} \sum_{s=1}^M F_{P_i}^{-1}(t_s) F_{Q_j}^{-1}(t_s),$$

for $t_1, \dots, t_M \stackrel{\text{i.i.d.}}{\sim} \text{Unif}_{[0,1]}$ for $M = 10,000$. Now, since P_i and Q_j are approximated as in (2.C.1), their quantile functions can be computed directly from their atoms and weights by sorting the atoms in increasing order and accumulating their weights accordingly.

Hence, if we consider the matrices $\mathbf{S}_{P,\mathbf{t}}$ and $\mathbf{S}_{Q,\mathbf{t}}$ defined as $(\mathbf{S}_{P,\mathbf{t}})_{i,s} = F_{P_i}^{-1}(t_s)$ and $(\mathbf{S}_{Q,\mathbf{t}})_{i,s} = F_{Q_i}^{-1}(t_s)$ for $i = 1, \dots, n$ and $s = 1, \dots, M$, we can approximate $\hat{\Sigma}_{\tilde{P},\tilde{Q}}^{\dagger}$ as $\mathbf{S}_{P,\mathbf{t}}^t \mathbf{S}_{Q,\mathbf{t}} / M$. Similarly, we can approximate $\hat{\Sigma}_{\tilde{P},\tilde{P}}^{\dagger}$ and $\hat{\Sigma}_{\tilde{Q},\tilde{Q}}^{\dagger}$.

Kernel mean embedding For the kernel mean indices, we take $\mathcal{S} = \mathcal{P}(\mathbb{X})$, $\mathbb{H} = \mathbb{H}_k$ the RKHS generated by a kernel k , and ι as the kernel mean embedding $\mu_k : P \rightarrow \int_{\mathbb{X}} k(x, \cdot) dP(x)$. Consequently,

$$\langle \mu_k(P_i), \mu_k(Q_j) \rangle_{\mathbb{H}} = \iint_{\mathbb{X} \times \mathbb{X}} k(x, y) dP_i(x) dQ_j(y) \approx \sum_{r=1}^R \sum_{s=1}^R \pi_{i,r} \omega_{j,s} k(\theta_{i,r}, \eta_{j,s}). \quad (2.C.2)$$

If there exist sequences $(\theta_1^*, \dots, \theta_R^*)$ and $(\eta_1^*, \dots, \eta_R^*)$ such that $\theta_{i,r} = \theta_r^*$ and $\eta_{i,r} = \eta_r^*$ for $r = 1, \dots, R$ for any $i = 1, \dots, n$, this method can be easily vectorized. Indeed, if we consider the matrices $\mathbf{K}$, $\mathbf{\Pi}$ and $\mathbf{\Omega}$ defined as $(\mathbf{K})_{r,s} = k(\theta_r^*, \eta_s^*)$, $(\mathbf{\Pi})_{i,r} = \pi_{i,r}$, and $(\mathbf{\Omega})_{i,r} = \omega_{i,r}$ for $i = 1, \dots, n$ and $r, s = 1, \dots, R$, we can approximate $\hat{\Sigma}_{\tilde{P},\tilde{Q}}^{\dagger}$ as $\mathbf{\Pi K \Omega}^t$. Similarly, we can approximate $\hat{\Sigma}_{\tilde{P},\tilde{P}}^{\dagger}$ and $\hat{\Sigma}_{\tilde{Q},\tilde{Q}}^{\dagger}$.

Alternatively, we project all P_i 's onto a common support by building a histogram with

a bin width of $h > 0$. To do so, we first consider the average measure across $P_1, \dots, P_n$,

$$\bar{P} = \frac{1}{n} \sum_{i=1}^n P_i = \frac{1}{n} \sum_{i=1}^n \sum_{r=1}^R \pi_{i,r} \delta_{\theta_{i,r}}$$

and take m to be the median of $\bar{P}$. We choose L to minimize the quantity

$$\#\{\theta_{i,r} \text{ s.t. } |\theta_{i,r} - m| > L\} + \frac{2L}{h},$$

balancing the count of atoms outside $[m - L, m + L]$ against the number of bins of width h inside the interval. We then build the histogram of $\bar{P}$ over $[m - L, m + L]$ and take the bin midpoints of the bins as new atoms, as well as every original atom $\theta_{i,r}$ lying outside $[m - L, m + L]$. Let us call the points in the new support as $\{\psi_1, \dots, \psi_N\}$. Finally, we approximate each P_i onto this shared support as

$$P_i \approx \sum_{s=1}^N \pi'_{i,s} \delta_{\psi_s},$$

where each $\pi'_{i,s}$ is given as the sum of all $\pi_{i,r}$ such that the closest point to $\theta_{i,r}$ is ψ_s .

This ensures all the P_i 's share the same set of atoms. A similar procedure can be followed to approximate all the Q_i 's with measures with equal discrete support. Then we can apply the same approximation of $\hat{\Sigma}_{\bar{P}, \bar{Q}}^\dagger$, $\hat{\Sigma}_{\bar{P}, \bar{P}}^\dagger$ and $\hat{\Sigma}_{\bar{Q}, \bar{Q}}^\dagger$, as above.

Feature map embedding For the kernel feature map indices, we take $\mathcal{S} = \mathcal{P}(\mathbb{X})$ for $\mathbb{X}$ compact, $\mathbb{H}_K$ the RKHS generated by the kernel K and ι as the feature map $\phi_K(P) = K(P, \cdot)$. Consequently,

$$\langle \phi_K(P_i), \phi_K(Q_j) \rangle_{\mathbb{H}} = K(P_i, Q_j) = \exp \left(-\frac{1}{2\gamma^2} \|\mu_k(P_i) - \mu_k(Q_j)\|_{\mathbb{H}_k}^2 \right).$$

Hence, recalling that

$$\|\mu_k(P_i) - \mu_k(Q_j)\|_{\mathbb{H}_k}^2 = \langle \mu_k(P_i), \mu_k(P_i) \rangle_{\mathbb{H}_k} - 2\langle \mu_k(P_i), \mu_k(Q_j) \rangle_{\mathbb{H}_k} + \langle \mu_k(Q_j), \mu_k(Q_j) \rangle_{\mathbb{H}_k},$$

we can apply again (2.C.2) and the methods outlined in the previous paragraph to each term to approximate this quantity.

Chapter 3

Measuring Exchangeability with Reproducing Kernel Hilbert Spaces¹

In this chapter, we apply the Trace Correlation with kernel mean embedding, presented in [Chapter 2](#) as an index of dependence between random probability measures. This index detects exchangeability for common models, is invariant under reparametrization, can be estimated through samples, and, crucially, is well-suited for posteriors. Moreover, its use leads to many convenient properties and tractable expressions, especially a priori and under mixing. We apply our general framework to

- (i) investigate the dependence a posteriori for the hierarchical Dirichlet process, retrieving a parametric convergence rate under very mild assumptions on the data;
- (ii) eliciting the dependence structure of a parametric model for a principled comparison with a nonparametric alternative.

In this chapter, we will refer to the Trace Correlation with kernel mean embedding as *Kernel Correlation*, and we will denote our random probability measures of interest as $\tilde{P}_1$ and $\tilde{P}_2$, rather than $\tilde{P}$ and $\tilde{Q}$ for ease of notation and tractability.

¹The content of this chapter can be found in [Catalano et al. \(2025\)](#).

In [Section 3.1](#), we study a hierarchical parametric model for partially exchangeable data and highlight the properties we desire from a measure of partial exchangeability in Bayesian multilevel models, discussing the limitations of the current proposals in the literature.

In [Section 3.2](#), we recollect the definition of the Kernel Correlation along with its main properties.

In [Section 3.3](#), we identify settings under which the kernel correlation detects exchangeability, i.e., $\text{Corr}_k(\tilde{P}_1, \tilde{P}_2) = 1$ if and only if $\tilde{P}_1 = \tilde{P}_2$ a.s. Particularly, the kernel must induce an injective kernel mean embedding on the set of bounded signed measures.

In [Section 3.4](#), we express the kernel correlation in terms of partially exchangeable observables. Remarkably, we show that it can be determined by two observables for each group. This inspires a natural asymptotically normal estimator for kernel correlation through independent copies.

In [Section 3.5](#), we identify a key structural assumption on the random measures for which the kernel correlation does not depend on the choice of kernel, and we underline that this assumption holds for multivariate species sampling models. Since the set-wise correlation is a specific choice of kernel, this implies

- (i) the known result that the set-wise correlation does not depend on the set A for multivariate species sampling models ([Franzolini et al., 2025](#));
- (ii) that our kernel correlation coincides with the set-wise correlation for this class of models, thus recovering the interpretable and tractable expressions known in the literature.

In [Section 3.6](#), we apply Corr_k to the study of mixture models ([Ferguson, 1983](#); [Lo, 1984](#); [Escobar and West, 1995](#)), which are widely used in Bayesian inference for density estimation and clustering. We give conditions to express the kernel correlation of the mixture in terms of the mixing measures with respect to an updated version of the

kernel. Our results imply that the kernel correlation remains invariant under mixing for multivariate species sampling models and detects exchangeability for mixture models. We extend this property to parametric models by reinterpreting them as a special case of mixture models.

In [Section 3.7](#), we investigate the posterior behavior of Corr_k for the archetype of Bayesian nonparametric hierarchical models, the Hierarchical Dirichlet Process ([Teh et al., 2006](#)). Under a *non-degeneracy* assumption on the data, which depends on the kernel and is mild only for injective kernels, we prove that our kernel correlation goes to 0 with a parametric rate. Our theoretical results are also confirmed by numerical simulations in [Section 3.8](#). We apply a Gibbs sampling scheme to compute the kernel correlation for different choices of the kernel, and we empirically demonstrate how the Gaussian and Laplace kernels are robust to the choice of hyperparameters, whereas the set A has a significant impact on the value of the set-wise correlation.

Finally, in [Section 3.9](#), we apply the kernel correlation to perform a model comparison between a Gaussian model and the hierarchical Dirichlet Process. Specifically, we show how the conclusions of the models depend greatly on the value of the kernel correlation, and that one should set its value to be the same in both models for a fair model comparison.

The code to reproduce all the numerical experiments in this chapter is written in **R** and is available at <https://github.com/francescomascari/kernel-correlation>.

3.1 Introduction

Let us consider the following simple parametric model for two groups of partially exchangeable observations. Details are provided in [Appendix 3.A](#).

Example 3.1.1. Consider $X_{i,j}|\theta_1, \theta_2 \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_i, s^2)$ for $s > 0$, $j \in \mathbb{N}$, and $i = 1, 2$, with $(\theta_1, \theta_2) \sim \mathcal{N}(\mathbf{0}, \tau^2 \Sigma)$, where $\tau > 0$, $\Sigma_{11} = \Sigma_{22} = 1$ and $\Sigma_{12} = \Sigma_{21} = \rho \in [-1, 1]$. Then, the prior correlation between θ_1 and θ_2 is $\text{Corr}(\theta_1, \theta_2) = \rho$, while a version of the posterior

correlation after n_i observations in group i for $i = 1, 2$ is

$$\text{Corr}(\theta_1, \theta_2 | \mathbf{X}^{(n_1, n_2)} = \mathbf{x}^{(n_1, n_2)}) = \frac{\rho}{\sqrt{1 + n_1 \frac{\tau^2}{s^2} (1 - \rho^2)} \sqrt{1 + n_2 \frac{\tau^2}{s^2} (1 - \rho^2)}},$$

where, here and in the rest of the chapter, we make use of the compact and convenient notations $\mathbf{X}^{(n_1, n_2)} := ((X_{1,j})_{j=1}^{n_1}, (X_{2,j})_{j=1}^{n_2})$ and $\mathbf{x}^{(n_1, n_2)} := ((x_{1,j})_{j=1}^{n_1}, (x_{2,j})_{j=1}^{n_2})$.

Since the parameters θ_i are real-valued, an intuitive measure of partial exchangeability is Pearson's linear correlation between the parameters. When the parameters have the same first and second moment, this measure detects exchangeability, in the sense that $\text{Corr}(\theta_1, \theta_2) = 1$ if and only if $(X_{1,j})_{j \in \mathbb{N}}$ and $(X_{2,j})_{j \in \mathbb{N}}$ are fully exchangeable.

If we move to a more general Bayesian model for partially exchangeable data ([de Finetti, 1938](#)) of the form

$$X_{i,j} | (\theta_1, \theta_2) \stackrel{\text{i.i.d.}}{\sim} P_{\theta_i}, \quad (\theta_1, \theta_2) \sim Q, \quad (3.1)$$

where $X_{i,j}$ is the j -th observation in group i , Q is the prior distribution for the parameter vector, and P_θ is the data distribution parametrized by θ , the Pearson correlation coefficient is not satisfactory as a measure of partial exchangeability. This is due to the fact that this index is based on the too-strong assumption of having the same first two moments, is not invariant under reparametrization, and cannot be applied when θ is higher- or infinite-dimensional.

Nonetheless, [Example 3.1.1](#) shows that the posterior correlation converges to 0 as at least one of the sample sizes n_1, n_2 diverges, independently of the *true* data-generating process. Interestingly, it differs from the typical assumptions of asymptotic analyses, such as data being independent and identically distributed (i.i.d.) from an unknown distribution or generated by the model. The asymptotic behavior of the correlation aligns with our intuition of borrowing information a posteriori: as the amount of data from each group increases, we rely on more evidence and thus the borrowing becomes weaker. The convergence rate of the posterior correlation is $\mathcal{O}((n_1 n_2)^{-1/2})$ as $\max(n_1, n_2) \rightarrow +\infty$.

whenever the prior correlation ρ is different from ± 1 , hinting to a fast decay of the borrowing of information. One of the main objectives of this chapter is to reproduce this asymptotic investigation in a nonparametric setting as well.

The first crucial idea to avoid the lack of invariance under reparametrization and extend the measure to nonparametric partially exchangeable models is to use de Finetti's Theorem (de Finetti, 1938). This foundational theorem ensures that, for any partially exchangeable sequence, the law of the parameters becomes unique once it is embedded in the space of probabilities. More precisely, models in (3.1) are uniquely characterized by the joint law of $\tilde{P}_1$ and $\tilde{P}_2$. In other terms, the law of $(\tilde{P}_1, \tilde{P}_2)$ does not change under reparametrization of the model. Thus, in the sequel, we work with the model

$$X_{i,j} | \tilde{P}_1, \tilde{P}_2 \stackrel{\text{i.i.d.}}{\sim} \tilde{P}_i, \quad (\tilde{P}_1, \tilde{P}_2) \sim Q, \quad (3.2)$$

for $j \in \mathbb{N}$ and $i = 1, 2$, rather than with (3.1). The following conceptual step is to measure dependence at the level of the random probabilities $\tilde{P}_1$ and $\tilde{P}_2$. Since full exchangeability of the observations is recovered when $\tilde{P}_1 = \tilde{P}_2$ a.s., we need an index of dependence $I(Q)$ that detects almost sure equality, i.e., $I(Q) = 1$ if and only if $\tilde{P}_1 = \tilde{P}_2$ a.s..

In the Bayesian nonparametric literature, there are two primary methods for measuring dependence between random measures on general Polish spaces. The first is the set-wise correlation $\text{Corr}(\tilde{P}_1(A), \tilde{P}_2(A))$, for any measurable set A . Although it is based only on the first two moments of the random measures, it is the most used in practice because its expression a priori stands out for tractability and interpretability for the majority of Bayesian nonparametric models (see, e.g., Rodríguez et al. (2008); Leisen et al. (2013); Griffin and Leisen (2017); Camerlenghi et al. (2019); Beraha et al. (2021); Ascolani et al. (2023); Denti et al. (2023); Lijoi et al. (2023); Horiguchi et al. (2024); Colombi et al. (2025)). In most of these settings, the set-wise correlation does not depend on the set A . This has been recently shown to be a property of the general class of multivariate species

sampling models (Franzolini et al., 2025), which includes nearly all priors mentioned above. However, we show that tractability can fail for random probabilities that do not belong to this class, such as those that arise from parametric models or a posteriori. In such cases, not only does the set-wise correlation depend on the choice of the set A , but its value for different sets can also change dramatically from 1 to -1 .

A second method that has been recently proposed is to use the Wasserstein distance to measure the discrepancy between Q and the law in the same Fréchet class inducing full exchangeability (Catalano et al., 2021, 2024). This method considers the full distribution of the random measures, detecting both exchangeability and independence, and can be naturally extended to a multi-group scenario. However, its tractability is limited to completely random vectors (Catalano et al., 2021), which are the natural multivariate extension of completely random measures (Kingman, 1967). Though they are commonly used to build nonparametric priors (Lijoi and Prünster, 2010), parametric priors and posteriors rarely belong to this class.

Summing up, the current proposals in the literature are not satisfactory for measuring dependence between parametric priors and a posteriori. A major objective for this work is to define a new index of dependence that detects exchangeability for common models, is well-suited for parametric and posterior random probability measures, and maintains the tractability of $\text{Corr}(\tilde{P}_1(A), \tilde{P}_2(A))$ for Bayesian nonparametric priors. We achieve this by generalizing the set-wise correlation through Kernel Correlation (Corr_k).

The theory of Reproducing Kernel Hilbert Spaces (RKHS) plays an important role in Bayesian modelling as it allows the specification of smooth priors over function spaces, e.g., through Gaussian processes (Rasmussen and Williams, 2006) and other Bayesian kernel-based models (see, e.g., Tipping (2001); Sollich (2002); Pillai et al. (2007); MacLehose and Dunson (2009); Chakraborty et al. (2012)), with important applications to nonparametric Bayesian modelling and functional data analysis. An interesting research direction uses RKHS embeddings to facilitate Bayesian computation, e.g., through Stein and maximum

mean discrepancy (see, e.g., [Fukumizu et al. \(2013\)](#); [Park et al. \(2016\)](#); [Liu and Wang \(2016\)](#); [Chen et al. \(2019\)](#); [Legramanti et al. \(2025\)](#)).

It is important to notice the need to build a measure of dependence directly on the random probability measures and not on the conditionally independent observables. For instance, if we consider the Kernelized Canonical Correlation ([Bach and Jordan, 2002](#)), the Constrained Covariance ([Gretton et al., 2005b](#)), and the Centered Kernel Alignment ([Kornblith et al., 2019](#)) at the level of the observables, we cannot detect almost sure equality of the two random probability measures, i.e., exchangeability of the observations. For instance, the observations in [Example 3.1.1](#) are exchangeable for $\rho = 1$. However, none of the indices above evaluated between observations X_{1,j_1} and X_{2,j_2} from distinct groups are equal to 1, as shown in [Table 3.1](#). More generally, we cannot expect to detect exchangeability by using standard measures of dependence between observations.

3.2 Kernel Correlation

In this section, we recollect some concepts from the theory of RKHS, as in [Section 2.3.2](#), as well as the definition of Kernel Correlation, which corresponds to the Trace Correlation with kernel mean embedding presented in [Chapter 2](#). To this end, we first recall some preliminaries on kernel mean embedding, and then we define the kernel covariance and kernel correlation between random probability measures.

Let $\mathbb{X}$ be a Polish Space and $\mathcal{X}$ its Borel σ -algebra. Once again, we denote by $\mathcal{P}(\mathbb{X})$ the space of probability measures on $(\mathbb{X}, \mathcal{X})$ and by $\mathcal{M}_b(\mathbb{X})$ the space of bounded signed

Table 3.1: Estimated values of Pearson correlation, Kernelized Canonical Correlation (KCC) with regularization $\varepsilon = 10^{-2}$, Constrained Covariance (COCO), and Centered Kernel Alignment (CKA) between X_{1,j_1} and X_{2,j_2} with samples of size 1000 from the model in [Example 3.1.1](#) for $s^2 = \tau^2 = 1$ and $\rho = 1$.

Pearson correlation	KCC for $\varepsilon = 10^{-2}$	COCO	CKA
0.4907	0.5024	0.1504	0.1010

measures on $(\mathbb{X}, \mathcal{X})$. A random probability measure $\tilde{P}$ on $\mathbb{X}$ is a random element on $\mathcal{P}(\mathbb{X})$, while its mean measure $\mathbb{E}[\tilde{P}]$ is the element in $\mathcal{P}(\mathbb{X})$ satisfying $\int f(x) d\mathbb{E}[\tilde{P}](x) = \mathbb{E}[\int f(x) d\tilde{P}(x)]$ for any bounded and measurable function $f : \mathbb{X} \rightarrow \mathbb{R}$. Most random probabilities that we mention in this work are defined on $\mathbb{X}$. We will not repeat this assumption unless needed.

A kernel $k : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$ is a measurable, symmetric, positive-definite function, i.e., $k(x, y) = k(y, x)$ for any $x, y \in \mathbb{X}$ and

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

for any $m \in \mathbb{N}$, any $x_1, \dots, x_m \in \mathbb{X}$ and $a_1, \dots, a_m \in \mathbb{R}$. Moreover, we will always assume k to be bounded, i.e., there exists $B > 0$ such that $|k(x, y)| \leq B$ for any $x, y \in \mathbb{X}$. We denote as $\mathbb{H}_k$ the RKHS associated to a kernel k , with inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}_k}$ and norm $\|\cdot\|_{\mathbb{H}_k}$. The reproducing property of RKHSs guarantees that for any $h \in \mathbb{H}_k$, $h(x) = \langle h(\cdot), k(x, \cdot) \rangle_{\mathbb{H}_k}$.

Firstly, we recall the definition of kernel mean embedding, already presented in (2.2) and its main properties. Every bounded signed measure $\xi \in \mathcal{M}_b(\mathbb{X})$ can be associated with a unique element $\mu_k(\xi) \in \mathbb{H}_k$ through its kernel mean embedding

$$\mu_k(\xi)(y) := \int_{\mathbb{X}} k(x, y) d\xi(x).$$

The reproducing property ensures the following identities for $\xi, \xi_1, \xi_2 \in \mathcal{M}_b(\mathbb{X})$ and $f \in \mathbb{H}_k$,

$$\langle f, \mu_k(\xi) \rangle_{\mathbb{H}_k} = \int_{\mathbb{X}} f(x) d\xi(x), \quad \langle \mu_k(\xi_1), \mu_k(\xi_2) \rangle_{\mathbb{H}_k} = \iint_{\mathbb{X} \times \mathbb{X}} k(x, y) d\xi_1(x) d\xi_2(y).$$

A kernel k is c_0 if it vanishes at infinity, that is, if the set $\{x : |k(x, y)| \geq \varepsilon\}$ is compact for every $\varepsilon > 0, y \in \mathbb{X}$; it is injective if the feature map $x \mapsto k(x, \cdot)$ is injective; it is characteristic if the kernel mean embedding is injective on $\mathcal{P}(\mathbb{X})$. Since $k(x, \cdot) = \mu_k(\delta_x)$, a characteristic kernel is always injective. For a c_0 -kernel, the kernel mean embedding is

injective on $\mathcal{M}_b(\mathbb{X})$ if and only if $\mathbb{H}_k$ is dense in $C_0(\mathbb{X})$, the space of continuous functions over $\mathbb{X}$ vanishing at infinity endowed with the uniform norm (Sriperumbudur et al., 2011). In this case, we call k c_0 -universal, which implies that k is characteristic since $\mathcal{P}(\mathbb{X}) \subset \mathcal{M}_b(\mathbb{X})$.

Any kernel k induces a squared pseudo-metric

$$d_k^2(x, y) := k(x, x) - 2k(x, y) + k(y, y),$$

which is a squared distance whenever the feature map $x \mapsto k(x, \cdot)$ is injective, i.e., when k is injective. In the rest of the chapter, we will often rewrite the following integral expression in terms of the pseudo-metric d_k as

$$\int k(x, x) dP_0(x) - \iint k(x, y) dP_0(x) dP_0(y) = \frac{1}{2} \iint d_k^2(x, y) dP_0(x) dP_0(y), \quad (3.3)$$

for $P_0 \in \mathcal{P}(\mathbb{X})$. We refer to Muandet et al. (2017) for a complete reference on standard kernels on Euclidean spaces and their properties.

In this work, we focus on some of the most prominent examples of kernels in the literature: the linear kernel $k(x, y) = \langle x, y \rangle_{\mathbb{H}}$, defined on a bounded subset $\mathbb{X}$ of a Hilbert space $\mathbb{H}$, the Gaussian kernel $k(x, y) = \exp(-\|x - y\|_{\mathbb{X}}^2 / (2\sigma^2))$ for some $\sigma > 0$, defined on any Hilbert space $\mathbb{X}$, and the Laplace kernel $k(x, y) = \exp(-\|x - y\|_1 / \beta)$ for some $\beta > 0$,

Table 3.2: Useful properties of kernels: injectivity (INJ), continuity (CONT), characteristic (CHAR), and c_0 -universality (c_0 -UNIV). The set-wise kernel is introduced in Section 3.5.

Type of kernel			Property of kernel			
Name	Expression for $k(x, y)$	Space $\mathbb{X}$	CONT	INJ	CHAR	c_0 -UNIV
Linear	$\langle x, y \rangle_{\mathbb{H}}$	Bounded subset of $\mathbb{H}$ Hilbert	✓	✓	✗	✗
Gaussian	$\exp(-\ x - y\ _{\mathbb{X}}^2 / (2\sigma^2)), \sigma > 0$	Hilbert	✓	✓	✓	✓
Laplace	$\exp(-\ x - y\ _1 / \beta), \beta > 0$	$\mathbb{R}^m$	✓	✓	✓	✓
Set-wise	$\mathbb{1}_A(x) \mathbb{1}_A(y), A \in \mathcal{X}$	Measure Space	✗	✗	✗	✗

defined on $\mathbb{X} = \mathbb{R}^m$, where $\|x - y\|_1 = |x_1 - y_1| + \cdots + |x_m - y_m|$. We refer to [Muandet et al. \(2017\)](#) for a complete reference on standard kernels on Euclidean spaces and [Guella \(2022\)](#) for an analysis of Gaussian kernels on Hilbert spaces. Both Gaussian and Laplace kernels are injective and c_0 -universal, while the linear kernel is injective and continuous, but not c_0 -universal. [Table 3.2](#) provides a summary of the main properties of the kernels presented above and the set-wise kernel, which is introduced later in [Section 3.5](#).

Given two random probabilities $\tilde{P}_1$ and $\tilde{P}_2$, their kernel mean embeddings $\mu_k(\tilde{P}_1)$ and $\mu_k(\tilde{P}_2)$ are random elements on $\mathbb{H}_k$. We define the kernel covariance between $\tilde{P}_1$ and $\tilde{P}_2$ as

$$\mathbb{Cov}_k(\tilde{P}_1, \tilde{P}_2) := \mathbb{E} \left[\left\langle \mu_k(\tilde{P}_1) - \mathbb{E}[\mu_k(\tilde{P}_1)], \mu_k(\tilde{P}_2) - \mathbb{E}[\mu_k(\tilde{P}_2)] \right\rangle_{\mathbb{H}_k} \right],$$

which is well-defined since k is bounded and thus $\mathbb{E}[\|\mu_k(\tilde{P}_i)\|_{\mathbb{H}_k}^2] < +\infty$ for $i = 1, 2$.

Explicit calculations of the kernel covariance are often possible thanks to the following integral representation.

Proposition 3.2.1. *Let $\tilde{P}_1, \tilde{P}_2$ be random probabilities with $P_{0,i} = \mathbb{E}[\tilde{P}_i]$ for $i = 1, 2$. Then,*

$$\mathbb{Cov}_k(\tilde{P}_1, \tilde{P}_2) = \mathbb{E} \left[\iint k(x, y) d\tilde{P}_1(x) d\tilde{P}_2(y) \right] - \iint k(x, y) dP_{0,1}(x) dP_{0,2}(y).$$

The definition of $\mathbb{Cov}_k$ implies properties similar to those of the standard covariance between real-valued random variables. In particular, the kernel covariance is a symmetric bilinear form, and it inherits the law of total covariance of Hilbert spaces: for a random variable Z defined on the same probability space of $\tilde{P}_1$ and $\tilde{P}_2$,

$$\mathbb{Cov}_k(\tilde{P}_1, \tilde{P}_2) = \mathbb{E}_Z \left[\mathbb{Cov}_k(\tilde{P}_1, \tilde{P}_2 | Z) \right] + \mathbb{Cov}_k \left(\mathbb{E}[\tilde{P}_1 | Z], \mathbb{E}[\tilde{P}_2 | Z] \right).$$

From the definition of kernel covariance, we derive the notion of kernel variance of a random probability measure $\tilde{P}$ as $\mathbb{Var}_k(\tilde{P}) := \mathbb{Cov}_k(\tilde{P}, \tilde{P})$. As with usual random variables,

a zero variance characterizes deterministic objects. We recall that a random probability $\tilde{P}$ is deterministic if $\tilde{P} = P$ a.s. for some $P \in \mathcal{P}(\mathbb{X})$.

Lemma 3.2.2. *For any random probability $\tilde{P}$, $\text{Var}_k(\tilde{P}) \geq 0$. If the kernel k is characteristic, then $\text{Var}_k(\tilde{P}) = 0$ if and only if $\tilde{P}$ is deterministic.*

We now have all the main ingredients to recall the definition of kernel correlation.

Definition 3.2.1. Let $\tilde{P}_1, \tilde{P}_2$ be random probabilities such that $\text{Var}_k(\tilde{P}_i) > 0$ for $i = 1, 2$. The *kernel correlation* between $\tilde{P}_1$ and $\tilde{P}_2$ induced by the kernel k is defined as

$$\text{Corr}_k(\tilde{P}_1, \tilde{P}_2) := \frac{\text{Cov}_k(\tilde{P}_1, \tilde{P}_2)}{\sqrt{\text{Var}_k(\tilde{P}_1)}\sqrt{\text{Var}_k(\tilde{P}_2)}}.$$

As already noted in [Chapter 2](#), the kernel correlation takes values in $[-1, 1]$ and is null for independent random probabilities.

Proposition 3.2.3. *The kernel correlation $\text{Corr}_k(\tilde{P}_1, \tilde{P}_2)$ takes values in $[-1, 1]$. Moreover, if $\tilde{P}_1$ and $\tilde{P}_2$ are independent, $\text{Corr}_k(\tilde{P}_1, \tilde{P}_2) = 0$.*

We observe that the kernel correlation can be considered as a generalization of Pearson's correlation between parameters, as considered in [Example 3.1.1](#). Indeed, by identifying any random real parameter θ with the random probability δ_θ , then $\text{Corr}(\theta_1, \theta_2) = \text{Corr}_k(\delta_{\theta_1}, \delta_{\theta_2})$ for $k(x, y) = xy$ the linear kernel on $\mathbb{X} = \Theta \subset \mathbb{R}$ bounded. However, as shown in the next section, the linear kernel is unable to detect the full exchangeability of the observations, which corresponds to $\tilde{P}_1 = \tilde{P}_2$ a.s.. We devote the next section to showing that other kernels do not suffer from this limitation.

3.3 Detecting Full Exchangeability

We now investigate under which assumptions our kernel correlation can detect the almost sure equality of two random probability measures. We illustrate two different conditions,

one that holds for many Bayesian nonparametric priors and the other that typically holds for the corresponding posteriors.

Lemma 3.3.1. *Let k be a c_0 -universal kernel. Then $\text{Corr}_k(\tilde{P}_1, \tilde{P}_2) = \pm 1$ if and only if $\tilde{P}_1 - \mathbb{E}[\tilde{P}_1] = \alpha(\tilde{P}_2 - \mathbb{E}[\tilde{P}_2])$ a.s. for some $\alpha \in \mathbb{R} \setminus \{0\}$. The sign of α and the one of $\text{Corr}_k(\tilde{P}_1, \tilde{P}_2)$ coincide.*

This result, which is a corollary to [Proposition 1.2.2](#) in this setting, is pivotal for the kernel correlation to identify full exchangeability both a priori and a posteriori for a.s. discrete random probability measures. This result is pivotal in the proof of [Corollary 2.3.2](#), which is restated here as [Theorem 3.3.2](#).

Theorem 3.3.2. *Let $\tilde{P}_i$ be an a.s. discrete random probability such that $\mathbb{E}[\tilde{P}_i]$ is atomless, for $i = 1, 2$. If k is c_0 -universal, then $\text{Corr}_k(\tilde{P}_1, \tilde{P}_2) = 1$ if and only if $\tilde{P}_1 = \tilde{P}_2$ a.s..*

[Theorem 3.3.2](#) does not put assumptions on the dependence between the random probabilities, but only on their marginal distribution. This will be crucial in [Section 3.6](#), where we will prove that the kernel correlation detects exchangeability also for mixtures and parametric models.

The assumptions of [Theorem 3.3.2](#) hold for several classes of nonparametric priors, including the most common specifications of normalized random measures with independent increments ([Regazzini et al., 2003](#)) or species sampling processes ([Pitman, 1996](#)). [Example 3.A.1](#) shows that the assumption of c_0 -universality for k cannot be removed by considering the linear kernel. On the other hand, the assumption of atomless mean measure can be relaxed, which is extremely useful when dealing with posterior distributions. Indeed, for many Bayesian nonparametric models, the posterior mean measure is a convex combination of an atomless measure and a discrete measure supported on the observed values. Recall that $z \in \mathbb{X}$ is a fixed atom for the random measure $\tilde{P}$ if $\mathbb{P}(\tilde{P}(\{z\}) > 0) > 0$.

Theorem 3.3.3. *Let $\tilde{P}_i$ be an a.s. discrete random probability such that for any fixed atom z_i and for any $\varepsilon > 0$, $\mathbb{P}(\tilde{P}_i(\{z_i\}) < \varepsilon) > 0$, for $i = 1, 2$. If k is c_0 -universal, then $\text{Corr}_k(\tilde{P}_1, \tilde{P}_2) = 1$ if and only if $\tilde{P}_1 = \tilde{P}_2$ a.s..*

The assumption on the fixed atoms guarantees that the realizations of the corresponding jumps can be arbitrarily small. It holds, for example, for the posterior of normalized random measures with independent increments (James et al., 2009). Example 3.A.2 shows that the hypothesis cannot be removed.

3.4 Estimation from Samples

In this section, we express the kernel correlation in terms of the partially exchangeable observables in (3.2) with an unforeseen take-home message. Whereas recovering the marginal distribution of the random probabilities requires an infinite sample from a partially exchangeable sequence, computing their kernel correlation needs only four observations, two for each group. This characterization also allows us to estimate the kernel correlation from samples generated by the model using a convenient asymptotically normal estimator.

Theorem 3.4.1. *Let $(X_{1,1}, X_{2,1}, X_{1,2}, X_{2,2})$ be partially exchangeable observations from (3.2). Then, we can write $\text{Cov}_k(\tilde{P}_1, \tilde{P}_2) = \text{Cov}_{\mathbb{H}_k}(k(X_{1,1}, \cdot), k(X_{2,1}, \cdot))$ and $\text{Var}_k(\tilde{P}_i) = \text{Cov}_{\mathbb{H}_k}(k(X_{i,1}, \cdot), k(X_{i,2}, \cdot))$ for $i = 1, 2$. In particular,*

$$\text{Corr}_k(\tilde{P}_1, \tilde{P}_2) = \frac{\text{Cov}_{\mathbb{H}_k}(k(X_{1,1}, \cdot), k(X_{2,1}, \cdot))}{\sqrt{\text{Cov}_{\mathbb{H}_k}(k(X_{1,1}, \cdot), k(X_{1,2}, \cdot))} \sqrt{\text{Cov}_{\mathbb{H}_k}(k(X_{2,1}, \cdot), k(X_{2,2}, \cdot))}}.$$

Remark 3. If $\mathbb{X} \subset \mathbb{R}$ bounded and $k(x, y) = xy$ is the linear kernel, then $\text{Corr}_k(\tilde{P}_1, \tilde{P}_2) = \text{Cov}(X_{1,1}, X_{2,1}) / \sqrt{\text{Cov}(X_{1,1}, X_{1,2}) \text{Cov}(X_{2,1}, X_{2,2})}$, which differs from $\text{Corr}(X_{1,1}, X_{2,1})$, Pearson's linear correlation between $X_{1,1}$ and $X_{2,1}$. As an example, if $\tilde{P}_1 = \tilde{P}_2 = \tilde{P}$ a.s. then $\text{Corr}(X_{1,1}, X_{2,1}) < 1$ if $\tilde{P} \neq \delta_X$ for some random variable X , as $(X_{1,2}, X_{2,1})$ are conditionally independent and not a.s. equal. In contrast, $\text{Corr}_k(\tilde{P}_1, \tilde{P}_2) = 1$ as we showed in

[Section 3.3](#). Hence, we need to go beyond $\mathbb{C}\text{orr}(X_{1,1}, X_{2,1})$ to detect exchangeability of the model, as underlined in [Section 3.1](#).

We note that $\mathbb{C}\text{ov}_k(\tilde{P}_1, \tilde{P}_2)$ only requires the joint law of one observation for each group, while $\mathbb{V}\text{ar}_k(\tilde{P}_i)$ requires the joint law of two observations in group i . Thus, [Theorem 3.4.1](#) suggests an estimator of the kernel correlation through independent 2×2 samples of a partially exchangeable sequence. In practice, these samples can be easily obtained whenever one can sample from the law of $(\tilde{P}_1, \tilde{P}_2)$ directly, as with most parametric models. When $\tilde{P}_1$ and $\tilde{P}_2$ are infinite-dimensional, one can either find a finite-dimensional approximation of the law of $(\tilde{P}_1, \tilde{P}_2)$ or find the predictive distribution by integrating out the random probabilities. This can be seen as the partially exchangeable generalization of the Blackwell-MacQueen urn scheme ([Blackwell and MacQueen, 1973](#)). Most partially exchangeable models in the literature have explicit expressions for the predictive distribution, both a priori and a posteriori.

Proposition 3.4.2. *Let $(X_{1,1}^{(t)}, X_{2,1}^{(t)}, X_{1,2}^{(t)}, X_{2,2}^{(t)})$ be independent partially exchangeable observations from [\(3.2\)](#), for $t = 1, \dots, M$. Then,*

$$\begin{aligned}\widehat{\mathbb{C}\text{ov}}_{k,M}(\tilde{P}_1, \tilde{P}_2) &:= \frac{1}{M-1} \sum_{t=1}^M k(X_{1,1}^{(t)}, X_{2,1}^{(t)}) - \frac{1}{(M-1)M} \sum_{t=1}^M \sum_{s=1}^M k(X_{1,1}^{(t)}, X_{2,1}^{(s)}), \\ \widehat{\mathbb{V}\text{ar}}_{k,M}(\tilde{P}_i) &:= \frac{1}{M-1} \sum_{t=1}^M k(X_{i,1}^{(t)}, X_{i,2}^{(t)}) - \frac{1}{(M-1)M} \sum_{t=1}^M \sum_{s=1}^M k(X_{i,1}^{(t)}, X_{i,2}^{(s)})\end{aligned}$$

are unbiased estimators of $\mathbb{C}\text{ov}_k(\tilde{P}_1, \tilde{P}_2)$ and $\mathbb{V}\text{ar}_k(\tilde{P}_i)$ respectively, for $i = 1, 2$.

Combining these two quantities, we obtain an estimator of the correlation, which notably preserves both the rate and the asymptotic normality of the parametric case.

Proposition 3.4.3. *With the notations of [Proposition 3.4.2](#), if $\mathbb{V}\text{ar}_k(\tilde{P}_i) > 0$ for $i = 1, 2$,*

$$\widehat{\mathbb{C}\text{orr}}_{k,M}(\tilde{P}_1, \tilde{P}_2) := \frac{\widehat{\mathbb{C}\text{ov}}_{k,M}(\tilde{P}_1, \tilde{P}_2)}{\sqrt{\widehat{\mathbb{V}\text{ar}}_{k,M}(\tilde{P}_1)} \sqrt{\widehat{\mathbb{V}\text{ar}}_{k,M}(\tilde{P}_2)}}$$

is an asymptotically normal estimator of the kernel correlation, i.e., $\sqrt{M}(\widehat{\text{Corr}}_{k,M}(\tilde{P}_1, \tilde{P}_2) - \text{Corr}_k(\tilde{P}_1, \tilde{P}_2))$ converges in distribution to a centered Gaussian distribution as $M \rightarrow +\infty$.

3.5 From Set-Wise to Kernel Correlation

In this section, we interpret the kernel correlation as a generalization of the widely used set-wise correlation $\text{Corr}(\tilde{P}_1(A), \tilde{P}_2(A))$, for a measurable set $A \in \mathcal{X}$. Despite having some undesirable behaviors due to the lack of continuity, we show that its expression is equal to the one for injective kernels for a large class of Bayesian nonparametric priors.

Proposition 3.5.1. *For $A \in \mathcal{X}$, the set-wise kernel $k_A(x, y) := \mathbb{1}_A(x)\mathbb{1}_A(y)$ defines a kernel such that $\text{Corr}_{k_A}(\tilde{P}_1, \tilde{P}_2) = \text{Corr}(\tilde{P}_1(A), \tilde{P}_2(A))$.*

As a direct consequence, note that $\text{Var}_{k_A}(\tilde{P}) = 0$ if and only if $\tilde{P}(A)$ is deterministic. In contrast to most of the standard kernels, k_A is not continuous, injective, or characteristic (cf. Table 3.2). In particular, it is not c_0 -universal, and thus it does not fulfil the conditions of Theorem 3.3.2 and Theorem 3.3.3. Unsurprisingly, we may find $\tilde{P}_1$ and $\tilde{P}_2$ that are not a.s. equal such that $\text{Corr}(\tilde{P}_1(A), \tilde{P}_2(A)) = 1$ for some A , as shown in Example 3.A.3. Even more strikingly, the correlation may change from -1 to $+1$ for the *same* pair of random measures simply by changing the set A .

Example 3.5.1. For $W \sim \text{Unif}_{[0,1]}$, $P \in \mathcal{P}(\mathbb{X})$ an atomless probability, and $x_1 \neq x_2 \in \mathbb{X}$, we define $\tilde{P}_i := W\delta_{x_i} + (1 - W)P$ for $i = 1, 2$. If we take $A \in \mathcal{X}$ such that $x_1, x_2 \notin A$ and $P(A) \neq 0$, then $\tilde{P}_1(A) = \tilde{P}_2(A) = (1 - W)P(A)$ a.s.. Thus $\text{Corr}(\tilde{P}_1(A), \tilde{P}_2(A)) = 1$. If we take $B \in \mathcal{X}$ such that $P(B) \notin \{0, 1\}$, $x_1 \in B$, $x_2 \notin B$, then $\tilde{P}_1(B) = P(B) + W(1 - P(B))$ while $\tilde{P}_2(B) = (1 - W)P(B)$. It follows that $\text{Corr}(\tilde{P}_1(B), \tilde{P}_2(B)) = -1$.

Moreover, the lack of continuity of the set-wise kernel leads to a lack of continuity of the kernel correlation, which, in turn, compromises the stability in the assessment of the measure of partial exchangeability, as shown by a slight modification of the above

example in [Appendix 3.A](#) ([Example 3.A.4](#)). These examples strongly advocate for the use of a continuous kernel, such as the Gaussian or the Laplace, which provides more stable measures, beyond the already proved property of detecting full exchangeability.

Nevertheless, there is a large class of Bayesian nonparametric priors where these advantages are not necessary. We now identify a structural assumption on the random probabilities, which holds for most commonly used priors in Bayesian Nonparametrics, where the kernel correlation does not depend on the choice of the kernel. Thus, for this class of random probabilities, the kernel correlation coincides with the set-wise correlation for any choice of an injective kernel.

Proposition 3.5.2. *Let $\tilde{P}_1, \tilde{P}_2$ be random probabilities with same mean measure $\mathbb{E}[\tilde{P}_1] = \mathbb{E}[\tilde{P}_2] = P_0$. Then, the following conditions are equivalent for any $\eta \in \mathbb{R}$ and imply that $\eta \in [-1, 1]$:*

- (i) $\text{Cov}(\tilde{P}_1(A), \tilde{P}_2(A)) = \eta P_0(A)(1 - P_0(A))$ for any measurable set $A \in \mathcal{X}$;
- (ii) $\text{Cov}_k(\tilde{P}_1, \tilde{P}_2) = \eta \left(\int k(x, x) dP_0(x) - \iint k(x, y) dP_0(x) dP_0(y) \right)$ for any kernel k .

Applying [Proposition 3.5.2](#) to $\tilde{P}_1 = \tilde{P}_2 = \tilde{P}$ a.s., we deduce the following corollary for the variance.

Corollary 3.5.3. *Let $\tilde{P}$ be a random probability with $P_0 = \mathbb{E}[\tilde{P}]$. Then, the following conditions are equivalent for any $\lambda \in \mathbb{R}$ and imply that $\lambda \in [0, 1]$:*

- (i) $\text{Var}(\tilde{P}(A)) = \lambda P_0(A)(1 - P_0(A))$ for any measurable set $A \in \mathcal{X}$;
- (ii) $\text{Var}_k(\tilde{P}) = \lambda \left(\int k(x, x) dP_0(x) - \iint k(x, y) dP_0(x) dP_0(y) \right)$ for any kernel k .

If the conditions in [Proposition 3.5.2](#) and [Corollary 3.5.3](#) are met, then the kernel correlation does not depend on the kernel as long as it is well defined, that is, if the kernel variances are strictly positive (cf. [Lemma 3.2.2](#)). As an easy corollary, we deduce the following fundamental theorem.

Theorem 3.5.4. *Let $\tilde{P}_1, \tilde{P}_2$ be non-deterministic random probabilities that satisfy the conditions in [Proposition 3.5.2](#) for some $\eta \in [-1, 1]$, and the ones in [Corollary 3.5.3](#) for some $\lambda_i \in (0, 1]$, for $i = 1, 2$. Then for any injective kernel k and any set $A \in \mathcal{X}$ such that $\tilde{P}_i(A)$ is not deterministic for $i = 1, 2$,*

$$\text{Corr}_k(\tilde{P}_1, \tilde{P}_2) = \text{Corr}(\tilde{P}_1(A), \tilde{P}_2(A)) = \frac{\eta}{\sqrt{\lambda_1 \lambda_2}}.$$

Remark 4. In [Theorem 3.5.4](#), we only need the kernel k to be injective so that $\text{Var}_k(\tilde{P}_i)$ is non-null for $i = 1, 2$, whereas [Lemma 3.2.2](#) requires the stronger assumption that k is characteristic. However, since $\text{Var}_k(\tilde{P}_i)$ can be written as in [Corollary 3.5.3](#), $\tilde{P}_i$ is deterministic if and only if there exists $z \in \mathbb{X}$ such that $\tilde{P}_i = \delta_z$ a.s.. The details are in the proof of [Theorem 3.5.4](#) in [Appendix 3.B](#).

Remarkably, the results in [Franzolini et al. \(2025\)](#) guarantee that [Theorem 3.5.4](#) holds for multivariate species sampling processes. This class of multivariate priors includes most of the partially exchangeable models studied in the Bayesian nonparametric literature. For example, in the context of hierarchical models, it includes the hierarchical Dirichlet process ([Teh et al., 2006](#)), hierarchical normalized completely random measures ([Camerlenghi et al., 2019](#)), the semi-hierarchical Dirichlet process ([Beraha et al., 2021](#)), and the hidden hierarchical Dirichlet process ([Lijoi et al., 2023](#)), to name a few. We refer to [Franzolini et al. \(2025\)](#) for the complete list.

However, the framework of [Theorem 3.5.4](#) is not enough to analyze parametric and posterior random measures of nonparametric models. Indeed, even if $\tilde{P}_1$ and $\tilde{P}_2$ are multivariate species sampling priors, the posteriors do not belong to this class.

3.6 Kernel Correlation for Mixture Models

Mixture models are widely used in Bayesian inference for density estimation and clustering. The use of a nonparametric a.s. discrete prior as a mixing distribution allows for a potentially infinite number of components, with an evident gain in flexibility. In this section, we show that we can express the kernel correlation between mixture models in terms of a kernel correlation between mixing distributions with an updated kernel. Moreover, we show that for multivariate species sampling models, the kernel correlation between mixture models coincides with the set-wise correlation between the mixing measures. This validates a standard procedure in applied analyses with mixture models, where the prior elicitation of the borrowing of information is performed through the dependence between the mixing measures. Remarkably, our analysis of mixture models can be used to treat the kernel correlation of parametric models, showing that it detects exchangeability also in these settings.

In this section, $(\mathbb{X}, \mathcal{X})$ is a latent space, and $(\mathbb{Y}, \mathcal{Y})$ is the space of observations endowed with a kernel k . For simplicity, we endow $(\mathbb{Y}, \mathcal{Y})$ with a reference σ -finite measure denoted by dy , and we will consider only measures having a density with respect to this reference measure. The reader can think of $\mathbb{Y}$ being $\mathbb{R}^d$, endowed with the Lebesgue measure. Let $f : \mathbb{Y} \times \mathbb{X} \rightarrow [0, 1]$ be a probability density kernel. We define the partially exchangeable mixture model with probability density kernel f and mixing distribution Q as

$$Y_{i,j} | X_{i,j} \sim f(\cdot; X_{i,j}), \quad X_{i,j} | \tilde{P}_1, \tilde{P}_2 \stackrel{\text{i.i.d.}}{\sim} \tilde{P}_i, \quad (\tilde{P}_1, \tilde{P}_2) \sim Q, \quad (3.4)$$

for $j \in \mathbb{N}$ and $i = 1, 2$. We observe that $\{Y_{i,j}\}$ are partially exchangeable, and the model can be rewritten as $Y_{i,j} | \tilde{P}_1, \tilde{P}_2 \sim f_{\tilde{P}_i}$, where, for any $P \in \mathcal{P}(\mathbb{X})$, the *mixture density* is defined as

$$f_P(\cdot) := \int_{\mathbb{X}} f(\cdot; x) dP(x).$$

Remark 5. Consider a density f proportional to a kernel k on $\mathbb{Y} = \mathbb{X}$. Then, f_P coincides with the kernel mean embedding up to a multiplicative constant. In particular, if f is the density of $\mathcal{N}(\cdot, \sigma^2)$, f_P is the kernel mean embedding of the Gaussian kernel $k(x, y) = \exp(-(x-y)^2/(2\sigma^2))$. However, there is not a one-to-one correspondence between mixture densities and kernel mean embedding since not all densities f are symmetric functions, and their parameter space does not always coincide with the observation space, i.e., $\mathbb{X} \neq \mathbb{Y}$. Vice versa, not all kernel functions can be rewritten as unnormalized densities, since positivity may fail, as is the case with the linear kernel.

We say that the parametric family $\{f(\cdot; x)\}_x$ is *identifiable* if $f_{P_1} = f_{P_2}$ a.e. implies $P_1 = P_2$ for any $P_1, P_2 \in \mathcal{P}(\mathbb{X})$. This definition has been used, e.g., in [Nguyen \(2013\)](#), and it is a slight modification of the original one in [Teicher \(1961\)](#).

Theorem 3.6.1. *Let f be a probability density kernel on $\mathbb{Y} \times \mathbb{X}$. Then, for any kernel k on $\mathbb{Y} \times \mathbb{Y}$,*

$$k_f(x_1, x_2) := \iint_{\mathbb{Y} \times \mathbb{Y}} k(y_1, y_2) f(y_1; x_1) f(y_2; x_2) dy_1 dy_2,$$

is a kernel on $\mathbb{X} \times \mathbb{X}$ such that $\mathbb{Cov}_k(f_{\tilde{P}_1}, f_{\tilde{P}_2}) = \mathbb{Cov}_{k_f}(\tilde{P}_1, \tilde{P}_2)$. Moreover, if $\{f(\cdot; x)\}_x$ is identifiable and k is characteristic, then k_f is characteristic.

Consequently, under the assumptions of [Theorem 3.6.1](#), we can extend [Theorem 3.5.4](#) to mixture models. Remarkably, these imply that for multivariate species sampling models, the kernel correlation between mixture models coincides with the set-wise correlation.

Corollary 3.6.2. *Let $\tilde{P}_1, \tilde{P}_2$ satisfy the assumptions of [Theorem 3.5.4](#) and let f be a probability density kernel on $\mathbb{Y} \times \mathbb{X}$ such that $\{f(\cdot; x)\}_x$ is identifiable. Then, for every characteristic kernel k on $\mathbb{Y} \times \mathbb{Y}$ and set A such that $\tilde{P}_1(A)$ and $\tilde{P}_2(A)$ are not deterministic,*

$$\mathbb{Corr}_k(f_{\tilde{P}_1}, f_{\tilde{P}_2}) = \mathbb{Corr}_{k_f}(\tilde{P}_1, \tilde{P}_2) = \mathbb{Corr}(\tilde{P}_1(A), \tilde{P}_2(A)).$$

In the general case, $\mathbb{Corr}_{k_f}(\tilde{P}_1, \tilde{P}_2)$ will depend on the choice of k and some kernels

can be more tractable than others. When $\mathbb{Y} = \mathbb{R}^m$, it is convenient to consider a translation invariant kernel $k(y_1, y_2) = \psi(y_1 - y_2)$ for some positive continuous function $\psi : \mathbb{R}^m \rightarrow (0, +\infty)$. Since the kernel correlation is invariant with respect to scalar multiplication of the kernel, without loss of generality, we can assume $\psi(0) = 1$. For a probability distribution $\nu \in \mathcal{P}(\mathbb{R}^m)$, we denote by $\hat{\nu}(x) := \int e^{-i\langle x, z \rangle} d\nu(z)$ its Fourier transform, for $x \in \mathbb{R}^m$. Bochner's Theorem (Bochner, 1959) guarantees $\psi = \hat{\nu}$ for some $\nu \in \mathcal{P}(\mathbb{R}^m)$.

Proposition 3.6.3. *Let $k(y_1, y_2) = \psi(y_1 - y_2)$ be a translation invariant kernel for a continuous ψ such that $\psi(0) = 1$, and let $\nu \in \mathcal{P}(\mathbb{R}^m)$ such that $\psi = \hat{\nu}$. Then, k_f in Theorem 3.6.1 is equal to*

$$k_f(x_1, x_2) = \int_{\mathbb{R}^m} \hat{f}(z; x_1) \overline{\hat{f}(z; x_2)} d\nu(z) = \left\langle \hat{f}(\cdot; x_1), \hat{f}(\cdot; x_2) \right\rangle_{L^2(\nu; \mathbb{C})}.$$

Example 3.6.1. The Gaussian kernel $k(y_1, y_2) = \exp\left(-(y_1 - y_2)^2/(2\sigma^2)\right)$ is a translation invariant kernel on $\mathbb{R}$ with $\psi(z) = \exp(-z^2/(2\sigma^2))$, for $\sigma > 0$. We notice that $\psi(z) = \hat{\nu}(z)$, where ν is a normal distribution with mean 0 and variance σ^{-2} . For a Gaussian mixture model with $f(\cdot; x)$ being the density of $\mathcal{N}(x, \sigma_0^2)$ for some $\sigma_0 > 0$. In such case, $\hat{f}(y; x) = \exp(ixy - \sigma_0^2 y^2/2)$. Thus $k_f(x_1, x_2)$ is equal to

$$\sqrt{\frac{\sigma^2}{2\pi}} \int_{\mathbb{R}} \exp\left(ix_1 z - ix_2 z - \frac{1}{2} (2\sigma_0^2 + \sigma^2) z^2\right) dz = \sqrt{\frac{\sigma^2}{2\sigma_0^2 + \sigma^2}} \exp\left(-\frac{1}{2} \frac{(x_1 - x_2)^2}{2\sigma_0^2 + \sigma^2}\right),$$

which is a Gaussian kernel with updated parameters.

The construction for mixture models presented in this section enables us to revisit the parametric case of Example 3.1.1 within this new framework.

Example 3.6.2 (Example 3.1.1 – Revisited). We can revisit Example 3.1.1 as a mixture model as in (3.4) with $f(\cdot; x)$ being the density of $\mathcal{N}(x, s^2)$, and $\tilde{P}_1 = \delta_{\theta_1}$, $\tilde{P}_2 = \delta_{\theta_2}$ having joint law determined by $(\theta_1, \theta_2) \sim \mathcal{N}(\mathbf{0}, \tau^2 \Sigma)$, where $s, \tau > 0$, $\Sigma_{11} = \Sigma_{22} = 1$ and

$\Sigma_{12} = \Sigma_{21} = \rho \in [-1, 1]$.

We take the Gaussian kernel $k(y_1, y_2) = \exp\left(-(y_1 - y_2)^2/(2\sigma^2)\right)$ on $\mathbb{Y} \times \mathbb{Y}$. By [Example 3.6.1](#) the kernel k_f over $\mathbb{X} \times \mathbb{X}$ is $k_f(x_1, x_2) = \sqrt{\sigma^2/(2s^2 + \sigma^2)} \exp(-(x_1 - x_2)^2/(2(2s^2 + \sigma^2)))$, which is proportional to a Gaussian kernel. The computation of $\text{Corr}_{k_f}(\tilde{P}_1, \tilde{P}_2)$ only involves Gaussian integrals and can be done with [\(3.E.1\)](#). We obtain, thanks to [Corollary 3.6.2](#),

$$\text{Corr}_k(f_{\tilde{P}_1}, f_{\tilde{P}_2}) = \text{Corr}_{k_f}(\tilde{P}_1, \tilde{P}_2) = \frac{\sqrt{\frac{\sigma^2}{2\tau^2(1-\rho)+2s^2+\sigma^2}} - \sqrt{\frac{\sigma^2}{2\tau^2+2s^2+\sigma^2}}}{\sqrt{\frac{\sigma^2}{2s^2+\sigma^2}} - \sqrt{\frac{\sigma^2}{2\tau^2+2s^2+\sigma^2}}}.$$

Since k_f is c_0 -universal, and since $\tilde{P}_i = \delta_{\theta_i}$ are discrete a.s. with atomless $\mathbb{E}[\tilde{P}_i] = \mathcal{N}(0, \tau^2)$, we can apply [Theorem 3.3.2](#). Hence, it holds that $\text{Corr}_k(f_{\tilde{P}_1}, f_{\tilde{P}_2}) = 1$ if and only if $\tilde{P}_1 = \tilde{P}_2$ a.s..

The strategy of taking $\tilde{P}_1 = \delta_{\theta_1}$ and $\tilde{P}_2 = \delta_{\theta_2}$ can be extended to other parametric models as well. Hence, the previous example shows that kernel correlation can detect exchangeability for parametric models by reinterpreting them as mixture models over a vector of a.s. discrete random probabilities that do not belong to the class of multivariate species sampling models. The same rewriting shows that it detects exchangeability for mixture models independently of their dependence structure, thus extending [Theorem 3.3.2](#) to this setting.

Corollary 3.6.4. *Let $\tilde{P}_i$ be an a.s. discrete random probability such that $\mathbb{E}[\tilde{P}_i]$ is atomless, for $i = 1, 2$, and let f be a probability density kernel on $\mathbb{Y} \times \mathbb{X}$. If k_f is c_0 -universal, then $\text{Corr}_k(f_{\tilde{P}_1}, f_{\tilde{P}_2}) = 1$ if and only if $f_{\tilde{P}_1} = f_{\tilde{P}_2}$ a.s..*

3.7 Hierarchical Dirichlet Process

In this section, we turn to one of the most popular models for partially exchangeable data, the hierarchical Dirichlet Process ([Teh et al., 2006](#)), and we study the behavior of

the kernel correlation both a priori and a posteriori. Remarkably, we are able to recover the same rate of convergence of the parametric model in [Example 3.1.1](#) at the cost of adding an extra assumption on the non-degeneracy of the data sequence with respect to the kernel. The set-wise correlation is not the best choice in this context: the verification of the assumption depends heavily on the choice of the set, and there is no sensible way to choose it before looking at the specific dataset.

To fix the notation, let us recall the definitions of both the Dirichlet Process ([Ferguson, 1973](#)) and the hierarchical Dirichlet Process ([Teh et al., 2006](#)), already introduced in [Section 2.4](#). A random probability $\tilde{P} \sim \text{DP}(c, P_0)$ follows a Dirichlet Process with concentration $c > 0$ and base measure $P_0 \in \mathcal{P}(\mathbb{X})$ if

$$(\tilde{P}(A_1), \tilde{P}(A_2), \dots, \tilde{P}(A_m)) \sim \text{Dir}(cP_0(A_1), cP_0(A_2), \dots, cP_0(A_m))$$

for any partition $\{A_1, A_2, \dots, A_m\}$ of $\mathbb{X}$, where Dir indicates the finite-dimensional Dirichlet distribution. Starting from the definition of a Dirichlet Process, the hierarchical Dirichlet Process (hDP) of [Teh et al. \(2006\)](#) provides a natural way of building a dependent nonparametric prior for a vector $(\tilde{P}_1, \tilde{P}_2)$, which can then be used to define partially exchangeable models as in [\(3.2\)](#). Specifically, $(\tilde{P}_1, \tilde{P}_2) \sim \text{hDP}(c, c_0, P_0)$ for $c, c_0 > 0$ and P_0 an atomless probability measure on $\mathbb{X}$, if

$$\tilde{P}_1, \tilde{P}_2 | \tilde{P}_0 \stackrel{\text{i.i.d.}}{\sim} \text{DP}(c, \tilde{P}_0), \quad \tilde{P}_0 \sim \text{DP}(c_0, P_0). \quad (3.5)$$

The calculations in [Example 3.A.5](#) show that the hDP a priori satisfies [Corollary 3.5.3](#) with $\lambda_1 = \lambda_2 = (1 + c + c_0)(1 + c)^{-1}(1 + c_0)^{-1}$ and [Proposition 3.5.2](#) with $\eta = (1 + c_0)^{-1}$. Thus by [Theorem 3.5.4](#),

$$\text{Corr}_k(\tilde{P}_1, \tilde{P}_2) = \text{Corr}(\tilde{P}_1(A), \tilde{P}_2(A)) = \frac{1 + c}{1 + c + c_0},$$

for any injective kernel k and any measurable set A such that $P_0(A) \notin \{0, 1\}$. The expression of $\text{Corr}(\tilde{P}_1(A), \tilde{P}_2(A))$ had already been derived in [Camerlenghi et al. \(2019\)](#): our contribution is to show that it coincides with the kernel correlation of *any* injective kernel.

We now study the rate of convergence of the kernel correlation of the hDP a posteriori to zero as the number of observations diverges. First, we define a notion of non-degeneracy for a sequence, which in our setting will be the sequence of observations within each group.

Definition 3.7.1. We say that a sequence $(x_j)_{j \in \mathbb{N}}$ is non-degenerate with respect to a kernel k if

$$\liminf_{n \rightarrow +\infty} \frac{1}{n^2} \sum_{j=1}^n \sum_{h=1}^n d_k^2(x_j, x_h) = \liminf_{n \rightarrow +\infty} \int_{\mathbb{X}} \int_{\mathbb{X}} d_k^2(x, y) d\hat{P}_n(x) d\hat{P}_n(y) > 0,$$

where $\hat{P}_n = n^{-1} \sum_{j=1}^n \delta_{x_j}$ and $d_k^2(x, y) = k(x, x) - 2k(x, y) + k(y, y)$.

We note that d_k is a pseudo-metric, and it is a distance if k is injective. Its role in the study of the kernel correlation can be understood thanks to [\(3.3\)](#).

The notion of non-degeneracy depends heavily on the choice of the kernel. We gain a better understanding by considering $(x_j)_{j \in \mathbb{N}}$ a realization of an infinitely exchangeable sequence $X_j | \tilde{P} \stackrel{\text{i.i.d.}}{\sim} \tilde{P}$, with $\tilde{P} \sim Q$ its de Finetti measure. By de Finetti's Representation Theorem ([de Finetti, 1937](#)), the empirical measure converges weakly to $\tilde{P}$ a.s. as the number of observations n diverges, thus

$$\liminf_{n \rightarrow +\infty} \iint d_k^2(x, y) d\hat{P}_n(x) d\hat{P}_n(y) = \iint d_k^2(x, y) d\tilde{P}(x) d\tilde{P}(y) \quad \text{a.s.} \quad (3.6)$$

If the kernel is injective, then d_k is a distance. Hence, the right-hand side is zero if and only if $\tilde{P} = \delta_Z$ a.s. for some random variable Z on $\mathbb{X}$. This means that the sequence $(X_j)_{j \in \mathbb{N}}$ is a.s. constant, which is arguably an intuitive notion of degeneracy. However,

when $k(x, y) = \mathbb{1}_A(x)\mathbb{1}_A(y)$ for some A , the right hand side in (3.6) is zero whenever $\tilde{P}(A) \in \{0, 1\}$ a.s.. Summarizing, the non-degeneracy assumption is not restrictive when the kernel k is injective, and in such cases, it does not depend on k . For instance, linear, Gaussian, and Laplace kernels induce the same notion of non-degeneracy. In contrast, for the set-wise kernel, the notion of non-degeneracy depends heavily on the choice of the set, which cannot be chosen before seeing the data.

The following theorem states that if we assume that the sequences of observables are non-degenerate for both groups, there is a regular version of the posterior, derived in Camerlenghi et al. (2019) and reported in (3.C.1), for which we can recover the parametric rate of convergence. For observed data $\mathbf{x}^{(n_1, n_2)} = ((x_{1,j})_{j=1}^{n_1}, (x_{2,j})_{j=1}^{n_2})$, we denote by $\mathcal{L}(\tilde{P}_1, \tilde{P}_2 | \mathbf{X}^{(n_1, n_2)} = \mathbf{x}^{(n_1, n_2)})$ the point-wise evaluation of the corresponding Markov kernel.

Theorem 3.7.1. *Consider a partially exchangeable model as in (3.2) for $(\tilde{P}_1, \tilde{P}_2) \sim \text{hDP}(c, c_0, P_0)$ for $c, c_0 > 0$ and P_0 an atomless probability measure on $\mathbb{X}$. If $(x_{i,j})_{j \in \mathbb{N}}$ is a non-degenerate sequence with respect to a kernel k for $i = 1, 2$, then as $\max(n_1, n_2) \rightarrow +\infty$,*

$$\text{Corr}_k(\tilde{P}_1, \tilde{P}_2 | \mathbf{X}^{(n_1, n_2)} = \mathbf{x}^{(n_1, n_2)}) = \mathcal{O}\left(\frac{1}{\sqrt{n_1 n_2}}\right).$$

Our results on the hierarchical Dirichlet process rely on a fine understanding of its posterior structure, which we revise in the Supplementary Material. These results are also helpful to design the simulations in Section 3.8.

3.8 Numerical Simulations

This section aims to numerically validate the convergence rate obtained in Theorem 3.7.1 and analyze the stability of the kernel correlation with respect to the choice of kernel. We devise two different strategies to approximate the kernel correlation a posteriori. The

sampling-based method utilizes the estimator presented in [Section 3.4](#) and can be applied whenever it is possible to generate 2×2 samples from the model a posteriori. The *analytics-based* method is built on the quasi-conjugacy property of the hDP ([Teh et al., 2006](#); [Camerlenghi et al., 2019](#)). We empirically show that the *ad hoc* construction of the analytics-based estimator has a lower variance, underlining the importance of analytic calculations when possible.

3.8.1 Sampling-based vs Analytics-based Method for the Hierarchical Dirichlet Process

Both the sampling-based method and the analytics-based method rely on the quasi-conjugacy property of the augmented hDP model we obtain after the introduction of a specific sequence of latent random variables $\mathbf{T}^{(n_1, n_2)}$, commonly referred to as *tables* in the restaurant franchise metaphor ([Teh et al., 2006](#); [Camerlenghi et al., 2019](#); [Catalano et al., 2023](#)). To generate a 2×2 sample from the posterior distribution, as in the sampling-based algorithm, we generate $\mathbf{T}^{(n_1, n_2)}$ conditionally on $\mathbf{X}^{(n_1, n_2)}$ and apply the predictive distribution of the quasi-conjugate scheme. Similarly, in the analytics-based algorithm, we find the exact expression of the kernel correlation a posteriori conditionally on $\mathbf{T}^{(n_1, n_2)}$, and then we generate R copies of $\mathbf{T}^{(n_1, n_2)} | \mathbf{X}^{(n_1, n_2)}$ to marginalize the tables out. We refer to [Appendix 3.C](#) for a detailed description of the two algorithms.

We test the two algorithms by computing the posterior kernel correlation for the Gaussian kernel with $\sigma = 1$ coming from an hDP prior [\(3.5\)](#) with $c_0 = c = 1$ and $P_0 \sim \text{Unif}_{[0,1]}$ when the data are $n_1 = n_2 = 10$ observations from the model. We run both algorithms on the same data 100 times. In the analytics-based algorithm, we run the Gibbs sampler $R = 10$ times to approximate the expectations in the law of total covariance. In contrast, the sampling-based algorithm uses the empirical covariance estimator, as described in [Proposition 3.4.2](#), with $M = 10,000$ independent samples. The box plots on

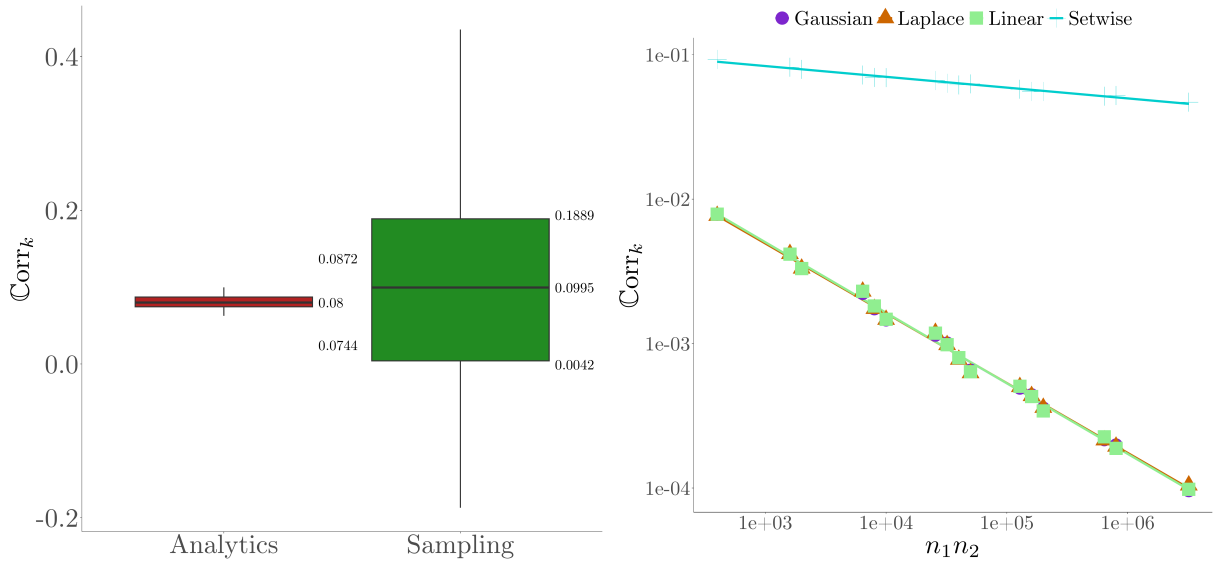


Figure 3.1: **Left:** Box plots for *Analytics*-based and *Sampling*-based algorithms for 100 realizations of the estimate of the posterior kernel correlation for $(\tilde{P}_1, \tilde{P}_2) \sim \text{hDP}(c = 1, c_0 = 1, P_0 = \text{Unif}_{[0,1]})$ when we condition on $n_1 = n_2 = 10$ data points sampled from the model and use a Gaussian kernel with $\sigma = 1$. **Right:** Log-log plot of the kernel correlation a posteriori as a function of $n_1 n_2$ with $n_1 = 4^i$ and $n_2 = 5^j$ for $i, j = 2, 3, 4, 5$ for $(\tilde{P}_1, \tilde{P}_2) \sim \text{hDP}(c = 1, c_0 = 1, P_0 = \text{Unif}_{[0,1]})$ and data points sampled from the model. The kernel correlation is computed for different kernels: Gaussian with $\sigma = 1$, Laplace with $\beta = 1$, set-wise with $A = [0, 0.95]$, and linear.

the left of Figure 3.1 show the estimated posterior kernel covariance for the two methods, annotating the first, second, and third quartiles. We observe that the sampling-based method exhibits significantly more variability than the analytics-based method. A possible explanation is that the former, despite the vast number of generated samples, does not account for the additional knowledge about the posterior, whereas the latter is built *ad hoc* for this model. There is an evident trade-off between generality and variability: when possible, analytic computations can reduce the variability.

3.8.2 Convergence Rate

Our aim is to empirically recover the convergence rate $\mathcal{O}((n_1 n_2)^{-1/2})$ of Theorem 3.7.1. We recover this for several injective kernels, while empirically showing that the set-wise kernel has a slower convergence rate.

We compute the posterior kernel correlation with Gaussian, Laplace, linear, and set-wise kernels for $(\tilde{P}_1, \tilde{P}_2) \sim \text{hDP}(c = 1, c_0 = 1, P_0 = \text{Unif}_{[0,1]})$, when the observations are sampled from the model for $n_1 = 4^i$ and $n_2 = 5^j$ for $i, j = 2, 3, 4, 5$. We choose these values to have a grid of different values for $n_1 n_2$. The Gaussian and the Laplace kernel have their parameters set to $\sigma = \beta = 1$, while the set-wise kernel is taken for $A = [0, 0.95]$. We use the analytics-based method because of its lower variance, with $R = 1000$. The log-log plot on the right of [Figure 3.1](#) reports the value of the estimated posterior kernel correlation for the different values of $n_1 n_2$ and different kernel choices. Moreover, we report the regression line of the log-kernel correlation as a function of $\log(n_1 n_2)$. The estimated value of the slope is approximately $-1/2$ for Gaussian, Laplace, and linear kernels, as it is noticeable in [Figure 3.1](#). This result, in the logarithmic domain, confirms our theoretical findings since the sequence of observables is non-degenerate for injective kernels. However, the set-wise correlation in [Theorem 3.7.1](#) fails to capture this rate of convergence and has a much slower decrease due to the choice of $A = [0, 0.95]$. Since the marginal distribution of the observables is $\text{Unif}_{[0,1]}$, almost all the values will be contained in A , making the sequence *almost* degenerate for the chosen set.

3.8.3 Stability of the Kernel

We now investigate the stability of the posterior kernel correlation with respect to the hyperparameters of the kernel. As expected, the choice of hyperparameters does not have a significant impact on the value of the kernel correlation for Gaussian or Laplace kernels, while it dramatically changes the value of the set-wise correlation.

As shown in [Example 3.5.1](#), the set-wise kernel can depend heavily on the choice of the set A , and we expect this to be the case for the posterior since it has a similar structure. We investigate the same question for the parameters of Gaussian and Laplace kernels by considering a similar setting to the one in the previous section. We compute the posterior kernel correlation with Gaussian, Laplace, and set-wise kernels for $(\tilde{P}_1, \tilde{P}_2) \sim$

$\text{hDP}(c = 1, c_0 = 1, P_0 = \text{Unif}_{[0,1]})$, when the observations are sampled from the model for $n_1 = n_2 \in \{0, 10, 100, 1000\}$. We choose the parameters for Gaussian, Laplace, and set-wise kernels to take one of three values as follows: $\sigma, \beta \in \{10^{-3}, 1, 10^3\}$, and $A \in \{[0, .1], [0, .5], [0, .9]\}$. We use the analytics-based algorithm with $R = 1000$. Table 3.3 shows the values of the kernel correlation for different kernels and parameters as the sizes of the observables, n_1 and n_2 , increase. Both Gaussian and Laplace kernels show little variability for any sample size. In contrast, for different choices of A , the variability of the set-wise correlation rises dramatically as n_1, n_2 increase, suggesting a lack of robustness of the index a posteriori. This is especially problematic as it is challenging to elicit A without prior knowledge of the data since, for a sensible measurement, A must not contain too many or too few observations.

3.9 Model Comparison

Having established an RKHS-based index for measuring partial exchangeability for both parametric and nonparametric models, we use it to calibrate prior parameters across different models to match the same index value. We consider the Gaussian parametric model in Example 3.1.1 and the hierarchical Dirichlet Process in (3.5) with $P_0 = \mathcal{N}(0, t^2)$ with $t^2 = s^2 + \tau^2$, so that both models share the same prior predictive distribution. We set the remaining parameters to ensure the same value of marginal kernel correlation a priori and study the implications for posterior inference.

Table 3.3: Value of kernel correlation for Gaussian, Laplace, and set-wise kernels for different choices of parameters and sample sizes n_1, n_2 .

n_1, n_2	Gaussian, $\sigma > 0$			Laplace, $\beta > 0$			Set-wise, $A = [0, b]$		
	$\sigma = 10^{-3}$	$\sigma = 1$	$\sigma = 10^3$	$\beta = 10^{-3}$	$\beta = 1$	$\beta = 10^3$	$b = 0.1$	$b = 0.5$	$b = 0.9$
0	$6.7 \cdot 10^{-1}$	$6.7 \cdot 10^{-1}$	$6.7 \cdot 10^{-1}$	$6.7 \cdot 10^{-1}$	$6.7 \cdot 10^{-1}$	$6.7 \cdot 10^{-1}$	$6.7 \cdot 10^{-1}$	$6.7 \cdot 10^{-1}$	$6.7 \cdot 10^{-1}$
10	$1.8 \cdot 10^{-2}$	$1.7 \cdot 10^{-2}$	$1.7 \cdot 10^{-2}$	$1.8 \cdot 10^{-2}$	$1.7 \cdot 10^{-2}$	$1.7 \cdot 10^{-2}$	$1.2 \cdot 10^{-1}$	$1.6 \cdot 10^{-2}$	$1.2 \cdot 10^{-1}$
100	$1.8 \cdot 10^{-3}$	$1.7 \cdot 10^{-3}$	$1.7 \cdot 10^{-3}$	$1.8 \cdot 10^{-3}$	$1.6 \cdot 10^{-3}$	$1.7 \cdot 10^{-3}$	$6.9 \cdot 10^{-2}$	$1.6 \cdot 10^{-3}$	$6.9 \cdot 10^{-2}$
1000	$1.9 \cdot 10^{-4}$	$1.7 \cdot 10^{-4}$	$1.7 \cdot 10^{-4}$	$1.7 \cdot 10^{-4}$	$1.8 \cdot 10^{-4}$	$1.7 \cdot 10^{-4}$	$5.0 \cdot 10^{-2}$	$1.8 \cdot 10^{-4}$	$5.0 \cdot 10^{-2}$

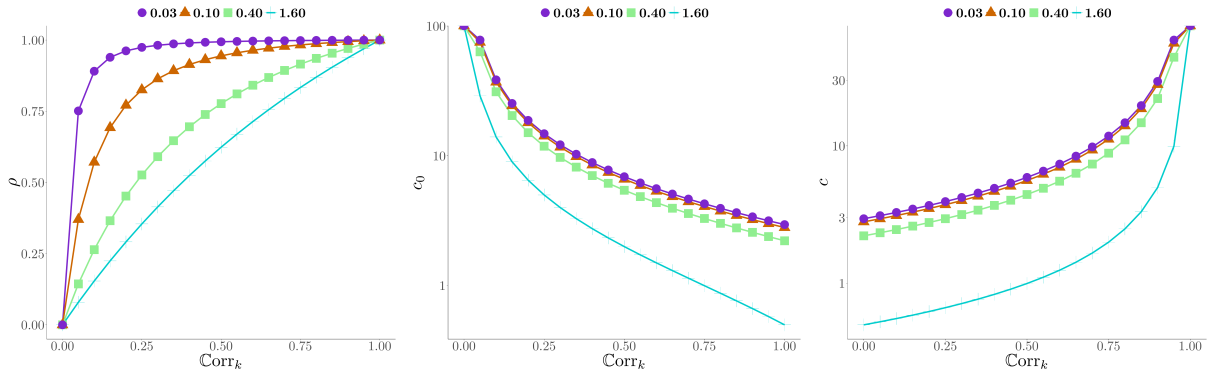


Figure 3.2: Values of different parameters for the Gaussian model and the hDP for $v = 1/4$, $t^2 = 2$, and different values of σ as the kernel correlation varies. **Left:** Value of ρ for the Gaussian model. **Center:** Value of c_0 for the hDP. **Right:** Value of c for the hDP.

As shown in [Appendix 3.E](#), given a fixed marginal variance of the observables ($\text{Var}(X_i) = t^2$) and fixed kernel variances a priori ($\text{Var}_k(\tilde{P}_i) = v$), we can compute the model parameters (s, τ, ρ for the Gaussian model, c_0, c for the hDP) for any value of the kernel correlation $\text{Corr}_k(\tilde{P}_1, \tilde{P}_2)$, provided that the parameter σ of the Gaussian kernel satisfies $\sigma < \sigma^* := \sqrt{2}t/\sqrt{1/(1-v)^2 - 1}$. In [Figure 3.2](#), we study how the values of the parameters change as the kernel correlation varies for $t^2 = 2$, $v = 1/4$, and $\sigma^2 = (\sigma^*/4^i)^2/2$ for $i = 0, 1, 2, 3$. For the Gaussian model, the value of ρ (on the left) increases as the kernel correlation increases, with $\rho = 0$ corresponding to independence between the two groups and a null kernel correlation, and $\rho = 1$ corresponding to full exchangeability between the two groups and a kernel correlation equal to 1. For the hDP, we notice that c_0 (in the middle) diverges as the kernel correlation approaches 0; this can be explained by the fact that we obtain almost sure equality of the two random probability measures whenever $\tilde{P}_0$ converges to P_0 a.s.. Conversely, c (on the right) diverges as the kernel correlation approaches 1 since both $\tilde{P}_1$ and $\tilde{P}_2$ converge to $\tilde{P}_0$ a.s.. We observe that a high value of σ yields the same correlation value for smaller values of c and c_0 , resulting in greater numerical stability. Thus, we fix $\sigma = \sigma^*/\sqrt{2}$ for the rest of the section.

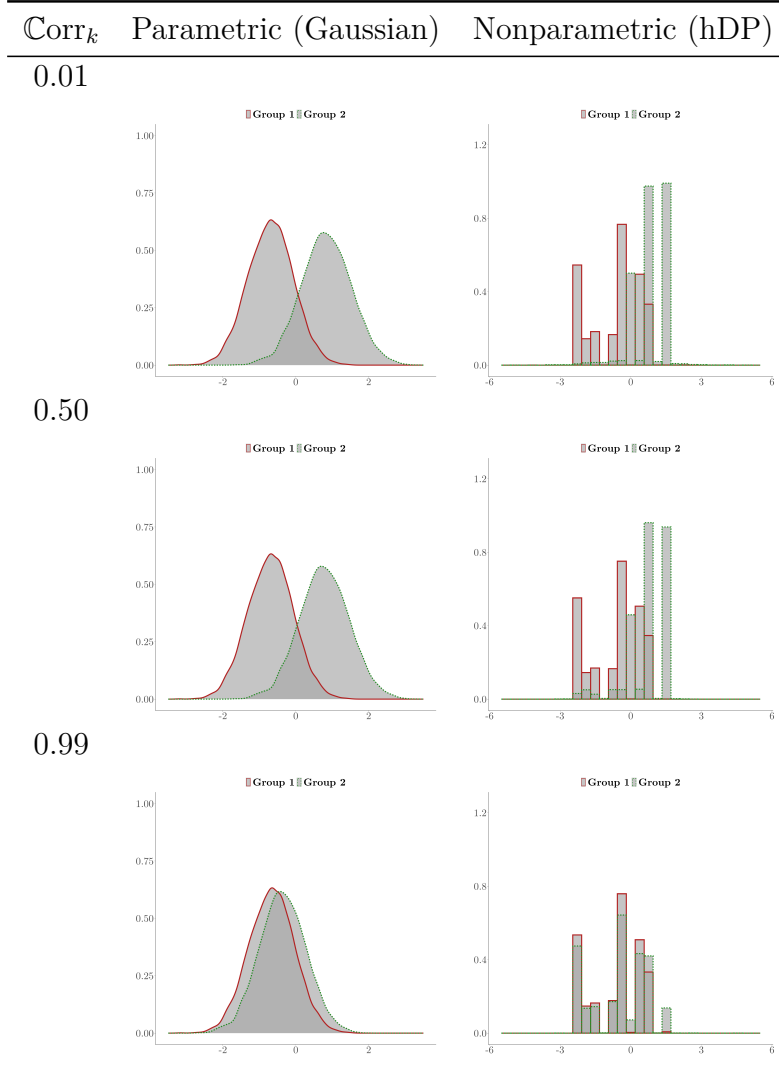
We provide two illustrations of how the dependence a priori impacts posterior inference, on simulated and real data, highlighting the necessity of matching dependencies for accurate

model comparisons. First, we generate the observations in each group as independent sequences $(X_{1,j})_{j=1}^{n_1}$ and $(X_{2,j})_{j=1}^{n_2}$ according to the following sampling scheme,

$$\begin{aligned} X_{1,1}, \dots, X_{1,n_1} | \tilde{P}_1 &\stackrel{\text{i.i.d.}}{\sim} \tilde{P}_1 & \tilde{P}_1 &\sim \text{hDP}(c = 10, c_0 = 10, P_0 = \mathcal{N}(-1, 2)), \\ X_{2,1}, \dots, X_{2,n_2} | \tilde{P}_2 &\stackrel{\text{i.i.d.}}{\sim} \tilde{P}_2 & \tilde{P}_2 &\sim \text{hDP}(c = 10, c_0 = 10, P_0 = \mathcal{N}(1, 2)). \end{aligned} \quad (3.7)$$

We consider unbalanced groups, as borrowing information is particularly useful in this setting, with $n_1 = 200$ and $n_2 = 5$. We consider three different values of the index,

Table 3.4: Predictive distributions for different values of the kernel correlation for the Gaussian and the hDP cases, conditionally on $(X_{1,j})_{j=1}^{n_1}$ and $(X_{2,j})_{j=1}^{n_2}$ as in (3.7) for $n_1 = 200$ and $n_2 = 5$.



namely $\text{Corr}_k \in \{0.01, 0.50, 0.99\}$, corresponding to the situation of almost independence, intermediate dependence, and almost exchangeability, respectively. In [Table 3.4](#), we show the empirical distribution of a sample of size $M = 10,000$ from the posterior mean measure, which coincides with the posterior one-step-ahead predictive distribution, for the two groups for both models. We set $v = 1/4$, $t^2 = 2$, $\sigma = \sigma^*/\sqrt{2}$, and the values of ρ , c_0 , c determined by the value of the kernel correlation, as shown in [Figure 3.3](#). We notice that the two distributions are more similar for values of the kernel correlation close to 1, as expected, and the second group is heavily affected by the larger number of observations in the first group. This intuition can be quantified through the absolute difference of their means, which we estimate through the sample means. The results are shown in the top row of [Figure 3.3](#). On the left, we plot the absolute mean differences for different values of the kernel correlations for the Gaussian model; at the center, we reproduce the same analysis for the hDP. As expected, the estimates tend to be closer for similar values of the index; in other words, the values near the main diagonal tend to be smaller. On the right, we compare the absolute mean differences between the Gaussian model and the hDP. We notice that this distance tends to be bigger between two instances of the same model (both Gaussian or hDP) with different kernel correlations, rather than between a Gaussian and an hDP model with the same kernel correlation.

We conclude with a similar type of analysis on the Palmer Archipelago (Antarctica) penguin data ([Horst et al., 2020](#)) with the goal of making predictions on the flipper length for male and female penguins. We consider all male penguins in the dataset (168) and a subsample of 5 female penguins to benefit from the borrowing of information across groups. The presence of ties could be ascribed to rounding error, leading to a dominated hierarchical model such as the parametric Gaussian hierarchical model in [Example 3.1.1](#), or to the presence of latent subpopulations, leading to an a.s. discrete hierarchical model such as the hierarchical Dirichlet process in [\(3.5\)](#). We show that setting the same kernel correlation a priori leads to a more meaningful model comparison: indeed, the predictions

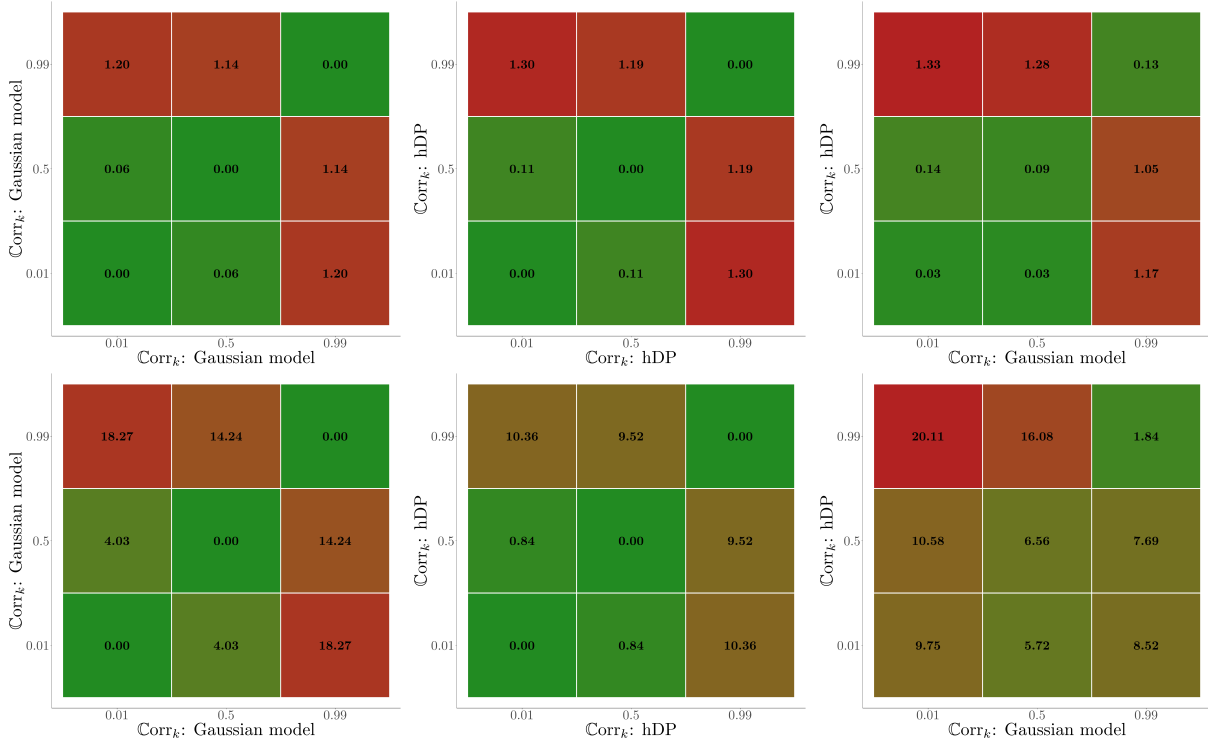


Figure 3.3: Absolute difference between the empirical averages of two samples of size $M = 10,000$ from the predictive distribution of Group 2 for different values of kernel correlation for Gaussian vs. Gaussian (**left**), hDP vs. hDP (**center**), and Gaussian vs. hDP (**right**) for $v = 1/4$, $t^2 = 2$, and $\sigma = t/\sqrt{1/(1-v)^2 - 1}$. **Top row**: the data are simulated from (3.7); **Bottom row**: the data comes from the Palmer Penguins dataset Horst et al. (2020).

using the same model with different kernel correlations can lead to differences between mean predictions that are larger than those obtained by using different models with the same value of kernel correlation. This is pictured in the bottom row of Figure 3.3, where, e.g., the difference between the hierarchical Gaussian model with kernel correlation 0.5 and the hDP model with kernel correlation 0.99 dominates the difference between the two models with the same kernel correlation 0.5. This analysis provides evidence of the importance of fixing the same prior kernel correlation, when possible, to mitigate the effect of the choice of the hyperparameters in the model comparison.

Discussion

In this work, we have introduced a measure of partial exchangeability by quantifying the dependence of random probability measures through Reproducing Kernel Hilbert Spaces. A distinctive feature of our index, termed *kernel correlation*, is to detect exchangeability for a broad class of models, including almost surely discrete random probabilities, their posterior updates, mixture models, and standard parametric models. We have identified some mild conditions on the marginal distributions that are remarkably agnostic of the dependence structure and that we expect to hold in many other settings not considered in this work.

The kernel correlation extends the widely used set-wise correlation and coincides with the former for multivariate species sampling models (Franzolini et al., 2025), which encompass most discrete priors in the Bayesian nonparametric literature; hence, it is easily computable in such settings. We have shown that they coincide on mixture models as well. For other random probabilities, such as those arising from posterior and parametric models, our kernel correlation shows better properties in terms of detection of exchangeability, robustness, and posterior convergence rates of the dependence structure. We have provided a simple and efficient estimator that only uses four observables of the partially exchangeable sequence. To our knowledge, this is the first framework that enables a fair model comparison between parametric and nonparametric models by fixing the same amount of prior dependence. We were also able to investigate the behavior of the dependence structure a posteriori. Remarkably, we have found that the kernel correlation for the hierarchical Dirichlet process (Teh et al., 2006) goes to zero at a parametric rate of convergence. Since the dependence structure drives the borrowing of information, our results show that as the sample size grows, each additional datapoint contributes progressively less to the borrowing of information across groups. This work lays the groundwork for analyzing the dependence a posteriori for other hierarchical models (Camerlenghi et al., 2019) or more

general dependent priors ([Quintana et al., 2022](#); [Wade and Inácio, 2025](#)), complementing well-established frequentist asymptotic analyses on the recovery of the true distribution in partially exchangeable settings ([Nguyen, 2016](#); [Catalano et al., 2022](#)).

Appendices

Below, we provide additional materials that were omitted or summarized in [Chapter 3](#). The content is organized as follows.

- In [Appendix 3.A](#), we provide some examples related to the statements in [Chapter 3](#).
- In [Appendix 3.B](#), we prove the statements in [Chapter 3](#).
- In [Appendix 3.C](#), we recall some distributional properties of the hierarchical Dirichlet process that are useful for the proofs and the simulations in [Chapter 3](#).
- In [Appendix 3.D](#), we describe the algorithms used to estimate the kernel correlation a posteriori for the hierarchical Dirichlet process.
- In [Appendix 3.E](#), we provide additional details on the computations for the model comparison in [Section 3.9](#).

3.A Examples

Example 3.1.1 (Computations) Consider $X_{i,j}|\theta_1, \theta_2 \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\theta_i, s^2)$ for $j \in \mathbb{N}$ and $i = 1, 2$, with $(\theta_1, \theta_2) \sim \mathcal{N}(\mathbf{0}, \tau^2 \Sigma)$ for $s, \tau > 0$, where $\Sigma_{11} = \Sigma_{22} = 1$ and $\Sigma_{12} = \Sigma_{21} = \rho \in [-1, 1]$.

The negative log-likelihood can be expressed, up to an additive constant, as

$$-\log \mathbb{P}(\mathbf{X}^{(n_1, n_2)} | \theta_1, \theta_2) = \frac{1}{2s^2} \sum_{j=1}^{n_1} (X_{1,j} - \theta_1)^2 + \frac{1}{2s^2} \sum_{j=1}^{n_2} (X_{2,j} - \theta_2)^2.$$

Thus, by Bayes' rule, $\theta_1, \theta_2 | \mathbf{X}^{(n_1, n_2)}$ has density proportional to

$$\mathbb{P}(\theta_1, \theta_2 | \mathbf{X}^{(n_1, n_2)}) \propto \exp \left(-\frac{1}{2s^2} \sum_{j=1}^{n_1} (X_{1,j} - \theta_1)^2 - \frac{1}{2s^2} \sum_{j=1}^{n_2} (X_{2,j} - \theta_2)^2 - \frac{1}{2\tau^2} \boldsymbol{\theta}^\top \Sigma^{-1} \boldsymbol{\theta} \right).$$

By standard algebraic manipulations, one can prove that $\theta_1, \theta_2 | \mathbf{X}^{(n_1, n_2)}$ follows a Gaussian distribution $\mathcal{N}(\boldsymbol{\theta}^*, \Sigma^*)$ with

$$\Sigma^* = \frac{s^2 \tau^2}{s^4 + (n_1 + n_2)s^2 \tau^2 + n_1 n_2 \tau^4 (1 - \rho^2)} \begin{pmatrix} s^2 + n_2 \tau^2 (1 - \rho^2) & s^2 \rho \\ s^2 \rho & s^2 + n_1 \tau^2 (1 - \rho^2) \end{pmatrix} \quad (3.A.1)$$

and

$$\boldsymbol{\theta}^* = \frac{\tau^2}{s^4 + (n_1 + n_2)s^2 \tau^2 + n_1 n_2 \tau^4 (1 - \rho^2)} \begin{pmatrix} (s^2 + n_2 \tau^2 (1 - \rho^2))n_1 \bar{X}_1 + s^2 \rho n_2 \bar{X}_2 \\ (s^2 + n_1 \tau^2 (1 - \rho^2))n_2 \bar{X}_2 + s^2 \rho n_1 \bar{X}_1 \end{pmatrix}, \quad (3.A.2)$$

where $\bar{X}_i = \sum_{j=1}^{n_i} X_{i,j} / n_i$ for $i = 1, 2$ is the empirical mean.

The correlation between θ_1 and θ_2 a posteriori depends only on Σ^* but not on $\mathbf{X}^{(n_1, n_2)}$.

A direct computation gives

$$\text{Corr}(\theta_1, \theta_2 | \mathbf{X}^{(n_1, n_2)} = \mathbf{x}^{(n_1, n_2)}) = \frac{\Sigma_{12}^*}{\sqrt{\Sigma_{11}^*} \sqrt{\Sigma_{22}^*}} = \frac{\rho}{\sqrt{1 + n_1 \frac{\tau^2}{s^2} (1 - \rho^2)} \sqrt{1 + n_2 \frac{\tau^2}{s^2} (1 - \rho^2)}}.$$

Example 3.A.1. Consider the linear kernel $k(x, y) = xy$ on $\mathbb{X} = [0, 2]$. For $X \sim \text{Unif}_{[0,1]}$, let $\tilde{P}_1 := \delta_X$ and $\tilde{P}_2 := \delta_{X+1}$. It follows that $\tilde{P}_1, \tilde{P}_2$ are a.s. discrete with atomless mean measures $\mathbb{E}[\tilde{P}_1] = \text{Unif}_{[0,1]}$ and $\mathbb{E}[\tilde{P}_2] = \text{Unif}_{[1,2]}$. However, $\text{Corr}_k(\tilde{P}_1, \tilde{P}_2) = \text{Corr}(X, X + 1) = 1$.

Example 3.A.2. Let $\tilde{P}_i = \omega_i \delta_{x_i} + (1 - \omega_i) \tilde{P}$, where $\tilde{P}$ is an a.s. discrete random probability, $\omega_i \in (0, 1)$, and $x_1 \neq x_2 \in \mathbb{X}$, for $i = 1, 2$. Note that $\tilde{P}_1$ and $\tilde{P}_2$ have fixed jumps at deterministic points and, thus, do not satisfy the assumptions of [Theorem 3.3.3](#). Since

$\mathbb{V}\text{ar}_k(\tilde{P}_i) = (1 - \omega_i)^2 \mathbb{V}\text{ar}_k(\tilde{P})$ and $\mathbb{C}\text{ov}_k(\tilde{P}_1, \tilde{P}_2) = (1 - \omega_1)(1 - \omega_2) \mathbb{V}\text{ar}_k(\tilde{P})$, it follows that $\mathbb{C}\text{orr}_k(\tilde{P}_1, \tilde{P}_2) = 1$.

Example 3.A.3. For $X \sim \text{Unif}_{[0,1]}$, let $\tilde{P}_1 := \delta_X$ and $\tilde{P}_2 := \delta_{1-X}$, which have atomless mean measures. If we consider the set $A = [1/4, 3/4]$, then $\tilde{P}_1(A) = \tilde{P}_2(A)$ a.s.. Thus, $\mathbb{C}\text{orr}(\tilde{P}_1(A), \tilde{P}_2(A)) = 1$.

Example 3.A.4. Let $W \sim \text{Unif}_{[0,1]}$ and let $P \in \mathcal{P}(\mathbb{X})$. For any closed set A there exists a sequence $(x_n)_{n \in \mathbb{N}} \subset A^c$ such that $x_n \rightarrow x_\infty \in \partial A$ as $n \rightarrow \infty$. The sequence of random probabilities $\tilde{P}_n := W\delta_{x_n} + (1 - W)P$ converges weakly to $\tilde{P}_\infty := W\delta_{x_\infty} + (1 - W)P$ a.s.. If k is continuous and bounded, the Dominated Convergence Theorem easily implies that $\mathbb{C}\text{orr}_k(\tilde{P}_n, \tilde{P}_\infty) \rightarrow 1$ as $n \rightarrow +\infty$, as one would expect. However, by taking the set-wise kernel for A , [Example 3.5.1](#) shows that $\mathbb{C}\text{orr}(\tilde{P}_n(A), \tilde{P}_\infty(A)) = -1$ for every $n \in \mathbb{N}$.

Example 3.A.5. Firstly, we observe that $\mathbb{V}\text{ar}(\tilde{P}_0(A)) = P_0(A)(1 - P_0(A))/(1 + c_0)$ since $\tilde{P}_0 \sim \text{DP}(c_0, P_0)$, leading to $\tilde{P}_0(A) \sim \text{Beta}(c_0 P_0(A), c_0(1 - P_0(A)))$. Secondly, to find the variance of $\tilde{P}_i(A)$, we apply the law of total variance conditioning on $\tilde{P}_0$. With some algebraic manipulations, we obtain

$$\mathbb{V}\text{ar}(\tilde{P}_i(A)) = \frac{1 + c + c_0}{(1 + c)(1 + c_0)} P_0(A)(1 - P_0(A)).$$

Thus the hDP a priori satisfies [Corollary 3.5.3](#) with $\lambda_1 = \lambda_2 = (1 + c + c_0)(1 + c)^{-1}(1 + c_0)^{-1}$.

With similar calculations based on the law of total covariance and conditional independence,

$$\mathbb{C}\text{ov}(\tilde{P}_1(A), \tilde{P}_2(A)) = \mathbb{V}\text{ar}(\tilde{P}_0(A)) = \frac{P_0(A)(1 - P_0(A))}{1 + c_0}.$$

This proves that the hDP also satisfies [Proposition 3.5.2](#) with $\eta = (1 + c_0)^{-1}$. Thus by [Theorem 3.5.4](#),

$$\mathbb{C}\text{orr}_k(\tilde{P}_1, \tilde{P}_2) = \mathbb{C}\text{orr}(\tilde{P}_1(A), \tilde{P}_2(A)) = \frac{1 + c}{1 + c + c_0},$$

for any injective kernel k and any measurable set A such that $P_0(A) \notin \{0, 1\}$.

3.B Proofs of the Statements

3.B.1 Proof of Proposition 3.2.1

Since $\mathbb{E}[\mu_k(\tilde{P}_i)] = \mu_k(P_{0,i})$ for $i = 1, 2$, we have

$$\mathbb{Cov}_k(\tilde{P}_1, \tilde{P}_2) = \mathbb{E} \left[\left\langle \mu_k(\tilde{P}_1), \mu_k(\tilde{P}_2) \right\rangle_{\mathbb{H}_k} \right] - \left\langle \mu_k(P_{0,1}), \mu_k(P_{0,2}) \right\rangle_{\mathbb{H}_k}.$$

Thus, the result derives from the application of (2.3).

3.B.2 Proof of Lemma 3.2.2

Note that

$$\mathbb{Var}_k(\tilde{P}) = \mathbb{E} \left[\left\| \mu_k(\tilde{P}) - \mathbb{E}[\mu_k(\tilde{P})] \right\|_{\mathbb{H}_k}^2 \right].$$

Thus $\mathbb{Var}_k(\tilde{P}) \geq 0$ with equality if and only if $\left\| \mu_k(\tilde{P}) - \mathbb{E}[\mu_k(\tilde{P})] \right\|_{\mathbb{H}_k}^2 = 0$ almost surely, that is, $\mu_k(\tilde{P}) = \mathbb{E}[\mu_k(\tilde{P})]$ almost surely. If k is characteristic, and as μ_k is linear, it happens if and only if $\tilde{P} = \mathbb{E}[\tilde{P}]$ almost surely, which means $\tilde{P}$ is deterministic.

3.B.3 Proof of Proposition 3.2.3

The correlation belongs to $[-1, 1]$ from an application of Cauchy-Schwarz:

$$\begin{aligned} \mathbb{Cov}_k^2(\tilde{P}_1, \tilde{P}_2) &= \mathbb{E} \left[\left\langle \mu_k(\tilde{P}_1) - \mathbb{E}[\mu_k(\tilde{P}_1)], \mu_k(\tilde{P}_2) - \mathbb{E}[\mu_k(\tilde{P}_2)] \right\rangle_{\mathbb{H}_k} \right]^2 \\ &\leq \mathbb{E} \left[\left\| \mu_k(\tilde{P}_1) - \mathbb{E}[\mu_k(\tilde{P}_1)] \right\|_{\mathbb{H}_k}^2 \right] \mathbb{E} \left[\left\| \mu_k(\tilde{P}_2) - \mathbb{E}[\mu_k(\tilde{P}_2)] \right\|_{\mathbb{H}_k}^2 \right] \\ &= \mathbb{Var}_k(\tilde{P}_1) \mathbb{Var}_k(\tilde{P}_2). \end{aligned}$$

If $\tilde{P}_1$ and $\tilde{P}_2$ are independent, so are $\mu_k(\tilde{P}_1)$ and $\mu_k(\tilde{P}_2)$, and thus $\mathbb{E}[\langle \mu_k(\tilde{P}_1), \mu_k(\tilde{P}_2) \rangle_{\mathbb{H}_k}] = \langle \mathbb{E}[\mu_k(\tilde{P}_1)], \mathbb{E}[\mu_k(\tilde{P}_2)] \rangle_{\mathbb{H}_k}$.

3.B.4 Proof of Lemma 3.3.1

Let $(\mathbb{H}, \langle \cdot, \cdot \rangle_{\mathbb{H}})$ be a Hilbert space and let X, Y be $\mathbb{H}$ -valued random variables with finite first and second moments. Without loss of generality, we assume that X and Y are centered. Then, $\text{Corr}_{\mathbb{H}}(X, Y) = \pm 1$ if and only if $\mathbb{E}[\langle X, Y \rangle_{\mathbb{H}}]^2 = \mathbb{E}[\|X\|_{\mathbb{H}}^2] \mathbb{E}[\|Y\|_{\mathbb{H}}^2]$, which means that there is equality in Cauchy-Schwarz. Thus X and Y are collinear, in the sense that $X = \alpha Y$ a.s. for $\alpha = \mathbb{E}[\langle X, Y \rangle_{\mathbb{H}}] / \mathbb{E}[\|Y\|_{\mathbb{H}}^2]$. In particular, α has the same sign as $\text{Corr}_{\mathbb{H}}(X, Y)$.

Applying this reasoning to $X = \mu_k(\tilde{P}_1)$ and $Y = \mu_k(\tilde{P}_2)$, we obtain

$$\mu_k(\tilde{P}_1) - \mathbb{E}[\mu_k(\tilde{P}_1)] = \alpha (\mu_k(\tilde{P}_2) - \mathbb{E}[\mu_k(\tilde{P}_2)]) \quad \text{a.s.}$$

for some $\alpha \in \mathbb{R} \setminus \{0\}$ having the same sign as $\text{Corr}_{\mathbb{H}_k}(\mu_k(\tilde{P}_1), \mu_k(\tilde{P}_2))$. Since the kernel mean embedding is linear and injective for bounded signed measures as k is c_0 -universal, the equality reads $\tilde{P}_1 - \mathbb{E}[\tilde{P}_1] = \alpha (\tilde{P}_2 - \mathbb{E}[\tilde{P}_2])$ a.s..

3.B.5 Proof of Theorem 3.3.2

It is trivial to prove that $\text{Corr}_k(\tilde{P}_1, \tilde{P}_2) = 1$ whenever $\tilde{P}_1 = \tilde{P}_2$ a.s..

Let us prove the converse implication. By Lemma 3.3.1, there exists some $\alpha > 0$ such that $\tilde{P}_1 - \mathbb{E}[\tilde{P}_1] = \alpha (\tilde{P}_2 - \mathbb{E}[\tilde{P}_2])$ a.s., which we rewrite as

$$\tilde{P}_1 - \alpha \tilde{P}_2 = \mathbb{E}[\tilde{P}_1] - \alpha \mathbb{E}[\tilde{P}_2] \quad \text{a.s.}$$

Since the random measure on the left is a.s. discrete, while the measure on the right is atomless, both sides must be a.s. null. In other words, $\tilde{P}_1 = \alpha \tilde{P}_2$ a.s.. Evaluating both

sides of this equality at $\mathbb{X}$, we obtain that $\alpha = 1$ as $\tilde{P}_1(\mathbb{X}) = \tilde{P}_2(\mathbb{X}) = 1$ a.s..

3.B.6 Proof of Theorem 3.3.3

It is trivial to prove that $\text{Corr}_k(\tilde{P}_1, \tilde{P}_2) = 1$ whenever $\tilde{P}_1 = \tilde{P}_2$ a.s..

Let us now prove the converse implication. Firstly, let us notice that the mean measures of $\tilde{P}_1$ and $\tilde{P}_2$, as any other probability measure, can be decomposed as

$$\mathbb{E}[\tilde{P}_i] = \omega_i Q_i^{\text{d}} + (1 - \omega_i) Q_i^{\text{a}},$$

where $\omega_i \in [0, 1]$, Q_i^{d} is a discrete probability measure on $\mathbb{X}$, and Q_i^{a} is an atomless probability measure on $\mathbb{X}$. By Lemma 3.3.1, there exist $\alpha > 0$ such that $\tilde{P}_1 - \mathbb{E}[\tilde{P}_1] = \alpha(\tilde{P}_2 - \mathbb{E}[\tilde{P}_2])$ a.s., which can be rewritten as

$$\tilde{P}_1 - \omega_1 Q_1^{\text{d}} - \alpha(\tilde{P}_2 - \omega_2 Q_2^{\text{d}}) = (1 - \omega_1) Q_1^{\text{a}} - \alpha(1 - \omega_2) Q_2^{\text{a}} \quad \text{a.s.}$$

Since the random measure on the left is a.s. discrete, while the measure on the right is atomless, both sides must be a.s. null. In particular, focusing on the left-hand side, we have

$$\tilde{P}_1 = \alpha \tilde{P}_2 + \xi \quad \text{a.s.},$$

where $\xi := \omega_1 Q_1^{\text{d}} - \alpha \omega_2 Q_2^{\text{d}}$ is a discrete bounded signed measure. Let z be any atom of ξ . On the one hand, if $\xi(\{z\}) > 0$, then $\tilde{P}_1(\{z\}) \geq \xi(\{z\}) > 0$ a.s.: it implies that z is a fixed atom of $\tilde{P}_1$, but it contradicts that $\mathbb{P}(\tilde{P}_1(\{z\}) < \varepsilon) > 0$ for any $\varepsilon > 0$. On the other hand, if $\xi(\{z\}) < 0$, then $\tilde{P}_2(\{z\}) \geq -\xi(\{z\})/\alpha > 0$: again, it implies that z is a fixed atom of $\tilde{P}_2$, but it contradicts that $\mathbb{P}(\tilde{P}_2(\{z\}) < \varepsilon) > 0$ for any $\varepsilon > 0$. Thus $\xi(\{z\}) = 0$ for any z , which implies that $\xi = 0$ as it is a purely discrete measure. From $\xi = 0$ it is easy to conclude that $\tilde{P}_1 = \alpha \tilde{P}_2$ a.s., thus $\tilde{P}_1 = \tilde{P}_2$ as they are both probability measures a.s..

3.B.7 Proof of Theorem 3.4.1

By linearity of the kernel mean embedding, $\mu_k(\tilde{P}_i) = \mathbb{E}[k(X_{i,1}, \cdot) | \tilde{P}_i]$ for $i = 1, 2$. Moreover, $X_{1,1}$ and $X_{2,1}$ are independent given $\tilde{P}_1, \tilde{P}_2$. Using the law of total covariance

$$\begin{aligned} \mathbb{Cov}_{\mathbb{H}_k}(k(X_{1,1}, \cdot), k(X_{2,1}, \cdot)) &= \mathbb{Cov}_{\mathbb{H}_k}\left(\mathbb{E}[k(X_{1,1}, \cdot) | \tilde{P}_1, \tilde{P}_2], \mathbb{E}[k(X_{2,1}, \cdot) | \tilde{P}_1, \tilde{P}_2]\right) \\ &= \mathbb{Cov}_{\mathbb{H}_k}(\mu_k(\tilde{P}_1), \mu_k(\tilde{P}_2)) \\ &= \mathbb{Cov}_k(\tilde{P}_1, \tilde{P}_2). \end{aligned}$$

Similarly, for the second term, we use that $\mu_k(\tilde{P}_i) = \mathbb{E}[k(X_{i,1}, \cdot) | \tilde{P}_i] = \mathbb{E}[k(X_{i,2}, \cdot) | \tilde{P}_i]$ and $X_{i,1}$ and $X_{i,2}$ are independent given $\tilde{P}_i$ for $i = 1, 2$. Thus, for $i = 1, 2$,

$$\begin{aligned} \mathbb{Cov}_{\mathbb{H}_k}(k(X_{i,1}, \cdot), k(X_{i,2}, \cdot)) &= \mathbb{Cov}_{\mathbb{H}_k}\left(\mathbb{E}[k(X_{i,1}, \cdot) | \tilde{P}_i], \mathbb{E}[k(X_{i,2}, \cdot) | \tilde{P}_i]\right) \\ &= \mathbb{Cov}_{\mathbb{H}_k}(\mu_k(\tilde{P}_i), \mu_k(\tilde{P}_i)) \\ &= \mathbb{Var}_k(\tilde{P}_i). \end{aligned}$$

3.B.8 Proof of Proposition 3.4.2

If Y, Z are two random variables valued in a Hilbert space $\mathbb{H}$ and $(Y^{(t)}, Z^{(t)})_{t=1}^M$ are i.i.d. samples from them, then an unbiased estimator of the covariance between Y and Z is

$$\widehat{\mathbb{Cov}}_{\mathbb{H}, M}(Y, Z) = \frac{1}{M-1} \sum_{t=1}^M \langle Y^{(t)} - \bar{Y}, Z^{(t)} - \bar{Z} \rangle_{\mathbb{H}} \quad \text{for } \bar{Y} = \frac{1}{M} \sum_{t=1}^M Y^{(t)}, \bar{Z} = \frac{1}{M} \sum_{t=1}^M Z^{(t)}.$$

Expanding the formula, this expression can be rewritten as

$$\widehat{\mathbb{Cov}}_{\mathbb{H}, M}(Y, Z) = \frac{1}{M-1} \sum_{t=1}^M \langle Y^{(t)}, Z^{(t)} \rangle_{\mathbb{H}} - \frac{M}{M-1} \langle \bar{Y}, \bar{Z} \rangle_{\mathbb{H}} \quad (3.B.1)$$

$$= \frac{1}{M-1} \sum_{t=1}^M \langle Y^{(t)}, Z^{(t)} \rangle_{\mathbb{H}} - \frac{1}{M(M-1)} \sum_{t=1}^M \sum_{s=1}^M \langle Y^{(t)}, Z^{(s)} \rangle_{\mathbb{H}}. \quad (3.B.2)$$

For the unbiased estimator of $\text{Cov}_k(\tilde{P}_1, \tilde{P}_2)$, we apply (3.B.2) with $Y = k(X_{1,1}, \cdot)$ and $Z = k(X_{2,1}, \cdot)$, as the covariance between Y and Z is the kernel covariance (see Theorem 3.4.1). We recover the statement due to the reproducing property $\langle k(x, \cdot), k(y, \cdot) \rangle_{\mathbb{H}_k} = k(x, y)$ for any x, y .

For the unbiased estimator of $\text{Var}_k(\tilde{P}_i)$, we use (3.B.2) with $Y = k(X_{i,1}, \cdot)$ and $Z = k(X_{i,2}, \cdot)$.

3.B.9 Proof of Proposition 3.4.3

The proof relies on the delta method. We only sketch it, as we do not aim to find the exact formula for the asymptotic covariance. Let us introduce the random vector $\mathbf{Z}$ in $\mathbb{R}^3 \times \mathbb{H}_k^4$:

$$\mathbf{Z} = \frac{1}{M} \sum_{t=1}^M \mathbf{Z}^{(t)}, \quad \text{with} \quad \mathbf{Z}^{(t)} = \begin{pmatrix} k(X_{1,1}^{(t)}, X_{2,1}^{(t)}) \\ k(X_{1,1}^{(t)}, X_{1,2}^{(t)}) \\ k(X_{2,1}^{(t)}, X_{2,2}^{(t)}) \\ k(X_{1,1}^{(t)}, \cdot) \\ k(X_{1,2}^{(t)}, \cdot) \\ k(X_{2,1}^{(t)}, \cdot) \\ k(X_{2,2}^{(t)}, \cdot) \end{pmatrix}.$$

By the central limit theorem for Hilbert spaces (see e.g. Hoffmann-Jorgensen and Pisier (1976)), and as k is bounded, this vector is asymptotically normal, in the sense that $\sqrt{M}(\mathbf{Z} - \mathbb{E}[\mathbf{Z}])$ converges to a normal distribution. On the other hand, from the formula (3.B.1) in the proof of Proposition 3.4.2,

$$\widehat{\text{Corr}}_{k,M}(\tilde{P}_1, \tilde{P}_2) = \frac{\mathbf{Z}_1 - \langle \mathbf{Z}_4, \mathbf{Z}_6 \rangle_{\mathbb{H}_k}}{\sqrt{\mathbf{Z}_2 - \langle \mathbf{Z}_4, \mathbf{Z}_5 \rangle_{\mathbb{H}_k}} \sqrt{\mathbf{Z}_3 - \langle \mathbf{Z}_6, \mathbf{Z}_7 \rangle_{\mathbb{H}_k}}}.$$

As the inner product is Fréchet differentiable in $\mathbb{H}_k$, thus Hadamard differentiable, we can apply the delta method (van der Vaart and Wellner, 1996, Theorem 3.9.4) to the

differentiable function $\phi(\mathbf{z}) = (\mathbf{z}_1 - \langle \mathbf{z}_4, \mathbf{z}_6 \rangle_{\mathbb{H}_k})(\mathbf{z}_2 - \langle \mathbf{z}_4, \mathbf{z}_5 \rangle_{\mathbb{H}_k})^{-1/2}(\mathbf{z}_3 - \langle \mathbf{z}_6, \mathbf{z}_7 \rangle_{\mathbb{H}_k})^{-1/2}$, and conclude that $\sqrt{M}(\phi(\mathbf{Z}) - \phi(\mathbb{E}[\mathbf{Z}]))$ converges to a Gaussian distribution, which is the conclusion of the theorem.

3.B.10 Proof of Proposition 3.5.1

The kernel k_A is symmetric and bounded by definition. To prove that it is positive semi-definite consider $x_1, \dots, x_m \in \mathbb{X}$ and $a_1, \dots, a_m \in \mathbb{R}$. Firstly, we notice that for $i, j \in \{1, \dots, m\}$, $k_A(x_i, x_j) = 1$ if $x_i, x_j \in A$ and $k_A(x_i, x_j) = 0$ otherwise. Hence, if we denote $I_A = \{i \in \{1, \dots, m\} \text{ s.t. } x_i \in A\}$,

$$\sum_{i=1}^m \sum_{j=1}^m a_i a_j k_A(x_i, x_j) = \sum_{i \in I_A} \sum_{j \in I_A} a_i a_j = \left(\sum_{i \in I_A} a_i \right)^2 \geq 0.$$

Let us write $P_{0,1} := \mathbb{E}[\tilde{P}_1]$ and $P_{0,2} := \mathbb{E}[\tilde{P}_2]$. Then, by Proposition 3.2.1 and Fubini's Theorem, it holds

$$\begin{aligned} \mathbb{Cov}_{k_A}(\tilde{P}_1, \tilde{P}_2) &= \mathbb{E} \left[\iint \mathbb{1}_A(x) \mathbb{1}_A(y) d\tilde{P}_1(x) d\tilde{P}_2(y) \right] - \iint \mathbb{1}_A(x) \mathbb{1}_A(y) dP_{0,1}(x) dP_{0,2}(y) \\ &= \mathbb{E}[\tilde{P}_1(A) \tilde{P}_2(A)] - P_{0,1}(A) P_{0,2}(A) = \mathbb{Cov}(\tilde{P}_1(A), \tilde{P}_2(A)). \end{aligned}$$

3.B.11 Proof of Proposition 3.5.2

By taking $k(x, y) = k_A(x, y) = \mathbb{1}_A(x) \mathbb{1}_A(y)$, Proposition 3.5.1 guarantees that (ii) implies (i). To show the converse, we prove that (ii) holds for an increasingly larger class of kernels.

Step 1. We observe that by linearity, (ii) holds for any kernel function k that can be written as a linear combination of kernel functions k_A .

Step 2. For a measurable function $f : \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{R}$, let us define its “symmetrized” version $\mathcal{S}[f]$ by $\mathcal{S}[f](x, y) = (f(x, y) + f(y, x))/2$. If A, B measurable then (ii) holds for

$k = \mathcal{S}[\mathbb{1}_{A \times B}]$. This comes from linearity as

$$\mathcal{S}[\mathbb{1}_{A \times B}] = \frac{1}{2}(k_{A \cup B} + k_{A \cap B} - k_{A \setminus B} - k_{B \setminus A}).$$

Step 3. Thanks to the monotone class lemma, the property (ii) holds for any $\mathcal{S}[f]$ where f is the indicator function of a measurable set in $\mathcal{X} \otimes \mathcal{X}$. Then from stability by linear combination and the monotone convergence theorem, (ii) holds for any $\mathcal{S}[f]$, where f is a measurable and bounded function over $\mathbb{X} \times \mathbb{X}$. The conclusion follows as $k = \mathcal{S}[k]$ since k is symmetric.

Lastly, we use (i) to prove $\eta \in [-1, 1]$. Indeed, we have

$$\eta = \frac{\text{Cov}(\tilde{P}_1(A), \tilde{P}_2(A))}{P_0(A) - P_0(A)^2} = \frac{\mathbb{E}[\tilde{P}_1(A)\tilde{P}_2(A)] - P_0(A)^2}{P_0(A) - P_0(A)^2}.$$

On the one hand, from $\tilde{P}_2(A) \leq 1$ we see that $\eta \leq (\mathbb{E}[\tilde{P}_1(A)] - P_0(A)^2)/(P_0(A) - P_0(A)^2) =$

1. On the other hand,

$$\eta \geq -\frac{P_0(A)^2}{P_0(A) - P_0(A)^2} = -\frac{P_0(A)}{P_0(A^c)},$$

being A^c the complement of A . Up to exchanging the role of A and A^c , we can always find a set A such that $P_0(A) \leq P_0(A^c)$, hence $\eta \geq -1$.

3.B.12 Proof of [Corollary 3.5.3](#)

We apply [Proposition 3.5.2](#) to $\tilde{P}_1 = \tilde{P}_2 = \tilde{P}$. The only new element is that $\lambda \geq 0$, which is easily deduced from the fact that the variance is always non-negative, see [Lemma 3.2.2](#).

3.B.13 Proof of [Theorem 3.5.4](#)

The result comes from a direct application of [Proposition 3.5.2](#) and [Corollary 3.5.3](#), provided that both variances are non-null for both correlations.

Firstly, $\text{Var}(\tilde{P}_1(A)), \text{Var}(\tilde{P}_2(A)) \neq 0$ since $\tilde{P}_i(A)$ is not deterministic for $i = 1, 2$.

Secondly, $\text{Var}_k(\tilde{P}_1), \text{Var}_k(\tilde{P}_2) \neq 0$. If by contradiction,

$$\int k(x, x) dP_0(x) - \iint k(x, y) dP_0(x) dP_0(y) = \frac{1}{2} \iint d_k^2(x, y) dP_0(x) dP_0(y) = 0,$$

it implies that $d_k^2(x, y) = 0$ for $P_0 \times P_0$ -almost every $x, y \in \mathbb{X}$. Now, $d_k^2(x, y)$ is a distance since k is injective. Hence, $x = y$ for $P_0 \times P_0$ -almost every $x, y \in \mathbb{X}$. This implies that there exists $z \in \mathbb{X}$ such that $P_0 = \delta_z$. Hence, $\tilde{P}_1 = \tilde{P}_2 = \delta_z$, which contradicts the assumption that $\tilde{P}_1$ and $\tilde{P}_2$ are non-deterministic almost surely.

3.B.14 Proof of Theorem 3.6.1

Firstly, k_f is symmetric due to the symmetry of k . With Fubini's Theorem and given the definition of k_f , for any signed bounded measures ξ_1, ξ_2 on $\mathbb{X}$,

$$\begin{aligned} & \iint_{\mathbb{Y} \times \mathbb{Y}} k(y_1, y_2) f_{\xi_1}(y_1) f_{\xi_2}(y_2) dy_1 dy_2 \\ &= \iint_{\mathbb{X} \times \mathbb{X}} \left(\iint_{\mathbb{Y} \times \mathbb{Y}} k(y_1, y_2) f(y_1; x_1) f(y_2; x_2) dy_1 dy_2 \right) d\xi_1(x_1) d\xi_2(x_2) \\ &= \iint_{\mathbb{X} \times \mathbb{X}} k_f(x_1, x_2) d\xi_1(x_1) d\xi_2(x_2). \end{aligned} \quad (3.B.3)$$

Applying this formula with $\xi_1 = \xi_2$, we see that the last expression in (3.B.3) is non-negative (because k is positive definite), and thus k_f is positive definite.

Then, let us write $P_{0,1} := \mathbb{E}[\tilde{P}_1]$ and $P_{0,2} := \mathbb{E}[\tilde{P}_2]$. Thus, $\mathbb{E}[f_{\tilde{P}_i}] = f_{P_{0,i}}$. Applying (3.B.3) with $(\xi_1, \xi_2) = (\tilde{P}_1, \tilde{P}_2)$ and then $(P_{0,1}, P_{0,2})$, we have

$$\begin{aligned} \mathbb{E} \left[\iint_{\mathbb{Y} \times \mathbb{Y}} k(y_1, y_2) f_{\tilde{P}_1}(y_1) f_{\tilde{P}_2}(y_2) dy_1 dy_2 \right] &= \mathbb{E} \left[\iint_{\mathbb{X} \times \mathbb{X}} k_f(x_1, x_2) d\tilde{P}_1(x_1) d\tilde{P}_2(x_2) \right], \\ \iint_{\mathbb{Y} \times \mathbb{Y}} k(y_1, y_2) f_{P_{0,1}}(y_1) f_{P_{0,2}}(y_2) dy_1 dy_2 &= \iint_{\mathbb{X} \times \mathbb{X}} k_f(x_1, x_2) dP_{0,1}(x_1) dP_{0,2}(x_2). \end{aligned}$$

Hence by subtracting these two equalities we obtain $\text{Cov}_k(f_{\tilde{P}_1}, f_{\tilde{P}_2}) = \text{Cov}_{k_f}(\tilde{P}_1, \tilde{P}_2)$.

For the second part of the statement, we assume identifiability of the family $\{f(\cdot; x) :$

$x \in \mathbb{X}\}$ and that k is characteristic, and we want to show that k_f is characteristic. By [Lemma 3.B.1](#) below, which is an easy variant of [Sriperumbudur et al. \(2011, Proposition 4\)](#), we need to show that k_f is conditionally integrally strictly positive definite. Take $\xi \in \mathcal{M}_b(\mathbb{X})$ with $\xi(\mathbb{X}) = 0$, assume $\iint k_f(x_1, x_2) d\xi(x_1) d\xi(x_2) = 0$ and we want to show $\xi = 0$. With [\(3.B.3\)](#),

$$\iint_{\mathbb{X} \times \mathbb{X}} k_f(x_1, x_2) d\xi(x_1) d\xi(x_2) = \iint_{\mathbb{Y} \times \mathbb{Y}} k(y_1, y_2) f_\xi(y_1) f_\xi(y_2) dy_1 dy_2,$$

and, thus, if the left-hand side vanishes, so does the right-hand side. By Fubini, we can check that $f_\xi(y) dy$ is a measure with zero total mass, and as k itself is characteristic, [Lemma 3.B.1](#) implies that, as a measure, $f_\xi(y) dy = 0$. We write $\xi = \xi_+ - \xi_-$ for the Hahn-Jordan decomposition of ξ with $\xi_\pm$ non-negative measures. Calling $a = \xi_+(\mathbb{X}) = \xi_-(\mathbb{X})$ and $P_1 = \xi_+/a$, $P_2 = \xi_-/a$, we write $\xi = a(P_1 - P_2)$ with P_1, P_2 probability distributions over $\mathbb{X}$. By linearity, $f_\xi = a(f_{P_1} - f_{P_2})$, but we already know that $f_\xi = 0$ a.e., thus, $a = 0$ or $f_{P_1} = f_{P_2}$ a.e., which implies $P_1 = P_2$ by identifiability. It means, in any case, that $\xi = 0$, and the proof is concluded.

Lemma 3.B.1. *Let k be a bounded kernel on a space $\mathbb{X}$. Then k is characteristic if and only if it is conditionally integrally strictly positive definite, that is, if and only if, for any $\xi \in \mathcal{M}_b(\mathbb{X})$ with $\xi(\mathbb{X}) = 0$ and $\xi \neq 0$, we have*

$$\iint k(x_1, x_2) d\xi(x_1) d\xi(x_2) > 0.$$

Proof. From the reproducing property [\(2.3\)](#), we have

$$\iint k(x_1, x_2) d\xi(x_1) d\xi(x_2) = \|\mu_k(\xi)\|_{\mathbb{H}_k}^2,$$

and thus the left-hand side vanishes if and only if $\mu_k(\xi) = 0$.

First, assume that k is characteristic. By the Hahn-Jordan decomposition any ξ with

$\xi(\mathbb{X}) = 0$ but $\xi \neq 0$ can be written as $\xi = a(P_1 - P_2)$ with $a > 0$ for P_1, P_2 two distinct probability distributions. Thus, $\mu_k(P_1) \neq \mu_k(P_2)$ and by linearity $\mu_k(\xi) \neq 0$, which implies $\|\mu_k(\xi)\|_{\mathbb{H}_k}^2 > 0$.

Conversely, assume that k is not characteristic, and let P_1 and P_2 be two distinct probability distributions with $\mu_k(P_1) = \mu_k(P_2)$. With $\xi = P_1 - P_2$, we have $\xi \neq 0$ and $\xi(\mathbb{X}) = 0$. But $\mu_k(\xi) = 0$, so $\|\mu_k(\xi)\|_{\mathbb{H}_k}^2 = 0$. It shows that k is not conditionally integrally strictly positive definite. $\square$

3.B.15 Proof of Corollary 3.6.2

From Theorem 3.6.1, it holds that $\mathbb{C}\text{orr}_k(f_{\tilde{P}_1}, f_{\tilde{P}_2}) = \mathbb{C}\text{orr}_{k_f}(\tilde{P}_1, \tilde{P}_2)$. Now, since k_f is an injective kernel, the result follows by Theorem 3.5.4.

3.B.16 Proof of Proposition 3.6.3

The result comes directly by Fubini's Theorem, since $k_f(x_1, x_2)$ is equal to

$$\begin{aligned} & \iint \left(\int e^{-i\langle y_1 - y_2, z \rangle} d\nu(z) \right) f(y_1; x_1) f(y_2; x_2) dy_1 dy_2 \\ &= \int \left(\int e^{-i\langle y_1, z \rangle} f(y_1; x_1) dy_1 \int e^{i\langle y_2, z \rangle} f(y_2; x_2) dy_2 \right) d\nu(z) = \int \hat{f}(z; x_1) \overline{\hat{f}(z; x_2)} d\nu(z). \end{aligned}$$

3.B.17 Proof of Corollary 3.6.4

We use $\mathbb{C}\text{orr}_k(f_{\tilde{P}_1}, f_{\tilde{P}_2}) = \mathbb{C}\text{orr}_{k_f}(\tilde{P}_1, \tilde{P}_2)$ from Theorem 3.6.1 and then apply Theorem 3.3.2 to $\tilde{P}_1, \tilde{P}_2$ with kernel k_f .

3.B.18 Proof of Theorem 3.7.1

For this proof, we rely on the quasi-conjugacy property of the augmented hDP model, which is recalled in Appendix 3.C. This augmented model relies on the introduction of a

specific sequence of latent random variables $\mathbf{T}^{(n_1, n_2)}$ (Teh et al., 2006; Camerlenghi et al., 2019; Catalano et al., 2023), commonly referred to as “tables” in the restaurant franchise metaphor. For the sake of compactness, we write $\mathbf{X}$ and $\mathbf{T}$ instead of $\mathbf{X}^{(n_1, n_2)}$ and $\mathbf{T}^{(n_1, n_2)}$. The key result is that $(\tilde{P}_1, \tilde{P}_2)$, given $\mathbf{X}$ and $\mathbf{T}$, follows a hierarchical Dirichlet process with updated parameters, see (3.C.2) below. We start with an auxiliary lemma before moving to the core of the proof.

Lemma 3.B.2. *In the augmented hDP model in (3.C.2), we have*

$$\mathbb{Cov}_k(\tilde{P}_1, \tilde{P}_2 | \mathbf{X}, \mathbf{T}, \tilde{P}_0) = 0,$$

and, for $i = 1, 2$, with $\hat{P}_i = n_i^{-1} \sum_{j=1}^{n_i} \delta_{X_{i,j}}$,

$$\begin{aligned} \mathbb{Var}_k(\tilde{P}_i | \mathbf{X}, \mathbf{T}, \tilde{P}_0) &= \frac{1}{c + n_i + 1} \left(\frac{1}{2} \frac{c^2}{(c + n_i)^2} \iint d_k^2(x, y) d\tilde{P}_0(x) d\tilde{P}_0(y) \right. \\ &\quad \left. + \frac{cn_i}{(c + n_i)^2} \iint d_k^2(x, y) d\tilde{P}_0(x) d\hat{P}_i(y) + \frac{1}{2} \frac{n_i^2}{(c + n_i)^2} \iint d_k^2(x, y) d\hat{P}_i(x) d\hat{P}_i(y) \right). \end{aligned}$$

Proof. The covariance result follows from observing that in the augmented model (3.C.2), $\tilde{P}_1$ and $\tilde{P}_2$ are independent given $\mathbf{X}$, $\mathbf{T}$ and $\tilde{P}_0$.

The variance result comes from the fact that, conditionally on $\mathbf{X}, \mathbf{T}, \tilde{P}_0$, as in (3.C.2), $\tilde{P}_i$ follows a Dirichlet Process prior with baseline measure $\tilde{P}_i^* = c/(c + n_i) \tilde{P}_0 + n_i/(c + n_i) \hat{P}_i$ and concentration parameter $c + n_i$. Hence,

$$\mathbb{Var}_k(\tilde{P}_i | \mathbf{X}, \mathbf{T}, \tilde{P}_0) = \frac{1}{2} \frac{1}{c + n_i + 1} \iint d_k^2(x, y) d\tilde{P}_i^*(x) d\tilde{P}_i^*(y).$$

The result comes from the expansion of the integral on the right-hand side. $\square$

We now move to the proof of Theorem 3.7.1. To bound the correlation from above, we need to bound the covariance from above and bound the variance from below. We condition on the tables $\mathbf{T}$ and use Lemma 3.B.2 that we just proved. By the variance

decomposition, for $i = 1, 2$,

$$\begin{aligned}\mathbb{V}\text{ar}_k(\tilde{P}_i|\mathbf{X}) &\geq \mathbb{E}\left[\mathbb{V}\text{ar}_k(\tilde{P}_i|\mathbf{X}, \mathbf{T}, \tilde{P}_0)|\mathbf{X}\right] \\ &\geq \frac{1}{2} \frac{n_i^2}{(c + n_i + 1)(c + n_i)^2} \iint d_k^2(x, y) d\hat{P}_i(x) d\hat{P}_i(y),\end{aligned}$$

where, crucially, we use that $\hat{P}_i = n_i^{-1} \sum_{j=1}^{n_i} \delta_{X_{i,j}}$ is deterministic conditionally on $\mathbf{X}$ and does not depend on $\mathbf{T}$.

From the covariance decomposition,

$$\mathbb{C}\text{ov}_k(\tilde{P}_1, \tilde{P}_2|\mathbf{X}) = \mathbb{C}\text{ov}_k(\mathbb{E}[\tilde{P}_1|\mathbf{X}, \mathbf{T}, \tilde{P}_0], \mathbb{E}[\tilde{P}_2|\mathbf{X}, \mathbf{T}, \tilde{P}_0]|\mathbf{X}).$$

Now, recalling that for $i = 1, 2$,

$$\mathbb{E}[\tilde{P}_i|\mathbf{X}, \mathbf{T}, \tilde{P}_0] = \frac{c}{c + n_i} \tilde{P}_0 + \frac{n_i}{c + n_i} \hat{P}_i, \quad (3.B.4)$$

and as $\hat{P}_i$ is deterministic conditionally on $\mathbf{X}$, we have by bilinearity

$$\mathbb{C}\text{ov}_k(\tilde{P}_1, \tilde{P}_2|\mathbf{X}) = \frac{c^2}{(c + n_1)(c + n_2)} \mathbb{V}\text{ar}_k(\tilde{P}_0|\mathbf{X}).$$

With $B > 0$ bound on the kernel, we have $\mathbb{V}\text{ar}_k(\tilde{P}_0|\mathbf{X}) \leq B$, thus

$$0 \leq \mathbb{C}\text{ov}_k(\tilde{P}_1, \tilde{P}_2|\mathbf{X}) \leq \frac{c^2 B}{(c + n_1)(c + n_2)}.$$

Putting everything together, we deduce,

$$0 \leq \mathbb{C}\text{orr}_k(\tilde{P}_1, \tilde{P}_2|\mathbf{X}) \leq 2c^2 B \prod_{i=1}^2 \left(\frac{\sqrt{c + n_i + 1}}{n_i} \left(\iint d_k^2(x, y) d\hat{P}_i(x) d\hat{P}_i(y) \right)^{-1/2} \right).$$

By the assumption of non-degeneracy, the last terms in the product are larger than a strictly positive constant in the limit. The conclusion follows.

Remark 6. As can be seen in the proof, the result relies only on the conditional distribution of $(\tilde{P}_1, \tilde{P}_2)$ given $\mathbf{X}^{(n_1, n_2)}$, $\mathbf{T}^{(n_1, n_2)}$ and $\tilde{P}_0$, and not on the marginal distribution of $\mathbf{T}^{(n_1, n_2)}$ given $\mathbf{X}^{(n_1, n_2)}$.

3.C Distributional Properties of the Hierarchical Dirichlet Process for Proofs and Simulations

The partially exchangeable model with an hDP prior, as in (3.5), is quasi-conjugate a posteriori, as shown in Teh et al. (2006) and Camerlenghi et al. (2019). To explain this structure, we introduce an augmented model with additional latent variables, the *tables* $T_{i,j}$ (Teh et al., 2006), using the formulation a priori in Catalano et al. (2023).

$$(X_{i,j}, T_{i,j}) \big| \tilde{P}_{1,\text{XT}}, \tilde{P}_{2,\text{XT}} \stackrel{\text{i.i.d.}}{\sim} \tilde{P}_{i,\text{XT}}, \quad \tilde{P}_{1,\text{XT}}, \tilde{P}_{2,\text{XT}} \big| \tilde{P}_0 \stackrel{\text{i.i.d.}}{\sim} \text{DP}(c, \tilde{P}_0 \times H), \quad \tilde{P}_0 \sim \text{DP}(c_0, P_0),$$

for $j \in \mathbb{N}$ and $i = 1, 2$, where H is an atomless probability measure. Clearly, the sequence $\{X_{i,j}\}_{i,j}$ generated by this augmented model has the same marginal law as the partially exchangeable model with hDP prior as in (3.5). Moreover, we observe that, marginally, $\{T_{1,j}\}_j, \{T_{2,j}\}_j$ are two independent exchangeable sequences directed by the same Dirichlet Process prior with concentration parameter c and baseline measure H .

From this augmented model, following Camerlenghi et al. (2018, 2019), we can specify the joint one-step-ahead predictive distribution for the observations and the tables, which can be used to sample from the model both a priori and a posteriori. Let us denote with $X_1^*, \dots, X_K^*$ the unique values in $\{X_{i,j}\}_{i,j}$. For $h \in \{1, \dots, K\}$ let ℓ_h be the number of

unique values in $\{T_{i,j} : X_{i,j} = X_h^*\}_{i,j}$ and $|\ell| := \ell_1 + \dots + \ell_K$. Then, for $i = 1, 2$,

$$\begin{aligned} \mathbb{P}(X_{i,n_i+1} \in A, T_{i,n_i+1} \in B | \mathbf{X}^{(n_1,n_2)}, \mathbf{T}^{(n_1,n_2)}) &= \\ &= \sum_{j=1}^{n_i} \frac{1}{c + n_i} \delta_{X_{i,j}}(A) \delta_{T_{i,j}}(B) + \frac{c}{c + n_i} \left(\frac{c_0}{c_0 + |\ell|} P_0(A) + \sum_{h=1}^K \frac{\ell_h}{c_0 + |\ell|} \delta_{X_h^*}(A) \right) H(B), \end{aligned} \quad (3.C.1)$$

where $\mathbf{X}^{(n_1,n_2)} = ((X_{1,j})_{j=1}^{n_1}, (X_{2,j})_{j=1}^{n_2})$ and $\mathbf{T}^{(n_1,n_2)} = ((T_{1,j})_{j=1}^{n_1}, (T_{2,j})_{j=1}^{n_2})$.

This formula is usually interpreted through a *restaurant franchise* metaphor ([Teh et al., 2006](#)): two restaurants of a franchise share the same menu, made of infinitely many dishes sampled from the atomless baseline measure P_0 . Each restaurant has potentially infinitely many tables and serves only one dish per table. The first customer enters one of the two restaurants, say restaurant i , and sits at a new table, whose label is randomly generated from H , and eats the unique dish served at that table. Each of the next customers entering restaurant i sits at the same table as one of the other n_i customers with probability $1/(c + n_i)$ and, thus, eats their same dish. Alternatively, they sit at a new table, randomly generated from H , with probability $c/(c + n_i)$. There, they choose one of the other $|\ell|$ tables across the franchise with probability $1/(c_0 + |\ell|)$ and eat the same dish that is being eaten at that table. Alternatively, they eat a new dish from the menu with probability $c_0/(c_0 + |\ell|)$. In this metaphor, $X_{i,j}$ represents the dish eaten by the j -th customer in the i -th restaurant, while $T_{i,j}$ is the label of the table where they sit.

Remark 7. The total number of tables is $|\ell| = \ell_1 + \dots + \ell_K$, and $K \leq |\ell| \leq n_1 + n_2$. It is K when there is a unique table for each dish. It is $n_1 + n_2$ when there is one table per customer.

Remark 8. The predictive distribution above forces the sequences $\{X_{i,j}\}_{i,j}$ and $\{T_{i,j}\}_{i,j}$ to have some compatibility properties, which illustrate their dependence. Firstly, $T_{1,j_1} \neq T_{2,j_2}$ a.s. for every $j_i = 1, \dots, n_i$ for $i = 1, 2$. Secondly, if $T_{i,j_1} = T_{i,j_2}$ a.s. for some $j_i \in \{1, \dots, n_i\}$, then $X_{i,j_1} = X_{i,j_2}$ a.s.. Both these conditions can be interpreted in light

of the restaurant franchise metaphor. The former states that a table cannot be shared across different restaurants. The latter says that if two customers are seated at the same table, they must eat the same dish.

For a posterior characterization of the process, we report the quasi-conjugacy result of [Camerlenghi et al. \(2019\)](#). Conditionally on the latent tables,

$$\begin{aligned} \tilde{P}_i | \mathbf{X}^{(n_1, n_2)}, \mathbf{T}^{(n_1, n_2)}, \tilde{P}_0 &\stackrel{\text{ind.}}{\sim} \text{DP} \left(c + n_i, \frac{c}{c + n_i} \tilde{P}_0 + \frac{n_i}{c + n_i} \hat{P}_i \right), \\ \tilde{P}_0 | \mathbf{X}^{(n_1, n_2)}, \mathbf{T}^{(n_1, n_2)} &\sim \text{DP} \left(c_0 + |\ell|, \frac{c_0}{c_0 + |\ell|} P_0 + \frac{|\ell|}{c_0 + |\ell|} \hat{P}_0 \right), \end{aligned} \quad (3.C.2)$$

where $\hat{P}_0 = |\ell|^{-1} \sum_{h=1}^K \ell_h \delta_{X_h^*}$ and $\hat{P}_i = n_i^{-1} \sum_{j=1}^{n_i} \delta_{X_{i,j}} = n_i^{-1} \sum_{h=1}^K n_{i,h} \delta_{X_h^*}$, independently for $i = 1, 2$. In particular, this tells us that conditionally on the tables, the posterior can be interpreted as an hDP with unequal marginals. It follows that we can reproduce the type of calculations in [Example 3.A.5](#) to find explicit expressions of $\text{Cov}_k(\tilde{P}_1, \tilde{P}_2 | \mathbf{X}^{(n_1, n_2)}, \mathbf{T}^{(n_1, n_2)}, \tilde{P}_0)$ and $\text{Var}_k(\tilde{P}_i | \mathbf{X}^{(n_1, n_2)}, \mathbf{T}^{(n_1, n_2)}, \tilde{P}_0)$. These are key to the proof of [Theorem 3.7.1](#) and can be found in [Lemma 3.B.2](#).

To conclude the posterior characterization, we need to provide the conditional law of $\mathbf{T}^{(n_1, n_2)}$ given $\mathbf{X}^{(n_1, n_2)}$. However, in practice, we only need the conditional law of $(\ell_1, \dots, \ell_K)$ given $\mathbf{X}^{(n_1, n_2)}$. We observe that if we define $\ell_{i,h}$ the number of unique values in $\{T_{i,j} : X_{i,j} = X_h^*\}_j$ for fixed $i = 1, 2$, since there is no intersection between the tables at different restaurants, $\ell_h := \ell_{1,h} + \ell_{2,h}$ for $h = 1, \dots, K$. If we denote as $n_{i,h}$ the number of observations equal to X_h^* in group i , then the joint law of $\{\ell_{i,h}\}_{i,h}$ conditionally on $\mathbf{X}^{(n_1, n_2)}$ is proportional to

$$\frac{c_0^k}{(c)_{n_1}(c)_{n_2}} \frac{c^{|\ell|}}{(c_0)^{|\ell|}} \prod_{h=1}^K (\ell_h - 1)! |\mathfrak{s}(n_{1,h}, \ell_{1,h})| |\mathfrak{s}(n_{2,h}, \ell_{2,h})| \mathbb{1}_{\{1, \dots, n_{1,h}\}}(\ell_{1,h}) \mathbb{1}_{\{1, \dots, n_{2,h}\}}(\ell_{2,h}), \quad (3.C.3)$$

where $|\mathfrak{s}(n, \ell)|$ are the signless Stirling number of the first kind and $(a)_q = \Gamma(a + q)/\Gamma(a)$ is the rising factorial. The proof follows from specializing the partially exchangeable

partition probability function (pEPPF) [Camerlenghi et al. \(2019\)](#) for a given configuration of $(\ell_1, \dots, \ell_K)$.

In principle, we could use the expression in [\(3.C.3\)](#) to compute law of the latent tables $\mathbf{T}^{(n_1, n_2)}$ given $\mathbf{X}^{(n_1, n_2)}$. However, in practice, it becomes rapidly prohibitive since we have an unnormalized distribution and the normalization step can be time-consuming due to the size of the support, which increases with the number of observations. The most popular workaround is to implement a Gibbs sampler ([Camerlenghi et al., 2019](#)). After fixing an initial allocation of the tables $\mathbf{T}^{(n_1, n_2)}$ that satisfies the compatibility properties mentioned in [Remark 8](#), for every $j = 1, \dots, n_i$ and $i = 1, 2$ we remove $T_{i,j}$ and sample another value for it from the following discrete distribution,

$$\mathbb{P}(T_{i,j} = T_{i,j^*} | \mathbf{X}^{(n_1, n_2)}, \mathbf{T}_{-(i,j)}) \propto q_{i,j^*}, \quad \mathbb{P}(T_{i,j} = T^* | \mathbf{X}^{(n_1, n_2)}, \mathbf{T}_{-(i,j)}) \propto \frac{c}{c_0 + |\ell|} \ell_h, \quad (3.C.4)$$

where $\mathbf{T}_{-(i,j)}$ is the set $\mathbf{T}^{(n_1, n_2)}$ without $T_{i,j}$, q_{i,j^*} is the frequency of the table T_{i,j^*} in $\mathbf{T}_{-(i,j)}$, h is such that $X_{i,j} = X_h^*$, ℓ_h is the number of unique values in $\mathbf{T}_{-(i,j)}$ associated to X_h^* , and $|\ell|$ is the number of unique values in $\mathbf{T}_{-(i,j)}$. Finally, $T^* \sim H$ is a new value for the table.

3.D Algorithms for Numerical Simulations

Once we have set all the distributional properties, we can estimate the kernel correlation a posteriori for an hDP model in two different ways: either a *sampling-based* algorithm or an *analytics-based* algorithm. We write $\mathbf{X}$ and $\mathbf{T}$ instead of $\mathbf{X}^{(n_1, n_2)}$ and $\mathbf{T}^{(n_1, n_2)}$ for compactness.

3.D.1 Sampling-Based Algorithm

The first method consists of using the estimator defined in [Section 3.4](#). This estimator only uses our ability to generate samples from the model a posteriori.

Given the sequence of observables $\mathbf{X}$, we can initialize the sequence $\mathbf{T}$ to be i.i.d. from H . Then, we use the joint one-step-ahead predictive distribution in (3.C.1) twice to generate a 2×2 sample for the observations and the tables. By discarding the future tables, we get a sample from $\mathcal{L}(X_{1,n_1+1}, X_{2,n_2+1}, X_{1,n_1+2}, X_{2,n_2+2} | \mathbf{X}, \mathbf{T})$. Once this routine is completed, we update $\mathbf{T}$ conditionally on $\mathbf{X}$ using the Gibbs sampler introduced above.

The sampling procedure is repeated M times to generate as many independent and identically distributed 2×2 samples $(X_{1,n_1+1}^{(t)}, X_{2,n_2+1}^{(t)}, X_{1,n_1+2}^{(t)}, X_{2,n_2+2}^{(t)})_{t=1}^M$ and compute the sampling-based estimator as in Proposition 3.4.3.

See Algorithm 1 for a thorough step-by-step description of the sampling-based method to compute the kernel correlation.

Algorithm 1. Sampling-Based Algorithm for Kernel Correlation

require: $\mathbf{X}$, P_0 , H , $c_0 \geq 0$, $c \geq 0$, $M \in \mathbb{N}$.

Initialize $\mathbf{T}$ as an i.i.d. sample from H .

for $t = 1, \dots, M$ **do**

Sample $(X_{i,n_i+1}^{(t)}, T_{i,n_i+1}^{(t)}), (X_{i,n_i+2}^{(t)}, T_{i,n_i+2}^{(t)})$ for $i = 1, 2$ according to (3.C.1).

Update $\mathbf{T}$ through the Gibbs updating scheme (3.C.4), conditionally on $\mathbf{X}$.

end

$\text{varX}_1 \leftarrow \sum_{t=1}^M k(X_{1,n_1+1}^{(t)}, X_{1,n_1+2}^{(t)}) / (M-1) - \sum_{t=1}^M \sum_{s=1}^M k(X_{1,n_1+1}^{(t)}, X_{1,n_1+2}^{(s)}) / (M(M-1))$

$\text{varX}_2 \leftarrow \sum_{t=1}^M k(X_{2,n_2+1}^{(t)}, X_{2,n_2+2}^{(t)}) / (M-1) - \sum_{t=1}^M \sum_{s=1}^M k(X_{2,n_2+1}^{(t)}, X_{2,n_2+2}^{(s)}) / (M(M-1))$

$\text{covX} \leftarrow \sum_{t=1}^M k(X_{1,n_1+1}^{(t)}, X_{2,n_2+1}^{(t)}) / (M-1) - \sum_{t=1}^M \sum_{s=1}^M k(X_{1,n_1+1}^{(t)}, X_{2,n_2+1}^{(s)}) / (M(M-1))$

return $\text{corrX} \leftarrow \text{covX} / \sqrt{\text{varX}_1 \text{varX}_2}$.

3.D.2 Analytics-Based Algorithm

The second method uses our knowledge of the distributional properties of the posterior.

Computation of the variances. If we apply the law of total variance, conditionally on $\mathbf{T}$ and $\tilde{P}_0$, using in particular (3.B.4) and the bilinearity of the covariance, we may write

for $i = 1, 2$,

$$\mathbb{V}\text{ar}_k(\tilde{P}_i|\mathbf{X}) = \mathbb{E}\left[\mathbb{V}\text{ar}_k(\tilde{P}_i|\mathbf{X}, \mathbf{T}, \tilde{P}_0)|\mathbf{X}\right] + \frac{c^2}{(c + n_i)^2} \mathbb{V}\text{ar}_k(\tilde{P}_0|\mathbf{X}). \quad (3.D.1)$$

For the first summand, we use the expression of $\mathbb{V}\text{ar}_k(\tilde{P}_i|\mathbf{X}, \mathbf{T}, \tilde{P}_0)$ in [Lemma 3.B.2](#), and obtain

$$\mathbb{E}\left[\mathbb{V}\text{ar}_k(\tilde{P}_i|\mathbf{X}, \mathbf{T}, \tilde{P}_0)|\mathbf{X}\right] = \mathbb{E}\left[V_{i,1}(\mathbf{X}, \mathbf{T}) + V_{i,2}(\mathbf{X}, \mathbf{T}) + V_{i,3}(\mathbf{X}, \mathbf{T})|\mathbf{X}\right], \quad (3.D.2)$$

with $V_{i,1}$, $V_{i,2}$ and $V_{i,3}$ defined below. Indeed, the first term in [\(3.D.2\)](#) is

$$V_{i,1}(\mathbf{X}, \mathbf{T}) := \frac{1}{2} \frac{1}{(c + n_i + 1)(c + n_i)^2} \mathbb{E}\left[\iint d_k^2(x, y) d\tilde{P}_0(x) d\tilde{P}_0(y) \middle| \mathbf{X}, \mathbf{T}\right].$$

Let us introduce

$$P_0^* := \mathbb{E}\left[\tilde{P}_0|\mathbf{X}, \mathbf{T}\right] = W_0(\mathbf{T})P_0 + \sum_{h=1}^K W_h(\mathbf{T})\delta_{X_h^*}, \quad (3.D.3)$$

for $W_0(\mathbf{T}) = c_0/(c_0 + |\ell|)$ and $W_h(\mathbf{T}) = \ell_h/(c_0 + |\ell|)$ for $h = 1, \dots, K$. By quasi-conjugacy in [\(3.C.2\)](#) and the computations a priori,

$$\begin{aligned} \mathbb{V}\text{ar}_k(\tilde{P}_0|\mathbf{X}, \mathbf{T}) &= \mathbb{E}\left[\iint k(x, y) d\tilde{P}_0(x) d\tilde{P}_0(y) \middle| \mathbf{X}, \mathbf{T}\right] - \iint k(x, y) dP_0^*(x) dP_0^*(y) \\ &= \frac{1}{2(c_0 + |\ell| + 1)} \iint d_k^2(x, y) dP_0^*(x) dP_0^*(y). \end{aligned}$$

Thus, we see that

$$V_{i,1}(\mathbf{X}, \mathbf{T}) = \frac{1}{2} \frac{1}{(c + n_i + 1)(c + n_i)^2} \left(1 - \frac{1}{c_0 + |\ell| + 1}\right) \iint d_k^2(x, y) dP_0^*(x) dP_0^*(y). \quad (3.D.4)$$

The second term in (3.D.2) is rewritten by the linearity of the expectation:

$$\begin{aligned} V_{i,2}(\mathbf{X}, \mathbf{T}) &:= \mathbb{E} \left[\frac{cn_i}{(c+n_i+1)(c+n_i)^2} \iint d_k^2(x, y) d\tilde{P}_0(x) d\hat{P}_i(y) \middle| \mathbf{X}, \mathbf{T} \right] \\ &= \frac{cn_i}{(c+n_i+1)(c+n_i)^2} \iint d_k^2(x, y) dP_0^*(x) d\hat{P}_i(y) \end{aligned} \quad (3.D.5)$$

with P_0^* as in (3.D.3). Lastly, the third term in (3.D.2) is

$$V_{i,3}(\mathbf{X}, \mathbf{T}) := \frac{1}{2} \frac{n_i^2}{(c+n_i+1)(c+n_i)^2} \iint d_k^2(x, y) d\hat{P}_i(x) d\hat{P}_i(y), \quad (3.D.6)$$

as the expectation is discarded, being the integrand completely determined by $\mathbf{X}$.

For the second summand in (3.D.1), we need $\mathbb{V}\text{ar}_k(\tilde{P}_0 | \mathbf{X})$. We start from the definition

$$\mathbb{V}\text{ar}_k(\tilde{P}_0 | \mathbf{X}) = \mathbb{E} \left[\iint k(x, y) d\tilde{P}_0(x) d\tilde{P}_0(y) \middle| \mathbf{X} \right] - \iint k(x, y) d\mathbb{E}[\tilde{P}_0 | \mathbf{X}](x) d\mathbb{E}[\tilde{P}_0 | \mathbf{X}](y)$$

For both expectations, we apply the tower law, conditioning to $\mathbf{T}$. That yields, with the functions $V_{0,1}$ and $V_{0,2}$ defined below,

$$\mathbb{V}\text{ar}_k(\tilde{P}_0 | \mathbf{X}) = \mathbb{E}[V_{0,1}(\mathbf{X}, \mathbf{T}) | \mathbf{X}] - V_{0,2}(\mathbf{X}). \quad (3.D.7)$$

Following the computations we did for $\mathbb{V}\text{ar}_{i,1}(\mathbf{X}, \mathbf{T})$, the function $V_{0,1}$ is defined and rewritten as

$$\begin{aligned} V_{0,1}(\mathbf{X}, \mathbf{T}) &:= \mathbb{E} \left[\iint k(x, y) d\tilde{P}_0(x) d\tilde{P}_0(y) \middle| \mathbf{X}, \mathbf{T} \right] \\ &= \iint \left(k(x, y) + \frac{1}{2(c_0 + |\ell| + 1)} d_k^2(x, y) \right) dP_0^*(x) dP_0^*(y). \end{aligned} \quad (3.D.8)$$

On the other hand

$$V_{0,2}(\mathbf{X}) := \iint k(x, y) d\mathbb{E}[P_0^* | \mathbf{X}](x) d\mathbb{E}[P_0^* | \mathbf{X}](y), \quad (3.D.9)$$

where, with the notations of (3.D.3),

$$\mathbb{E}[P_0^* | \mathbf{X}] = \mathbb{E}[W_0(\mathbf{T}) | \mathbf{X}] P_0 + \sum_{h=1}^K E[W_h(\mathbf{T}) | \mathbf{X}] \delta_{X_h^*}. \quad (3.D.10)$$

Computation of the covariance. For the covariance between $\tilde{P}_1$ and $\tilde{P}_2$, if we apply the law of total covariance, conditionally on $\mathbf{T}$ and $\tilde{P}_0$, and Lemma 3.B.2 we may write

$$\text{Cov}_k(\tilde{P}_1, \tilde{P}_2 | \mathbf{X}) = \frac{c^2}{(c + n_1)(c + n_2)} \mathbb{V}\text{ar}_k(\tilde{P}_0 | \mathbf{X}). \quad (3.D.11)$$

Note that we have used (3.B.4) again and the bilinearity of the covariance. The expression for $\mathbb{V}\text{ar}_k(\tilde{P}_0 | \mathbf{X})$ has already been expanded, see (3.D.7) above.

Running the Gibbs sampler. Given the sequence of observable $\mathbf{X}$, we can initialize the sequence $\mathbf{T}$ to be i.i.d. from H . Then, we can update $\mathbf{T}$ R times conditionally on $\mathbf{X}$ using the Gibbs sampler introduced above. Hence, we obtain a sequence of $(\mathbf{T}_1, \dots, \mathbf{T}_R)$. We approximate the expression in (3.D.2) as

$$\mathbb{E}[\mathbb{V}\text{ar}_k(\tilde{P}_i | \mathbf{X}, \mathbf{T}, \tilde{P}_0) | \mathbf{X}] \approx \frac{1}{R} \sum_{r=1}^R V_{i,1}(\mathbf{X}, \mathbf{T}_r) + V_{i,2}(\mathbf{X}, \mathbf{T}_r) + V_{i,3}(\mathbf{X}, \mathbf{T}_r),$$

using for that the expressions (3.D.4), (3.D.5), and (3.D.6). The first term appearing in the expression of $\mathbb{V}\text{ar}_k(\tilde{P}_0 | \mathbf{X})$ in (3.D.7) is approximated as

$$\mathbb{E}[V_{0,1}(\mathbf{X}, \mathbf{T}) | \mathbf{X}] \approx \frac{1}{R} \sum_{r=1}^R V_{0,1}(\mathbf{X}, \mathbf{T}_r),$$

using (3.D.8). Lastly, to approximate $V_{0,2}(\mathbf{X})$ we use (3.D.9) together with (3.D.10), where

in the last formula we estimate the expectation of the weights $W_0(\mathbf{T}), W_1(\mathbf{T}), \dots, W_K(\mathbf{T})$ conditionally on $\mathbf{X}$ as

$$\mathbb{E}[W_h(\mathbf{T})|\mathbf{X}] \approx \frac{1}{R} \sum_{r=1}^R W_h(\mathbf{T}_r)$$

for $h = 0, \dots, K$.

Using these expressions, we can approximate the expression for the covariance in (3.D.11) and the variances in (3.D.1) a posteriori to obtain the corresponding correlation. See Algorithm 2 for a detailed, step-by-step description of the analytics-based method for computing the kernel correlation.

Approximating integrals with respect to P_0 . Notice that we need to approximate integrals with respect to P_0 . To that end we approximate P_0 as $\sum_{t=1}^M \delta_{Z_t}/M$, where $Z_1, \dots, Z_M \stackrel{\text{i.i.d}}{\sim} P_0$. Hence, all the integrals above can be rewritten as finite sums since we are integrating with respect to discrete measures. Specifically, we apply the approximation

$$\begin{aligned} \mathcal{I}_{P_0} &:= \iint k(x, y) dP_0(x) dP_0(y) \approx \frac{1}{M^2} \sum_{t=1}^M \sum_{r=1}^M k(Z_t, Z_r), \\ \mathcal{I}_{P_0, h} &:= \iint k(x, y) dP_0(x) d\delta_{X_h^*}(y) \approx \frac{1}{M} \sum_{t=1}^M k(Z_t, X_h^*) \quad \text{for } h = 1, \dots, K. \end{aligned}$$

Algorithm 2. Analytics-Based Algorithm for Kernel Correlation

require: $\mathbf{X}$, P_0 , H , $c_0 \geq 0$, $c \geq 0$, $R \in \mathbb{N}$, $M \in \mathbb{N}$.

Compute $\mathbf{X}_1^*, \dots, \mathbf{X}_K^*$, $\mathbf{n}_{1,1}, \dots, \mathbf{n}_{1,K}$, and $\mathbf{n}_{2,1}, \dots, \mathbf{n}_{2,K}$.

// Precompute integrals with respect to P_0 .

Sample $\mathbf{Z}_1, \dots, \mathbf{Z}_M$ i.i.d. from P_0 .

$I_{P_0} \leftarrow \sum_{t=1}^M \sum_{s=1}^M k(\mathbf{Z}_t, \mathbf{Z}_s) / M^2$

for $h = 1, \dots, K$ **do**

$I_{P_0,h} \leftarrow \sum_{t=1}^M k(\mathbf{Z}_t, \mathbf{X}_h^*) / M$

end

// Run the Gibbs sampler.

Initialize $\mathbf{T}$ as an i.i.d. sample from H .

for $r = 1, \dots, R$ **do**

 Compute $\mathbf{l}_1, \dots, \mathbf{l}_K$.

 Compute $\mathbf{W}_0^{(r)}, \mathbf{W}_1^{(r)}, \dots, \mathbf{W}_K^{(r)}$.

$\text{VarXT}_1^{(r)} \leftarrow \mathbf{V}_{1,1}(\mathbf{X}, \mathbf{T}) + \mathbf{V}_{1,2}(\mathbf{X}, \mathbf{T}) + \mathbf{V}_{1,3}(\mathbf{X}, \mathbf{T}) + c^2 \mathbf{V}_{0,1}(\mathbf{X}, \mathbf{T}) / (c + \mathbf{n}_1)^2$

$\text{VarXT}_2^{(r)} \leftarrow \mathbf{V}_{2,1}(\mathbf{X}, \mathbf{T}) + \mathbf{V}_{2,2}(\mathbf{X}, \mathbf{T}) + \mathbf{V}_{2,3}(\mathbf{X}, \mathbf{T}) + c^2 \mathbf{V}_{0,1}(\mathbf{X}, \mathbf{T}) / (c + \mathbf{n}_2)^2$

$\text{CovXT}^{(r)} \leftarrow c^2 \mathbf{V}_{0,1}(\mathbf{X}, \mathbf{T}) / ((c + \mathbf{n}_1)(c + \mathbf{n}_2))$

 Update $\mathbf{T}$ through the Gibbs updating scheme (3.C.4), conditionally on $\mathbf{X}$.

end

// Evaluate $\mathbb{E}[P_0^* | \mathbf{X}]$ and $V_{0,2}(\mathbf{X})$.

Compute $\bar{\mathbf{W}}_0, \bar{\mathbf{W}}_1, \dots, \bar{\mathbf{W}}_K$.

$\mathbf{V}_{0,2} \leftarrow \bar{\mathbf{W}}_0^2 I_{P_0} + 2\bar{\mathbf{W}}_0 \sum_{h=1}^K \bar{\mathbf{W}}_h I_{P_0,h} + \sum_{h=1}^K \sum_{j=1}^K \bar{\mathbf{W}}_h \bar{\mathbf{W}}_j k(\mathbf{X}_h^*, \mathbf{X}_j^*)$

// Merge all computations together and return the correlation.

$\text{varX}_1 \leftarrow \overline{\text{VarXT}_1} - c^2 \mathbf{V}_{0,2} / (c + \mathbf{n}_1)^2$

$\text{varX}_2 \leftarrow \overline{\text{VarXT}_2} - c^2 \mathbf{V}_{0,2} / (c + \mathbf{n}_2)^2$

$\text{covX} \leftarrow \overline{\text{CovXT}} - c^2 \mathbf{V}_{0,2} / ((c + \mathbf{n}_1)(c + \mathbf{n}_2))$

return $\text{corrX} \leftarrow \text{covX} / \sqrt{\text{varX}_1 \text{varX}_2}$.

3.E Computations for Model Comparison

In [Section 3.9](#), we compare the Gaussian case in [Example 3.1.1](#) and the hDP model in [\(3.5\)](#) with $P_0 = \mathcal{N}(0, t^2)$ for a Gaussian kernel $k(x, y) = \exp(-(x - y)^2/(2\sigma^2))$ with parameter $\sigma > 0$.

To perform simulations for different values of the kernel correlation, we set the parameters so that the observables have the same marginal distributions. Since the marginal distribution of the observables is $\mathcal{N}(0, s^2 + \tau^2)$ for the Gaussian model, while it is $\mathcal{N}(0, t^2)$ for the hDP model, we need to set the constraint $t^2 = s^2 + \tau^2$.

In a similar line of reasoning, we set the kernel variances a priori to be equal to the same value $v > 0$ for each group for each case. To compute the kernel variances, we use the identity: if $\kappa(x, y) = a \exp(-(x - y)^2/(2b^2))$ while P is $\mathcal{N}(0, c^2)$, then

$$\int \kappa(x, x) dP(x) - \iint \kappa(x, y) dP(x) dP(y) = a \left(1 - \sqrt{\frac{b^2}{2c^2 + b^2}} \right). \quad (3.E.1)$$

We deduce the kernel variance for the parametric model using [Theorem 3.6.1](#) and the explicit expression of [Proposition 3.2.1](#): we apply [\(3.E.1\)](#) with $c^2 = \tau^2$; while, from [Example 3.6.1](#), we have $a^2 = \sqrt{\sigma^2/(2s^2 + \sigma^2)}$ and $b^2 = 2s^2 + \sigma^2$. We also deduce the kernel variance for the nonparametric model from [Corollary 3.5.3](#) and [Example 3.A.5](#): in the case we use [\(3.E.1\)](#) with $a = 1$, $b^2 = \sigma^2$ and $c^2 = t^2$. By imposing that both variances are equal to v , we obtain the constraint

$$v = \sqrt{\frac{\sigma^2}{2s^2 + \sigma^2}} - \sqrt{\frac{\sigma^2}{2\tau^2 + 2s^2 + \sigma^2}} = \frac{1 + c + c_0}{(1 + c)(1 + c_0)} \left(1 - \sqrt{\frac{\sigma^2}{2t^2 + \sigma^2}} \right).$$

Now, if we set the kernel correlation to be equal to a value $\xi \in [0, 1]$, we can determine the parameters for both the Gaussian case and the hDP case. In other words, once we fix v , ξ and t^2 we can determine s^2 , τ^2 , and ρ for the Gaussian case, and c_0 and c for the

hDP case.

For the Gaussian case, we have to solve the following system.

$$\begin{cases} t^2 &= \tau^2 + s^2, \\ v &= \sqrt{\frac{\sigma^2}{2s^2 + \sigma^2}} - \sqrt{\frac{\sigma^2}{2\tau^2 + 2s^2 + \sigma^2}}, \\ v\xi &= \sqrt{\frac{\sigma^2}{2\tau^2(1-\rho) + 2s^2 + \sigma^2}} - \sqrt{\frac{\sigma^2}{2\tau^2 + 2s^2 + \sigma^2}}, \end{cases}$$

which leads to

$$\begin{cases} s^2 &= \frac{\sigma^2}{2} \left(\left(v + \sqrt{\frac{\sigma^2}{2t^2 + \sigma^2}} \right)^{-2} - 1 \right) \\ \tau^2 &= t^2 - s^2 \\ \rho &= \frac{t^2 - \frac{\sigma^2}{2} \left(\left(v\xi + \sqrt{\frac{\sigma^2}{2t^2 + \sigma^2}} \right)^{-2} - 1 \right)}{t^2 - \frac{\sigma^2}{2} \left(\left(v + \sqrt{\frac{\sigma^2}{2t^2 + \sigma^2}} \right)^{-2} - 1 \right)}, \end{cases}$$

which is solvable with $s^2, \tau^2 > 0$ and $\rho \in [0, 1]$ for $v \in (0, 1)$ and $t^2 > \sigma^2/2(1/(1-v)^2 - 1)$.

For the hDP case, we have to solve the system

$$\begin{cases} v &= \frac{1+c+c_0}{(1+c)(1+c_0)} \left(1 - \sqrt{\frac{\sigma^2}{2t^2 + \sigma^2}} \right), \\ v\xi &= \frac{1}{1+c_0} \left(1 - \sqrt{\frac{\sigma^2}{2t^2 + \sigma^2}} \right), \end{cases}$$

which leads to

$$\begin{cases} c_0 &= \frac{1}{v\xi} \left(1 - \sqrt{\frac{\sigma^2}{2t^2 + \sigma^2}} \right) - 1, \\ c &= \frac{1}{1-\xi} \left(\frac{1}{v} \left(1 - \sqrt{\frac{\sigma^2}{2t^2 + \sigma^2}} \right) - 1 \right), \end{cases}$$

which, again, is solvable with $c_0, c > 0$ for $v \in (0, 1)$ and $t^2 > \sigma^2/2(1/(1-v)^2 - 1)$.

To complete the explanation of the model comparison in [Section 3.9](#), we need to characterize the one-step-ahead posterior predictive for both the Gaussian case in [Example 3.1.1](#) and the hDP model in [\(3.5\)](#) with $P_0 = \mathcal{N}(0, t^2)$. For the Gaussian case, we have from

Example 3.1.1 in Appendix 3.A that the posterior distribution of $\boldsymbol{\theta} = (\theta_1, \theta_2)$ given $\mathbf{X}^{(n_1, n_2)}$ is $\mathcal{N}(\boldsymbol{\theta}^*, \Sigma^*)$ with $\boldsymbol{\theta}^*$ and Σ^* are as in (3.A.1) and (3.A.2), respectively. Consequently, the one-step-ahead predictive distribution for the i -th group is $\mathcal{N}(\theta_i^*, s^2 + \Sigma_{i,i}^*)$ for $i = 1, 2$. For the hDP model, we generate a sample from the posterior predictive distribution using a Gibbs sampler similar to the one described in Appendix 3.C. For each sampled value, we update the allocation of the tables $\mathbf{T}^{(n_1, n_2)}$ conditionally on $\mathbf{X}^{(n_1, n_2)}$. Then, we generate a new data point from the posterior augmented model using the one-step-ahead predictive distribution in (3.C.1).

References

- Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 44(2):139–177.
- Alvarez-Melis, D. and Fusi, N. (2020). Geometric Dataset Distances via Optimal Transport. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21428–21439. Curran Associates, Inc.
- Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Birkhäuser Basel.
- Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons.
- Aronszajn, N. (1950). Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68(3):337–404.
- Ascolani, F., Franzolini, B., Lijoi, A., and Prünster, I. (2023). Nonparametric Priors with Full-Range Borrowing of Information. *Biometrika*, 111(3):945–969.
- Bach, F. R. and Jordan, M. I. (2002). Kernel Independent Component Analysis. *Journal of Machine Learning Research*, 3:1–48.
- Bellazzi, R., Codegoni, A., Gualandi, S., Nicora, G., and Vercesi, E. (2021). The Gene Mover’s Distance: Single-cell similarity via Optimal Transport. *arXiv:2102.01218*.

- Beraha, M., Guglielmi, A., and Quintana, F. A. (2021). The Semi-Hierarchical Dirichlet Process and Its Application to Clustering Homogeneous Distributions. *Bayesian Analysis*, 16(4):1187–1219.
- Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer.
- Bigot, J., Gouet, R., Klein, T., and López, A. (2017). Geodesic PCA in the Wasserstein Space by Convex PCA. *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, 53(1):1–26.
- Blackwell, D. and MacQueen, J. B. (1973). Ferguson Distributions Via Pólya Urn Schemes. *The Annals of Statistics*, 1(2):353–355.
- Blanchard, G., Lee, G., and Scott, C. (2011). Generalizing from Several Related Classification Tasks to a New Unlabeled Sample. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 24, pages 2178–2186. Curran Associates, Inc.
- Bochner, S. (1959). *Lectures on Fourier Integrals; with an Author’s Supplement on Monotonic Functions, Stieltjes Integrals, and Harmonic Analysis*. Princeton University Press.
- Bonet, C., Vauthier, C., and Korba, A. (2025). Flowing Datasets with Wasserstein over Wasserstein Gradient Flows. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J., editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 4911–4967. PMLR.
- Breiman, L. and Friedman, J. H. (1985). Estimating Optimal Transformations for Multiple Regression and Correlation. *Journal of the American Statistical Association*, 80(391):580–598.
- Buccianti, A., Mateu-Figueras, G., and Pawlowsky-Glahn, V. (2006). *Compositional Data Analysis in the Geosciences: From Theory to Practice*. Geological Society of London.
- Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. (2019). Distribution Theory for Hierarchical Processes. *The Annals of Statistics*, 47(1):67–92.

- Camerlenghi, F., Lijoi, A., and Prünster, I. (2018). Bayesian Nonparametric Inference Beyond the Gibbs-Type Framework. *Scandinavian Journal of Statistics*, 45(4):1062–1091.
- Carrière, M., Cuturi, M., and Oudot, S. (2017). Sliced Wasserstein Kernel for Persistence Diagrams. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 664–673. PMLR.
- Catalano, M., De Blasi, P., Lijoi, A., and Prünster, I. (2022). Posterior Asymptotics for Boosted Hierarchical Dirichlet Process Mixtures. *Journal of Machine Learning Research*, 23(80):1–23.
- Catalano, M., Del Sole, C., Lijoi, A., and Prünster, I. (2023). A Unified Approach to Hierarchical Random Measures. *Sankhya A*, 86:255–287.
- Catalano, M. and Lavenant, H. (2025). Measures of Dependence Based on Wasserstein distances. *arXiv:2510.06034*.
- Catalano, M., Lavenant, H., Lijoi, A., and Prünster, I. (2024). A Wasserstein Index of Dependence for Random Measures. *Journal of the American Statistical Association*, 119(547):2396–2406.
- Catalano, M., Lavenant, H., and Mascari, F. (2025). Measuring Partial Exchangeability with Reproducing Kernel Hilbert Spaces. *arXiv:2509.20221*.
- Catalano, M., Lijoi, A., and Prünster, I. (2021). Measuring Dependence in the Wasserstein Distance for Bayesian Nonparametric Models. *The Annals of Statistics*, 49(5):2916–2947.
- Chakraborty, S., Ghosh, M., and Mallick, B. K. (2012). Bayesian Nonlinear Regression for Large p Small n Problems. *Journal of Multivariate Analysis*, 108:28–40.
- Chatterjee, S. (2021). A New Coefficient of Correlation. *Journal of the American Statistical Association*, 116(536):2009–2022.
- Chen, W. Y., Barp, A., Briol, F. X., Gorham, J., Girolami, M., Mackey, L., and Oates, C. (2019). Stein Point Markov Chain Monte Carlo. In Chaudhuri, K. and Salakhutdinov, R.,

editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 1011–1021. PMLR.

Chewi, S., Niles-Weed, J., and Rigollet, P. (2025). Statistical Optimal Transport . In Morel, J. M. and Teissier, B., editors, *École d’Été de Probabilités de Saint-Flour XLIX – 2019*, pages 1–260. Springer.

Christmann, A. and Steinwart, I. (2010). Universal Kernels on Non-Standard Input Spaces. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 23, pages 406–414. Curran Associates, Inc.

Colombi, A., Argiento, R., Camerlenghi, F., and Paci, L. (2025). Hierarchical Mixture of Finite Mixtures. *Bayesian Analysis*, Advance Publication:1–29.

Cortes, C., Mohri, M., and Rostamizadeh, A. (2012). Algorithms for Learning Kernels Based on Centered Alignment. *Journal of Machine Learning Research*, 13(28):795–828.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley & Sons.

Cristianini, N., Shawe-Taylor, J., Elisseeff, A., and Kandola, J. (2001). On Kernel-Target Alignment. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems*, volume 14, pages 367–373. MIT Press.

Dall’Aglia, G., Kotz, S., and Salinetti, G. (1991). *Advances in Probability Distributions with Given Marginals*. Springer Dordrecht.

Darbellay, G. and Vajda, I. (1999). Estimation of the Information by an Adaptive Partitioning of the Observation Space. *IEEE Transactions on Information Theory*, 45(4):1315–1321.

de Finetti, B. (1937). La Prévision : ses Lois Logiques, ses Sources Subjectives. *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques*, 7(1):1–68.

de Finetti, B. (1938). Sur la Condition d’Equivalence Partielle. *Actualités Scientifiques et Industrielles*, 739:5–18.

- Delicado, P. (2011). Dimensionality Reduction when Data are Density Functions. *Computational Statistics & Data Analysis*, 55(1):401–420.
- Denti, F., Camerlenghi, F., Guindani, M., and Mira, A. (2023). A Common Atoms Model for the Bayesian Nonparametric Analysis of Nested Data. *Journal of the American Statistical Association*, 118(541):405–416.
- Deshmukh, A. A., Dogan, U., and Scott, C. (2017). Multi-Task Learning for Contextual Bandits. In Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 4849–4857. Curran Associates, Inc.
- Dette, H., Siburg, K. F., and Stoimenov, P. A. (2013). A Copula-Based Non-parametric Measure of Regression Dependence. *Scandinavian Journal of Statistics*, 40(1):21–41.
- Dubey, A. and Pentland, A. (2020). Kernel Methods for Cooperative Multi-Agent Contextual Bandits. In Daumé III, H. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 2740–2750. PMLR.
- Efron, B. and Morris, C. (1973). Stein’s Estimation Rule and Its Competitors – An Empirical Bayes Approach. *Journal of the American Statistical Association*, 68(341):117–130.
- Egozcue, J. J., Díaz-Barrero, J. L., and Pawlowsky-Glahn, V. (2006). Hilbert Space of Probability Density Functions Based on Aitchison Geometry. *Acta Mathematica Sinica. English Series*, 22(4):1175–1182.
- Escobar, M. D. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588.
- Escoufier, Y. (1973). Le Traitement des Variables Vectorielles. *Biometrics*, 29(4):751–760.
- Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230.

- Ferguson, T. S. (1983). Bayesian Density Estimation by Mixtures of Normal Distributions. In Rizvi, M. H., Rustagi, J. S., and Siegmund, D., editors, *Recent Advances in Statistics*, pages 287–302. Academic Press.
- Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*. Springer.
- Filzmoser, P., Hron, K., Martín-Fernández, J. A., and Palarea-Albaladejo, J. (2021). *Advances in Compositional Data Analysis: Festschrift in Honour of Vera Pawlowsky-Glahn*. Springer Cham.
- Franzolini, B., Lijoi, A., Prünster, I., and Rebaudo, G. (2025). Multivariate Species Sampling Models. *arXiv:2503.24004*.
- Friedman, J. H. and Rafsky, L. C. (1983). Graph-Theoretic Measures of Multivariate Association and Prediction. *The Annals of Statistics*, 11(2):377–391.
- Fréchet, M. (1951). Sur les Tableaux de Corrélation dont les Marges sont Données. *Annales de l'Université de Lyon, Section A*, 9:53–77.
- Fukumizu, K., Song, L., and Gretton, A. (2013). Kernel Bayes’ Rule: Bayesian Inference with Positive Definite Kernels. *Journal of Machine Learning Research*, 14(118):3753–3783.
- Gamboa, F., Klein, T., and Lagnoux, A. (2018). Sensitivity Analysis Based on Cramér–von Mises Distance. *SIAM–ASA Journal on Uncertainty Quantification*, 6(2):522–548.
- Gangbo, W. and McCann, R. J. (1996). The Geometry of Optimal Transportation. *Acta Mathematica*, 177(2):113–161.
- Gebelin, H. (1941). Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. *Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis. Second Edition*. Chapman & Hall.

- Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.
- Geuter, J., Bonet, C., Korba, A., and Alvarez-Melis, D. (2025). DDEQs: Distributional Deep Equilibrium Models Through Wasserstein Gradient Flows. *arXiv:2503.01140*.
- Ghodrati, L. and Panaretos, V. M. (2022). Distribution-on-Distribution Regression via Optimal Transport Maps. *Biometrika*, 109(4):957–974.
- Ghosal, R., Varma, V. R., Volfson, D., Hillel, I., Urbanek, J., Hausdorff, J. M., Watts, A., and Zipunnikov, V. (2023). Distributional Data Analysis via Quantile Functions and Its Application to Modeling Digital Biomarkers of Gait in Alzheimer’s Disease. *Biostatistics*, 24(3):539–561.
- Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., and Jenkinson, M. (2013). The Minimal Preprocessing Pipelines for the Human Connectome Project. *NeuroImage*, 80:105–124.
- Greenacre, M. (2018). *Compositional Data Analysis in Practice*. CRC Press.
- Greenacre, M. (2021). Compositional Data Analysis. *Annual Review of Statistics and Its Application*, 8(1):271–299.
- Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(25):723–773.
- Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005a). Measuring Statistical Dependence with Hilbert-Schmidt Norms. In Jain, S., Simon, H. U., and Tomita, E., editors, *Proceedings of the 16th International Conference on Algorithmic Learning Theory*, pages 63–77. Springer.
- Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. (2007). A Kernel Statistical Test of Independence. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*, volume 20, pages 984–991. Curran Associates, Inc.

- Gretton, A., Herbrich, R., Bousquet, O., Smola, A., and Schölkopf, B. (2005b). Kernel Methods for Measuring Independence. *Journal of Machine Learning Research*, 6(70):2075–2129.
- Griffin, J. E. and Leisen, F. (2017). Compound Random Measures and Their Use in Bayesian Non-Parametrics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):525–545.
- Guella, J. C. (2022). On Gaussian Kernels on Hilbert Spaces and Kernels on Hyperbolic Spaces. *Journal of Approximation Theory*, 279(105765).
- Haviv, D., Remšík, J., Gatie, M., Snopkowski, C., Takizawa, M., Pereira, N., Bashkin, J., Jovanovich, S., Nawy, T., Chaligne, R., Boire, A. K., Hadjantonakis, A. K., and Pe’er, D. (2025). The Covariance Environment Defines Cellular Niches for Spatial Inference. *Nature Biotechnology*, 43:269–280.
- Hein, M. and Bousquet, O. (2005). Hilbertian Metrics and Positive Definite Kernels on Probability Measures. In Cowell, R. G. and Ghahramani, Z., editors, *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, volume R5, pages 136–143. PMLR.
- Hirschfeld, H. O. (1935). A Connection Between Correlation and Contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31(4):520–524.
- Hoeffding, W. (1948). A Non-Parametric Test of Independence. *The Annals of Mathematical Statistics*, 19(4):546–557.
- Hoffmann-Jorgensen, J. and Pisier, G. (1976). The Law of Large Numbers and the Central Limit Theorem in Banach Spaces. *The Annals of Probability*, 4(4):587–599.
- Horiguchi, A., Chan, C., and Ma, L. (2024). A Tree Perspective on Stick-Breaking Models in Covariate-Dependent Mixtures. *Bayesian Analysis*, Advance Publication:1–28.
- Horst, A. M., Hill, A. P., and Gorman, K. B. (2020). *palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. R package version 0.1.0.

- Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer.
- Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika*, 28(3–4):321–377.
- Hron, K., Menafooglio, A., Templ, M., Hrušová, K., and Filzmoser, P. (2016). Simplicial Functional Principal Component Analysis. *Computational Statistics & Data Analysis*, 101:216–230.
- Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons.
- Ishwaran, H. and James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association*, 96(453):161–173.
- James, L. F., Lijoi, A., and Prünster, I. (2009). Posterior Analysis for Normalized Random Measures with Independent Increments. *Scandinavian Journal of Statistics*, 36(1):76–97.
- Josse, J. and Holmes, S. (2016). Measuring Multivariate Association and Beyond. *Statistics Surveys*, 10(none):132–167.
- Kallenberg, O. (2017). *Random Measures, Theory and Applications*. Springer.
- Kantorovich, L. V. (1941). On the Translocation of Masses. *Doklady Akademii Nauk SSSR*, 37(7–8):227–229.
- Kendall, M. G. (1938). A New Measure of Rank Correlation. *Biometrika*, 30(1–2).
- Kingman, J. F. C. (1967). Completely Random Measures. *Pacific Journal of Mathematics*, 21(1):59–78.
- Kneip, A. and Utikal, K. J. (2001). Inference for Density Families Using Functional Principal Component Analysis. *Journal of the American Statistical Association*, 96(454):519–542.
- Kokoszka, P., Miao, H., Petersen, A., and Shang, H. L. (2019). Forecasting of Density Functions with an Application to Cross-Sectional and Intraday Returns. *International Journal of Forecasting*, 35(4):1304–1317.

- Kokoszka, P. and Reimherr, M. (2017). *Introduction to Functional Data Analysis*. CRC Press.
- Kolouri, S., Zou, Y., and Rohde, G. K. (2016). Sliced Wasserstein Kernels for Probability Distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267. IEEE.
- Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of Neural Network Representations Revisited. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 3519–3529. PMLR.
- Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating Mutual Information. *Physical Review E*, 69(6):066138.
- Kratsios, A., Debarnot, V., and Dokmanić, I. (2023). Small Transformers Compute Universal Metric Embeddings. *Journal of Machine Learning Research*, 24(170):1–48.
- Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From Word Embeddings To Document Distances. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 957–966. PMLR.
- Kvålseth, T. O. (1987). Entropy and Correlation: Some Comments. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 17(3):517–519.
- Kvålseth, T. O. (2017). On Normalized Mutual Information: Measure Derivations and Properties. *Entropy*, 19(11):631.
- Legramanti, S., Durante, D., and Alquier, P. (2025). Concentration of Discrepancy-Based Approximate Bayesian Computation via Rademacher Complexity. *The Annals of Statistics*, 53(1):37–60.
- Leisen, F., Lijoi, A., and Spanò, D. (2013). A Vector of Dirichlet Processes. *Electronic Journal of Statistics*, 7:62–90.
- Leurgans, S. E., Moyeed, R. A., and Silverman, B. W. (1993). Canonical Correlation Analysis When the Data are Curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 55(3):725–740.

- Levina, E. and Bickel, P. (2001). The Earth Mover’s Distance is the Mallows Distance: some Insights from Statistics. In *Proceedings of the Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 251–256. IEEE.
- Lijoi, A. and Prünster, I. (2010). Models Beyond the Dirichlet Process. In Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G., editors, *Bayesian Nonparametrics*, pages 80–136. Cambridge University Press.
- Lijoi, A., Prünster, I., and Rebaudo, G. (2023). Flexible Clustering via Hidden Hierarchical Dirichlet Priors. *Scandinavian Journal of Statistics*, 50(1):213–234.
- Lindley, D. V. and Smith, A. F. M. (1972). Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 34(1):1–41.
- Liu, Q. and Wang, D. (2016). Stein Variational Gradient Descent: a General Purpose Bayesian Inference Algorithm. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29, pages 2378–2386. Curran Associates, Inc.
- Lo, A. Y. (1984). On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. *The Annals of Statistics*, 12(1):351–357.
- Lopez-Paz, D., Hennig, P., and Schölkopf, B. (2013). The Randomized Dependence Coefficient. In Burges, C. J., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 26, pages 1–9. Curran Associates, Inc.
- Lyons, R. (2013). Distance Covariance in Metric Spaces. *The Annals of Probability*, 41(5):3284 – 3305.
- MacLehose, R. F. and Dunson, D. B. (2009). Nonparametric Bayes Kernel-Based Priors for Functional Data Analysis. *Statistica Sinica*, 19(2):611–629.

- Maier, E. M., Stöcker, A., Fitzenberger, B., and Greven, S. (2025). Additive Density-on-Scalar Regression in Bayes Hilbert Spaces with an Application to Gender Economics. *The Annals of Applied Statistics*, 19(1):680–700.
- Marti, G., Andler, S., Nielsen, F., and Donnat, P. (2017). Exploring and Measuring Non-Linear Correlations: Copulas, Lightspeed Transportation and Clustering. In Anava, O., Khaleghi, A., Cuturi, M., Kuznetsov, V., and Rakhlin, A., editors, *Proceedings of the Time Series Workshop at NeurIPS 2016*, volume 55, pages 59–69. PMLR.
- Mascari, F., Lavenant, H., and Catalano, M. (2026+). Hilbert-Based Correlation Indices for Distributional Data. Working Paper.
- Matabuena, M., Petersen, A., Vidal, J. C., and Gude, F. (2021). Glucodensities: A New Representation of Glucose Profiles Using Distributional Data Analysis. *Statistical Methods in Medical Research*, 30(6):1445–1464.
- Menafoglio, A., Guadagni, A., and Secchi, P. (2014). A Kriging Approach Based on Aitchison Geometry for the Characterization of Particle-Size Curves in Heterogeneous Aquifers. *Stochastic Environmental Research and Risk Assessment*, 28(7):1835–1851.
- Meunier, D., Pontil, M., and Ciliberto, C. (2022). Distribution Regression with Sliced Wasserstein Kernels. In Chaudhuri, K., Jegelka, S., Le Song, C., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 15501–15523. PMLR.
- Monge, G. (1781). Mémoire sur la Théorie des Déblais et des Remblais. *Histoire de l’Académie Royale des Sciences de Paris*, 177:666–704.
- Moon, Y. I., Rajagopalan, B., and Lall, U. (1995). Estimation of Mutual Information Using Kernel Density Estimators. *Physical Review E*, 52(3):2318–2321.
- Muandet, K., Fukumizu, K., Dinuzzo, F., and Schölkopf, B. (2012). Learning from Distributions via Support Measure Machines. In Pereira, F., Burges, C. J., Bottou, L., and Weinberger, K.,

- editors, *Advances in Neural Information Processing Systems*, volume 25, pages 10–18. Curran Associates, Inc.
- Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017). Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends in Machine Learning*, 10(1–2):1–141.
- Móri, T. F. and Székely, G. J. (2020). The Earth Mover’s Correlation. *Annales Universitatis Scientiarum Budapestinensis de Rolando Eötvös Nominatae. Sectio Computatorica*, 50:249–268.
- Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer.
- Nerini, D. and Ghattas, B. (2007). Classifying Densities Using Functional Regression Trees: Applications in Oceanology. *Computational Statistics & Data Analysis*, 51(10):4994–5007.
- Nguyen, X. L. (2013). Convergence of Latent Mixing Measures in Finite and Infinite Mixture Models. *The Annals of Statistics*, 41(1):370–400.
- Nguyen, X. L. (2016). Borrowing Strength in Hierarchical Bayes: Posterior Concentration of the Dirichlet Base Measure. *Bernoulli*, 22(3):1535–1571.
- Nies, T. G., Staudt, T., and Munk, A. (2025). Transport Dependency: Optimal Transport Based Dependency Measures. *The Annals of Applied Probability*, 35(4):2292–2362.
- Nitzan, M., Karaïskos, N., Friedman, N., and Rajewsky, N. (2019). Gene Expression Cartography. *Nature*, 576:132–137.
- Orlova, D. Y., Zimmerman, N., Meehan, S., Meehan, C., Waters, J., Ghosn, E. E. B., Filatenkov, A., Kolyagin, G. A., Gernez, Y., Tsuda, S., Moore, W., Moss, R. B., Herzenberg, L. A., and Walther, G. (2016). Earth Mover’s Distance (EMD): A True Metric for Comparing Biomarker Expression Levels in Cell Populations. *PLOS ONE*, 11(3):1–14.
- Panaretos, V. M. and Zemel, Y. (2020). *An Invitation to Statistics in Wasserstein Space*. Springer.

- Paninski, L. (2003). Estimation of Entropy and Mutual Information. *Neural Computation*, 15(6):1191–1253.
- Park, J. Y. and Qian, J. (2012). Functional Regression of Continuous State Distributions. *Journal of Econometrics*, 167(2):397–412.
- Park, M., Jitkrittum, W., and Sejdinovic, D. (2016). K2-ABC: Approximate Bayesian Computation with Kernel Embeddings. In Gretton, A. and Robert, C. C., editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 398–407. PMLR.
- Pawlowsky-Glahn, V. and Buccianti, A. (2011). *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons.
- Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. John Wiley & Sons.
- Pearson, K. (1920). Notes on the History of Correlation. *Biometrika*, 13(1):25–45.
- Petersen, A. and Müller, H. G. (2016). Functional Data Analysis for Density Functions by Transformation to a Hilbert Space. *The Annals of Statistics*, 44(1):183–218.
- Petersen, A. and Müller, H. G. (2019). Wasserstein Covariance for Multiple Random Densities. *Biometrika*, 106(2):339–351.
- Petersen, A., Zhang, C., and Kokoszka, P. (2022). Modeling Probability Densities Functions as Data Objects. *Econometrics and Statistics*, 21:159–178.
- Peyré, G. and Cuturi, M. (2019). Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607.
- Pillai, N. S., Wu, Q., Liang, F., Mukherjee, S., and Wolpert, R. L. (2007). Characterizing the Function Space for Bayesian Kernel Models. *Journal of Machine Learning Research*, 8(62):1769–1797.

- Pitman, J. (1996). Some Developments of the Blackwell-MacQueen Urn Scheme. In Ferguson, T. S., Shapley, L. S., and MacQueen, J. B., editors, *Statistics, Probability and Game Theory: Papers in Honor of David Blackwell*, pages 245–268. Institute of Mathematical Statistics.
- Quintana, F. A., Müller, P., Jara, A., and MacEachern, S. N. (2022). The Dependent Dirichlet Process and Related Models. *Statistical Science*, 37(1):24–41.
- Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2012). Wasserstein Barycenter and Its Application to Texture Mixing. In Bruckstein, A. M., ter Haar Romeny, B. M., Bronstein, A. M., and Bronstein, M. M., editors, *Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer.
- Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer.
- Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.
- Regazzini, E., Lijoi, A., and Prünster, I. (2003). Distributional Results for Means of Normalized Random Measures with Independent Increments. *The Annals of Statistics*, 31(2):560–585.
- Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. (2011). Detecting Novel Associations in Large Data Sets. *Science*, 334(6062):1518–1524.
- Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The Nested Dirichlet Process. *Journal of the American Statistical Association*, 103(483):1131–1154.
- Rényi, A. (1959). On Measures of Dependence. *Acta Mathematica Academiae Scientiarum Hungaricae*, 10(3):441–451.
- Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians*. Birkhäuser.
- Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X. N., Holmes, A. J., Eickhoff, S. B., and Yeo, B. T. T. (2017). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114.

- Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., Lee, L., Chen, J., Brumbaugh, J., Rigollet, P., Hochedlinger, K., Jaenisch, R., Regev, A., and Lander, E. S. (2019). Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*, 176(4):928–943.
- Schweizer, B. and Wolff, E. F. (1981). On Nonparametric Measures of Dependence for Random Variables. *The Annals of Statistics*, 9(4):879–885.
- Schölkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press.
- Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell Labs Technical Journal*, 27(3):379–423.
- Sklar, A. (1959). Fonctions de Répartition à n Dimensions et Leurs Marges. *Publications de l'Institut Statistique de l'Université de Paris*, 8:229–231.
- Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D. A., Griffanti, L., Harms, M. P., Kelly, M., Laumann, T., Miller, K. L., Moeller, S., Petersen, S., Power, J., Salimi-Khorshidi, G., Snyder, A. Z., Vu, A. T., Woolrich, M. W., Xu, J., Yacoub, E., Uğurbil, K., Van Essen, D. C., and Glasser, M. F. (2013). Resting-State fMRI in the Human Connectome Project. *NeuroImage*, 80:144–168.
- Sollich, P. (2002). Bayesian Methods for Support Vector Machines: Evidence and Predictive Class Probabilities. *Machine Learning*, 46:21–52.
- Spearman, C. (1904). The Proof and Measurement of Association Between Two Things. *The American Journal of Psychology*, 15(1):72–101.
- Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. G. (2011). Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research*, 12(70):2389–2410.

- Srivastava, A., Jermyn, I., and Joshi, S. (2007). Riemannian Analysis of Probability Density Functions with Applications in Vision. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 1664–1671. IEEE.
- Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer.
- Strehl, A. and Ghosh, J. (2002). Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research*, 3:583–617.
- Szabó, Z., Sriperumbudur, B. K., Póczos, B., and Gretton, A. (2016). Learning Theory for Distribution Regression. *Journal of Machine Learning Research*, 17(152):1–40.
- Székely, G. J. and Rizzo, M. L. (2009). Brownian Distance Covariance. *The Annals of Applied Statistics*, 3(4):1236–1265.
- Székely, G. J., Rizzo, M. L., and Nail, K. B. (2007). Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics*, 35(6):2769–2794.
- Talská, R., Hron, K., and Grygar, T. M. (2021). Compositional Scalar-on-Function Regression with Application to Sediment Particle Size Distributions. *Mathematical Geosciences*, 53:1667–1695.
- Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581.
- Teicher, H. (1961). Identifiability of Mixtures. *The Annals of Mathematical Statistics*, 32(1):244–248.
- Tikhonov, A. N. (1963). On the Solution of Incorrectly Put Problems and the Regularisation Method. In *Outlines of the Joint Soviet-American Symposium on Partial Differential Equations*, pages 261–265. Siberian Branch of the Russian Academy of Sciences.
- Tipping, M. E. (2001). Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244.

- Tjøstheim, D., Håkon Otneim, H., and Støve, B. (2022). Statistical Dependence: Beyond Pearson's ρ . *Statistical Science*, 37(1):90–109.
- Tukey, J. W. (1972). Data Analysis, Computation and Mathematics. *Quarterly of Applied Mathematics*, 30(1):51–65.
- van den Boogart, K. G., Egozcue, J. J., and Pawlowsky-Glahn, V. (2010). Bayes Linear Space. *SORT: Statistics and Operations Research Transactions*, 34(1):201–222.
- van den Boogart, K. G., Egozcue, J. J., and Pawlowsky-Glahn, V. (2014). Bayes Hilbert Space. *Australian & New Zealand Journal of Statistics*, 56(2):171–194.
- van der Vaart, A. W. and Wellner, J. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer.
- Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., and Ugurbil, K. (2013). The WU-Minn Human Connectome Project: An Overview. *NeuroImage*, 80:62–79.
- Villani, C. (2008). *Optimal Transport: Old and New*. Springer.
- Wade, S. and Inácio, V. (2025). Bayesian Dependent Mixture Models: A Predictive Comparison and Survey. *Statistical Science*, 40(1):81–108.
- Wang, J. L., Chiou, J. M., and Müller, H. G. (2016). Functional Data Analysis. *Annual Review of Statistics and Its Application*, 3:257–295.
- Yanagimoto, T. (1970). On Measures of Association and a Related Problem. *Annals of the Institute of Statistical Mathematics*, 22:57–63.
- Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H., and Buckner, R. L. (2011). The Organization of the Human Cerebral Cortex Estimated by Intrinsic Functional Connectivity. *Journal of Neurophysiology*, 106(3):1125–1165.
- Zhang, Z. and Müller, H. G. (2011). Functional Density Synchronization. *Computational Statistics & Data Analysis*, 55:2234–2249.

Zhou, H., Lin, Z., and Yao, F. (2026). Intrinsic Wasserstein Correlation Analysis. *Statistica Sinica*, 36(2):1–22.

Ziegel, J., Ginsbourger, D., and Dümbgen, L. (2024). Characteristic Kernels on Hilbert Spaces, Banach Spaces, and on Sets of Measures. *Bernoulli*, 30(2):1441–1457.