Differentiability techniques in Statistics with applications

Universidad Autónoma de Madrid

Departamento de Matemáticas

Luis Alberto Rodríguez[2]

Supervisors: Javier Cárcamo[1] and Antonio Cuevas[2]

[1] Departamento de Matemáticas, University of the Basque Country,

[2] Departamento de Matemáticas, Universidad Autónoma de Madrid

December 14, 2023

Agradecimientos

Antes de comenzar con el cuerpo y las artes oscuras de este trabajo, me voy a permitir el lujo de dedicar unas palabras a aquellos que lo han hecho posible. En primer lugar los jefes: Javi y Antonio. Creo que ellos más que nadie me han aguantado y motivado a seguir en los momentos más delicados de este proceso de llegar a doctor. Especialmente en esa conversión, como lo llama Antonio, de matemático teórico a aplicado. Esas duras simulaciones que nunca terminaban de estar y que hacían perder el pelo a Javi. Gracias, bosses, desde el cariño por haberme guiado estos años.

Junto a ellos, la gente del grupo: Amparo, Bojan, Carmen, José Enrique, José Luis y José Ramón. De ellos he aprendido estadística tanto en el aula como fuera de ella. Especial mención a Carmen y José Luis por preocuparse como auténticos hermanos mayores. Y a Bojan, mi compañero de fatigas computacionales del que tanto `R` he aprendido (aunque él lo niegue). No me puedo olvidar en esta parte de Patxi, nuestro amigo vasco sin el que esta tesis se habría retrasado unos 6 meses más. Finalmente, gracias a Ana Justel, Julián de la Horra y Patricio por mostrarme otras formas de aprender y enseñar estadística; y a los antes profesores, ahora también compañeros, del Departamento.

En segundo lugar, a los amigos y familia. A mis padres y mi hermano por aguantarme fuera de la universidad. Todos sabemos que no soy fácil. A los de toda la vida. Empezando, como no, por el *boom*: Dede. A ti no hay que decirte nada porque ya te lo sabes, has estado ahí cada segundo. A Andrea Ezquerra, por nuestras quedadas de pascuas a ramos de cotilleos. A Ixchel, reciente redescubrimiento. A Juan y Arturo, por ser mis matemáticos puristas. Y a Yue y Ángela por ser mis compañeras de barrio. Gracias por acompañarme. Pero sobre todo a ti, Elena M. P., que te has convertido en un apoyo indispensable en mi vida y en mi trabajo.

Además, los amigos de doctorado. Los que ya se fueron: Adri, Álex, Diego, Julio e Ismael. La hermana mayor Bea. El alemán que parla català Florian. A Jan, mi mejor alumno de castellano y con quien más horas he pasado. Recuerda no llegar con retraso nunca, que eres alemán. Los dis-funcionales: Antonio y Paula por ser unos coordinadores maravillosos y mejores compañeros. Los que están por irse: Celia, David, Javi y Sergio. Y, por último, a los franceses: Alberto, Paola, Lukas, Fu-Hsuan y Clèment por acogerme como a un castellano más. Gracias, Alberto, por contar conmigo para *hacer la estadística grande de nuevo.*

Para terminar, a esos revisores y lectores anónimos que han citado y mejorado el trabajo de estos años.

Contents

Abstract

This thesis aims at exploring some advanced versions of the so-called "Delta method" (Chapter 2), with a particular focus on applications in non-parametric inference (Chapter 3), two-sample problems (Chapter 4) and clustering (Chapter 5). A few global conclusion and suggestions for future work are presented in Chapter 6. Let us now develop the main contributions in more detail.

Some context and historical perspective

The classical, elementary version of the Delta method (Rice (1995, p. 149)) is a standard topic in undergraduate courses of mathematical statistics. The basic idea is to use differentiability techniques to get first or second-order expansions of some statistic of interest in order to study its asymptotic behaviour; in particular, the classical Delta method is a common tool to prove asymptotic normality of many estimators.

As it turns out, in many relevant statistical problems, the parameter of interest can be expressed as a functional $\varphi(\mathrm{P})$ of the underlying distribution P. In such cases a natural estimator of $\varphi(\mathrm{P})$ is just $\varphi(\mathbb{P}_n)$, obtained by plugging-in the empirical distribution $\mathbb{P}_n$ on the functional φ. Such a simple idea was already pointed out by Fisher (1922). In this functional framework it is quite natural to obtain "functional local expansions" of φ around P using suitable, stochastic, versions of the classical concepts of functional differentiation (in Gateaux, Fréchet or Hadamard sense). Such idea was successfully exploited by von Mises (1947), in a classical paper, later developed in his posthumous book von Mises (1964). Further earlier developments and applications of these ideas are due to Kallianpur and Rao (1955) and Filippova (1962). See also the book by Serfling (2009).

In the 1980's this methodology based on differentiation of functionals experienced an additional boost, due to its applications to robust statistics and bootstrap theory. In the robustness field, the "directional" (Gateaux) derivative of a statistical functional is interpreted in terms of the *influence curve*, a measure of the local sensitivity of the corresponding estimator against outlying observations. Also, the integral of the square influence function coincides with the asymptotic variance under some quite general conditions. An early, nice account of these ideas can be found in the book by Huber (2004); see also Hampel et al. (2011). In the bootstrap field, the functional differentiation techniques provides a simple, elegant method to derive the asymptotic validity of the bootstrap approximations to the sampling distribution; see, e.g., Parr (1985) for an early example.

The main research lines, contents and contributions of this PhD thesis

As shown in the previous paragraphs, the differentiation method has a wealth of fruitful statistical applications. Whatever the application we have in mind, an obvious strategy is to look for weaker differentiability notions, valid under broader conditions but still keeping some essentials, such as first-order local approximations and chain rule. One of these extensions was the Hadamard directional differentiability, introduced by Shapiro (1990). This notion is defined and commented in **Chapter 1**, which is devoted to summarize some important auxiliary tools we use throughout the work; these include as well some essentials of Empirical Processes Theory and Reproducing Kernel Hilbert Spaces (RKHS).

In **Chapter 2** of this thesis we show that a wide class of relevant statistical functionals, defined in terms of a supremum, satisfy Shapiro's weaker notion of differentiability, for which the Delta method is still applicable. This result is remarkable, as supremum-type transformations lead typically to a loss of (most notions of) differentiability.

Some applications of this result are developed in **Chapter 3**. In particular, we obtain the asymptotic distribution of the two-sample Kolmogorov-Smirnov statistic, a key tool in nonparametric statistics, under the alternative hypothesis. Our result improves on a classical theorem proved by Raghavachari (1973), as we are able to drop the assumption of continuity imposed on the involved distributions, a condition that has been echoed in subsequent works in the field. The proof lies on the differentiability of the supremum functional in the Skorohod space (see Neuhaus (1971) and Seijo and Sen (2011)).

Three additional applications are also included in Chapter 3. First, the asymptotic behavior of the Berk-Jones statistic is obtained under the alternative hypothesis of unequal distributions, thus solving an open problem proposed in Jager and Wellner (2004). Second, the asymptotic distribution of the statistic for the goodness-of-fit test based on the supremum metric for copulas is also derived. In particular, if the empirical copula process is used for estimation, we provide a further extension of the results in Fermanian (2013). Third, and perhaps, more important, we derive the asymptotic distribution of the so-called *maximum mean discrepancy* (MMD), an increasingly popular method to measure discrepancies between distributions which includes, as a particular case, the kernel metrics based on the distance between the embeddings of distributions in a suitable Reproducing Kernel Hilbert Space (RKHS).

In **Chapter 4** a test for the classical two-sample problem is proposed (i.e., testing the equality for two distributions based on independent samples of them). This new proposal for testing homogeneity is based on kernel distances, a particular type of MMD. The theoretical and practical aspects of this test are analyzed with a special focus on high-dimensional and functional two-sample problems. In particular, an empirical study is included to compare the proposed test with other popular alternatives.

Chapter 5 is an application of the main mentioned differentiability result, obtained in Chapter 2, to the problem of uniqueness of the k-means set (principal points of a distribution). More specifically, it is proved that the uniqueness of the k-means minimizing set (the set of k-means or principal points) is equivalent to the asymptotic normality of the empirical risk statistic used to calculate the sample k-means sets. Also, a consistency result, adapted to the case of multiple k-means is established in terms of the Gromov-Hausdorff metric. While the k-means method is arguably the most popular clustering methodology, the non-uniqueness of the population (theoretical) k-means is a somewhat enigmatic problem. No simple condition for such uniqueness is available so far in the literature though the problem has some relevance in order to ensure the stability of the k-means algorithms (Caponnetto and Rakhlin (2006)) and the validity of the asymptotic results on which this method relies; see Cuesta and Matrán (1988), Pollard (1981), and Pollard (1982). As a consequence of the mentioned asymptotic characterization of the k-means uniqueness, a statistical test is proposed to check the null hypothesis that the underlying distribution has a unique set of k-means. An empirical study is also included.

Some publications derived from this thesis

The contents of Chapters 2 and 3 are essentially included in Cárcamo et al. (2020). This paper was positively reviewed in the mathematical data base MathScinet, [Review MR4091104] where the reviewer points out

> *The proposed methodology is interesting and could find a lot of applications since the supremum norm is very common in statistics to quantify the deviation between an observed phenomenon and a theoretical model (e.g. the famous Kolmogorov-Smirnov test). The whole paper is well written.*

At the time of writing these lines, this paper has achieved 27 citations in Google Scholar and 6 citations in the Web of Science database.

The contents of Chapter 4 are included in a paper (under second revision) submitted to the *Journal of Multivariate Analysis.*

Finally, the materials of **Chapter 5** will be included in a manuscript in preparation.

Resumen

En esta tesis doctoral se han explorado diferentes aplicaciones del conocido “Método delta” (Capítulo 2). En concreto, se han investigado aplicaciones a inferencia no-paramétrica (Capítulo 3), a los problemas de dos muestras u homogeneidad (Capítulo 4) y a la metodología de k-medias (Capítulo 5). Finalmente, se presentan en el Capítulo 7 las conclusiones del trabajo realizado. A continuación, se desarrolla en más detalle la temática de cada capítulo.

Contexto previo y perspectiva histórica

La versión clásica del Método delta (Rice (1995, p.149)) es contenido de los cursos de estadística del grado de Matemáticas. La idea fundamental es la utilización de un desarrollo de Taylor de primer o segundo orden para determinar el comportamiento asintótico del estadístico de interés. En particular, se utiliza para demostrar la normalidad asintótica de un estimador.

Es sabido que en una amplia cantidad de problemas relevantes en Estadística, el parámetro a estimar se puede expresar como un funcional $\varphi(\mathrm{P})$ de la medida subyacente P. En esos casos, un estimador natural de $\varphi(\mathrm{P})$ es $\varphi\left(\mathbb{P}_n\right)$, obtenido mediante la metodología *plug-in*, que consiste en sustituir P por $\mathbb{P}_n$, la medida de probabilidad empírica, en el argumento del funcional φ. Esta idea fue apuntada ya en la década de los años 20 del siglo pasado en Fisher (1922). En este marco funcional, pues, es natural tratar de proceder mediante el desarrollo de Taylor de φ en torno a P utilizando la noción apropiada de diferenciabilidad (Gâteaux, Frèchet o Hadamard). Estas ideas ya fueron introducidas en von Mises (1947), posteriormente recogidas en su obra póstuma von Mises (1964). Como desarrollos posteriores de esta idea podríamos mencionar Kallianpur and Rao (1955), Filippova (1962) o el libro Serfling (2009).

En la década de los 80, esta metodología se vio impulsada, en gran medida, por sus aplicaciones a la estadística robusta y la teoría sobre el bootstrap. Respecto a los avances en robustez, la derivada de Gâteaux de un estadístico se puede interpretar en términos de la *curva de influencia*, una medida local de la sensiblidad del estimador ante la presencia de atípicos. Además, la integral del cuadrado de la función de influencia resulta ser la varianza de la distribución asintótica del estimador bajo condiciones razonablemente generales. Una recopilación de estos resultados puede encontrarse en Huber (2004) o Hampel et al. (2011). Respecto al estudio del bootstrap, la diferenciación de funcionales provee de una maquinaria simple y elegante para demostrar la validez de las aproximaciones

bootstrap. Uno de los primeros ejemplos puede encontrarse en Parr (1985).

Temática principal, contenidos y contribuciones de esta tesis doctoral

Como se comentaba en los párrafos anteriores, la diferenciabilidad es de gran interés en estadística. Para cualquier aplicación que tengamos en mente que encaje en el paradigma anterior, una aproximación razonable es verificar las condiciones de diferenciablidad (débil), aplicable bajo condiciones lo más generales posibles; manteniendo las propiedades fundamentales de la aproximación de primer orden, como la regla de la cadena. Una noción que encaja en esta descripción es la diferenciabilidad Hadamard direccional, introducida en Shapiro (1990). Este concepto es desarrollado en el **Capítulo 1**. Además, se tratan otras herramientas matemáticas necesarias para la elaboración de esta tesis, como la teoría de procesos empíricos y los espacios de Hilbert de núcleo reproductor (RKHS).

En el **Capítulo 2** se prueba que una clase muy amplia de funcionales, expresados en términos de supremos, satisface la noción de diferenciabilidad débil introducida por Shapiro, bajo la cual el Método delta es aplicable. Esto es destacable en tanto a que este tipo de funcionales presentan, habitualmente, falta de suavidad.

Algunas aplicaciones de este resultado se presentan en el **Capítulo 3**. En particular, se obtiene la distribución asintótica del estadístico de Kolmogorov-Smirnov para los problemas de dos muestras, un herramienta básica de la estadística no paramétrica, bajo la hipótesis alternativa. Nuestro resultado mejora sustancialmente el trabajo de Raghavachari (1973) pues se eliminan restricciones de continuidad que han sido replicadas en la literatura desde entonces. La demostración se apoya en el cálculo de la derivada de Hadamard direccional para el supremo en el espacio de Skorohod (Neuhaus (1971) y Seijo and Sen (2011)).

Tres aplicaciones adicionales de los resultados de diferenciabilidad del Capítulo 2 se incluyen en el Capítulo 3. En primer lugar se calcula la distribución asintótica de estadísticos de tipo Berk-Jones bajo la hipótesis alternativa, es decir, cuando las distribuciones son diferentes. De esta manera, se resuelve una pregunta abierta en Jager and Wellner (2004). En segundo lugar, también se calcula la distribución asintótica para algunos problemas de bondad de ajuste basados en la norma del supremo. En particular este resultado supone una extensión de los resultados presentados en Fermanian (2013) sobre el estimador cópula empírico. Finalmente, quizá el más importante de este capítulo, es el resultado asintótico sobre las distancias de discrepancia máxima (*Maximum Mean Discrepancy (MMD)* en inglés). Este método, cada vez más popular para cuantificar diferencias entre distribuciones de probabilidad, se basa en métricas de probabilidad integrales. Entre otros casos, se incluyen las métricas de tipo núcleo (*kernel metrics*), en las que la discrepancia se mide sobre la bola unidad de un espacio de Hilbert de núcleo reproductor (RKHS).

El **Capítulo 4** se centra en una propuesta de un test para problemas de dos muestras, es decir, un contraste de hipótesis en el que la hipótesis nula es la igualdad en distribución de dos poblaciones dadas dos muestras independientes. Dicho test se fundamenta en el uso de las distancias de tipo kernel, un caso particular especial de

MMD. Los aspectos teóricos, así como las aplicaciones, son desarrollados a lo largo del capítulo, centrándonos especialmente en datos de alta dimensión y funcionales. Además, se presenta un estudio de simulación para comparar esta nueva propuesta con las ya existentes en la literatura.

Finalmente, en el **Capítulo 5** se aplica el teorema de diferenciabilidad del Capítulo 2 al problema sobre la unicidad del conjunto de k-medias. Concretamente, se demuestra que la unicidad del conjunto de k-medias es equivalente a la normalidad asintótica del estimador plug-in del riesgo empírico, es decir, de la suma de cuadrados dentro de los grupos promediada. Además, se proporciona un resultado de consistencia, adaptado al caso de no unicidad del conjunto de k-medias, en términos de la distancia de Gromov-Hausdorff. Es destacable que, aunque k-medias es uno de los métodos de clustering más utilizados, la no unicidad de las k-medias poblacionales (de la medida subyacente) es una cuestión aún por explorar. No se conocía hasta ahora una condición equivalente, ni suficiente; tratable. No obstante, la unicidad supone una cierta garantía de estabilidad de los algoritmos de aproximación (Caponnetto and Rakhlin (2006)) y la validez de los resultados asintóticos, como en Cuesta and Matrán (1988), Pollard (1981) y Pollard (1982). Como aplicación de esa caracterización sobre la unicidad de las k-medias, se propone también un test para contrastar la hipótesis nula de unicidad de dicho conjunto para la medida de probabilidad subyacente. Se incluye un estudio empírico al respecto.

Publicaciones derivadas de esta tesis doctoral

Los Capítulos 2 y 3 componen el artículo Cárcamo et al. (2020). Este trabajo ha obtenido una valoración positiva en MathScinet: [Review MR4091104].

En el momento de escribir este resumen de la tesis, este artículo ha sido citado 27 veces según Google Scholar y 6 veces según la base de datos de Web of Science.

El contenido correspondiente al Capítulo 4 forma parte de un trabajo bajo segunda revisión en el *Journal of Multivariate Analysis*.

Finalmente, los contenidos del **Capítulo 5** formarán parte de un trabajo en estado de elaboración muy avanzado.

Chapter 1

Introduction and preliminaries

In this Chapter, we provide a brief introduction to Hadamard directional differentiability and the Delta method (Section 1.1). Additionally, we offer a summary of the basics of empirical processes and plug-in estimation (Section 1.2).

The theory of Reproducing Kernel Hilbert Spaces (RKHS) is another important auxiliary tool in this PhD thesis. Such theory is briefly reviewed (Section 1.3). The results presented in this Section regarding the integrability of elements of the RKHS and the mean embedding are extensively employed in Chapter 4. While most of this background is well-known or can be found in the literature, it is included here to introduce the necessary notation and make this thesis as self-contained as possible.

1.1 Hadamard directional differentiability and Delta method

In many situations, it is common to face the problem of estimating a transformation, $\varphi(\theta)$, of a (possibly infinite-dimensional) parameter θ. Typically, θ is unknown but can be estimated by means of T_n and φ is a map defined in a metric space. If φ is smooth enough in a local neighborhood of θ –for instance, differentiable at θ in a precise sense– the asymptotic distribution of (the normalized version) of $\varphi(T_n)$ can be determined by expanding φ around θ and using an invariance principle for T_n in the underlying metric space. Of course, this is the key idea behind the *(functional) Delta method.* At this point, several notions of differentiabilit arise. We start with the notion of Gâteaux directional differentiability.

Definition 1. Let $\mathcal{D}$ and $\mathcal{E}$ be real Banach spaces with norms $\|\cdot\|_{\mathcal{D}}$ and $\|\cdot\|_{\mathcal{E}}$, respectively. A map $\varphi : \mathcal{D} \longrightarrow \mathcal{E}$ is said to be *Gâteaux directionally differentiable* at $\theta \in \mathcal{D}$ tangentially to a set $\mathcal{D}_0 \subset \mathcal{D}$ if there exists a map $\varphi'_\theta : \mathcal{D}_0 \longrightarrow \mathcal{E}$ such that

$$\left\| \frac{\varphi(\theta + t_n h) - \varphi(\theta)}{t_n} - \varphi'_\theta(h) \right\|_{\mathcal{E}} \longrightarrow 0, \tag{1.1}$$

for all $h \in \mathcal{D}_0$ and all sequences $(t_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ such that $t_n \searrow 0$.

It is well-known that Gâteaux differentiability is too weak for the Delta method to hold Huber (2011, Section 2.5). To solve this problem, the directions along which we approach to $\varphi(\theta)$ in (1.1) have to be allowed to change with n. This naturally leads to the concept of Hadamard directional differentiability. Shapiro (1990) has been followed for the next definition.

Definition 2. In the context of the previous definition, we say that $\varphi : \mathcal{D} \longrightarrow \mathcal{E}$ is *Hadamard directionally differentiable* at $\theta \in \mathcal{D}$ tangentially to a set $\mathcal{D}_0 \subset \mathcal{D}$ if there exists a map $\varphi'_\theta : \mathcal{D}_0 \longrightarrow \mathcal{E}$ such that

$$\left\| \frac{\varphi\left(\theta + t_n h_n\right) - \varphi(\theta)}{t_n} - \varphi'_\theta(h) \right\|_{\mathcal{E}} \longrightarrow 0, \tag{1.2}$$

for all $h \in \mathcal{D}_0$ and all sequences $(h_n)_{n\in\mathbb{N}} \in \mathcal{D}^{\mathbb{N}}$, $(t_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$ such that $t_n \searrow 0$ and $\|h_n - h\|_{\mathcal{D}} \to 0$.

Obviously, Hadamard directional differentiability implies the Gâteaux one. The only difference between the directional and the usual differentiability is that the derivative φ'_θ is no longer required to be linear in Definitions 1 and 2. Nevertheless, if equation (1.2) is satisfied, then φ'_θ is continuous and homogeneous of degree 1 Shapiro (1990, Proposition 3.1).

Remark 3. If φ is as in the preliminaries of Definitions 1 and 2, and additionally φ is locally Lipschitz, i.e., there exists a constant $C > 0$ such that $\|\varphi(f) - \varphi(g)\|_{\mathcal{E}} \leq C\|f - g\|_{\mathcal{D}}$, for all $f, g \in \mathcal{D}$ in a neighborhood of each point of $\mathcal{D}$, then Hadamard directional differentiability is equivalent to the Gâteaux one (see Shapiro (1990, Proposition 3.5)).

Two useful properties of the Hadamard differentiability are established in the following result for posterior use (see (2.1) and Remark 19). The proofs can be found in Shapiro (1990).

Theorem 4. *Let $\mathcal{D}$, $\mathcal{E}$ and $\mathcal{G}$ be real Banach spaces, $\varphi : \mathcal{D} \longrightarrow \mathcal{E}$ and $\psi : \mathcal{D} \longrightarrow \mathcal{E}$ Hadamard directional differentiable at θ tangentially to a set $\mathcal{D}_0 \subseteq \mathcal{D}$ and $\kappa : \mathcal{E} \longrightarrow \mathcal{G}$ Hadamard directional differentiable at $\varphi(\theta)$ tangentially to $\mathcal{E}_0 \subseteq \mathcal{E}$. Additionally assume that $\varphi'_\theta(\mathcal{D}_0) \subseteq \mathcal{E}_0$. Then:*

1. *$\varphi + \psi$ is Hadamard directional differentiable tangentially at θ to $\mathcal{D}_0$ and $\varphi'_\theta + \psi'_\theta$.*

2. *$\kappa \circ \varphi$ is Hadamard directional differentiable at θ tangentially to $\mathcal{D}_0$ and $(\kappa \circ \varphi)'_\theta = \kappa'_{\varphi_\theta} \circ \varphi'_\theta$.*

Furthermore, it is worth mentioning that there exist implicit and inverse function theorems for Hadamard directional differentiability. These are particularly relevant when dealing with M and Z-estimators (see A. van der Vaart and Wellner (1996, Sections 3.2 and 3.3)). For a discussion about these properties, refer to Fernholz (1983, Chapter 3).

Finally, the important fact about Hadamard directional differentiability is that it allows the application of the *extended (functional) Delta method*.

Proposition 5 (Delta method)**.** *Let $\mathcal{D}$ and $\mathcal{E}$ be Banach spaces and $\varphi : \mathcal{D}_\phi \subset \mathcal{D} \longrightarrow \mathcal{E}$, where $\mathcal{D}_\varphi$ is the domain of φ. Assume that φ is Hadamard directionally differentiable at $\theta \in \mathcal{D}_\varphi$ tangentially to a set $\mathcal{D}_0 \subset \mathcal{D}$. For some sample spaces Ω_n, let $T_n : \Omega_n \longrightarrow \mathcal{D}_\varphi$ be maps such that $r_n\left(T_n - \theta\right)$ converges weakly to T $r_n\left(T_n - \theta\right) \rightsquigarrow T$, for some sequence of numbers $r_n \longrightarrow \infty$ and a random element T that takes values in $\mathcal{D}_0$. Then, $r_n\left(\varphi\left(T_n\right) - \varphi(\theta)\right) \rightsquigarrow \varphi'_\theta(T)$. If additionally φ'_θ can be continuously extended to $\mathcal{D}$, then we have that $r_n\left(\varphi\left(T_n\right) - \varphi(\theta)\right) = \varphi'_\theta\left(r_n\left(T_n - \theta\right)\right) + o_{\mathrm{P}}(1)$.*

Remark 6. The detailed proof of Proposition 5 can be found in Shapiro (1991, Theorem 2.1) (see also Römisch (2004, Theorem 1) or Fang and Santos (2019, Theorem 2.1)), but it is essentially the same one as for the traditional Delta method A. W. van der Vaart (2000, Theorem 20.8). The key idea is to apply the *extended Continuous Mapping Theorem* A. van der Vaart and Wellner (1996, Theorem 1.11.1) to the sequence of functionals defined by $\varphi_n(h) = r_n\left(\varphi\left(\theta + r_n^{-1} h\right) - \varphi(\theta)\right)$, $n \in \mathbb{N}$.

In the present context, let us assume that $\theta_n \to \theta$ and $r_n\left(T_n - \theta_n\right) \rightsquigarrow T$, and we want to determine conditions so that $r_n\left(\varphi\left(T_n\right) - \varphi\left(\theta_n\right)\right) \rightsquigarrow \varphi'_\theta(T)$. As it is pointed out in A. van der Vaart and Wellner (1996, p. 375), a stronger form of differentiability is needed to obtain such a “uniform” version of the Delta method.

Definition 7. In the context of Definition 1, we say that $\varphi : \mathcal{D} \longrightarrow \mathcal{E}$ is *uniformly Hadamard differentiable* at $\theta \in \mathcal{D}$ tangentially to a set $\mathcal{D}_0 \subset \mathcal{D}$ if there exists a map $\varphi'_\theta : \mathcal{D}_0 \longrightarrow \mathcal{E}$ such that

$$\left\| \frac{\varphi\left(\theta_n + t_n h_n\right) - \varphi\left(\theta_n\right)}{t_n} - \varphi'_\theta(h) \right\|_{\mathcal{E}} \to 0,$$

for all $h \in \mathcal{D}_0$ and all sequences $(t_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$, $(\theta_n)_{n\in\mathbb{N}}$, $(h_n)_{n\in\mathbb{N}} \in \mathcal{D}^{\mathbb{N}}$ such that $t_n \searrow 0$, $\|\theta_n - \theta\|_{\mathcal{D}} \to 0$, and $\|h_n - h\|_{\mathcal{D}} \to 0$.

If φ is uniformly Hadamard differentiable at θ, $\theta_n \to \theta$ and $r_n\left(T_n - \theta_n\right) \rightsquigarrow T$, we still have that $r_n\left(\varphi\left(T_n\right) - \varphi\left(\theta_n\right)\right) \rightsquigarrow \varphi'_\theta(T)$; see A. van der Vaart and Wellner (1996, Theorem 3.9.5).

1.2 Empirical processes and plug-in estimation

Let $X_1, \ldots, X_n$ be a sample of independent random variables following the distribution of P on a measurable space $(\mathcal{X}, \mathcal{S})$. It is frequent for $\mathcal{X}$ to be a metric space, a subset of a topological vector space or a subset of $\mathbb{R}^d$ endowed with the Borel topology σ-algebra $\mathcal{S}$. With this idea in mind, a new measure, called the empirical measure, is built. In mathematical terms, we can think the sample points as a (random) discretization of P.

Definition 8. The *empirical distribution* associated to a sample $X_1, \ldots, X_n$ from the measure P is defined as $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, where δ_x is the measure that gives mass 1 to $\{x\}$. The *empirical process* associated to that sample is defined as $\mathbb{G}_n = \sqrt{n}\left(\mathbb{P}_n - \mathrm{P}\right)$.

Given a measurable function $f : \mathcal{X} \longrightarrow \mathbb{R}$ we use the following functional notation for integrals:

$$\mathrm{P}(f) = \int_{\mathcal{X}} f(x)\, \mathrm{dP}(x), \quad \mathbb{P}_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

Let $\mathcal{F}$ be a class of measurable functions over $\mathcal{X}$ and $f_1, \ldots, f_l \in \mathcal{F}$, by the multivariate Central Limit Theorem we have that $(\mathbb{G}_n(f_1), \ldots, \mathbb{G}_n(f_l))$ is asymptotically normal distributed with measure and covariance Σ, $\Sigma_{ij} = \mathbb{C}\mathrm{ov}_{\mathrm{P}}(f_i, f_j) = \mathrm{P}(f_i f_j) - \mathrm{P}(f_i)\,\mathrm{P}(f_j) = \mathrm{P}((f_i - \mathrm{P}(f_i))\,(f_j - \mathrm{P}(f_j)))$ (provided that the second moment exists). When the empirical process is considered over the whole class of measurable functions $\mathcal{F}$, the notation $\{\mathbb{G}_n(f) : f \in \mathcal{F}\}$ is used. In the following Subsection 1.2.1 pre-Gaussian and Donsker classes are introduced. The goal of the Donsker's theorem is making the statement of the Central Limit Theorem uniform in the class $\mathcal{F}$. In this direction, it is usually assumed that

$$\sup_{f \in \mathcal{F}} (|f(x) - \mathrm{P}(f)|) < \infty, \quad \text{for all } x \in \mathcal{X}.$$

So, the weak convergence of $\{\mathbb{G}_n(f) : f \in \mathcal{F}\}$ (uniform in $\mathcal{F}$) essentially amounts to state conditions for the weak convergence in $\ell^\infty(\mathcal{F})$.

1.2.1 Pre-Gaussian and Donsker classes

In this section we follow A. van der Vaart and Wellner (1996, Part 1). Some additional details, omitted here, can be found there. The starting point for the uniform weak convergence of $\{\mathbb{G}_n(f) : f \in \mathcal{F}\}$ is determining the candidate for the limit process. By simplicity, call it $\mathbb{G}_{\mathrm{P}}$. By the properties of weak convergence $\rightsquigarrow$ of random variables, $\mathbb{G}_{\mathrm{P}}$ must be tight in $\ell^\infty(\mathcal{F})$. Further, by the multidimensional Central Limit Theorem, it is a tight Gaussian process in $\ell^\infty(\mathcal{F})$. At this point, the following concept arises concerning the existence of the limit $\mathbb{G}_{\mathrm{P}}$. Firstly, recall that a process $\{T(f) : f \in \mathcal{F}\}$ is Gaussian if and only if for every finite subset J of $\mathcal{F}$, the vector $\{T(f) : f \in J\}$ is Gaussian.

Definition 9 (Pre-Gaussian class)**.** The class $\mathcal{F}$ is called *pre-Gaussian* (or P-pre-Gaussian) if and only if there exists a version of $\mathbb{G}_{\mathrm{P}}$ whose sample paths are uniformly continuous P-almost surely respect to the intrinsic pseudometric ρ_{P}, defined as

$$\rho_{\mathrm{P}}(f,g) = \mathbb{E}\left(|\mathbb{G}_{\mathrm{P}}(f) - \mathbb{G}_{\mathrm{P}}(g)|^2\right)^{1/2}, \quad f,g \in \mathcal{F}.$$

Further, any P-pre-Gaussian class is totally bounded respect to the intrinsic pseudometric ρ_{P}. Actually, in Definition 9 any p-mean pseudodistance can be taken. The goal of taking the exponent 2 relies in the fact that, given the covariance structure of $\mathbb{G}_{\mathrm{P}}$, ρ_{P} is expressed as

$$\rho_{\mathrm{P}}(f,g) = \mathrm{P}((f - \mathrm{P}(f))\,(g - \mathrm{P}(g)))^{1/2} = \left(\mathrm{P}\left((f-g)^2\right) - (\mathrm{P}(f-g))^2\right)^{1/2} \quad f,g \in \mathcal{F}.$$

Moreover, if the class $\mathcal{F}$ is contained $\mathrm{L}^1(\mathcal{X}, \mathcal{S}, \mathrm{P}) \equiv \mathrm{L}^1(\mathrm{P})$-bounded, then ρ_{P} is equivalent to the $\mathrm{L}^2(\mathrm{P})$ semimetric $\rho_{\mathrm{L}^2(\mathrm{P})}= \left(\mathrm{P}\left((f-g)^2\right)\right)^{1/2}$. The process $\mathbb{G}_{\mathrm{P}}$ is known in the literature as P-*Brownian bridge*.

Note that $\mathbb{G}_n$ is linear when acting on elements of $\mathcal{F}$. At this point, it is worth asking *how much* of this linearity is preserved by $\mathbb{G}_{\mathrm{P}}$. Let us introduce the following notion of linearity.

Definition 10. If $\mathfrak{X}$ is a subset of a vector space, a function $g : \mathfrak{X} \longrightarrow \mathbb{R}$ is said to be *prelinear* on $\mathfrak{X}$ if $\sum_{i=1}^{r} \lambda_i \, g(x_i) = 0$ whenever $\sum_{i=1}^{r} \lambda_i \, x_i = 0$, for $r \in \mathbb{N}$, $\lambda_i \in \mathbb{R}$ and $x_i \in \mathfrak{X}$ $(i = 1, \ldots, r)$.

Let $\{\mathbb{G}_{\mathrm{P}}(f) : f \in \mathcal{F}\}$ be a P-Brownian bridge indexed by $\mathcal{F}$. Observe that if $\sum_{i=1}^{r} \lambda_i f_i = 0$, with $\lambda_i \in \mathbb{R}$ and $f_i \in \mathcal{F}$ $(i = 1, \ldots, r)$, we have that

$$\mathbb{E}\left(\sum_{i=1}^{r} \lambda_i \, \mathbb{G}_{\mathrm{P}}(f_i)\right)^2 = \mathrm{P}\left(\left(\sum_{i=1}^{r} \lambda_i f_i\right)^2\right) - \left(\mathrm{P}\left(\sum_{i=1}^{r} \lambda_i f_i\right)\right)^2 = 0. \tag{1.3}$$

From (1.3), and using the Karhunen-Loève expansion of the P-Brownian bridge, it can be shown that $\mathbb{G}_{\mathrm{P}}$ has prelinear sample paths a.s. (see the proof of Giné and Nickl (2021, Theorem 3.7.28) for details). Actually, this is true if $\mathcal{F}$ is P-pre-Gaussian (see Giné and Nickl (2021, Definition 3.7.26, p. 251) and Giné and Nickl (2021, Remark 3.7.27)).

Nevertheless, the existence of $\mathbb{G}_{\mathrm{P}}$ is not a sufficient condition to derive the weak convergence of $\mathbb{G}_n$.

Definition 11. The class $\mathcal{F}$ is said to be P-*Donsker* if and only if the process $\{\mathbb{G}_n(f) : f \in \mathcal{F}\}$ converge weakly (uniformly in $\ell^\infty(\mathcal{F})$) to $\mathbb{G}_{\mathrm{P}}$. The class $\mathcal{F}$ is said to be *universal* or *uniform* Donsker if and only if it is P-Donsker for every probability measure P on the sample space $(\mathcal{X}, \mathcal{S})$.

At first glance, it might look like previous definition does not give any information. Several sufficient conditions for the Donsker property to hold are known. Briefly speaking, this conditions are related to *how many balls of a fixed radius are needed to cover a uniformly bounded class $\mathcal{F}$ and how this number behaves when the radius of the balls tends to 0.* Details about these characterizations of the Donsker property are used in Chapter 5. For further reading see Giné and Nickl (2021), Ledoux and Talagrand (1991), A. W. van der Vaart (2000), and A. van der Vaart and Wellner (1996), among others. In Section 4.6, a Donsker's theorem for particular sets of RKHS is proved. Also, in Chapter 5, devoted to the uniqueness of k-means sets, the Donsker property plays a relevant role to derive the asymptotics results and a test for uniqueness.

1.2.2 Related topics

In this subsection we just focus on two particular, simple cases, of empirical process in this work. In the final paragraphs, we outline the plug-in methodology that plays a central role in what follows.

Classical empirical process

Usual classes of functions considered in literature are $\{\mathbf{1}_{(-\infty,x_1]\times\ldots\times(-\infty,x_d]} : x = (x_1,\ldots,x_d) \in \mathbb{R}^d\}$ or $\{\mathbf{1}_{(-\infty,x_1]\times\ldots\times(-\infty,x_d]} : x = (x_1,\ldots,x_d) \in \overline{\mathbb{R}}^d\}$, where $\mathbf{1}_A$ is the indicator function of the set A. Note that $\mathrm{P}\left(\mathbf{1}_{(-\infty,x_1]\times\ldots\times(-\infty,x_d]}\right) = F(x)$ for any $x \in \mathbb{R}^d, \overline{\mathbb{R}}^d$, where F is the distribution function of the measure P. Analogously, the *empirical distribution function* (associated to a sample $X_1,\ldots,X_n$) $\mathbb{F}_n$ is defined as $\mathbb{P}_n\left(\mathbf{1}_{(-\infty,x_1]\times\ldots\times(-\infty,x_d]}\right)$. As the reader can infer, these classes of functions are usually used to derive Central Limit Theorems for real multivariate random variables.

In this context, the empirical process $\mathbb{G}_n$ takes the following form: $\mathbb{G}_n\left(\mathbf{1}_{(-\infty,x]}\right) = \sqrt{n}\,(\mathbb{F}_n(x) - F(x))$. Other usual notations are $\mathbb{G}_n(x)$ or $\mathbb{G}_{n,x}$. The class of indicators is universally Donsker for measures with finite second moment ($\int_{\mathbb{R}^d} \|x\|_2^2 \,\mathrm{d}F(x) < \infty$). It is also quite common finding that the limit is denoted by $\mathbb{B}_F$. The "B" comes from the word bridge, as this process was formerly known as *F-Brownian bridge*. With the same notation, we have that the covariance structure of $\mathbb{B}_F$ is $\mathbb{E}\left(\mathbb{B}_F(x)\mathbb{B}_F(y)\right) = F(x \wedge y) - F(x)F(y)$, where $x \wedge y = (\min(x_1,y_1),\ldots,\min(x_d,y_d))$. This notation is extensively used in Chapter 3 since the majority of the situations posed there belongs to this classic framework. For the sake of completeness, the process $\mathbb{B} = \{\mathbb{B}_x : x \in [0,1]\}$, in other words, when $F \equiv \mathbf{1}_{[0,1]}$; is known as the *standard Brownian bridge*. Note that $\mathbb{B}_F = \mathbb{B} \circ F$.

Independent empirical processes

Throughout the former section the focus was on the empirical process given one data sample $X_1,\ldots,X_n$ of i. i. d. following the distribution of P. In the homogeneity problem, given two unknown measures P and Q, the null hypothesis $\mathrm{H}_0 : \mathrm{P} = \mathrm{Q}$ (against the alternative $\mathrm{H}_1 : \mathrm{P} \neq \mathrm{Q}$) is tested. Hence, we have two samples: $X_1,\ldots,X_n$ coming from P, and $Y_1,\ldots,Y_m$, from Q. These samples are assumed to be independent. This framework usually appears in science when two *identical* experiments are done under the same conditions (null hypothesis) and the goal is validating the results. Therefore, this mathematical scenario is in the heart of science as it provides the necessary tools in a quite general scenario.

Given the two samples described above, we consider

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}, \quad \mathbb{Q}_m = \frac{1}{m}\sum_{i=1}^{m}\delta_{Y_i},$$

and the respective empirical processes $\mathbb{G}_n^{\mathrm{P}} = \sqrt{n}\,(\mathbb{P}_n - \mathrm{P})$ and $\mathbb{G}_n^{\mathrm{Q}} = \sqrt{m}\,(\mathbb{Q}_m - \mathrm{Q})$. Further, given that in this problem the interest is focus on discriminating P and Q, the process $\mathbb{G}_{n,m} = \sqrt{\frac{n\,m}{n+m}}\,(\mathbb{P}_n - \mathbb{Q}_m - \mathrm{P} + \mathrm{Q})$. Provided that the class $\mathcal{F}$ is P and Q-Donsker and that $\frac{n}{n+m} \overset{(n,m)\to\infty}{\longrightarrow} \xi \in [0,1]$,

$$\mathbb{G}_{n,m} \rightsquigarrow \mathbb{G} = \sqrt{1-\xi}\,\mathbb{G}_{\mathrm{P}} - \sqrt{\xi}\,\mathbb{G}_{\mathrm{Q}}, \tag{1.4}$$

holds, where $\mathbb{G}_{\mathrm{P}}$ and $\mathbb{G}_{\mathrm{Q}}$ are independent P and Q-Brownian bridges, respectively; and independent. As it can be observed, $\mathbb{G}$ is Gaussian process with continuous sample paths respect to the pseudometric $\rho = \max(\rho_{\mathrm{P}}, \rho_{\mathrm{Q}})$ (or other equivalent metric such as the sum $\rho_{\mathrm{P}} + \rho_{\mathrm{Q}}$), with mean 0 and covariance

$$\mathbb{E}(\mathbb{G}(f)\,\mathbb{G}(g)) = \xi\,\mathbb{C}\mathrm{ov}_{\mathrm{P}}(f,g) + (1-\xi)\,\mathbb{C}\mathrm{ov}_{\mathrm{Q}}(f,g), \quad f,g \in \mathcal{F},$$

with $\mathbb{C}\mathrm{ov}_{\nu}(f,g) = \nu(f\,g) - \nu(f)\,\nu(g)$, $\nu \in \{\mathrm{P},\mathrm{Q}\}$.

Plug-in estimation

In statistical problems we often deal with functionals $\varphi(\mathrm{P})$ of the underlying distribution P. A natural estimator of $\varphi(\mathrm{P})$, often called "plug-in estimator", is just $\varphi(\mathbb{P}_n)$. So, it is obtained by just replacing P with the empirical probability measure $\mathbb{P}_n$ corresponding to a random sample of size n. The use of differentiability techniques, as those considered in this work, is particularly relevant analyze how $\varphi(\mathbb{P}_n)$ approximates $\varphi(\mathrm{P})$.

The plug-in methodology underlies our approaches to the problems we consider in Chapters 4 and 5.

1.3 Reproducing Kernel Hilbert Spaces

The theory of RKHS is relevant in this thesis, particularly in Chapter 4. This is a classical and well-known topic; see Janson (1997, Appendix F) for a brief account of the RKHS theory and Berlinet and Thomas-Agnan (2011) or Hsing and Eubank (2015) for a statistical perspective. As it can be inferred from the introduction of Berlinet and Thomas-Agnan (2011), RKHS are important because they provide an environment to define transformations to solve classical problems in statistics. An outstanding example is the support vector machines algorithm, a linear classification method where the data is carried to a space of larger dimension (see Vapnik (1999, Chapter 5)). In this work we use the same idea, known as the *kernel trick*, to perform an homogeneity test in the same way as Gretton et al. (2007) and the references therein (see Chapter 4). But first, for the sake of completeness, let us remind some important features of these spaces.

Definition 12. Let $\mathcal{H}$ be a Hilbert space of real-valued functions on $\mathcal{X}$ with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. A function $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ is called a *reproducing kernel* of $\mathcal{H}$ if and only if

(a) For $y \in \mathcal{X}$, the function $k(\cdot, y) \in \mathcal{H}$;

(b) *Reproducing property:* For $y \in \mathcal{X}$ and $f \in \mathcal{H}$, we have that $\langle f, k(\cdot,y)\rangle_{\mathcal{H}} = f(y)$.

If the Hilbert space has a reproducing kernel k, $\mathcal{H}$ is called a *reproducing kernel Hilbert space* (RKHS in short). To highlight the underlying kernel, we will write $\mathcal{H} \equiv \mathcal{H}_k$ and $\langle\cdot,\cdot\rangle_{\mathcal{H}} \equiv \langle\cdot,\cdot\rangle_{\mathcal{H}_k}$.

By the reproducing property, every kernel is symmetric and positive definite function. Conversely, let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a *kernel*, that is, a symmetric and positive semi-definite function. Let us consider $\mathcal{H}_k^0$, the pre-Hilbert space of all finite linear combinations $g(\cdot) = \sum_{i=1}^n \alpha_i \, k(x_i, \cdot)$ (with $\alpha_i \in \mathbb{R}$, $n \in \mathbb{N}$ and $x_i \in \mathcal{X}$), endowed with the inner product

$$\left\langle \sum_{i=1}^n \alpha_i \, k(x_i, \cdot), \sum_{j=1}^m \beta_j \, k(x_j, \cdot) \right\rangle_{\mathcal{H}_k} = \sum_{i,j} \alpha_i \, \beta_j \, k(x_i, x_j) \,. \tag{1.5}$$

The space $\mathcal{H}_k$ is characterized as the pointwise limits of Cauchy sequences in this pre-Hilbert space. The statement of this paragraph is known as Moore-Aronszajn's theorem (Berlinet and Thomas-Agnan (2011, Theorem 3)). Additionally, the RKHS $\mathcal{H}_k$ is the completion of $\mathcal{H}_k^0$.

Definition 12 is equivalent to saying that $\mathcal{H}_k$ is a Hilbert space of functions on $\mathcal{X}$ such that for all $x \in \mathcal{X}$ the evaluation function $\mathrm{ev}_x : \mathcal{H}_k \longrightarrow \mathbb{R}$, defined for $h \in \mathcal{H}_k$ by $\mathrm{ev}_x(h) = h(x)$, is a continuous map (see Berlinet and Thomas-Agnan (2011, Theorem 1)). Therefore, by Riesz's representation theorem, for each $x \in \mathcal{X}$, there exists $\varphi_x \in \mathcal{H}_k$ such that for all $f \in \mathcal{H}_k$, $f(x) = \langle f, \varphi_x \rangle_{\mathcal{H}_k}$, or just $f = \langle f, \varphi \rangle_{\mathcal{H}_k}$. The function φ is often called *feature mapping*. By the reproducing property, we have that $\varphi_x = k(\cdot, x)$. In particular, $k(x,y) = \langle \varphi_x, \varphi_y \rangle_{\mathcal{H}_k}$.

1.3.1 The mean embedding

A priori, functions in $\mathcal{H}_k$ do not need to be integrable, not even measurable. Necessary and sufficient conditions for measurability are provided in Berlinet and Thomas-Agnan (2011, Theorem 90) when $\mathcal{H}_k$ is separable. Given a Borel probability measure ν, typical conditions imposed on k in this context are $\int_{\mathcal{X}} \sqrt{k(x,x)} \, \mathrm{d}\,\nu(x) < \infty$ or $\int_{\mathcal{X}} k(x,x) \, \mathrm{d}\,\nu(x) < \infty$. These conditions, satisfied for the families usually used in Statistics (see 4.3 and Sriperumbudur (2016)), are closely related to Bochner integrability of the feature mapping, that is $\int_{\mathcal{X}} k(x,x)^{\gamma/2} \, \mathrm{d}\,\nu(x) = \int_{\mathcal{X}} \|\varphi_x\|_{\mathcal{H}_k}^{\gamma} \, \mathrm{d}\,\nu(x)$ with $\gamma > 1$. In this subsection we deal with the *weak or Pettis integrability* of the feature mapping. Specifically, we stablish an equivalence between the continuity of the integral as a functional in $\mathcal{H}_k$ and Pettis integrability of the feature mapping. From now on, we assume that $\mathcal{X}$ is a separable topological space (metrizable if required).

Definition 13. Let $(\mathcal{H}_k, \langle \cdot, \cdot \rangle_{\mathcal{H}_k})$ be a RKHS on $\mathcal{X}$ and let ν be a Borel probability measure on $\mathcal{X}$. The *mean embedding* of ν is an element $\mu_\nu \in \mathcal{H}_k$ such that for all $f \in \mathcal{H}_k$, $\nu(f) = \int_{\mathcal{X}} f(x) \, \mathrm{d}\,\nu(x) = \langle f, \mu_\nu \rangle_{\mathcal{H}_k}$.

Note that we are denoting the integral of f respect to the measure ν as $\nu(f)$, as in the empirical processes theory (see A. van der Vaart and Wellner (1996)[Chapter 2.1]). By the Riesz's representation theorem (Conway (2019, Chapter 1.3)), necessary and sufficient condition for the existence of the mean-embedding is the continuity of the integral in $\mathcal{H}_k$. In fact, this statement is valid for every Hilbert space. In a RKHS, thanks to the reproducing property, we can give additional characterizations. The following definition has been taken from Pettis (1938, Definition 2.1).

Definition 14. Let $(\mathcal{H}_k, \langle \cdot, \cdot \rangle_{\mathcal{H}_k})$ be a RKHS on $\mathcal{X}$ and $F : \mathcal{X} \longrightarrow \mathcal{H}_k$ be a weakly measurable map, that is, for every $g \in \mathcal{H}_k$ the real function $\langle F, g \rangle_{\mathcal{H}_k}$ is measurable. We say that F is *Pettis* or *weakly integrable* with respect to a Borel probability measure ν on $\mathcal{X}$ if and only if

1. For every $h \in \mathcal{H}_k$, the map
$$\begin{array}{rccc} \langle F, h \rangle_{\mathcal{H}_k} : & \mathcal{X} & \longrightarrow & \mathbb{R} \\ & x & \mapsto & \langle F(x), h \rangle_{\mathcal{H}_k} \end{array}.$$
lies in $\mathrm{L}^1(\nu)$.

2. There exists $m_F \in \mathcal{H}_k$ such that for every $h \in \mathcal{H}_k$, $\langle m_F, h \rangle_{\mathcal{H}_k} = \nu\left(\langle F, h \rangle_{\mathcal{H}_k}\right)$.

The element m_F of $\mathcal{H}_k$ is called *the integral of F* (with respect to ν).

Remark 15. When $F(x) = k(\cdot, x) = \varphi_x$, the feature mapping Definition 14 can be rewritten as:

1. By the reproducing property, condition 1 is equivalent to $\mathcal{H}_k \subseteq \mathrm{L}^1(\nu)$.

2. By the Riesz's representation theorem and the reproducing property, condition 2 is equivalent to that the integral ν is a continuous functional on $\mathcal{H}_k$.

By Remark 15, the Pettis integrability of the feature mapping with respect to a measure is equivalent to the existence of the mean embedding of such measure. Let us now focus on the properties of Pettis integral to state necessary and sufficient conditions about the existence of the mean embedding. More specifically, we show that, in Definition 14, condition 1 implies condition 2.

Proposition 16. *Let $(\mathcal{H}_k, \langle \cdot, \cdot \rangle_{\mathcal{H}_k})$ be a RKHS on $\mathcal{X}$ and let ν be a Borel probability measure on $\mathcal{X}$. The following four conditions are equivalent:*

1. *There exists the mean embedding μ_ν of ν in $\mathcal{H}_k$.*

2. *The feature mapping φ is Pettis ν-integrable.*

3. *Let $\mathcal{F}_{\mathcal{H}_k}$ be the unit ball of $\mathcal{H}_k$. Then, $\sup\limits_{f \in \mathcal{F}_{\mathcal{H}_k}} (\nu(f)) < \infty$.*

4. *$\mathcal{H}_k \subseteq \mathrm{L}^1(\nu)$.*

In any, and hence all, of these situations, ν defines a continuous linear functional on $\mathcal{H}_k$ and
$$\|\mu_\nu\|_{\mathcal{H}_k} = \|\nu\|_{\mathcal{H}_k^*} = \left(\int_{\mathcal{X}} \int_{\mathcal{X}} k(x,y) \, \mathrm{d}\nu(y) \, \mathrm{d}\nu(x)\right)^{1/2}, \tag{1.6}$$
where $\mathcal{H}_k^*$ is the dual space of $\mathcal{H}_k$.

Proof. Let us prove the following equivalences: 1⇔2,2⇔3 and 2⇔4.

1⇔2. The proof of this equivalence is a formalization of the statement of Remark 15. By definition of the feature mapping φ, for every $f \in \mathcal{H}_k$ $\int_{\mathcal{X}} |f(x)| \,\mathrm{d}\,\nu(x) = \int_{\mathcal{X}} \left|\langle f, \varphi_x \rangle_{\mathcal{H}_k}\right| \,\mathrm{d}\,\nu(x)$. That is, condition 1 in Definition 14 and $\mathcal{H}_k \subseteq \mathrm{L}^1(\nu)$ are equivalent. Additionally, for every $f \in \mathcal{H}_k$ $\langle f, \mu_\nu \rangle_{\mathcal{H}_k} = \int_{\mathcal{X}} f(x) \,\mathrm{d}\,\nu(x) = \int_{\mathcal{X}} \langle f, \varphi_x \rangle_{\mathcal{H}_k} \,\mathrm{d}\,\nu(x)$, so condition 2 in Definition 14 and the existence of mean embedding are equivalent.

2⇔3. Condition 1 in Definition 14 means that ν is well defined (as a linear functional on $\mathcal{H}_k$). Additionally, by the Riesz's representation theorem (see Conway (2019, Chapter 1, 3.4)), condition 2 in Definition 14 and the continuity of ν are equivalent. Since ν is linear, continuity of ν and $\sup_{f \in \mathcal{F}_{\mathcal{H}_k}} (|\nu(f)|) < \infty$ are equivalent (see Conway (2019, Chapter 1, 3.1)).

2⇔4. It is clear that, by condition 1 in Definition 14, statement 2 implies claim 4. Conversely, let us assume that $\mathcal{H}_k \subseteq \mathrm{L}^1(\nu)$, which is condition 1 in Definition 14. By Hille and Phillips (1957, Theorem 3.7.1), there exists $h^{**} \in \mathcal{H}_k^{**}$ (the bidual space of $\mathcal{H}_k$) such that for every $f \in \mathcal{H}_k$,

$$h^{**}\left(\langle \cdot, f \rangle_{\mathcal{H}_k}\right) = \int_{\mathcal{X}} \langle \varphi_x, f \rangle_{\mathcal{H}_k} \,\mathrm{d}\,\nu(x) = \int_{\mathcal{X}} f(x) \,\mathrm{d}\,\nu(x),$$

where $\langle \cdot, f \rangle_{\mathcal{H}_k}$ stands for the functional associated with f by the Riesz's representation theorem. Such h^{**} is unique. Since every Hilbert space is reflexive, there exists $h \in \mathcal{H}_k$ such that for every $f \in \mathcal{H}_k$ satisfies $h^{**}\left(\langle \cdot, f \rangle_{\mathcal{H}_k}\right) = \langle h, f \rangle_{\mathcal{H}_k}$. Then condition 2 of Definition 14 holds. We conclude that h is the Pettis integral of φ with respect to ν.

Formula (1.6) is just the expression of the norm of μ_ν deduced from the properties of Pettis integral.

For example, conditions 1-4 in Proposition 16 always hold if $\int_{\mathcal{X}} \|\varphi_x\|_{\mathcal{H}_k} \,\mathrm{d}\nu(x) = \int_{\mathcal{X}} \sqrt{k(x,x)} \,\mathrm{d}\,\nu(x) < \infty$. Indeed, by Cauchy–Schwarz inequality,

$$|\nu(f)| = \left| \int_{\mathcal{X}} \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} \,\mathrm{d}\,\nu(x) \right| \leq \|f\|_{\mathcal{H}_k} \int_{\mathcal{X}} \sqrt{k(x,x)} \,\mathrm{d}\,\nu(x).$$

Hence, we conclude that there exists μ_ν satisfying Definition 13. In particular, any bounded kernel trivially fullfills this requirement. This property is known as *Bochner* or *strong integrability* (see Hille and Phillips (1957, Definition 3.7.3))

The mean embedding of ν can be seen as a smoothed representation of the distribution of ν using the kernel k within the RKHS. This becomes evident when ν is absolutely continuous with density f_ν and k is a translation-invariant kernel, i.e., $k(x,y) = \Phi(x-y)$, for some real function Φ (see Wendland (2004, Chapter 6) or Wynne and Duncan (2022, Appendix A.1)). In this scenario, μ_ν is the convolution of f_ν and Φ. Additionally, mean embeddings, also known as "potential functions", appear in other mathematical fields, such as functional analysis (see El-Fallah et al. (2014, p. 15)). Furthermore, the mean embedding plays a crucial role in hypothesis testing based on kernel distances and uniform kernel distances (see Sejdinovic et al. (2013), Section 4.2, and the references therein).

Chapter 2

Directional differentiability for supremum-type functionals

The aim of this chapter is to discuss the (directional) differentiability of the supremum norm –as well as various related functionals that commonly appear in statistics– viewed as a real functional from the space of bounded functions defined on an arbitrary set or a measure space. We consider the supremum norm, the supremum, the infimum, and the amplitude of a real function. The (usually non-linear) derivatives of these maps adopt simple expressions under suitable assumptions on the underlying space.

2.1 Introduction

The general framework

The supremum or uniform norm has been systematically used in statistics to quantify the deviation between an observed model and a theoretical one. A well-known case is the goodness-of-fit problem, where the Kolmogorov distance (i.e., the uniform distance between distribution functions) is one of the main tools to carry out the testing procedures. In this context, the prototypical example is the Kolmogorov-Smirnov test in which the supremum norm between the empirical distribution function of the sample and the reference distribution function is employed. The sup-norm has also been notably considered in the literature of almost all fields of statistics such as robustness, density estimation, regression and classification, among others. The reason for the extensive use of this distance might rely on different factors: it has a clear and simple interpretation; it takes into account the global behaviour of the functions; and, in general, it is easy to compute.

The problem under study

Throughout this chapter, $\mathfrak{X}$ is a nonempty set and $\ell^\infty(\mathfrak{X})$ is the real Banach space of bounded functions $f : \mathfrak{X} \longrightarrow \mathbb{R}$, equipped with the supremum norm, $\|f\|_\infty = \sup_{x\in\mathfrak{X}} (|f(x)|)$. Usually, we will omit the variable x in the supremum, denoting it as $\sup_{\mathfrak{X}} (f)$. If additionally $(\mathfrak{X}, \mathcal{S}, \nu)$ is a measure space, where $\mathcal{S}$ is a σ-algebra and ν a positive measure, we denote by $\mathrm{L}^\infty(\mathfrak{X},\mathcal{S},\nu)$ the set of classes of equivalence of measurable and essentially bounded functions $f:\mathfrak{X}\longrightarrow\mathbb{R}$ with the norm $\|f\|_{\mathrm{L}^\infty(\mathfrak{X},\mathcal{S},\nu)}=\operatorname*{ess\,sup}_{\mathfrak{X}}(|f|)$, where

$$\begin{aligned}\operatorname*{ess\,sup}_{\mathfrak{X}}(f)&=\sup(\{C\in\mathbb{R}:\nu(\{x\in\mathfrak{X}:f(x)\le C\})>0\})\\&=\inf(\{C\in\mathbb{R}:\nu(\{x\in\mathfrak{X}:f(x)>C\})=0\}).\end{aligned}$$

Important examples of this general setting are $\mathfrak{X}=\mathbb{R}^d$ or $\overline{\mathbb{R}}^d$ $(d\ge1)$, with $\overline{\mathbb{R}}\equiv[-\infty,+\infty]$ the extended real line, and $\mathfrak{X}=\mathcal{F}$, a class of real valued functions. In fact, $\mathrm{L}^\infty(\mathfrak{X},\mathcal{P}(\mathfrak{X}),\nu)=\ell^\infty(\mathfrak{X})$, where ν is the counting measure. From now on, unless specifically mentioned, we use this space for the sake of simplicity.

For $\theta\in\ell^\infty(\mathfrak{X})$, the quantity of interest that we want to estimate is $\phi(q)$, where ϕ is any of the following functionals:

$$\begin{aligned}&\delta(f)=\|f\|_\infty,\quad \sigma(f)=\sup_{\mathfrak{X}}(f),\quad \iota(f)=\inf_{\mathfrak{X}}(f),\quad\text{and}\\&\alpha(f)=\operatorname*{amp}_{\mathfrak{X}}(f),\quad\text{for } f\in\ell^\infty(\mathfrak{X}),\end{aligned}\tag{2.1}$$

with $\operatorname*{amp}_{\mathfrak{X}}(f)=\sup_{\mathfrak{X}}(f)-\inf_{\mathfrak{X}}(f)$, the amplitude of the function f.

We will assume that θ can be estimated by T_n, a random element taking values in $\ell^\infty(\mathfrak{X})$ a.s. satisfying

$$r_n\,(T_n-\theta)\rightsquigarrow T\quad\text{in }\ell^\infty(\mathfrak{X}),\quad\text{a.s. } n\to\infty,\tag{2.2}$$

where r_n is a sequence of real numbers such that $r_n\to\infty$, T is a tight Borel random variable in $\ell^\infty(\mathfrak{X})$, and we use the arrow "$\rightsquigarrow$" to denote the weak convergence of probability measures. The scaling r_n usually goes to infinity as $\sqrt{n}$, but its behaviour could be different in some examples.

For $\varphi\in\{\delta,\sigma,\iota,\alpha\}$ in (2.1), we are interested in analyzing the asymptotic behaviour of the normalized estimator of $\varphi(\theta)$, that is, the statistic given by

$$D_n(\varphi)\equiv D_\varphi(\theta,T_n,r_n)=r_n\,(\varphi(T_n)-\varphi(\theta)).\tag{2.3}$$

Background

By the Continuous Mapping Theorem (see A. van der Vaart and Wellner (1996, Theorem 1.3.6)), when $\theta=0$ (the null function), the weak convergence in (2.2) directly implies that $D_n(\varphi)\rightsquigarrow\varphi(T)$. (Note that in this case "$\rightsquigarrow$" is the usual convergence in distribution of random variables since φ is real valued.) This situation often corresponds to the case in which $D_n(\varphi)$ is a normalized discrepancy –usually measured in terms of the sup-norm– for testing the null hypothesis $\mathrm{H}_0:\theta=0$. In this setting, the limiting behaviour of $D_n(\varphi)$ if $\theta\ne0$ provides information regarding the asymptotic power of the underlying testing procedure. The classical result on the asymptotic distribution of the Kolmogorov-Smirnov statistic under the null hypothesis (see, e.g., A. W. van der Vaart (2000)) is a well-known example.

Finding the asymptotic distribution of $D_n(\varphi)$ in (2.3) when θ is not identically zero is a more challenging problem. So far, this problem has been tackled generally for the sup-norm and some particular choices of the function θ. To the best of our knowledge, the first remarkable result in this direction was obtained by Raghavachari (1973). This author found the asymptotic distribution of the normalized version of the plug-in estimator of $\varphi(F - G)$ (for $\varphi \in \{\delta, \sigma, \alpha\}$) in the one-sample and two-sample cases when F and G are continuous univariate distribution functions. (The results in Raghavachari (1973) have also been summarized in DasGupta (2008, Chapter 26).) Over the years, the ideas in Raghavachari (1973) have been used and replicated by several authors to obtain different results in similar settings. A non-exhaustive list of these references is: Álvarez-Esteban et al. (2012); Álvarez-Esteban et al. (2016); Freitag et al. (2006); Hjort (1990); Schmoyer (1988); among others. In Genest and Nešlehová (2014), the authors discussed a test of radial symmetry for copulas in which the key element is the estimation of $\left\|C - \overline{C}\right\|_\infty$, where C is a bivariate copula and $\overline{C}$ is its survival copula. Dette et al. (2018) used the same technique to find the asymptotic distribution of the estimator of $\|m_1(\beta_1) - m_2(\beta_2)\|_\infty$, where $m_1(\beta_1)$ and $m_2(\beta_2)$ are regression functions with parameters β_1 and β_2, respectively.

The proposed methodology

In all the previous references the same approach has been used to compute the limiting distributions: the direct probabilistic analysis of the considered statistics. For instance, the proofs in Raghavachari (1973) are essentially based on a careful analysis of the behaviour of the empirical process in the set of points around which the supremum in $\|F - G\|_\infty$ is attained. However, we explore here an alternative, more general, approach. It is based on the idea that the statistics in (2.3) have indeed the usual form suitable to apply the functional Delta method. Therefore, in light of (2.3), a direct and intuitive approach to find the asymptotic distribution of $D_n(\varphi)$ could be to analyze the differentiability of the maps in (2.1) and use the functional Delta method. In fact, as it will become evident in this chapter, looking at the behaviour and analytic properties of the underlying functional is much more enlightening than working directly with the probability distribution of the statistic.

Though there are many possible ways of defining the concept of differentiability of maps between metrics or normed spaces, within this context Hadamard differentiability is perhaps the most convenient as it is appropriate for applying the functional Delta method A. W. van der Vaart (2000, Section 20). However, there are many important maps which are *not* Hadamard differentiable. Thus, for example, the functionals in (2.1) are clearly continuous but non-differentiable. Despite not being fully differentiable, we will show that these maps are *Hadamard directionally differentiable*. This weaker notion of differentiability was introduced by Shapiro (1990). Shapiro (1991) and Dümbgen (1993) (see also Römisch (2004)) independently showed that the Delta method still holds for directional differentiable maps. Recently, this idea has been successfully exploited in Beare and Fang (2017) and Sommerfeld and Munk (2018). Fang and Santos (2019) also illustrate the applicability of the directional differentiability to a wide variety of problems in econometrics.

Structure of the chapter

In Section 2.2 we prove that the maps in (2.1) are Hadamard directional differentiable and determine their derivatives (see 1.1). In particular, this implies that an extended version of the functional Delta method can be applied for these mappings. As far as we know, in the statistical community the Hadamard directional differentiability of the infimum under no additional conditions on the underlying space was first obtained by Römisch (2004, Proposition 1), after a personal communication of P. Lachout. Fang and Santos Fang and Santos (2019, Lemma S.4.9), also obtained an expression for the Hadamard directional derivative of the supremum for continuous functions defined on a compact metric space.

If the space $\mathfrak{X}$ is endowed with additional structure, then simpler expressions for the derivatives can be obtained as well as exact conditions under which the maps are fully Hadamard differentiable. We specifically deal with the case where $\mathfrak{X}$ is a compact metric space (Section 2.3), a totally bounded metric space (Section 2.4), weakly compact subset of a Banach space (Section 2.5), and the union of unit balls of a family of reproducing kernel Hilbert spaces. We also consider in detail the situation in which $\mathfrak{X} = \overline{\mathbb{R}}^d$ and the functions belong to $\mathcal{D}\left(\overline{\mathbb{R}}^d\right)$, the extension of the Skorohod space in $[0,1]^d$ (introduced in Neuhaus (1971)) to the whole $\overline{\mathbb{R}}^d$. The space $\mathcal{D}\left(\overline{\mathbb{R}}^d\right)$ is an important subspace of $\ell^\infty\left(\overline{\mathbb{R}}^d\right)$ as it includes the paths of many well-known stochastic processes with jumps in their paths such as multivariate empirical and copula processes.

In the following sections we discuss the analytic properties of the functionals introduced according to the mathematical structure of $\mathfrak{X}$. Specifically, we show the Hadamard differentiability of them under a variety of situations, quite common in statistics.The versatility of the proposed methodology is illustrated in depth in Chapter 3, Chapter 4 and 5, where we derive the asymptotic distribution of various statistics with no additional effort. We base the results on the directional differentiability of the functionals and the weak convergence of the underlying stochastic processes. Hence, this unifying approach allows us to reduce a usually difficult statistical problem to a much simpler analytical question related to the directional differentiability of the corresponding functional.

2.2 A general result

In the next theorem we show that the maps introduced in Section 2.1 are directionally differentiable at every function of $\ell^\infty(\mathfrak{X})$, where $\mathfrak{X}$ is an arbitrary set. In the sequel $\operatorname{sgn}(\cdot)$ denotes the sign function:

$$\operatorname{sgn}(x) = \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

Theorem 17. *The maps* δ, σ, ι *and* α *in* (2.1) *are Hadamard directionally differentiable at every* $f \in \ell^\infty(\mathfrak{X})$. *For* $g \in \ell^\infty(\mathfrak{X})$, *their derivatives are respectively given by*

$$\begin{aligned} \delta'_f(g) &= \lim_{\varepsilon \searrow 0} \sup_{A_\varepsilon(|f|)} (g \operatorname{sgn}(f)), & \sigma'_f(g) &= \lim_{\varepsilon \searrow 0} \sup_{A_\varepsilon(f)} (g), \\ \iota'_f(g) &= \lim_{\varepsilon \searrow 0} \inf_{B_\varepsilon(f)} (g), & \alpha'_f(g) &= \lim_{\varepsilon \searrow 0} \left(\sup_{A_\varepsilon(f)} (g) - \inf_{B_\varepsilon(f)} (g) \right), \end{aligned} \tag{2.4}$$

where, for $\varepsilon > 0$ *and* $h \in \ell^\infty(\mathfrak{X})$, $A_\varepsilon(h)$ *and* $B_\varepsilon(h)$ *are the superlevel and sublevel sets of* h *defined by*

$$A_\varepsilon(h) = \left\{ x \in \mathfrak{X} : h(x) \geq \sup_{\mathfrak{X}} (h) - \varepsilon \right\} \quad \text{and} \quad B_\varepsilon(h) = \left\{ x \in \mathfrak{X} : h(x) \leq \inf_{\mathfrak{X}} (h) + \varepsilon \right\}.$$

Moreover, if $(\mathfrak{X}, \mathcal{S}, \nu)$ *is a measure space, the result still holds if we substitute the suprema (respectively infima) by essential suprema (respectively infima) with respect to* ν.

Proof. First of all, observed that δ, σ, ι, and α are locally Lipschitz funcionals on $\ell^\infty(\mathfrak{X})$. Then, by Remark 3, it suffices to prove that they are Gâteaux directional differentiable. Let us fix $f \in \ell^\infty(\mathfrak{X}) \setminus \{0\}$. We start with σ as the conclusion for the rest of the maps can be derived from this case. For $n \in \mathbb{N}$ and each sequence of real numbers $(s_n)_{n \in \mathbb{N}}$ such that $s_n \nearrow \infty$, we consider $\sigma_n(f) : \ell^\infty(\mathfrak{X}) \longrightarrow \mathbb{R}$ defined by

$$\sigma_n(f, g) = \sup_{\mathfrak{X}} (s_n f + g) - s_n \sup_{\mathfrak{X}} (f), \quad g \in \ell^\infty(\mathfrak{X}). \tag{2.5}$$

In order to proof that σ is Gâteaux differentiable, it suffices to show that $\sigma_n(f, g) \to \sigma'_f(g)$, as $n \to \infty$, with $\sigma'_f(g)$ defined in (2.4). As it can be observed, $s_n = \frac{1}{t_n}$ in Definition 1. For $\varepsilon > 0$ and $x \notin A_\varepsilon(f)$, we have that

$$s_n f(x) + g(x) - s_n \sup_{\mathfrak{X}} (f) \leq \sup_{\mathfrak{X}} (g) - s_n \varepsilon.$$

Hence, for all $\varepsilon > 0$, we obtain that

$$\begin{aligned} \limsup_{n \to \infty} \sigma_n(f, g) &= \limsup_{n \to \infty} \left(\sup_{A_\varepsilon(f)} (s_n f + g) - s_n \sup_{\mathfrak{X}} (f) \right) \\ &\leq \sup_{A_\varepsilon(f)} (g). \end{aligned} \tag{2.6}$$

Conversely, let us define

$$h(\varepsilon) = \sup_{A_\varepsilon(f)} (g), \quad \varepsilon > 0. \tag{2.7}$$

Observe that h is non-decreasing and thus the limit as ε decreases to 0 exists and, by definition, coincides with $\sigma'_f(g)$. For each $m \in \mathbb{N}$, there exists $x_m \in A_{1/m}(f)$ satisfying

$$g(x_m) \geq h\left(\frac{1}{m}\right) - \frac{1}{m} \quad \text{and} \quad f(x_m) \geq \sup_{\mathfrak{X}} (f) - \frac{1}{m}. \tag{2.8}$$

From (2.8), for each s_n, we have that

$$\begin{aligned} h\left(\frac{1}{m}\right) &\leq g\left(x_m\right) + \frac{1}{m} \\ &= s_n\, f\left(x_m\right) + g\left(x_m\right) - s_n\, f\left(x_m\right) + \frac{1}{m} \\ &\leq \sup_{\mathfrak{X}}\, \left(s_n\, f + g\right) - s_n \left(\sup_{\mathfrak{X}} (f) - \frac{1}{m}\right) + \frac{1}{m} \\ &= \sigma_n(f, g) + \frac{\left(s_n + 1\right)}{m}. \end{aligned} \tag{2.9}$$

Now (2.9) implies that, for all $n \in \mathbb{N}$,

$$\lim_{\varepsilon \searrow 0} \sup_{A_\varepsilon(f)} (g) = \lim_{m \to \infty} h\left(\frac{1}{m}\right) \leq \sigma_n(f, g). \tag{2.10}$$

The proof corresponding to σ follows from (2.6) and (2.10).

Now, we consider the map δ in (2.1). Assume that $f \in \ell^\infty(\mathfrak{X})$ with $\|f\|_\infty > 0$. For $g \in \ell^\infty(\mathfrak{X})$, we have to show that $\delta_n(f, g) \to \delta'_f(g)$, as $n \to \infty$, where $\delta_n(f, g) = \left\|s_n\, f + g\right\|_\infty - s_n\, \|f\|_\infty$ and $s_n \nearrow \infty$. First, for $\varepsilon < \frac{\|f\|_\infty}{2}$ and $s_n > \frac{2\, \|g\|_\infty}{\|f\|_\infty}$, it is readily checked that $s_n\, |f| + \operatorname{sgn}(f)\, g \geq 0$ globally on $A_\varepsilon(|f|)$. We hence conclude that

$$\lim_{n \to \infty} \delta_n(f, g) = \lim_{n \to \infty} \sigma_n(|f|, g \operatorname{sgn}(f)) = \sigma'_{|f|}(g \operatorname{sgn}(f)) = \delta'_f(g).$$

The proof for ι and α follows from the duality between supremum and infimum and linearity of differentiation (see Theorem 4).

Finally, the case in which $\mathfrak{X}$ is a measure space can be treated in a similar way so it is therefore omitted. □

Remark 18. From the proof of Theorem 17 it can be inferred that the (real) limits which appears in the definition of Hadamard directional differentiability can be computed for σ and ι in more general contexts. For instance, if $f \in \ell^\infty_+(\mathfrak{X})$, where

$$\ell^\infty_+(\mathfrak{X}) = \left\{ f : \mathfrak{X} \longrightarrow \mathbb{R} : \sup_{\mathfrak{X}} (f) < \infty \right\},$$

and $g_n \to g$ in $\ell^\infty(\mathfrak{X})$, we still have that

$$\left| \frac{\sigma\left(f + t_n\, g_n\right) - \sigma(f)}{t_n} - \sigma'_f(g) \right| \leq \|g_n - g\|_\infty + \left| \sigma\left(\frac{f}{t_n} + g\right) - \sigma\left(\frac{f}{t_n}\right) - \sigma'_f(g) \right|.$$

Therefore, the same lines as in the proof of Theorem 17 show that the limit fo the left hand side when $n \to \infty$ can be computed for σ when $f \in \ell^\infty_+(\mathfrak{X})$ and the value is $\sigma'_f(g)$ given in (2.4), which is well defined, for $g \in \ell^\infty(\mathfrak{X})$. Further, as in proof of the extended Delta method (see Proposition 5), the extended Continuous Mapping Theorem (see A. van der Vaart and Wellner (1996, Th. 1.11.1)) still works in this case because the argument in the previous proof only depends on the set $A_\varepsilon(f)$ (those points in $\mathfrak{X}$ that are close to $\sup_{\mathfrak{X}} (f)$). Analogously, the map ι is Hadamard directionally differentiable over the class

$$\ell^\infty_-(\mathfrak{X}) = \left\{ f : \mathfrak{X} \longrightarrow \mathbb{R} : -\infty < \inf_{\mathfrak{X}} (f) \right\}.$$

Remark 19. We observe that $\delta = \sigma \circ a$, where $a : \ell^\infty(\mathfrak{X}) \longrightarrow \ell^\infty(\mathfrak{X})$ is defined by $a(f) = |f|$. With similar techniques it can be seen that a is Hadamard directional differentiability and $a'_f(g) = \operatorname{sgn}(f)\, g + |g|\, \mathbf{1}_{\{f=0\}}$ where $\mathbf{1}_A$ denotes the characteristic function of the set A. Specifically, define $a_n(f,g) = |s_n\, f + g| - s_n\, |f|$ where $s_n \nearrow \infty$ when $n \to \infty$. It is easily proved that $a_n(f,g) \to a'_f(g)$ in $\ell^\infty(\mathfrak{X})$ when $n \to \infty$. By the chain rule (Theorem 4):

$$\delta'_f(g) = \sigma'_{|f|}\left(a'_f(g)\right) = \lim_{\varepsilon \searrow 0} \sup_{A_\varepsilon(|f|)} \left(g \operatorname{sgn}(f) + |g|\, \mathbf{1}_{\{f=0\}}\right).$$

Finally, observe that if $x \in A_\varepsilon(|f|)$, then $f(x) \neq 0$, so the second term of the sum inside the brackets of the right hand side of the previous equation drops out.

As pointed out in Section 2.1, Römisch (2004, Proposition 1), provides the same result as Theorem 17 for the infimum. Obviously, the derivatives of the supremum and amplitude of a function can be derived from the infimum by duality. The additional contribution of Theorem 17 is the differentiability of the supremum norm operator, δ; and the differentiability of the absolute value. Also, the proof we have included here is slightly different to the one in Römisch (2004). The expressions in (2.4) will be used throughout Sections 2.3–2.6 to obtain simplified expressions of the derivatives. The key point is that, under additional conditions such as compactness or total boundedness, convergent subsequences can be taken from the (maximizing) sequence of Equation (2.8).

Theorem 17 ensures that the functionals in (2.1) are Hadamard directionally differentiable. Nevertheless, in general these maps are *not* uniformly Hadamard differentiable (see Definition 7) as the following example shows.

Example 20. Let $\mathfrak{X}$ be the interval $[0,1]$ in $\mathbb{R}$ and we consider the function $f \equiv 1$. For $x \in [0,1]$ and $n \in \mathbb{N}$, let $f_n(x) = 1 + \frac{x}{n}$, $g(x) = 1 - x$, and $s_n = n$. We have that $f_n \to f$ in $\ell^\infty(\mathfrak{X})$ and it is easy to check that $\sigma_n\left(f_n, g\right) = 0$, where σ_n is given in (2.5). However, $\sigma'_f(g) = \sup_{[0,1]}(g) = 1$. We conclude that σ is not uniformly Hadamard differentiable, and therefore neither are the rest of the maps in (2.1).

Following the same ideas as in the proof of Theorem 17, the following partial result can be proved.

Corollary 21. *Let δ, σ, ι and α be as in* (2.1). *For each f, $g \in \ell^\infty(\mathfrak{X})$ and all sequences $(t_n)_{n\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}$, $(f_n)_{n\in\mathbb{N}}$, $(g_n)_{n\in\mathbb{N}} \in \ell^\infty(\mathfrak{X})^{\mathbb{N}}$ such that $t_n \searrow 0$, $f_n \to f$ and $g_n \to g$ in $\ell^\infty(\mathfrak{X})$, we have that*

$$\begin{aligned} \limsup_{n\to\infty} \frac{\delta\left(f_n + t_n\, g_n\right) - \delta\left(f_n\right)}{t_n} &\leq \delta'_f(g), & \limsup_{n\to\infty} \frac{\sigma\left(f_n + t_n\, g_n\right) - \sigma\left(f_n\right)}{t_n} &\leq \sigma'_f(g), \\ \liminf_{n\to\infty} \frac{\iota\left(f_n + t_n\, g_n\right) - \iota\left(f_n\right)}{t_n} &\geq \iota'_f(g), & \limsup_{n\to\infty} \frac{\alpha\left(f_n + t_n\, g_n\right) - \alpha\left(f_n\right)}{t_n} &\leq \alpha'_f(g), \end{aligned} \tag{2.11}$$

where δ'_f, σ'_f, ι'_f and α'_f are given in (2.4).

In general, the reverse inequalities in (2.11) fail to hold because it is not possible to control the term $\frac{(\varphi(f_n)-\varphi(f))}{t_n}$ (for $\varphi \in \{\delta, \sigma, \iota, \alpha\}$), for all sequences $(t_n)\, n \in \mathbb{N} \in \mathbb{R}$ and $(f_n)\, n \in \mathbb{N} \in \ell^\infty(\mathfrak{X})$ such that $t_n \searrow 0$ and $f_n \to f$.

2.3 Compact metric spaces

One of the most common occasions when the limit in ε of the derivatives in (2.4) can be removed is when $\mathfrak{X}$ is a compact metric space. The derivatives can be characterized by means of convergent sequences in $\mathfrak{X}$ as the following corollary shows.

Corollary 22. *In the context of Theorem 17, let us further assume that* $(\mathfrak{X}, d)$ *is a compact metric space. The derivatives in* (2.4) *can be expressed as*

$$\begin{aligned} \delta'_f(g) &= \sup_{A_0(|f|)} \left((g \operatorname{sgn}(f))^{\blacktriangle}_{|f|} \right), & \sigma'_f(g) &= \sup_{A_0(f)} \left(g^{\blacktriangle}_f \right), \\ \iota'_f(g) &= \inf_{B_0(f)} \left(g^{\blacktriangledown}_f \right), & \alpha'_f(g) &= \sup_{A_0(f)} \left(g^{\blacktriangle}_f \right) - \inf_{B_0(f)} \left(g^{\blacktriangledown}_f \right), \end{aligned} \tag{2.12}$$

where for $h, l \in \ell^\infty(\mathfrak{X})$,

$$\begin{aligned} A_0(h) &= \left\{ x \in \mathfrak{X} : \exists \, (x_n)_{n \in \mathbb{N}} \in \mathfrak{X}^{\mathbb{N}} \text{ such that } x_n \to x \text{ and } h(x_n) \to \sup_{\mathfrak{X}} (h) \right\}, \\ B_0(h) &= \left\{ x \in \mathfrak{X} : \exists \, (x_n)_{n \in \mathbb{N}} \in \mathfrak{X}^{\mathbb{N}} \text{ such that } x_n \to x \text{ and } h(x_n) \to \inf_{\mathfrak{X}} (h) \right\}, \end{aligned} \tag{2.13}$$

$$\begin{aligned} h^{\blacktriangle}_l(x) &= \sup \left(\left\{ \limsup_{n \to \infty} \left(h(x_n) \right) : x_n \to x \text{ and } l(x_n) \to \sup_{\mathfrak{X}} (l) \right\} \right), \quad x \in A_0(l), \\ h^{\blacktriangledown}_l(x) &= \inf \left(\left\{ \liminf_{n \to \infty} \left(h(x_n) \right) : x_n \to x \text{ and } l(x_n) \to \inf_{\mathfrak{X}} (l) \right\} \right), \quad x \in B_0(l). \end{aligned} \tag{2.14}$$

Proof. We only give a detailed proof for σ because the rest of the cases are analogous. We consider the sequence $(x_m) m \in \mathbb{N}$ satisfying (2.8) obtained in Theorem 17. As $(\mathfrak{X}, d)$ is compact, we can extract a convergent subsequence $x_{m_k} \to x$ in $\mathfrak{X}$, as $k \to \infty$. From (2.8), we have that $x \in A_0(f)$ and, recalling (2.7), from Theorem 17, we obtain that

$$\sigma'_f(g) = \lim_{k \to \infty} h\left(\frac{1}{m_k} \right) \le \limsup_{k \to \infty} g(x_{m_k}) \le g^{\blacktriangle}_f(x) \le \sup_{A_0(f)} \left(g^{\blacktriangle}_f \right). \tag{2.15}$$

In the other direction, let $x \in A_0(f)$ and $(x_n)_{n \in \mathbb{N}} \in \mathfrak{X}^{\mathbb{N}}$ such that $x_n \to x$ and $f(x_n) \to \sup_{\mathfrak{X}} (f)$. For each $\varepsilon > 0$, we have that $x_n \in A_\varepsilon(f)$, for n large enough. We therefore conclude that

$$\limsup_{n \to \infty} g(x_n) \le \sup_{A_\varepsilon(f)} (g), \qquad \text{for all } \varepsilon > 0. \tag{2.16}$$

The conclusion follows from (2.15), (2.16) and Theorem 17. □

Remark 23. From the proof of Corollary 22 we see that the result is still valid for sequentially compact topological spaces. Nevertheless, this extension is not relevant for the applications considered in this thesis and it is therefore omitted in what follows.

In the following, if $(\mathfrak{X}, d)$ is a metric space we denote by $\mathcal{C}_{\mathrm{b}}(\mathfrak{X}, d)$ the subset of $\ell^\infty(\mathfrak{X})$ constituted by continuous functions. By the extreme values theorem, if $(\mathfrak{X}, d)$ is compact, $\mathcal{C}_{\mathrm{b}}(\mathfrak{X}, d) = \mathcal{C}(\mathfrak{X}, d)$, where the latter is the space of continuous functions on $(\mathfrak{X}, d)$. We observe that if $g \in \mathcal{C}_{\mathrm{b}}(\mathfrak{X}, d)$, then $g^{\blacktriangle}_f(x) = g(x)$ $(x \in A_0(f))$ and $g^{\blacktriangledown}_f(x) = g(x)$

($x \in B_0(f)$), where $g_f^{\blacktriangle}$ and $g_f^{\blacktriangledown}$ are defined as in (2.14). If we further assume that $f \in \mathcal{C}_{\mathrm{b}}(\mathfrak{X}, d)$, we have that $A_0(|f|) = M^+(|f|)$, $A_0(f) = M^+(f)$ and $B_0(f) = M^-(f)$, where for $h \in \ell^\infty(\mathfrak{X})$,

$$M^+(h) = \left\{x \in \mathfrak{X} : h(x) = \sup_{\mathfrak{X}} (h)\right\} \quad \text{and} \quad M^-(h) = \left\{x \in \mathfrak{X} : h(x) = \inf_{\mathfrak{X}} (h)\right\}. \tag{2.17}$$

This observation yields the following corollary.

Corollary 24. *Let $(\mathfrak{X}, d)$ be a compact metric space and let δ, σ, ι and α be the maps defined in* (2.1). *The maps σ, ι and α are Hadamard directionally differentiable at any $f \in \ell^\infty(\mathfrak{X})$ tangentially to the set $\mathcal{C}(\mathfrak{X}, d)$ with derivatives, for $g \in \mathcal{C}(\mathfrak{X}, d)$,*

$$\sigma_f'(g) = \sup_{A_0(f)} (g), \quad \iota_f'(g) = \inf_{B_0(f)} (g) \quad \text{and} \quad \alpha_f'(g) = \sup_{A_0(f)} (g) - \inf_{B_0(f)} (g). \tag{2.18}$$

If additionally $f \in \mathcal{C}(\mathfrak{X}, d) \setminus \{0\}$, we have that

$$\begin{aligned} \delta_f'(g) &= \sup_{M^+(|f|)} (g \,\mathrm{sgn}(f)), & \sigma_f'(g) &= \sup_{M^+(f)} (g), \\ \iota_f'(g) &= \inf_{M^-(f)} (g), & \alpha_f'(g) &= \sup_{M^+(f)} (g) - \inf_{M^-(f)} (g), \end{aligned} \tag{2.19}$$

where $M^+(\cdot)$ and $M^-(\cdot)$ are defined in (2.17).

Observe that $M^+(|f|)$ (respectively, $M^+(f)$ and $M^-(f)$) in (2.17) is the set of extremal points corresponding to the sup-norm (respectively, the supremum and infimum) of f.

The expression of the derivative σ_f' in (2.19) for continuous functions defined on a compact metric space has been previously obtained in Fang and Santos (2019, Lemma S.4.9). Observe that characterizations in (2.18) are valid even when the function f is not continuous (as in the more general Corollary 22). Note also that $M^+(|f|)$ (respectively, $M^+(f)$ and $M^-(f)$) in (2.17) is the set of extremal points corresponding to the sup-norm (respectively, the supremum and infimum) of f.

Another interesting question is to find conditions under which the derivatives of the maps are linear, i.e., the cases in which the mappings are fully Hadamard differentiable. This kind of results can be traced back to Banach (1932) (see also Leonard and Taylor (1983), Leonard and Taylor (1985), and the references therein). In these works the supremum norm differentiability was investigated from the point of view of functional analysis within the space $\mathcal{C}(\mathfrak{X}, d)$, with $(\mathfrak{X}, d)$ a compact metric space. The following result, a direct consequence of Corollary 24, provides general outcomes in a different context. We denote by $\mathrm{Card}(A)$ the cardinal of the set A.

Corollary 25. *Assume that $(\mathfrak{X}, d)$ is a compact metric space and let $f \in \ell^\infty(\mathfrak{X}) \setminus \{0\}$. Let $A_0(\cdot)$ and $B_0(\cdot)$ be the sets in* (2.13). *For the maps defined in* (2.1) *we have that:*

(a) *The map δ is (fully) Hadamard differentiable at f tangentially to the set $\mathcal{C}(\mathfrak{X}, d)$ if and only if* $\mathrm{Card}\,(A_0(|f|)) = 1$ *and* $\left\{\limsup_{n\to\infty} \mathrm{sgn}\,(f\,(x_n)) : x_n \to x \text{ and } |f\,(x_n)| \to \|f\|_\infty\right\} = \{c\}$. *In such a case, $\delta_f'(g) = c\,g\,(x^*)$, where $A_0(|f|) = \{x^*\}$.*

(b) *The map σ is (fully) Hadamard differentiable at f tangentially to the set $\mathcal{C}(\mathfrak{X},d)$ if and only if* $\operatorname{Card}(A_0(f)) = 1$. *In such a case, $\sigma'_f(g) = g(x^+)$, where $A_0(f) = \{x^+\}$.*

(c) *The map ι is (fully) Hadamard differentiable at f tangentially to the set $\mathcal{C}(\mathfrak{X},d)$ if and only if* $\operatorname{Card}(B_0(f)) = 1$. *In such a case, $\iota'_f(g) = g(x^-)$, where $B_0(f) = \{x^-\}$.*

(d) *The map α is (fully) Hadamard differentiable at f tangentially to the set $\mathcal{C}(\mathfrak{X},d)$ if and only if* $\operatorname{Card}(A_0(f)) = \operatorname{Card}(B_0(f)) = 1$. *In such a case, $\alpha'_f(g) = g(x^+) - g(x^-)$, where $A_0(f) = \{x^+\}$ and $B_0(f) = \{x^-\}$.*

Note that when $f \in \mathcal{C}(\mathfrak{X},d)$, $A_0(|f|) = M^+(|f|)$ in (2.17) and the condition $\operatorname{Card}(A_0(|f|)) = 1$ means that f is a *peaking function*, that is, there exists $x^* \in \mathfrak{X}$ such that $|f(x^*)| = \|f\|_\infty$ and $|f(x^*)| > |f(x)|$, for all $x \in \mathfrak{X}$ with $x \neq x^*$.

From a statistical point of view, identifying the cases in which the maps are Hadamard differentiable has two important consequences when the limit in (2.2) is Gaussian: firstly, as the linear derivatives are (essentially) the evaluation at an appropriate point, by the extended Delta method (see Proposition 5), the asymptotic distribution of the statistic in (2.3) is normal; secondly, the standard bootstrap for (2.3) is consistent if and only if the underlying map φ is fully Hadamard differentiable (see Fang and Santos (2019)).

2.4 Totally bounded metric spaces

If T is a tight Borel measurable map into $\ell^\infty(\mathfrak{X})$ as in (2.2), then there is a pseudo-metric on $\mathfrak{X}$ such that the sample paths of T are uniformly continuous and $\mathfrak{X}$ is totally bounded (see A. van der Vaart and Wellner (1996, Lemma 1.5.9)). For statistical applications it is therefore important to determine conditions under which the derivatives in (2.4) have similar expressions as those in Corollary 24 when the underlying space is totally bounded.

We recall that if $(\mathfrak{X},d)$ is a totally bounded metric space, $(\overline{\mathfrak{X}},d)$ is a compact metric space, where $\overline{\mathfrak{X}}$ is the completion of $\mathfrak{X}$ with respect to d. Further, the space $\mathcal{C}_{\mathrm{bu}}(\mathfrak{X},d)$ of bounded and uniformly continuous functions $f : \mathfrak{X} \longrightarrow \mathbb{R}$ is isometric to $\mathcal{C}(\overline{\mathfrak{X}},d)$. Indeed, by the extreme values theorem, if $(\mathfrak{X},d)$ is a totally bounded metric space, then $\mathcal{C}_{\mathrm{bu}}(\mathfrak{X},d) = \mathcal{C}_{\mathrm{u}}(\mathfrak{X},d)$. Indeed, each $f \in \mathcal{C}_{\mathrm{u}}(\mathfrak{X},d)$ has a unique extension to a function $\overline{f} \in \mathcal{C}(\overline{\mathfrak{X}},d)$. For $x \in \overline{\mathfrak{X}} \setminus \mathfrak{X}$, this extension is defined by $\overline{f}(x) = \lim_{n\to\infty} f(x_n)$, with $(x_n)_{n\in\mathbb{N}} \in \mathfrak{X}^{\mathbb{N}}$ such that $x_n \to x$. (In fact, Cauchy-continuity is enough to check that $\overline{f}$ is well-defined, but uniform continuity suffices for our purposes.)

In this setting, it is straightforward to check that Corollary 22 still holds if we substitute the sets $A_0(\cdot)$ and $B_0(\cdot)$ by

$$\begin{aligned}
\overline{A}_0(h) &= \left\{x \in \overline{\mathfrak{X}} : \exists\,(x_n)_{n\in\mathbb{N}} \in \mathfrak{X}^{\mathbb{N}} \text{ such that } x_n \to x \text{ and } h(x_n) \to \sup_{\mathfrak{X}}(h)\right\},\\
\overline{B}_0(h) &= \left\{x \in \overline{\mathfrak{X}} : \exists\,(x_n)_{n\in\mathbb{N}} \in \mathfrak{X}^{\mathbb{N}} \text{ such that } x_n \to x \text{ and } h(x_n) \to \inf_{\mathfrak{X}}(h)\right\},
\end{aligned} \tag{2.20}$$

for $h \in \ell^\infty(\mathfrak{X})$. In particular, the following corollary, important for the statistical applications included in Section 3.4 and Chapter 5, holds.

Corollary 26. *Let* $(\mathfrak{X}, d)$ *be a totally bounded metric space and let* δ, σ, ι *and* α *be the maps defined in* (2.1).

(a) *The maps* σ, ι *and* α *are Hadamard directionally differentiable at* $f \in \ell^\infty(\mathfrak{X})$ *tangentially to the set* $\mathcal{C}_{\mathrm{u}}(\mathfrak{X}, d)$ *with derivatives, for* $g \in \mathcal{C}_{\mathrm{u}}(\mathfrak{X}, d)$,

$$\sigma'_f(g) = \sup_{\overline{A}_0(f)} (\overline{g}), \quad \iota'_f(g) = \inf_{\overline{B}_0(f)} (\overline{g}) \quad \text{and} \quad \alpha'_f(g) = \sup_{\overline{A}_0(f)} (\overline{g}) - \inf_{\overline{B}_0(f)} (\overline{g}),$$

where $\overline{A}_0(\cdot)$ *and* $\overline{B}_0(\cdot)$ *are defined in* (2.20).

(b) *If additionally* $f \in \mathcal{C}_{\mathrm{u}}(\mathfrak{X}, d) \setminus \{0\}$, *we have that*

$$\begin{aligned} \delta'_f(g) &= \sup_{\overline{M}^+(|f|)} \left(\overline{g}\,\mathrm{sgn}\left(\overline{f}\right)\right), & \sigma'_f(g) &= \sup_{\overline{M}^+(|f|)} (\overline{g}), \\ \iota' f(g) &= \inf_{\overline{M}^-(|f|)} (\overline{g}), & \alpha'_f(g) &= \sup_{\overline{M}^+(|f|)} (\overline{g}) - \inf_{\overline{M}^-(|f|)} (\overline{g}), \end{aligned}$$

where for $h \in \mathcal{C}_{\mathrm{u}}(\mathfrak{X}, d)$,

$$\overline{M}^+(h) = \left\{x \in \overline{\mathfrak{X}} : \overline{h}(x) = \sup_{\mathfrak{X}} (h)\right\} \quad \text{and} \quad \overline{M}^-(h) = \left\{x \in \overline{\mathfrak{X}} : \overline{h}(x) = \inf_{\mathfrak{X}} (h)\right\}. \tag{2.21}$$

Remark 27. Corollary 25 still holds if $(\mathfrak{X}, d)$ is a totally bounded metric space and we replace $\mathcal{C}(\mathfrak{X}, d)$, $A_0(\cdot)$ and $B_0(\cdot)$ with $\mathcal{C}_{\mathrm{u}}(\mathfrak{X}, d)$, $\overline{A}_0(\cdot)$ and $\overline{B}_0(\cdot)$ (defined in (2.20)), respectively.

2.5 Weakly compact sets

The compacteness assumption on $\mathfrak{X}$ in Corollaries 22 and 24 could be too demanding in some infinite-dimensional settings. A simple inspection of the proof of Corollary 22 shows that a similar result can be stated when $\mathfrak{X}$ is a weakly compact subset of a Banach space by using Eberlein–Šmulian's theorem (see Conway (2019, p. 163)). In such a case, Corollary 22 still holds by substituting the sets $A_0(h)$ and $B_0(h)$ in (2.13) and the quantities $h^{\blacktriangle}_f(x)$ and $h^{\blacktriangledown}_f(x)$ in (2.14) respectively by

$$\begin{aligned} A_0^w(h) &= \left\{x \in \mathfrak{X} : \exists\, (x_n)\, n \in \mathbb{N} \in \mathfrak{X}^{\mathbb{N}} \text{ such that } x_n \rightharpoonup x \text{ and } h(x_n) \to \sup_{\mathfrak{X}} (h)\right\}, \\ B_0^w(h) &= \left\{x \in \mathfrak{X} : \exists\, (x_n)\, n \in \mathbb{N} \in \mathfrak{X}^{\mathbb{N}} \text{ such that } x_n \rightharpoonup x \text{ and } h(x_n) \to \inf_{\mathfrak{X}} (h)\right\}, \end{aligned} \tag{2.22}$$

and

$$\begin{aligned} h_f^{\blacktriangle,w}(x) &= \sup\left(\left\{\limsup_{n\to\infty} h(x_n) : x_n \rightharpoonup x \text{ and } f(x_n) \to \sup_{\mathfrak{X}} (f)\right\}\right), \quad x \in A_0^w(f), \\ h_f^{\blacktriangle,w}(x) &= \inf\left(\left\{\liminf_{n\to\infty} h(x_n) : x_n \rightharpoonup x \text{ and } f(x_n) \to \inf_{\mathfrak{X}} (f)\right\}\right), \quad x \in A_0^w(f), \end{aligned} \tag{2.23}$$

where $x_n \rightharpoonup x$ stands for the weak convergence in the corresponding space. We recall that if $(x_n) \in \mathcal{B}$ with $\mathcal{B}$ a Banach space, $x_n \rightharpoonup x$ means that $\psi(x_n) \to \psi(x)$ for all $\psi \in \mathcal{B}^*$, the topological dual space of $\mathcal{B}$ formed by linear and continuous functionals from $\mathcal{B}$ to $\mathbb{R}$. If $\mathcal{B} = \mathcal{H}$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, by the Riesz's representation theorem (Conway (2019, Chapter 1.3)) the weak convergence amounts to $\langle x_n, y \rangle_{\mathcal{H}} \to \langle x, y \rangle_{\mathcal{H}}$, for all $y \in \mathcal{H}$.

In this context, we have analogous results as Corollaries 24 and 26 by changing the set of tangency points to the space of continuous and pre-linear functions $\mathcal{C}_{\mathrm{bpl}}(\mathfrak{X}, d)$. Every prelinear function g defined on $\mathfrak{X}$ (10) admits a unique extension to a linear function on $\operatorname{span}(\mathfrak{X})$, the linear span of $\mathfrak{X}$ (see Dudley (2014, Lemma 2.30, p. 88)). This extension is given by

$$\widetilde{g}\left(\sum_{i=1}^{r} \lambda_i x_i\right) = \sum_{i=1}^{r} \lambda_i g(x_i), \quad \text{with } x_i \in \mathfrak{X} \text{ and } \lambda_i \in \mathbb{R} \ (i = 1, \ldots, r). \tag{2.24}$$

Further, if $\mathcal{B}$ is a Banach space with norm $\|\cdot\|$, $d_{\mathcal{B}}$, the metric on $\mathcal{B}$, i.e., $d_{\mathcal{B}}(x,y) = \|x - y\|$ $(x, y \in \mathcal{B})$; is the usual metric considered in $\mathfrak{X}$.

Corollary 28. *Let $\mathcal{B}$ be a Banach space and let δ, σ, ι and α be the maps in* (2.1)*. Let us assume that the set $\mathfrak{X} \subset \mathcal{B}$ satisfies the following two conditions:*

(i) *$\mathfrak{X}$ is a weakly compact subset of $\mathcal{B}$.*

(ii) *For each $g \in \mathcal{C}_{\mathrm{bpl}}(\mathfrak{X}, d_{\mathcal{B}})$, its linear extension $\widetilde{g}$ in* (2.24) *is continuous on* $\operatorname{span}(\mathfrak{X})$.

Then, the maps σ, ι and α are Hadamard directionally differentiable at $f \in \ell^\infty(\mathfrak{X})$ tangentially to $\mathcal{C}_{\mathrm{bpl}}(\mathfrak{X}, d_{\mathcal{B}})$ with derivatives, for $g \in \mathcal{C}_{\mathrm{bpl}}(\mathfrak{X}, d_{\mathcal{B}})$,

$$\sigma'_f(g) = \sup_{A_0^w(f)}(g), \quad \iota'_f(g) = \inf_{B_0^w(f)}(g), \quad \text{and} \quad \alpha'_f(g) = \sup_{A_0^w(f)}(g) - \inf_{B_0^w(f)}(g),$$

where $A_0^w(\cdot)$ and $B_0^w(\cdot)$ are defined in (2.22)*.*

If additionally $f \in \mathcal{C}_{\mathrm{bpl}}(\mathfrak{X}, d_{\mathcal{B}}) \setminus \{0\}$, then the derivatives of δ, σ, ι and α are as in (2.19)*.*

Proof. As in the previous proofs, we only discuss the map σ. Let us consider $x \in A_0^w(f)$ (defined in (2.22)) and $g \in \mathcal{C}_{\mathrm{bpl}}(\mathfrak{X}, d_{\mathcal{B}})$. We consider a sequence $(x_n)_{n \in \mathbb{N}} \in \mathfrak{X}^{\mathbb{N}}$ such that $x_n \rightharpoonup x$ and $f(x_n) \to \sup_{\mathfrak{X}}(f)$ (the existence of such a sequence is guaranteed by condition (i) and 2.8). Condition (ii) and Hahn–Banach's theorem imply that there exists a linear and continuous map, say $\overline{g}$, defined on $\mathcal{B}$ such that $\overline{g} = \widetilde{g}$ on $\operatorname{span}(\mathfrak{X})$, and hence $\overline{g} = g$ on $\mathfrak{X}$. As $\overline{g} \in \mathcal{B}^*$ and $x_n \rightharpoonup x$, we conclude that $\lim_{n \to \infty} g(x_n) = g(x)$. This shows that $g_f^{\blacktriangle,w}(x) = g(x)$, with $g_f^{\blacktriangle,w}(x)$ defined as in (2.23), and the conclusion follows from the observation at the beginning of this section.

Finally, if $f \in \mathcal{C}_{\mathrm{bpl}}(\mathfrak{X}, d_{\mathcal{B}})$, the same argument used before shows that $A_0^w(f) = M^+(f)$, where the set $M^+(\cdot)$ is defined in (2.17). □

We observe that hypothesis (i) in the previous corollary is essential to extract a weakly convergent subsequence in $\mathfrak{X}$. We also observe that condition (ii) cannot be dropped as, in general, the linear extension $\widetilde{g}$ of a function $g \in \mathcal{C}_{\text{bpl}}(\mathfrak{X}, d_{\mathcal{B}})$ is not necessarily continuos in $\text{span}(\mathfrak{X})$ as the following example shows: Let $\mathcal{B}$ be an infinite-dimensional and Banach space with norm $\|\cdot\|$. We consider $\mathfrak{X} = (x_n)_{n\in\mathbb{N}} \in \mathcal{B}^{\mathbb{N}}$, where $x_0 = 0$ and $\{x_n\}_{n=1}^{\infty}$ is a linearly independent subset of $\mathcal{B}$ such that $\|x_n\| = \frac{1}{n}$ $(n \in \mathbb{N})$. It is easy to check that the function defined by $g(0) = 0$ and $g(x_n) = \frac{1}{\sqrt{n}}$ $(n \in \mathbb{N})$ belongs to $\mathcal{C}_{\text{bpl}}(\mathfrak{X}, d_{\mathcal{B}})$, but its linear extension $\widetilde{g}$ is *not* continuous because it is not bounded on the unit sphere since $\widetilde{g}\left(\frac{x_n}{\|x_n\|}\right) = \sqrt{n}$ $(n \in \mathbb{N})$.

The following proposition provides easy to check conditions guaranteeing that Corollary 28 (ii) is fulfilled.

Proposition 29. *Let $\mathcal{B}$ be a Banach space with norm $\|\cdot\|$ and $\mathfrak{X} \subset \mathcal{B}$. Let us assume that one of the following two conditions is satisfied:*

(a) *There exists $x \in \mathfrak{X}$ and $\delta > 0$ such that $B(x, \delta) = \{y \in \mathcal{B} : \|y - x\| \leq \delta\} \subset \mathfrak{X}$.*

(b) *$\mathcal{B}$ is a Hilbert space and there exists $\{x_i\}_{i\in I} \subset \mathfrak{X}$ (I arbitrary index set) such that $\text{span}(\mathfrak{X}) = \text{span}(\{x_i\}_{i\in I})$, $\{x_i\}_{i\in I}$ are pairwise orthogonal and $c = \inf_{i\in I}(\|x_i\|) > 0$.*

Then, for each $g \in \mathcal{C}_{\text{bpl}}(\mathfrak{X}, d_{\mathcal{B}})$, its linear extension $\widetilde{g}$ in (2.24) *is continuous on* $\text{span}(\mathfrak{X})$.

Proof. Let us assume that (a) holds. As $g \in \mathcal{C}(\mathfrak{X}, d_{\mathcal{B}})$, the condition $B(x, \delta) \subset \mathfrak{X}$ ensures that $\widetilde{g}$ in (2.24) is continuous at x, and, by linearity, continuous on $\text{span}(\mathfrak{X})$.

Assume now that (b) is satisfied. For $x \in \text{span}(\mathfrak{X})$, we can write $x = \sum_{i=1}^{r} \lambda_i x_i$, with $\lambda_i \in \mathbb{R}$ $(i = 1, \ldots, r)$. Taking into account that $\|x\| = \sum_{i=1}^{r} |\lambda_i| \, \|x_i\| \geq c \sum_{i=1}^{r} |\lambda_i|$, we finally obtain that

$$|\widetilde{g}(x)| \leq \|g\|_{\infty} \sum_{i=1}^{r} |\lambda_i| \leq \frac{\|g\|_{\infty} \, \|x\|}{c}.$$

The previous inequalities show that $\widetilde{g}$ is continuous on $\text{span}(\mathfrak{X})$ and the proof is complete. □

Closed bounded convex subsets of a reflexive Banach space are weakly compact (see Brezis and Brézis (2011, Corollary 3.22)). Therefore, the hypotheses of Corollary 28 are general enough to include many infinite-dimensional sets. Thanks to Proposition 29, an important example covered by Corollary 28 is when $\mathfrak{X}$ is the closed unit ball of a reflexive Banach space, and, in particular, the closed unit ball of a Hilbert space. On the other hand, working with prelinear functions could seem to be too restrictive. However, we point out that if P is a probability measure and a set $\mathfrak{X}$ is P-pre-Gaussian (Giné and Nickl (2021, Definition 3.7.26, p. 251)), there is a version of the P-bridge whose sample paths are prelinear (see Giné and Nickl (2021, Theorem 3.7.28, p. 252)). Such a version is usually called *suitable*.

Remark 30. Corollary 25 still holds with the obvious modifications if $\mathfrak{X}$ is in the conditions of Corollary 28. It is enough to replace convergence with weak convergence and $\mathcal{C}(\mathfrak{X}, d)$, $A_0(\cdot)$ and $B_0(\cdot)$ with $\mathcal{C}_{\text{bpl}}(\mathfrak{X}, d_{\mathcal{B}})$, $A_0^w(\cdot)$ and $B_0^w(\cdot)$, respectively.

2.6 The case $\mathfrak{X} = \overline{\mathbb{R}}^d$ and the Skorohod space $\mathcal{D}\left(\overline{\mathbb{R}}^d\right)$

Throughout this section $\mathfrak{X} = \overline{\mathbb{R}}^d$ $(d \geq 1)$ endowed with d_e, the metric corresponding to the Euclidean norm on $[0,1]^d$ through a given homeomorphism. Hence, $\left(\overline{\mathbb{R}}^d, d_e\right)$ is a compact metric space and we can apply Corollaries 22 and 24 in Section 2.3.

Many important stochastic processes take values in the one-dimensional Skorohod space, $\mathcal{D}\left(\overline{\mathbb{R}}\right)$, consisting of all the *càdlàg* functions, that is, right-continuous functions having limit from the left at every point (*continue à droite, limite à gauche*). This space provides a natural and convenient setting to analyze the behaviour of processes with unidimensional time parameter and jumps in their paths such as Poisson processes, Lévy processes, empirical processes or discretizations of stochastic processes, among others. Skorohod-type spaces are usually equipped with different norms to make them separable. However, we are only interested in a multidimensional extension of the Skorohod space viewed as a subset of $\ell^\infty\left(\overline{\mathbb{R}}^d\right)$ with the supremum norm. The final aim of this section is to provide alternative expressions for the directional derivatives in (2.12) when the involved functions belong to the d-dimensional Skorohod space.

The d-dimensional Skorohod space, introduced in Neuhaus (1971) (see also Bickel and Wichura (1971)) and more recently considered in Seijo and Sen (2011)), is usually defined in compact rectangles of $\mathbb{R}^d$. We will firstly extend this space to functions defined in $\overline{\mathbb{R}}^d$.

For $v \in \{-1,1\}$ and $x \in \overline{\mathbb{R}}$, let

$$I_v(x) = \begin{cases} [-\infty, x), & \text{if } v = -1,\ x \in \overline{\mathbb{R}}, \\ (x, +\infty], & \text{if } v = 1,\ x \in \overline{\mathbb{R}}, \end{cases}$$

and

$$\widetilde{I}_v(x) = \begin{cases} [-\infty, x), & \text{if } v = -1,\ x < \infty, \\ \overline{\mathbb{R}}, & \text{if } v = -1,\ x = -\infty, \\ \varnothing, & \text{if } v = 1,\ x = \infty, \\ [x, \infty], & \text{if } v = 1,\ x < \infty. \end{cases}$$

We consider $\mathcal{V} = \{-1,1\}^d$ the set of 2^d vertices of $[-1,1]^d$. For $\mathbf{v} = (v_1, \ldots, v_d) \in \mathcal{V}$ and $\mathbf{x} = (x_1, \ldots, x_d) \in \overline{\mathbb{R}}^d$, we define the $\mathbf{v}$-quadrants of $\mathbf{x}$ by

$$Q_{\mathbf{v}}(\mathbf{x}) = I_{v_1}(x_1) \times \cdots \times I_{v_d}(x_d) \quad \text{and} \quad \widetilde{Q}_{\mathbf{v}}(\mathbf{x}) = \widetilde{I}_{v_1}(x_1) \times \cdots \times \widetilde{I}_{v_d}(x_d).$$

Observe that $Q_{\mathbf{v}}(\mathbf{x}) \subset \widetilde{Q}_{\mathbf{v}}(\mathbf{x})$, $\widetilde{Q}_{\mathbf{v}}(\mathbf{x}) \cap \widetilde{Q}_{\mathbf{v}'}(\mathbf{x}) = \varnothing$ whenever $\mathbf{v}$, $\mathbf{v}' \in \mathcal{V}$ with $\mathbf{v} \neq \mathbf{v}'$, and $\bigcup_{\mathbf{v}\in\mathcal{V}} \widetilde{Q}_{\mathbf{v}}(\mathbf{x}) = \overline{\mathbb{R}}^d$, for all $\mathbf{x} \in \overline{\mathbb{R}}^d$. Additionally, for each $\mathbf{x} \in \overline{\mathbb{R}}^d$, there exists a unique $\mathbf{v}_{\mathbf{x}} \in \mathcal{V}$ such that $\mathbf{x} \in \widetilde{Q}_{\mathbf{v}_{\mathbf{x}}}(\mathbf{x})$. For instance, if $\mathbf{x} \in \mathbb{R}^d$, we have that $\mathbf{v}_{\mathbf{x}=\mathbf{1}}$, where $\mathbf{1} = (1, \ldots, 1)$.

With the previous concepts we can define the quadrant limits. Let us consider a function $f : \overline{\mathbb{R}}^d \longrightarrow \mathbb{R}$, $\mathbf{v} \in \mathcal{V}$ and $\mathbf{x} \in \overline{\mathbb{R}}^d$. We say that $l \in \mathbb{R}$ is the $\mathbf{v}$-limit of f at $\mathbf{x}$ if $Q_{\mathbf{v}}(\mathbf{x}) \neq \varnothing$ and for every sequence $(\mathbf{x}_n)_{n\in\mathbb{N}} \in Q_{\mathbf{v}}(\mathbf{x})^{\mathbb{N}}$ such that $\mathbf{x}_n \to \mathbf{x}$, we have that $f(\mathbf{x}_n) \to l$. In such a case, we denote $l \equiv f_{\mathbf{v}}(\mathbf{x})$. Additionally, it is said that f is continuous from above at $\mathbf{x} \in \overline{\mathbb{R}}^d$ if $f_{\mathbf{v}_{\mathbf{x}}}(\mathbf{x})$ exists and $f_{\mathbf{v}_{\mathbf{x}}}(\mathbf{x}) = f(\mathbf{x})$. We say that f is continuous from above if it is continuous from above at every $\mathbf{x} \in \overline{\mathbb{R}}^d$.

Definition 31. The *Skorohod space* on $\overline{\mathbb{R}}^d$, denoted by $\mathcal{D}\left(\overline{\mathbb{R}}^d\right)$, is the collection of all continuous from above real functions f defined in $\overline{\mathbb{R}}^d$ for which the $\mathbf{v}$-limit of f exists for every $\mathbf{v} \in \mathcal{V}$ and $\mathbf{x} \in \overline{\mathbb{R}}^d$ such that $Q_{\mathbf{v}}(\mathbf{x}) \neq \varnothing$.

When $d = 1$, $\mathcal{D}\left(\overline{\mathbb{R}}\right)$ is usual Skorohod space on $\overline{\mathbb{R}}$. The properties of the multidimensional Skorohod space in $[0,1]^d$ shown in Neuhaus (1971) can be extended with no difficulty to $\mathcal{D}\left(\overline{\mathbb{R}}^d\right)$. For instance, the elements in $\mathcal{D}\left(\overline{\mathbb{R}}^d\right)$ belong to $\mathcal{D}(\overline{\mathbb{R}})$ in each coordinate, have at most countably many discontinuities and all of them are of "finite-jump type". The fact that $\mathcal{D}\left(\overline{\mathbb{R}}^d\right) \subset \ell^\infty\left(\overline{\mathbb{R}}^d\right)$ follows from Neuhaus (1971, Corollary 1.6) by noting that functions in $\mathcal{D}\left(\overline{\mathbb{R}}^d\right)$ have finite quadrant limits at infinity points.

Remark 32. We observe that if $f \in \mathcal{D}\left(\overline{\mathbb{R}}^d\right)$ and $(\mathbf{x}_n) \in \widetilde{Q}_{\mathbf{v}}(\mathbf{x})$ such that $\mathbf{x}_n \to \mathbf{x}$, then $f(\mathbf{x}_n) \to f_{\mathbf{v}}(\mathbf{x})$. This follows from the fact that

$$\widetilde{Q}_{\mathbf{v}}(\mathbf{x}) = \left\{\mathbf{y} \in \overline{\mathbb{R}}^d : \mathbf{y} \in \overline{Q_{\mathbf{v}_{\mathbf{y}}}(\mathbf{y}) \cap Q_{\mathbf{v}}(\mathbf{x})}\right\},$$

where $\overline{A}$ denotes the closure of the set A. In other words, the functions in $\mathcal{D}\left(\overline{\mathbb{R}}^d\right)$ have quadrant limits in $\widetilde{Q}_{\mathbf{v}}(\mathbf{x})$.

We are now in position to see how the derivatives in (2.12) look like when $\mathfrak{X} = \overline{\mathbb{R}}^d$ and the functions on which they act belong to $\mathcal{D}\left(\overline{\mathbb{R}}^d\right)$.

Corollary 33. *For any $f \in \mathcal{D}\left(\overline{\mathbb{R}}^d\right) \setminus \{0\}$, the maps δ, σ, ι and α in* (2.1) *are Hadamard directionally differentiable at f tangentially to $\mathcal{D}\left(\overline{\mathbb{R}}^d\right)$. For $g \in \mathcal{D}\left(\overline{\mathbb{R}}^d\right)$, their derivatives are given by*

$$\begin{aligned}
\delta_f'(g) &= \max_{\mathbf{v}\in\mathcal{V}} \left(\sup_{M_{\mathbf{v}}^+(|f|)} \left(g_{\mathbf{v}} \operatorname{sgn}(f_{\mathbf{v}})\right)\right), \\
\sigma_f'(g) &= \max_{\mathbf{v}\in\mathcal{V}} \left(\sup_{M_{\mathbf{v}}^+(f)} (g_{\mathbf{v}})\right), \\
\iota_f'(g) &= \min_{\mathbf{v}\in\mathcal{V}} \left(\inf_{M_{\mathbf{v}}^-(f)} (g_{\mathbf{v}})\right), \\
\alpha_f'(g) &= \max_{\mathbf{v}\in\mathcal{V}} \left(\sup_{M_{\mathbf{v}}^+(f)} (g_{\mathbf{v}})\right) - \min_{\mathbf{v}\in\mathcal{V}} \left(\inf_{M_{\mathbf{v}}^-(f)} (g_{\mathbf{v}})\right),
\end{aligned} \tag{2.25}$$

where for $h \in \mathcal{D}\left(\overline{\mathbb{R}}^d\right)$,

$$\begin{aligned}
M_{\mathbf{v}}^+(h) &= \left\{\mathbf{x} \in \overline{\mathbb{R}}^d : Q_{\mathbf{v}}(\mathbf{x}) \neq \varnothing \text{ and } h_{\mathbf{v}}(\mathbf{x}) = \sup_{\mathbb{R}} (h)\right\}, \\
M_{\mathbf{v}}^-(h) &= \left\{\mathbf{x} \in \overline{\mathbb{R}}^d : Q_{\mathbf{v}}(\mathbf{x}) \neq \varnothing \text{ and } h_{\mathbf{v}}(\mathbf{x}) = \inf_{\mathbb{R}} (h)\right\}.
\end{aligned} \tag{2.26}$$

Proof. This corollary can be proved as Corollary 22 by taking into account Remark 32 and the following fact: As the number of non-empty quadrants of each point in $\overline{\mathbb{R}}^d$ is finite, each sequence converging to a point $\mathbf{x} \in \overline{\mathbb{R}}^d$ has a subsequence contained in $\widetilde{Q}_{\mathbf{v}}(\mathbf{x})$, for some $\mathbf{v} \in \mathcal{V}$. In particular, for every $h \in \mathcal{D}\left(\overline{\mathbb{R}}^d\right)$, it holds that $A_0(h) = \bigcup_{\mathbf{v}\in\mathcal{V}} M_{\mathbf{v}}^+(h)$ and $B_0(h) = \bigcup_{\mathbf{v}\in\mathcal{V}} M_{\mathbf{v}}^-(h)$, where $A_0(h)$ and $B_0(h)$ are defined in (2.13). □

The sets $M_{\mathbf{v}}^{+}(h)$ (respectively, $M_{\mathbf{v}}^{-}(h)$) in (2.26) might coincide for different $\mathbf{v} \in \mathcal{V}$. For instance, when f is continuous, $M_{\mathbf{v}}^{+}(|f|) = M^{+}(|f|)$, $M_{\mathbf{v}}^{+}(f) = M^{+}(f)$, and $M_{\mathbf{v}}^{-}(f) = M^{-}(f)$, for all $\mathbf{v} \in \mathcal{V}$, where $M^{+}(\cdot)$ and $M^{-}(\cdot)$ are defined in (2.17).

We emphasize that $g_{\mathbf{v}} \equiv g$, for all $\mathbf{v} \in \mathcal{V}$, whenever $g \in \mathcal{C}\left(\overline{\mathbb{R}}^d, d_e\right)$. The following corollary is important for applications because many stochastic processes that commonly appear as weak limits of other processes have continuous paths a.s.

Corollary 34. *For any $f \in \mathcal{D}\left(\overline{\mathbb{R}}^d\right) \setminus \{0\}$, the maps δ, σ, ι and α in* (2.1) *are Hadamard directionally differentiable at f tangentially to $\mathcal{C}\left(\overline{\mathbb{R}}^d, d_e\right)$. For $g \in \mathcal{C}\left(\overline{\mathbb{R}}^d, d_e\right)$, their derivatives are given by*

$$\delta_f'(g) = \max_{\mathbf{v}\in\mathcal{V}} \left(\sup_{M_{\mathbf{v}}^{+}(|f|)} \left(g \operatorname{sgn}\left(f_{\mathbf{v}}\right)\right)\right), \qquad \sigma_f'(g) = \max_{\mathbf{v}\in\mathcal{V}} \left(\sup_{M_{\mathbf{v}}^{+}(f)} (g)\right),$$

$$\iota_f'(g) = \min_{\mathbf{v}\in\mathcal{V}} \left(\inf_{M_{\mathbf{v}}^{-}(f)} (g)\right), \qquad \alpha_f'(g) = \max_{\mathbf{v}\in\mathcal{V}} \left(\sup_{M_{\mathbf{v}}^{+}(f)} (g)\right) - \min_{\mathbf{v}\in\mathcal{V}} \left(\inf_{M_{\mathbf{v}}^{-}(f)} (g)\right),$$

with $M_{\mathbf{v}}^{+}(\cdot)$ and $M_{\mathbf{v}}^{-}(\cdot)$ defined in (2.26).

If additionally $f \in \mathcal{C}\left(\overline{\mathbb{R}}^d, d_e\right)$, the derivatives are as in (2.19).

2.7 Unit balls of RKHS

Let $\mathcal{H}_k$ be a RKHS of real-valued functions on $\mathcal{X}$. We consider the maximum mean discrepancy with respect to the class $\mathcal{F}_{\mathcal{H}_k}$, the unit ball of $\mathcal{H}_k$

$$d_k(\mathrm{P}, \mathrm{Q}) = \mathrm{MMD}\left[\mathcal{F}_{\mathcal{H}_k}, \mathrm{P}, \mathrm{Q}\right] = \sup_{f\in\mathcal{F}_{\mathcal{H}_k}} \left(\mathrm{P}(f) - \mathrm{Q}(f)\right), \tag{2.27}$$

where P and Q are two probability measures on $\mathcal{X}$ (see (4.3) and (4.7) for further details). The quantity $d_k(\mathrm{P}, \mathrm{Q})$ is usually called *kernel distance* (see Definition 49 in Chapter 3 or Definition 4.3 in Chapter 4). We will restrict ourselves to measurable functions in $\mathcal{F}_{\mathcal{H}_k}$.

Assume that P and Q are two probability measures for which their mean embeddings μ_{P} and μ_{Q} exist (see Proposition 16). An important property of the kernel distance in (2.27) is that it can be expressed as

$$d_k(\mathrm{P}, \mathrm{Q}) = \left\| \mu_{\mathrm{P}} - \mu_{\mathrm{Q}} \right\|_{\mathcal{H}_k}, \tag{2.28}$$

(see Borgwardt et al. (2006), Gretton, Borgwardt, et al. (2012, Lemma 4), or (4.7)).

The statistical properties of the estimators of kernel distances are typically obtained by using the representation in (2.28) (see for instance Gretton, Borgwardt, et al. (2012) and the references therein). However, a more versatile approach is offered in Chapter 4. Thanks to the results in Theorem 17, we can discuss the asymptotic properties of the plug-in estimator $d_k\left(\mathbb{P}_n, \mathbb{Q}_m\right)$ in (2.27) using the Delta method (see Proposition 5), with $\mathbb{P}_n$ and $\mathbb{Q}_m$ being the empirical measures associated with random samples of P and Q, respectively.

The following result shows that the mapping $\sigma(f) = \sup_{\mathcal{F}_{\mathcal{H}_k}}(f)$ $(g \in \mathcal{C}_{\mathrm{b}}(\mathcal{F}_{\mathcal{H}_k}, d_{\mathcal{H}_k}))$ is in general fully Hadamard differentiable at $\mathrm{P}-\mathrm{Q}$. At a first glance this could seem surprising as the supremum is not usually fully differentiable (see Section 1.1). However, the intuition behind this result is the following: As the kernel distance in (2.27) can be alternatively expressed as the norm in the RKHS between the mean embeddings of the corresponding probability measures (see (2.28)) and the norm in a Hilbert space is fully differentiable, then σ should be also fully differentiable at $\mathrm{P}-\mathrm{Q}$. Moreover, taking into account Corollary 25 and Remarks 27 and 30, the derivative has to be the evaluation at an appropriate function of $\mathcal{F}_{\mathcal{H}_k}$. The previous ideas are formalized in the following corollary.

Corollary 35. *Let $\mathcal{F}_{\mathcal{H}_k}$ be the unit ball in $\mathcal{H}_k$, reproducing kernel Hilbert space, and let us consider two Borel probability measures* P *and* Q *on $\mathcal{X}$ for which their mean embeddings μ_{P} and μ_{Q} exist and $\mu_{\mathrm{P}} \neq \mu_{\mathrm{Q}}$. We consider* $\mathrm{P}-\mathrm{Q}$ *as an element of $\ell^\infty(\mathcal{F}_{\mathcal{H}_k})$. We have that the mapping $\sigma(f) = sup_{\mathcal{F}_{\mathcal{H}_k}}(g)$ $(g \in \ell^\infty(\mathcal{F}_{\mathcal{H}_k}))$ is (fully) Hadamard differentiable at* $\mathrm{P}-\mathrm{Q}$ *tangentially to $\mathcal{C}_b(\mathcal{F}_{\mathcal{H}_k}, d_{\mathcal{H}_k})$ with derivative*

$$\sigma'_{\mathrm{P}-\mathrm{Q}}(g) = g(h^+), \quad \text{with} \quad h^+ = \frac{\mu_{\mathrm{P}} - \mu_{\mathrm{Q}}}{\|\mu_{\mathrm{P}} - \mu_{\mathrm{Q}}\|_{\mathcal{H}_k}}. \tag{2.29}$$

Proof. We will first check that if $h_\varepsilon \in A_\varepsilon(\mathrm{P}-\mathrm{Q})$, where $A_\varepsilon(\mathrm{P}-\mathrm{Q})$ is defined in Theorem 17, then $h_\varepsilon \to h^+$ in $\mathcal{H}_k$ as $\varepsilon \to 0$, with h^+ in (2.29). To see this, we first note that

$$\|h_\varepsilon - h^+\|^2_{\mathcal{H}_k} = 1 + \|h_\varepsilon\|^2_{\mathcal{H}_k} - \frac{2}{\|\mu_{\mathrm{P}} - \mu_{\mathrm{Q}}\|_{\mathcal{H}_k}} \langle h_\varepsilon, \mu_{\mathrm{P}} - \mu_{\mathrm{Q}}\rangle_{\mathcal{H}_k}. \tag{2.30}$$

As $h_\varepsilon \in A_\varepsilon(\mathrm{P}-\mathrm{Q})$, from (2.28), we obtain that $\langle h_\varepsilon, \mu_{\mathrm{P}} - \mu_{\mathrm{Q}}\rangle_{\mathcal{H}_k} \geq \|\mu_{\mathrm{P}} - \mu_{\mathrm{Q}}\|_{\mathcal{H}_k} - \varepsilon$. Finally, from (2.30) and as $h_\varepsilon \in \mathcal{F}_{\mathcal{H}_k}$, we have that $\|h_\varepsilon - h^+\|^2_{\mathcal{H}_k} \leq \frac{2\varepsilon}{\|\mu_{\mathrm{P}}-\mu_{\mathrm{Q}}\|_{\mathcal{H}_k}}$.

Now, we will check that $\sigma'_{\mathrm{P}-\mathrm{Q}}(g) = g(h^+)$, for $g \in \mathcal{C}_{\mathrm{b}}(\mathcal{F}_{\mathcal{H}_k}, d_{\mathcal{H}_k})$. We firstly observe that $h^+ \in A_\varepsilon(\mathrm{P}-\mathrm{Q})$, for all $\varepsilon > 0$. Hence, from Theorem 17, Equation (2.4), we have that $g(h^+) \leq \sigma'_{\mathrm{P}-\mathrm{Q}}(g)$. On the other hand, as in the proof of Theorem 17, we can extract a sequence $h_m \in \mathcal{F}_{\mathcal{H}_k}$ satisfying that $\sup_{A_{1/m}(\mathrm{P}-\mathrm{Q})}(\mathrm{P}-\mathrm{Q})(g) \leq g(h_m) + \frac{1}{m}$. As g is continuous and $h_m \to h^+$ as $m \to \infty$, we obtain that $\sigma'_{\mathrm{P}-\mathrm{Q}}(g) = \lim_{m\to\infty} \sup_{A_{1/m}(\mathrm{P}-\mathrm{Q})}(g) \leq g(h^+)$, and the proof is complete. □

On one hand, a careful glance at Corollary 35 shows that the tangent space is $\mathcal{C}_b(\mathcal{F}_{\mathcal{H}_k}, d_{\mathcal{H}_k})$. On the other, the sample paths of $\mathbb{G}_{\mathrm{P}}$ belong to $\mathcal{C}_u(\mathcal{F}_{\mathcal{H}_k}, \rho_{\mathrm{P}})$ P-almost surely. Regarding to Delta method (see Proposition 5, or also Shapiro (1990, Theorem 2.1)), we need the tangent space $\mathcal{C}_b(\mathcal{F}_{\mathcal{H}_k}, d_{\mathcal{H}_k})$ to include the sample paths of $\mathbb{G}_{\mathrm{P}}$. The following lemma gives sufficient conditions for it.

Lemma 36. *Let* P *be a Borel probability measure defined on $\mathcal{X}$ and we consider the pseudo-metric ρ_{P}. We have that*

$$\rho_{\mathrm{P}}(f,g) \leq \left(\int_{\mathcal{X}} k(x,x)\,\mathrm{dP}(x)\right)^{1/2} \|f-g\|_{\mathcal{H}_k}, \quad f,g \in \mathcal{F}_{\mathcal{H}_k}. \tag{2.31}$$

In particular, if $\int_{\mathcal{X}} k(x,x)\ \mathrm{dP}(x) < \infty$*, it holds that* $\mathcal{C}_{\mathrm{u}}\left(\mathcal{F}_{\mathcal{H}_k}, \rho_{\mathrm{P}}\right) = \mathcal{C}_{\mathrm{u}}\left(\mathcal{F}_{\mathcal{H}_k}, \rho_{\mathrm{L}^2(\mathrm{P})}\right) \subset \mathcal{C}_{\mathrm{u}}\left(\mathcal{F}_{\mathcal{H}_k}, d_{\mathcal{H}_k}\right)$.

Proof. For $x \in \mathcal{X}$, by the reproducing property of k twice (see Definition 12) and Cauchy–Schwarz inequality, we have that

$$\begin{aligned} |f(x) - g(x)| &= \left|\langle f - g, k(x,\cdot)\rangle_{\mathcal{H}_k}\right| \\ &\leq \|f-g\|_{\mathcal{H}_k} \|k(x,\cdot)\|_{\mathcal{H}_k} \\ &= \|f-g\|_{\mathcal{H}_k}\sqrt{k(x,x)}. \end{aligned} \tag{2.32}$$

Therefore, from (2.32) we conclude that

$$\begin{aligned} \rho^2_{\mathrm{L}^2(\mathrm{P})}(f,g) &\leq \int_{\mathcal{X}} (f-g)^2\ \mathrm{dP} \\ &\leq \|f-g\|^2_{\mathcal{H}_k} \int_{\mathcal{X}} k(x,x)\ \mathrm{dP}(x). \end{aligned}$$

So (2.31) holds. In addition, again by the reproducing property and the Cauchy-Schwarz inequality

$$\mathrm{P}(|f-g|) \leq \|f-g\|_{\mathcal{H}_k} \int_{\mathcal{X}} \sqrt{k(x,x)}\ \mathrm{dP}(x).$$

By Jensen's inequality, $\int_{\mathcal{X}} \sqrt{k(x,x)}\ \mathrm{dP}(x) < \left(\int_{\mathcal{X}} k(x,x)\ \mathrm{dP}(x)\right)^{1/2}$. So, if $\int_{\mathcal{X}} k(x,x)\ \mathrm{dP}(x) < \infty$ then $\mathcal{F}_{\mathcal{H}_k}$ is bounded in $\mathrm{L}^1(\mathrm{P})$ norm by the reproducing property. Hence pseudometrics ρ_{P} and $\rho_{\mathrm{L}^2(\mathrm{P})}$ are equivalent. Further, by (2.31), $d_{\mathcal{H}_k}$ dominates both pseudometrics. □

In other words, Lemma 36 ensures that the Corollary 35 can be applied together with Delta method and Donsker type theorems whenever $\int_{\mathcal{X}} k(x,x)\ \mathrm{dP}(x) < \infty$.

Union of unit balls

In Chapter 4 a new distance related to kernel distances is introduced: *supremum kernel distance (SKD).* Given P and Q probability measures and a family of reproducing kernels $\{k_\lambda : \lambda \in \Lambda\}$ (possibly a mixture of parametric families), the SKD is defined as the maximum mean discrepancy over the union of balls $\mathcal{F}_{\mathcal{H}_{k,\Lambda}}$. That is

$$d_{k,\Lambda}(\mathrm{P},\mathrm{Q}) = \mathrm{MMD}\left[\mathcal{F}_{\mathcal{H}_{k,\Lambda}}, \mathrm{P}, \mathrm{Q}\right] = \sup_{f \in \mathcal{F}_{\mathcal{H}_{k,\Lambda}}} \left(\mathrm{P}(f) - \mathrm{Q}(f)\right)$$

(analogously to the kernel distance, see Definition 4.4). Further, it can be checked that identity (2.28) can be extended to this framework as

$$d_{k,\Lambda}(\mathrm{P},\mathrm{Q}) = \sup_{\lambda\in\Lambda} \left((d_{k,\lambda}(\mathrm{P},\mathrm{Q})\right) = \sup_{\lambda\in\Lambda}\left(\left\|\mu^\lambda_{\mathrm{P}} - \mu^\lambda_{\mathrm{Q}}\right\|_{\mathcal{H}_{k,\lambda}}\right),$$

where for each $\lambda \in \Lambda$, μ^λ_{P} and μ^λ_{Q} are the mean embeddings of P and Q, respectively, in each $\mathcal{H}_{k,\lambda}$ (the RKHS generated by k_λ).

The next goal is to prove a similar result to Corollary 35 for the union of balls. Just observe that there is not a canonical mechanism of giving a metric to $\mathcal{F}_{\mathcal{H}_{k,\Lambda}}$ such that each $\mathcal{F}_{\mathcal{H}_{k,\lambda}}$ is a natural topological subspace. Therefore, $\mathcal{C}_{\mathrm{b}}\left(\mathcal{F}_{\mathcal{H}_{k,\Lambda}},\rho\right)$, where ρ is the two-sample process joint pseudometric ρ defined in Subsection 1.2.2, is the natural tangent space for Hadamard directional differentiability in the two-sample problem. Furthermore, let us assume the following tecnical conditions. In what follows, k is a kernel. We use the standard notation in functional analysis and operator theory; for k_1 and k_2 positive definite kernels on $\mathcal{X}$, we denote $k_1 \ll k_2$ if and only if $k_2 - k_1$ is a positive definite kernel; see Aronszajn (1950, Part I.7).

(Dom) *Dominance assumption.* There exists a constant $c > 0$ such that $k_\lambda \ll c\,k$, for all $\lambda \in \Lambda$. Further, k is bounded on the diagonal, that is, $\sup_{x\in\mathcal{X}}\left(k(x,x)\right) < \infty$.

(Ide) *Identifiability assumption.* If $\mathrm{P} \neq \mathrm{Q}$, there exists $\lambda \in \Lambda$ such that $\mu_{\mathrm{P}}^{\lambda} \neq \mu_{\mathrm{Q}}^{\lambda}$.

(Par) *Continuous parametrization.* Λ is a compact subset of $\mathbb{R}^d$ (with $d \in \mathbb{N}$) and, for a fixed $(x,y) \in \mathcal{X} \times \mathcal{X}$, the function $\lambda \mapsto k_\lambda(x,y)$ is continuous from Λ to $\mathbb{R}$.

Assumption (Dom) is necessary for the application of the Dominated Convergence Theorem (DCT) in the proof of Corollary 37. In particular, it implies that for every λ, $\mathcal{H}_{k,\lambda}$ is constituted by continuous and bounded functions, therefore measurable and integrable. Moreover, under this condition the mean embedding $\mu_{\mathrm{P}}^{\lambda}$ exists (for each P and λ). In particular, the supremum kernel distance is well-defined.

Assumption (Ide) entails that $d_{k,\Lambda}(\mathrm{P},\mathrm{Q}) > 0$, whenever $\mathrm{P} \neq \mathrm{Q}$, i.e., the supremum kernel distance separates different probability measures. Therefore, $d_{k,\Lambda}$ in (4.3) is a proper metric on $\mathcal{M}_{\mathrm{p}}(\mathcal{X})$. Regarding this, we recall that a reproducing kernel k is said to be *characteristic* whenever $d_k(\mathrm{P},\mathrm{Q}) = 0$ if and only if $\mathrm{P} = \mathrm{Q}$, for all $\mathrm{P},\mathrm{Q} \in \mathcal{M}_{\mathrm{p}}(\mathcal{X})$; see Fukumizu et al. (2007). This is equivalent to "integrally strictly positive definiteness", see Sriperumbudur et al. (2010, Theorem 7). Hence, (Ide) could be understood as *the family* $\{k_\lambda : \lambda \in \Lambda\}$ *being characteristic* in the sense that for each pair of different measures there is a kernel in the family separating them. In infinite dimension, necessary and sufficient conditions for the Gaussian kernel to be characteristic are given in Wynne and Duncan (2022). From the perspective of computing the derivative of the supremum σ, it essentially means that derivation is done in a point away from zero.

Finally, (Par) is a technical requirement to derive the explicit formula of the Hadamard directional derivative using Theorem 17. Further discussion on these assumptions related to the two sample test and Donsker property can be found in Section 4.3.

Corollary 37. *Let us assume that the family of kernels* $\{k_\lambda : \lambda \in \Lambda\}$ *satisfies (Dom), (Ide) and (Par). Let* P *and* Q *be Borel probability measures on* $\mathcal{X}$ *such that* $\mathrm{P} \neq \mathrm{Q}$*, then the mapping* $\sigma(\nu) = \sup_{\mathcal{F}_{\mathcal{H}_{k,\Lambda}}}(\nu)$ *for* $\nu \in \ell^\infty\left(\mathcal{F}_{\mathcal{H}_{k,\Lambda}}\right)$ *is Hadamard directionally differentiable at* $\mathrm{P}-\mathrm{Q}$ *tangentially to* $\mathcal{C}_{\mathrm{b}}\left(\mathcal{H}_{k,\Lambda},\rho\right)$*. In such a case, the (directional) derivative of* σ *at the point* $\mathrm{P}-\mathrm{Q}$ *is given by*

$$\sigma'_{\mathrm{P}-\mathrm{Q}}(g) = \sup_{\lambda\in\Lambda_0}\left(g\left(h^{+,\lambda}\right)\right) = \sup_{L}(g), \quad g \in \mathcal{C}\left(\mathcal{F}_{\mathcal{H}_{k,\Lambda}},\rho\right), \tag{2.33}$$

where

$$h^{+,\lambda} = \frac{\mu_{\mathrm{P}}^{\lambda} - \mu_{\mathrm{Q}}^{\lambda}}{\left\|\mu_{\mathrm{P}}^{\lambda} - \mu_{\mathrm{Q}}^{\lambda}\right\|_{\mathcal{H}_{k,\lambda}}}, \tag{2.34}$$

and

$$\Lambda_0 = \left\{\lambda \in \Lambda : \left\|\mu_{\mathrm{P}}^{\lambda} - \mu_{\mathrm{Q}}^{\lambda}\right\|_{\mathcal{H}_{k,\lambda}} = d_{k,\Lambda}(\mathrm{P},\mathrm{Q})\right\} \quad \text{and} \quad L = \left\{h^{+,\lambda} : \lambda \in \Lambda_0\right\}. \tag{2.35}$$

Proof. Let us fix $g \in \mathcal{C}_{\mathrm{b}}(\mathcal{F}_{k,\Lambda},\rho)$. From Theorem 17, we have that σ is Hadamard directionally differentiable and

$$\sigma'_{\mathrm{P}-\mathrm{Q}}(g) = \lim_{\varepsilon\searrow 0} \sup_{A_{\varepsilon}(\mathrm{P}-\mathrm{Q})} (g), \tag{2.36}$$

where $A_{\varepsilon}(\mathrm{P}-\mathrm{Q}) = \left\{h \in \mathcal{F}_{\mathcal{H}_{k,\Lambda}} : (\mathrm{P}-\mathrm{Q})(h) \geq d_{k,\Lambda}(\mathrm{P},\mathrm{Q}) - \varepsilon\right\}$. For every $\varepsilon > 0$, it is clear that $L \subseteq A_{\varepsilon}(\mathrm{P}-\mathrm{Q})$, where L is defined in (2.35). Hence, we have that

$$\sup_{\lambda\in\Lambda_0} \left(g\left(h^{+,\lambda}\right)\right) \leq \sigma'_{\mathrm{P}-\mathrm{Q}}(g). \tag{2.37}$$

Conversely, we consider a maximizing sequence $(h_m)_{m\in\mathbb{N}}$ satisfying that $h_m \in A_{1/m}(\mathrm{P}-\mathrm{Q})$ and

$$\sup_{A_{1/m}(\mathrm{P}-\mathrm{Q})} (g) \leq g\left(h_m\right) + \frac{1}{m}. \tag{2.38}$$

Each $h_m \in \mathcal{F}_{\mathcal{H}_{k,\lambda_m}}$ $(m \in \mathbb{N})$, for some $\lambda_m \in \Lambda$. We consider the sequence $(h^{+,\lambda_m})_{m\in\mathbb{N}}$. Using (Par), by restricting if needed, to a subsequence we can assume that $\lambda_m \to \lambda^* \in \Lambda$. Next we prove the following facts:

(i) $\lambda^* \in \Lambda_0$, where Λ_0 is in (2.35), and

$$\left\|\mu_{\mathrm{P}}^{\lambda_m} - \mu_{\mathrm{Q}}^{\lambda_m}\right\|_{\mathcal{H}_{k,\lambda_m}} \to \left\|\mu_{\mathrm{P}}^{\lambda^*} - \mu_{\mathrm{Q}}^{\lambda^*}\right\|_{\mathcal{H}_{k,\lambda^*}} = d_{k,\Lambda}(\mathrm{P},\mathrm{Q}) > 0. \tag{2.39}$$

(ii) $\rho\left(h_m, h^{+,\lambda_m}\right) \to 0$, as $m \to \infty$.

(iii) $\rho\left(h_m, h^{+,\lambda^*}\right) \to 0$, as $m \to \infty$.

First, (2.39) is obtained by using the representation of the kernel distance as a double integral in (1.6) (with $\nu = \mathrm{P}-\mathrm{Q}$), together with (Dom), (Par) and the Dominated Convergence Theorem (DCT). Further, as $h_m \in \mathcal{F}_{\mathcal{H}_{k,\lambda_m}} \cap A_{1/m}(\mathrm{P}-\mathrm{Q})$, we obtain that

$$\left\|\mu_{\mathrm{P}}^{\lambda_m} - \mu_{\mathrm{Q}}^{\lambda_m}\right\|_{\mathcal{H}_{k,\lambda_m}} \geq \mathrm{P}\left(h_m\right) - \mathrm{Q}\left(h_m\right) \geq d_{k,\Lambda}(\mathrm{P},\mathrm{Q}) - \frac{1}{m}.$$

Hence, from (2.39) and by taking $m \to \infty$ we obtain that $\lambda^* \in \Lambda_0$. The fact that $d_{k,\Lambda}(\mathrm{P},\mathrm{Q}) > 0$ follows from (Ide) and the proof of (i) is complete.

To show (ii), using the same ideas as in the proof of equation (2.30) and (i), we obtain that

$$\left\|h_m - h^{+,\lambda_m}\right\|^2_{\mathcal{H}_{k,\lambda_m}} \leq 2 - \frac{2\, d_{k,\Lambda}(\mathrm{P},\mathrm{Q}) - \frac{1}{m}}{\left\|\mu_{\mathrm{P}}^{\lambda_m} - \mu_{\mathrm{Q}}^{\lambda_m}\right\|_{\mathcal{H}_{k,\lambda_m}}} \to 0, \quad \text{as } m \to \infty.$$

Now, from (2.32) and (Dom), we have that

$$\begin{aligned}\rho_{\mathrm{P}}^2\left(h_m, h^{+,\lambda_m}\right)^2 &\leq \rho_{\mathrm{L}^2(\mathrm{P})}^2\left(h_m, h^{+,\lambda_m}\right)^2 \\ &\leq \left\|h_m - h^{+,\lambda_m}\right\|_{\mathcal{H}_{k,\lambda_m}}^2 c \int_{\mathcal{X}} k(x,x)\, \mathrm{dP}(x) \to 0, \quad \text{as } m \to \infty.\end{aligned}$$

Therefore, $\rho\left(h_m, h^{+,\lambda_m}\right) \to 0$, and (ii) holds. To check (iii), by (ii), it is enough to see that $\rho\left(h^{+,\lambda_m}, h^{+,\lambda^*}\right) \to 0$, as $m \to \infty$. By (2.39) and repeatedly applying DCT (thanks to (Dom)), it can be checked that

$$h^{+,\lambda_m}(x) \to h^{+,\lambda^*}(x), \quad \text{as } m \to \infty \text{ and for all } x \in \mathcal{X}.$$

Furthermore, for m large enough, we have that

$$\left|h^{+,\lambda_m}(x)\right| \leq \frac{2\,c\,\left(|\mu_{\mathrm{P}}(x)| + |\mu_{\mathrm{Q}}(x)|\right)}{d_{k,\Lambda}(\mathrm{P},\mathrm{Q})} \in \mathrm{L}^2(\mathrm{P}),$$

where μ_{P} and μ_{Q} are the mean embeddings corresponding to the dominating kernel k in (Dom). Hence, we can apply one more time DCT to obtain that $\rho_{\mathrm{P}}\left(h^{+,\lambda_m}, h^{+,\lambda^*}\right) \to 0$. This implies that $\rho\left(h^{+,\lambda_m}, h^{+,\lambda^*}\right) \to 0$, as $m \to \infty$.

To finish, we use (2.38), (i), (iii), as well as the continuity of the functional g (with respect to the metric ρ) to obtain that

$$\sigma'_{\mathrm{P}-\mathrm{Q}}(g) = \lim_{m\to\infty} \sup_{A_{1/m}(\mathrm{P}-\mathrm{Q})}(g) \leq \lim_{m\to\infty} g\left(h_m\right) = g\left(h^{+,\lambda^*}\right) \leq \sup_{\lambda\in\Lambda_0}\left(g\left(h^{+,\lambda}\right)\right) = \sup_{L}(g).$$

The conclusion of this lemma follows from (2.37) and the previous inequalities.

Chapter 3

Asymptotic results for some classical problems

The aim of this chapter is to apply differentiability results of Chapter 2 to a collection of classical problems. We use an extended version of the functional Delta method to derive the asymptotic distribution of many statistics that can be expressed in terms of these maps. In this way, we provide a simple and unified approach and a suitable framework to deal with such type of statistics.

Using these ideas, we obtain the following applications that can be divided into two groups according to whether $\mathfrak{X} \subseteq \overline{\mathbb{R}}^d$ or $\mathfrak{X} = \mathcal{F}$ (a class of functions):

- **Case $\mathfrak{X} \subseteq \overline{\mathbb{R}}^d$:** In Section 3.1 we extend and give simpler and shorter proofs of the results in Raghavachari (1973) both in the one-sample and two-sample cases. The extension is carried out in different directions: Firstly, no assumption on the involved distribution functions is necessary to derive the asymptotic results. In contrast, in Raghavachari (1973) the continuity of the distribution functions is required. Secondly, the results are obtained in a multidimensional setting. We note that the proofs are very simple (compared with those in Raghavachari (1973)) because they just rely on the analysis of the differentiability of the functionals and the convergence of the associated processes separately. It should be further remarked that those works that have used the results and ideas in Raghavachari (1973) were forced to impose the continuity of the involved functions as an assumption in their statements (see for instance Álvarez-Esteban et al. (2016, Equation (11)), Freitag et al. (2006, Section 2) or Dette et al. (2018, Assumption 7.4.)). The regularity limitation of working with continuous functions is not mathematically aesthetic and it is in fact unnecessary, as we will show in this chapter. The extension of these kind of results to any dimension is important to include and be able to deal with multidimensional distribution functions such as copulas, considered in Section 3.2. In Section 3.3, we apply this technique to solve an open question by Jager and Wellner (2004) related to the Berk-Jones statistic.

- **Case $\mathfrak{X} = \mathcal{F}$:** In Section 3.4 we also derive similar results for the plug-in estimators of maximum mean discrepancies with respect to a Donsker class $\mathcal{F}$. Other results in a more concrete context, such as kernel distances (associated to the class of functions $\mathcal{F}_{\mathcal{H}_k}$) and the k-means problem (associated to the class of functions $\mathcal{F}_{V_k(B)}$) are detailed in Chapters 4 and 5.

In a wide variety of situations Theorem 17 and its subsequent corollaries, joint with the extended Delta method in Proposition 5, provide the right framework to obtain a number of significant examples in which the asymptotic distribution of a statistic of interest can be determined with ease. The combination of these results is summarized in the following theorem.

Theorem 38. *Let $\theta \in \ell^\infty(\mathfrak{X}) \setminus \{0\}$ and assume that there exists T_n taking values in $\ell^\infty(\mathfrak{X})$ a.s. such that $r_n(T_n - \theta) \rightsquigarrow T$, for a sequence of real numbers satisfying that $r_n \to \infty$ and a Borel random element T in $\ell^\infty(\mathfrak{X})$. Then, for $\varphi \in \{\delta, \sigma, \iota, \alpha\}$ in* (2.1)*, we have that*

$$r_n(\varphi(T_n) - \varphi(\theta)) \rightsquigarrow \varphi'_\theta(T), \tag{3.1}$$

where the derivatives φ'_θ are given in (2.4)*. Moreover, we have that $r_n(\varphi(T_n) - \varphi(\theta)) = \varphi'_\theta(r_n(T_n - \theta)) + o_{\mathrm{P}}(1)$.*

Theorem 38 is still valid for the maps σ, ι and α when $\theta = 0$ as $\sigma'_0(g) = \sup_{\mathfrak{X}}(g)$, $\iota'_0(g) = \inf_{\mathfrak{X}}(g)$ and $\alpha'_0(g) = \operatorname{amp}_{\mathfrak{X}}(g)$ are continuous maps. Further, for those $\theta \in \ell^\infty(\mathfrak{X})$ such that φ'_θ is linear, i.e., φ is fully Hadamard differentiable at θ (see Corollary 25 and Remarks 27 and 30), and when T is Gaussian, we conclude that $\varphi'_\theta(T)$ is normally distributed.

In the remaining of this chapter, we will apply the previous general result in different contexts to obtain the asymptotic distribution of several statistics.

3.1 Distribution functions

Let X and Y be two non-degenerate random vectors taking values on $\mathbb{R}^d$ ($d \geq 1$) with joint cumulative distribution functions $F(x) = \mathrm{P}(X \leq x)$ and $H(x) = \mathrm{P}(Y \leq x)$, $x \in \mathbb{R}^d$, where '$\leq$' stands for the coordinatewise order in $\mathbb{R}^d$. The goal in this section is to estimate $\varphi(F - H)$, where $\varphi \in \{\delta, \sigma, \alpha\}$ are defined in (2.1).

One-sample case

In this situation we have at our disposal a random sample $X_1, \dots, X_n$ from X. We estimate $F - H$ with $\mathbb{F}_n - H$, where $\mathbb{F}_n$ is the empirical distribution function of the observed sample, that is,

$$\mathbb{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{X_i \leq x\}}, \quad x \in \mathbb{R}^d,$$

and $\mathbf{1}_A$ stands for the indicator function of the set A (see Section 1.2.2).

The problem consists in finding the behavior, as $n \to \infty$, of

$$\begin{aligned} D_n(\delta) &= \sqrt{n}\,\left(\|\mathbb{F}_n - H\|_\infty - \|F - H\|_\infty\right), \\ D_n(\sigma) &= \sqrt{n}\left(\sup_{\mathbb{R}}\,(\mathbb{F}_n - H) - \sup_{\mathbb{R}}\,(F - H)\right) \\ D_n(\alpha) &= \sqrt{n}\left(\operatorname*{amp}_{\mathbb{R}}\,(\mathbb{F}_n - H) - \operatorname*{amp}_{\mathbb{R}}\,(F - H)\right). \end{aligned} \tag{3.2}$$

When $F \neq H$, the asymptotic distribution of the statistics $D_n(\delta)$, $D_n(\sigma)$ and $D_n(\alpha)$ in (3.2) can be viewed as the limit under the alternative hypothesis of the corresponding two-sided and one-sided Kolmogorov-Smirnov test statistics and Kuiper statistic, respectively.

In this example, for $\varphi \in \{\delta, \sigma, \alpha\}$, the statistics in (3.2) are $D_n(\varphi) \equiv D_\varphi(\theta, T_n, r_n)$ in (2.3) with $\theta = F - H$, $T_n = \mathbb{F}_n - H$, and $r_n = \sqrt{n}$. The underlying normalized process, i.e., $r_n\,(T_n - \theta)$, is nothing but the multivariate *empirical process* (indexed by points),

$$\mathbb{G}_{n,F}(x) = \sqrt{n}\,(\mathbb{F}_n(x) - F(x)), \quad n \in \mathbb{N}, \quad \mathbf{x} \in \mathbb{R}^d. \tag{3.3}$$

When there is no confusion with respect to the underlying distribution, we simply use the notation $\mathbb{G}_n$ for the empirical process in (3.3), as stated in Section 1.2.1. Remind that the collection of all indicator functions of lower (hyper)rectangles of $\overline{\mathbb{R}}^d$, $\left\{1_{(-\infty,x_1]\times\cdots\times(-\infty,x_d]} : (x_1, \ldots, x_d) \in \overline{\mathbb{R}}^d\right\}$, is universal Donsker (see Subsection 1.2.2 or A. van der Vaart and Wellner (1996, Example 2.1.3, p. 82)), the empirical process converges in law in $\ell^\infty\left(\overline{\mathbb{R}}^d\right)$. Recall, also, that the weak limit of $\mathbb{G}_n$ is denoted in this context by $\mathbb{B}_F$. If $d = 1$, the assertion "$\mathbb{G}_n \rightsquigarrow \mathbb{B}_F$ in $\ell^\infty\left(\overline{\mathbb{R}}\right)$" is nothing but the celebrated Donsker's theorem (Kolmogorov-Doob-Donsker-Dudley Central Limit Theorem). When $d \geq 2$, $\mathbb{B}_F$ is also called a tied-down or pinned F-Brownian sheet based on the measure with distribution function F.

In this particular case we have that $F - H \in \mathcal{D}\left(\overline{\mathbb{R}}^d\right)$, $\mathbb{G}_n \in \mathcal{D}\left(\overline{\mathbb{R}}^d\right)$ a.s., and $\mathbb{G}_n \rightsquigarrow \mathbb{B}_F$ in $\ell^\infty\left(\overline{\mathbb{R}}^d\right)$. Therefore, as a direct consequence of Theorem 38 and Corollary 33 we obtain the following result.

Proposition 39. *Assume that $F \neq H$ and let $\mathbb{B}_F$ be an F-Brownian bridge. For $\varphi \in \{\delta, \sigma, \alpha\}$, we consider the statistics $D_n(\varphi)$ defined in* (3.2). *We have that $D_n(\varphi) \rightsquigarrow \varphi'_{F-H}(\mathbb{B}_F)$, where the derivatives φ'_{F-H} are given as in (2.25).*

When $d = 1$, Proposition 39 improves Raghavachari (1973, Theorems 1, 2 and 3) as here F and H are not assumed to be continuous. If F is continuous, then $\mathbb{B}_F \in \mathcal{C}\left(\overline{\mathbb{R}}^d, d_e\right)$ a.s., and the limiting distributions in Proposition 39 have simpler expressions (see (2.19)). The following corollary provides a multidimensional extension of the results in Raghavachari (1973).

Corollary 40. *In the conditions of Proposition 39, let us further assume that $F, H \in \mathcal{C}\left(\overline{\mathbb{R}}^d, d_e\right)$ and we consider the sets $M^+(\cdot)$ and $M^-(\cdot)$ defined in* (2.17). *We have that:*

(i) $D_n(\delta) \rightsquigarrow \sup\limits_{M^+(|F-H|)} (\mathbb{B}_F\,\mathrm{sgn}(F - H))$.

(ii) $D_n(\sigma) \rightsquigarrow \sup_{M^+(F-H)} (\mathbb{B}_F)$.

(iii) $D_n(\alpha) \rightsquigarrow \sup_{M^+(F-H)} (\mathbb{B}_F) - \inf_{M^-(F-H)} (\mathbb{B}_F)$.

Remark 41. In the setting of the previous corollary, when $M^+(|F-H|)$ (respectively, $M^+(F-H)$; and $M^+(F-H)$ and $M^-(F-H)$) contains only one point, the mapping δ (respectively, σ and α) is fully Hadamard differentiable at $F-H$ (see Corollary 25). In particular, the asymptotic distribution of $D_n(\delta)$ (respectively, $D_n(\sigma)$ and $D_n(\alpha)$) is a zero mean Gaussian distribution. The asymptotic variance can be directly computed from the covariances of $\mathbb{B}_F$.

Two-sample case

Here, two (mutually independent) random samples are available, one of size n from F and another one of size m from H. Let $\mathbb{F}_n$ and $\mathbb{H}_m$ be the empirical distribution functions of the two samples, respectively, and set $N \equiv \frac{nm}{n+m}$. The two-sided, and one-sided Kolmogorov-Smirnov and Kuiper statistics in the two sample case are given by

$$\begin{aligned} D_{n,m}(\delta) &= \sqrt{N}\left(\|\mathbb{F}_n - \mathbb{H}_m\|_\infty - \|F-H\|_\infty\right), \\ D_{n,m}(\sigma) &= \sqrt{N}\left(\sup_{\mathbb{R}} (\mathbb{F}_n - \mathbb{H}_m) - \sup_{\mathbb{R}} (F-H)\right) \\ D_{n,m}(\alpha) &= \sqrt{N}\left(\operatorname{amp}_{\mathbb{R}} (\mathbb{F}_n - \mathbb{H}_m) - \operatorname{amp}(F-H)\right). \end{aligned} \tag{3.4}$$

In the general setting specified in (2.3), this situation matches to the case $\theta = F - H$, $T_{n,m} = \mathbb{F}_n - \mathbb{H}_m$ and $r_{n,m} = \sqrt{N}$. Hence, we have that

$$r_{n,m}\left(T_{n,m} - \theta\right) = \sqrt{\frac{m}{n+m}}\,\mathbb{G}_{n,F} - \sqrt{\frac{n}{n+m}}\,\mathbb{G}_{m,H}$$

with $\mathbb{G}_{n,F}$ and $\mathbb{G}_{m,H}$ independent empirical processes. We further observe that if the sampling scheme is balanced, that is, $\frac{n}{(n+m)} \to \xi$, with $0 \le \xi \le 1$ as $n, m \to \infty$, then $r_{n,m}\left(T_{n,m} - \theta\right) \rightsquigarrow \sqrt{1-\xi}\,\mathbb{B}_F - \sqrt{\xi}\,\widetilde{\mathbb{B}}_H$ in $\ell^\infty\left(\overline{\mathbb{R}}^d\right)$, where $\mathbb{B}_F$ and $\widetilde{\mathbb{B}}_H$ are two independent Brownian bridges associated with F and H, respectively. Hence, Theorem 38 and Corollary 34 directly imply the following result which improves and generalizes Raghavachari (1973, Theorems 4 and 5).

Proposition 42. *Let us consider a sampling scheme such that as n, $m \to \infty$, $\frac{n}{(n+m)} \to \xi$, with $0 \le \xi \le 1$ and let $\mathbb{B}_F$ and $\widetilde{\mathbb{B}}_H$ be two independent Brownian bridges associated with F and H, respectively. For $\varphi \in \{\delta, \sigma, \alpha\}$, we consider the statistics $D_{n,m}(\varphi)$ defined in (3.4). We have that $D_{n,m}(\varphi) \rightsquigarrow \varphi'_{F-H}\left(\sqrt{1-\xi}\,\mathbb{B}_F - \sqrt{\xi}\,\widetilde{\mathbb{B}}_H\right)$, where the derivatives φ'_{F-H} are given in (2.25). If we further have that $F, H \in \mathcal{C}\left(\overline{\mathbb{R}}^d, d_e\right)$, then the derivatives can be expressed as in (2.19).*

3.2 Copulas

In this section, for simplicity, we will assume that the involved distribution functions are continuous. Let us assume that the d-dimensional distribution function F has copula C and continuous marginal distribution functions $F_1, \ldots, F_d$. In other words, $F(x) = C(F_1(x_1), \ldots, F_d(x_d))$, for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. Let $\mathbb{F}_n$ and $\mathbb{F}_{n,i}$ $(i = 1, \ldots, d)$ be the empirical joint and i-th marginal distribution functions of a random sample of size n from F. The *empirical copula* is

$$\mathbb{C}_n(u) = \mathbb{F}_n\left(\mathbb{F}_{n,1}^{-1}(u_1), \ldots, \mathbb{F}_{n,d}^{-1}(u_d)\right), \quad u = (u_1, \ldots, u_d) \in [0,1]^d, \tag{3.5}$$

where $\mathbb{F}_{n,i}^{-1}$ stands for the generalized inverse of $\mathbb{F}_{n,i}$, i.e., the marginal quantile function of the i-th coordinate sample. The *empirical copula process* is defined by

$$\mathbb{G}_{n,C}(u) = \sqrt{n}\,(\mathbb{C}_n(u) - C(u)), \quad n \in \mathbb{N}, \quad \mathbf{u} \in [0,1]^d. \tag{3.6}$$

Empirical copula processes play the same role for copulas as empirical processes for distribution functions and they have been extensively used in goodness-of-fit testing problems for copulas (see Fermanian (2013) for an overview about this subject).

Several works have been devoted to discuss the asymptotic behavior of $\mathbb{G}_{n,C}$ in (3.6). For instance, in Segers (2012) (see also the references therein) it is shown that, under certain not very restrictive smoothness assumptions on the underlying copula C, $\mathbb{G}_{n,C}$ converges weakly in $\ell^\infty([0,1]^d)$. Specifically, let us assume that C satisfies the following regularity condition:

Condition 1. For each $i \in \{1, \ldots, d\}$, the i-th first order partial derivative of C, $\partial_i C$, exists and is continuous on the set $\{u = (u_1, \ldots, u_d) \in [0,1]^d : 0 < u_i < 1\}$.

If Condition 1 is satisfied, $\mathbb{G}_{n,C} \rightsquigarrow \mathbb{C}$ in $\ell^\infty([0,1]^d)$ (see Segers (2012, Proposition 3.1)), where $\mathbb{C}$ is a Gaussian process that can be represented as

$$\mathbb{G}_C(u) = \mathbb{B}_C(u) - \sum_{i=1}^{d} \partial_i C(u)\, \mathbb{B}_C^{(i)}(u_i), \quad u = (u_1, \ldots, u_d) \in [0,1]^d, \tag{3.7}$$

with $\mathbb{B}_C$ a C-Brownian bridge (see Subsection 1.2.2) and $\mathbb{B}_C^{(i)}(u_i) = \mathbb{B}_C(1, \ldots, 1, u_i, 1, \ldots, 1)$, the variable u_i appearing at the i-th entry.

Using Theorem 38 and Corollary 34, we immediately obtain the following result. Though details are omitted, similar results can be stated for the unilateral Kolmogorov-Smirnov and Kuiper statistics and the associated two sample problems.

Proposition 43. *Let C be a copula satisfying Condition 1 and let $\mathbb{C}_n$ be as in* (3.5)*. For any continuous copula $D \neq C$, the statistic*

$$L_n(C,D) = \sqrt{n}\,(\|\mathbb{C}_n - D\|_\infty - \|C - D\|_\infty),$$

converges in distribution to $\delta'_{C-D}(\mathbb{G}_C) = \sup_{M^+(|C-D|)} (\mathbb{G}_C \operatorname{sgn}(C-D))$, with $\mathbb{G}_C$ defined in (3.7) *and the set $M^+(\cdot)$ is given in* (2.17)*.*

Remark 44. Let C be a bivariate copula and we consider the survival copula

$$\overline{C}\left(u_1,u_2\right) = u_1 + u_2 - 1 + C\left(1-u_1, 1-u_2\right), \quad \left(u_1,u_2\right) \in [0,1]^2.$$

The statistics $L_n\left(C,\overline{C}\right)$ has been used in Genest and Nešlehová (2014) to derive a test of radial symmetry for bivariate copulas. Proposition 43 provides the asymptotic distribution of such statistic.

3.3 On a question by Jager and Wellner related to the Berk–Jones statistic

Let $\mathbb{F}_n$ be the empirical distribution function of a sample of size n from a univariate random variable with continuous distribution function F. Suppose that we want to test the null hypothesis $\mathrm{H}_0 : F = H$ versus the alternative $\mathrm{H}_1 : F \neq H$, where H is a fixed (and usually known) continuous distribution function. Berk and Jones (1979) (see also DasGupta (2008, Chapter 26.7)) introduced the test statistic

$$R\left(\mathbb{F}_n, H\right) = \sup_{x\in\mathbb{R}} \left(K\left(F_n(x), H(x)\right)\right), \tag{3.8}$$

where

$$K(x,y) = x\log\left(\frac{x}{y}\right) + (1-x)\,\log\left(\frac{1-x}{1-y}\right),$$

for $x \in [0,1]$ and $y \in (0,1)$. (The values of $K(x,y)$ when $x = 0,1$ are taken by continuity.)

For each $x \in \mathbb{R}$, $n\,K\left(\mathbb{F}_n(x), H(x)\right)$ is the log-likelihood ratio statistic for testing $\mathrm{H}_0 : F(x) = H(x)$ against $\mathrm{H}_1 : F(x) \neq H(x)$. Hence, $R\left(\mathbb{F}_n, H\right)$ in (3.8) is nothing but the supremum of these pointwise likelihood ratio tests statistics. Additionally, $K(x,y)$ is the Kullback-Leibler divergence between two Bernoulli distributions with means x and y. Hence, $K(x,y) \geq 0$ with equality if and only if $x = y$. In particular, $R\left(\mathbb{F}_n, H\right) = \left\|K\left(\mathbb{F}_n, H\right)\right\|_\infty$.

Berk and Jones (1979) computed the asymptotic distribution of (the normalized version of) $R\left(\mathbb{F}_n, F\right)$, i.e., the distribution of the statistic under the null hypothesis $F = G$. For a detailed proof, see Wellner and Koltchinskii (2003, Theorem 1.1) or Jager and Wellner (2007, Theorem 3.1). It holds that

$$n\,R\left(\mathbb{F}_n, F\right) - d_n \rightsquigarrow Y_4, \quad \text{as } n \to \infty, \tag{3.9}$$

where $\mathrm{P}\left(Y_4 \leq x\right) = \exp\left(-4\exp\left(-x\right)\right)$ for $x \in \mathbb{R}$, i.e., Y_4 has double-exponential extreme value distribution, and

$$d_n = \log_2(n) - \frac{1}{2}\log_3(n) - \frac{1}{2}\log\left(4\,\pi\right),$$

with $\log_2(n) = \log\left(\log(n)\right)$ and $\log_3(n) = \log\left(\log_2(n)\right)$.

In Jager and Wellner (2004, Question 2, p. 329), it was set out the open problem of finding the asymptotic behaviour of the Berk–Jones statistic under the alternative hypothesis. In other words, assuming that $F \neq H$, the question consists in finding conditions on F and H for which the statistic

$$B_n = \sqrt{n}\,\left(R\left(\mathbb{F}_n, H\right) - R(F,H)\right), \tag{3.10}$$

converges in distribution and, in such a case, identifying its weak limit, where $R\left(F_n, H\right)$ is given in (3.8) and $R(F,H) = \sup_{x\in\mathbb{R}}\left(K(F(x),H(x))\right)$.

Here we give a precise answer for the previous question. First, we note that B_n in (3.10) has the general form of (2.3). In other words,

$$B_n = D_\sigma\left(\theta = K(F,H), T_n = K\left(\mathbb{F}_n, H\right), r_n = \sqrt{n}\right), \tag{3.11}$$

where σ is defined in (2.1). As K is non-negative, it also holds that $B_n = D_\delta\left(K(F,H),\right.$ $\left.K\left(\mathbb{F}_n,H\right), \sqrt{n}\right)$ with δ in (2.1). Therefore, from (3.11) and Theorem 38, to obtain the asymptotic distribution of B_n in (3.10) it is enough to find the weak limit of the process $\mathbb{W}_n$ given by

$$\mathbb{W}_n = \sqrt{n}\,\left(K\left(\mathbb{F}_n, H\right) - K(F,H)\right). \tag{3.12}$$

This result is stated in the following theorem.

Theorem 45. *Let us assume that*

$$\int_{\mathbb{R}} \log^2\left(\frac{F(t)\,(1-H(t))}{H(t)\,(1-F(t))}\right) \mathrm{d}F(t) < \infty.$$

The process $\mathbb{W}_n$ defined in (3.12) *satisfies that $\mathbb{W}_n \rightsquigarrow \mathbb{W}$ in $\ell^\infty\left(\overline{\mathbb{R}}\right)$, where*

$$\mathbb{W} = \mathbb{B}_F\,\log\left(\frac{F\,(1-H)}{H\,(1-F)}\right), \tag{3.13}$$

and $\mathbb{B}_F$ is an F-Brownian bridge.

Proof. Using Taylor's theorem, we have that

$$K\left(\mathbb{F}_n, H\right) - K(F,H) = \left(\mathbb{F}_n - F\right)\,\log\left(\frac{F\,(1-H)}{H\,(1-F)}\right) + \frac{1}{2}\,\frac{\left(\mathbb{F}_n - F\right)^2}{F_n^*\,\left(1-F_n^*\right)}, \tag{3.14}$$

where F_n^* is between F and $\mathbb{F}_n$. We set

$$\widetilde{\mathbb{W}}_n = \sqrt{n}\,\left(\mathbb{F}_n - F\right)\,\log\left(\frac{F\,(1-H)}{H\,(1-F)}\right). \tag{3.15}$$

From (3.12) and (3.14), we have that

$$\left\|\mathbb{W}_n - \widetilde{\mathbb{W}}_n\right\|_\infty = \frac{\sqrt{n}}{2}\,\left\|\frac{\left(\mathbb{F}_n - F\right)^2}{F_n^*\,\left(1-F_n^*\right)}\right\|_\infty. \tag{3.16}$$

Now, from (3.16) and Wellner and Koltchinskii (2003, equation (2.2)) (see also Jager and Wellner (2007, equation (9))), we obtain that

$$\begin{aligned}\left\|\mathbb{W}_n - \widetilde{\mathbb{W}}_n\right\|_\infty &=_{\text{st}} \sqrt{n}\, R\left(\mathbb{F}_n, F\right) \\ &= \frac{1}{\sqrt{n}}\left(n\, R\left(\mathbb{F}_n, F\right) - d_n\right) + \frac{d_n}{\sqrt{n}},\end{aligned} \tag{3.17}$$

where $=_{\text{st}}$ stands for equality in distribution. From (3.9) and (3.17), we conclude that $\left\|\mathbb{W}_n - \widetilde{\mathbb{W}}_n\right\|_\infty \rightsquigarrow 0$. Hence, the processes $\mathbb{W}_n$ and $\widetilde{\mathbb{W}}_n$ have the same asymptotic behavior (see A. W. van der Vaart (2000, Theorem 18.10)). Finally, the conclusion follows from A. W. van der Vaart (2000, Example 19.12, p. 273). □

Remark 46. As it follows from the proof of Theorem 45, the process $\mathbb{W}_n$ behaves asymptotically as $\widetilde{\mathbb{W}}_n$ in (3.15), which is a *weighted empirical process*. Therefore, necessary and sufficient conditions for the convergence of the process $\mathbb{W}_n$ defined in (3.12) are given by the Chibisov-O'Reilly's theorem (see Shorack and Wellner (2009, p. 462)).

We are now in position to solve the question proposed in Jager and Wellner (2004).

Corollary 47. *In the conditions of Theorem 45, the statistic B_n in* (3.10) *satisfies that*

$$B_n \rightsquigarrow \sigma'_{K(F,H)}(\mathbb{W}) = \sup_{M^+(K(F,H))} (\mathbb{W}), \quad \text{as } n \to \infty,$$

where $\mathbb{W}$ is given in (3.13) *and the set $M^+(\cdot)$ is defined in* (2.17).

Remark 48. Similar results can be stated for the family of test statistics $S_n(s)$ based on φ-divergences introduced by Jager and Wellner (2007). Details are omitted.

3.4 Maximum mean discrepancies

3.4.1 Definition and examples

Let X and Y be two random variables taking values on a topological space $(\mathcal{X}, \tau)$ with Borel probability measures P and Q, respectively. We consider a statistic to measure the dissimilarity between P and Q (see Fortet and Mourier (1953) and Müller (1997)).

Definition 49. Let us consider a class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \longrightarrow \mathbb{R}$. The *maximum mean discrepancy* (MMD in short) between P and Q with respect to the class $\mathcal{F}$ is defined by

$$\text{MMD}[\mathcal{F}, \text{P}, \text{Q}] = \sup_{f \in \mathcal{F}} \left(\text{P}(f) - \text{Q}(f)\right). \tag{3.18}$$

To avoid indeterminate forms in the difference between expectations in (3.18), is it usually assumed that $\mathcal{F}$ is a subset of $\mathcal{C}_b(\mathcal{X}, \tau)$. The probability distribution of the variables is usually completely identified with the MMD with respect to $\mathcal{C}_b(\mathcal{X}, \tau)$. In fact, if $(\mathcal{X}, d)$ is a metric space, then $\text{P} = \text{Q}$ if and only if $\text{P}(f) = \text{Q}(f)$, for all $f \in \mathcal{C}_b(\mathcal{X}, d)$ (see Dudley (2002, Lemma 9.3.2)). However, the class $\mathcal{C}_b(\mathcal{X}, d)$ is in general too large to deal with it, so that suitable subsets are usually employed in practice. Another possibility is assuming that the functions $f \in \mathcal{F}$ satisfy that $\sup_{x\in\mathcal{X}} \left(\frac{|f(x)|}{b(x)}\right) < \infty$, for a measurable function $b : \mathcal{X} \longrightarrow [1,\infty)$ such that $\mathrm{P}(b), \mathrm{Q}(b) < \infty$. For simplicity, in the following we will not mention these necessary integrability requirements and we will assume that $\sup_{f\in\mathcal{F}} (\mathrm{P}(f)), \sup_{f\in\mathcal{F}} (\mathrm{Q}(f)) < \infty$. When this condition is satisfied, it is said that $\mathcal{F}$ is *integrally bounded.*

We observe that when $\mathcal{F}$ is symmetric, that is, $-f \in \mathcal{F}$ whenever $f \in \mathcal{F}$, we have that $\mathrm{MMD}[\mathcal{F}, \mathrm{P}, \mathrm{Q}] = \sup_{f\in\mathcal{F}} (|\mathrm{P}(f) - \mathrm{Q}(f)|)$. In other words, the MMD in (3.18) is the *integral probability metric* generated by $\mathcal{F}$ (see Müller (1997)). In Rachev et al. (2013, Section 4.4), it is also said that the metric has a ζ-structure (Zolotarev (1983)). In this section we will also assume that $\mathcal{F}$ is symmetric.

Some frequently used probability metrics can be expressed as $\mathrm{MMD}[\mathcal{F}, \mathrm{P}, \mathrm{Q}]$, for a suitable choice of the set of functions $\mathcal{F}$. In the following examples X and Y are two random variables with distribution functions F and H and associated probability measures P and Q, respectively.

1. *Kolmogorov metric.* This distance is $\|F - H\|_\infty$, which is the integral probability metric generated by $\mathcal{F} = \left\{\mathbf{1}_{(-\infty,x]} : x \in \mathbb{R}\right\}$. Further, it is also generated by the set of all functions of bounded variation 1 (see Müller (1997, Theorem 5.2)).

2. L^p *metrics.* For $1 \leq p < \infty$, this metric is defined by $d_p(F, H) = \|F - H\|_{\mathrm{L}^p}$ ($\|\cdot\|_{\mathrm{L}^p}$ being the usual L^p-norm). When X and Y are integrable, d_p admits the dual representation (see Rachev et al. (2013, p. 73)) $d_p(F, H) = \mathrm{MMD}\left[\mathcal{F}_p, \mathrm{P}, \mathrm{Q}\right]$, where $\mathcal{F}_p$ is the class of all Lebesgue a.e. differentiable functions f such that the derivative f' satisfies $\|f'\|_{\mathrm{L}^q} \leq 1$ (q being the conjugate of p, i.e., q is such that $\frac{1}{p} + \frac{1}{q} = 1$).

3. *Wasserstein metric.* This distance is a particular and important case of the L^p-metric with $p = 1$. Its generator is also the class $\mathcal{F}_{\mathrm{W}}$ of functions $f : \mathbb{R} \longrightarrow \mathbb{R}$ such that satisfies the Lipschitz condition $|f(x) - f(y)| \leq |x - y|$, for all $(x, y) \in \mathbb{R}^2$. By the Kantorovich–Rubinstein's theorem, $\|F - H\|_1 = \mathrm{MMD}\left[\mathcal{F}_{\mathrm{W}}, \mathrm{P}, \mathrm{Q}\right]$. The Wasserstein distance is also known as Kantorovich–Rubinstein distance. In the context of image processing, this metric is called the *earth mover's distance* (see Rubner et al. (2000)). The importance of the Wasserstein metric, as well as its relevance for optimal transport problems, has been summarized in Villani (2008, Section 6).

4. *Bounded Lipschitz metric.* This metric (see Huber (2011, p. 29)) is the integral probability metric generated by $\mathcal{F}_{\mathrm{BL}}$, the class of functions $f : \mathbb{R} \longrightarrow \mathbb{R}$ such that $\|f\|_{\mathrm{BL}} \leq 1$, where $\|f\|_{\mathrm{BL}} = \|f\|_{\mathrm{L}} + \|f\|_\infty$ and $\|\cdot\|_{\mathrm{L}}$ is the Lipschitz norm given by
$$\|f\|_{\mathrm{L}} = \sup_{x\neq y\in\mathbb{R}} \left(\frac{|f(x) - f(y)|}{|x - y|}\right).$$

5. *Zolotarev ideal metrics of order r.* For $r \in \mathbb{N}$, let $\mathcal{Z}_r$ be the class of $(r-1)$-times continuously differentiable functions $f : \mathbb{R} \longrightarrow \mathbb{R}$ satisfying the Lipschitz condition

$|f^{(r-1)}(x) - f^{(r-1)}(y)| \leq |x-y|$, for all $(x,y) \in \mathbb{R}^2$. (Here we use the notation $f^{(0)} \equiv f$.) The class $\mathcal{Z}_r$ can also be substituted by the set of functions f having r-th derivative $f^{(r)}$ a.e. and such that $|f^{(r)}| \leq 1$ a.e. The metric $\zeta_r = \mathrm{MMD}\left[\mathcal{Z}_r, \mathrm{P}, \mathrm{Q}\right]$ is called the *Zolotarev metric of order r* (see Rachev et al. (2013) for a general reference and properties of these distances). Convergence in ζ_r-metric implies weak convergence plus convergence of the r-th absolute moment. Zolotarev metrics have been used in Rao (1997) to obtain a CLT for independent, non-identically distributed random variables. As mentioned in Rachev et al. (2013, Section 15), the case $r = 2$ is appropriate for investigating some ageing properties of lifetime distributions. In Baíllo et al. (2019), ζ_2 has also been used to generate new distance measures for classifying X-ray astronomy data into stellar classes. The metric ζ_3 has been considered in the context of distributional recurrences (see Neininger and Rüschendorf (2004a) and Neininger and Rüschendorf (2004b)).

6. *Zolotarev metric of order r in* L^p: For $r \in \mathbb{N}$, and $1 \leq p \leq \infty$, the metric $\zeta_{r,p}$ is generated by $\mathcal{Z}_{r,p}$, the set of functions $f : \mathbb{R} \longrightarrow \mathbb{R}$ for which $f^{(r+1)}$ exists and satisfies $\|f^{(r+1)}\|_{\mathrm{L}^q} \leq 1$, where q is the conjugate of p ($\frac{1}{p} + \frac{1}{q} = 1 \Leftrightarrow q = \frac{p}{p-1}$). Note that $\zeta_{r,1} \equiv \zeta_{r+1}$ (the Zolotarev ideal metric of order $r+1$). In risk theory, the metrics $\zeta_{1,\infty}$ and $\zeta_{1,1}$ are respectively called the *stop-loss distance* and the *integrated stop-loss distance* (see Denuit et al. (2006)).

7. *Kernel distances*: when the class is $\mathcal{F}_{\mathcal{H}_k}$. This distances are treated in depth in Chapter 4 together with the new proposal *supremum kernel distances (SKD)* or *uniform kernel distances (UKD)*. In this second type of distances, the class of functions is the union of unit balls $\mathcal{F}_{\mathcal{H}_{k,\Lambda}}$.

3.4.2 A general asymptotic result for the MMD

The use of the empirical counterpart of the MMD was already considered in Fortet and Mourier (1953) and it has been extensively employed in machine learning when $\mathcal{F}$ is the unit ball in a reproducing kernel Hilbert space (RKHS) (see Chapter 4). In Sriperumbudur et al. (2012), the authors showed the consistency and rate of convergence of some estimators of various integral probability metrics. The asymptotic behaviour of an estimator of the Zolotarev metric of order r in L^p has been discussed in Cárcamo (2017). Here we provide a general result regarding the estimation of the MMD. We only consider the two sample case as this situation is the most frequently considered in the literature, but similar results can be obtained in the one sample case.

Let us consider $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ two independent random samples from X and Y with probability measures P and Q, respectively. We denote by $\mathbb{P}_n$ and $\mathbb{Q}_m$ the empirical measures associated with these samples (see Subsection 1.2.2). Given a class of functions $\mathcal{F}$, the empirical counterpart of $\mathrm{MMD}[\mathcal{F}, \mathrm{P}, \mathrm{Q}]$ in (3.18) is given by

$$\mathrm{MMD}\left[\mathcal{F}, \mathbb{P}_n, \mathbb{Q}_m\right] = \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \frac{1}{m} \sum_{j=1}^{m} f(Y_j) \right). \tag{3.19}$$

In this section we are interested in the asymptotic behavior of the quantity

$$M_{m,n} = \sqrt{N}\left(\text{MMD}\left[\mathcal{F}, \mathbb{P}_n, \mathbb{Q}_m\right] - \text{MMD}[\mathcal{F}, \text{P}, \text{Q}]\right), \quad \text{with} \quad N = \frac{n\,m}{n+m}. \tag{3.20}$$

We observe that $M_{m,n}$ is precisely $D_{n,m}(\sigma) = D_\sigma\left(\text{P} - \text{Q}, \mathbb{P}_n - \mathbb{Q}_m, \sqrt{N}\right)$ in (2.3), where the underlying space is $\mathfrak{X} = \mathcal{F}$. Therefore, from Theorem 38, to derive the asymptotic distribution of $M_{m,n}$ in (3.20) we only need to study the weak converge in $\ell^\infty(\mathcal{F})$ of the two sample empirical process $\mathbb{G}_{n,m} = \sqrt{N}\ (\mathbb{P}_n - \mathbb{Q}_m - \text{P} + \text{Q})$ (see Subsection 1.2.2).

With all these ingredients, the main result in this section that determines the asymptotic distribution of the statistic (3.19) is stated.

Theorem 50. *Let X and Y be two random variables with probability measures* P *and* Q*, respectively. Let us assume that*

(a) The sampling scheme is balanced, that is, $\frac{n}{n+m} \to \xi$, with $0 \le \xi \le 1$, as $n, m \to \infty$.

(b) The class $\mathcal{F}$ is simultaneously P *and* Q*-Donsker.*

We consider the pseudometric ρ on $\mathcal{F}$ given by

$$\rho(f,g)^2 = \rho_{\text{L}^2(\text{P})}(f,g)^2 + \rho_{\text{L}^2(\text{Q})}(f,g)^2, \quad f, g \in \mathcal{F}, \tag{3.21}$$

where $\rho_{\text{L}^2(\text{P})}$ and $\rho_{\text{L}^2(\text{Q})}$ are the intrinsic L^2*-pseudometrics of* P *and* Q *respectively. We have that $(\mathcal{F}, \rho)$ is a totally bounded pseudometric space, the functional* P − Q *belongs to $\mathcal{C}_u(\mathcal{F}, \rho)$ and the statistic $M_{n,m}$ defined in* (3.20) *satisfies that*

$$M_{n,m} \rightsquigarrow \sup_{\overline{M}^+(D,\rho)} (\mathbb{G}),$$

where $\mathbb{G} = \sqrt{1-\xi}\,\mathbb{G}_\text{P} - \sqrt{\xi}\,\mathbb{G}_\text{Q}$ is defined in (1.4) *and*

$$\overline{M}^+(D,\rho) = \left\{f \in \overline{\mathcal{F}} : \text{P}(f) - \text{Q}(f) = \text{MMD}[\mathcal{F}, \text{P}, \text{Q}]\right\},$$

with $\bar{\mathcal{F}}$ being the ρ-completion of $\mathcal{F}$.

Proof. First, from (a) and (b) we have that $\mathbb{G}_{n,m} \rightsquigarrow \mathbb{G}$, where $\mathbb{G}_{n,m} = \sqrt{\frac{nm}{n+m}}\ (\mathbb{P}_n - \mathbb{Q}_m - \text{P} + \text{Q})$ is the two-sample empirical process defined in Subsection 1.2.2. Hence, by Theorem 38, $M_{n,m} \rightsquigarrow \sigma'_D(\mathbb{G})$. Now, as $\mathcal{F}$ is P and Q-Donsker, the pseudo-metric spaces $(\mathcal{F}, \rho_\text{P})$ and $(\mathcal{F}, \rho_\text{Q})$ are totally bounded, where ρ_P and ρ_Q are the natural pseudo-metrics given in Subsection 1.2.1 (see also Giné and Nickl (2021, Remark 3.7.27)). Further, $\mathbb{G}_\text{P} \in \mathcal{C}_u\left(\mathcal{F}, \rho_\text{P}\right)$ and $\mathbb{G}_\text{Q} \in \mathcal{C}_u\left(\mathcal{F}, \rho_\text{Q}\right)$ a.s. Now, as by assumption $\underset{f\in\mathcal{F}}{|\text{P}(f)|}, \underset{f\in\mathcal{F}}{|\text{Q}(f)|} < \infty$, then we can conclude that $\left(\mathcal{F}, \rho_{\text{L}^2(\text{P})}\right)$ and $\left(\mathcal{F}, \rho_{\text{L}^2(\text{Q})}\right)$ are also totally bounded. It is easy to check that this implies that $(\mathcal{F}, \rho)$ is also totally bounded, where ρ is in (3.21). On the other hand, by Cauchy–Schwarz inequality, we have that $\text{P} - \text{Q} \in \mathcal{C}_u(\mathcal{F}, \rho)$ and also that the paths of $\mathbb{G}$ are in $\mathcal{C}_u(\mathcal{F}, \rho)$ a.s. since $\rho_\text{P}, \rho_\text{Q} \le \rho$. Therefore, the conclusion follows by applying Corollary 26 (b). □

Condition (b) in Theorem 50 is the key assumption to apply the previous result. In other words, we have to ensure that $\mathcal{F}$ is P,Q-Donsker. There are many results in the literature on empirical proceses guaranteeing that a class of functions is Donsker (see A. van der Vaart and Wellner (1996)). For instance, it is well-known that the set of indicators generating the Kolmogorov distance is universal Donsker. The unit ball for the Bounded Lipschitz metric is Donsker whenever the underlying distribution has some finite moments (see Nickl and Pötscher (2007, Corollary 5 and Remark 2)). In the same work, Nickl and Pötscher (2007) showed that bounded subsets of general function spaces defined over $\mathbb{R}^d$ are Donsker under some appropriate conditions. Examples include (weighted) Besov, Sobolev, Hölder, and Triebel type spaces. Some of these results have been extended in Sriperumbudur (2016).

Chapter 4

A two-sample test based on kernel distances

In this chapter, a suitable version of the so-called "kernel trick" is used to devise two-sample tests especially focused on high-dimensional and functional data. The proposal entails a simplification of the practical problem of selecting an appropriate kernel function. Specifically, we apply a uniform variant of the kernel trick which involves the supremum within a class of kernel-based distances. We obtain the asymptotic distribution of the test statistic under the null and alternative hypotheses. The proofs rely on empirical processes theory, combined with the Delta method and Hadamard directional differentiability techniques (see Chapter 1), and functional Karhunen-Loève-type expansions of the underlying processes. This methodology has some advantages over other standard approaches in the literature. We also give some experimental insight into the performance of our proposal compared to other popular approaches: in particular, we have considered the original kernel-based proposal by Gretton et al. (2007), as well as some variants of it based on splitting methods, and a test based on energy distances presented in Székely and Rizzo (2017).

4.1 Introduction

In this section we provide an extended summary including not only the main ideas of this chapter but, specially, the general setting, motivation and related literature, as well as the technical tools we use.

The kernel trick and some potential kernel traps

We focus on statistical problems where, essentially, the aim is to properly separate data coming from two different populations; this is the case of binary supervised classification and two-sample testing problems. In such situations, the *kernel trick* is a common paradigm. In a few words, the standard multivariate version (i.e., with data in $\mathbb{R}^d$) of the kernel trick lies in separating the data in both populations using a symmetric non-negative definite "kernel function". The values of the kernel can be seen as the inner product of transformed versions of the original observations in a different (usually higher-dimensional) space. It is expected that the groups can be better distinguished in the new final space; see Scholkopf and Smola (2018).

We are particularly interested in those situations in which the available data are high-dimensional or even functional (thus, infinite-dimensional). In such cases, the strategy of mapping the data into a higher-dimensional space does not seem to be so compelling. Still, the kernel trick remains meaningful in a sort of "second generation" version, whose point is to take the data to a more comfortable and flexible space. In this new space, the statistical methodology might be mathematically more tractable, and more easily implemented and interpreted. To be more precise, a probability distribution P on the sample space $\mathcal{X}$ is replaced with the function

$$\mu_{\mathrm{P}}(x) = \int_{\mathcal{X}} k(x,y)\ \mathrm{dP}(y), \quad x \in \mathcal{X}, \tag{4.1}$$

(its mean embedding, see Subsection 1.3.1) in an appropriate space of "nice functions" defined by means of the kernel k: the RKHS $\mathcal{H}_k$. In this way, the distance between two probability measures is computed in terms of the metric in the functional space (see also 3.4). As a matter of fact, one of the most appealing proposals in this direction relies on kernel-based distances, expressed in terms of the embedding transformation μ_{P} in (4.1); see Gretton et al. (2007).

The kernel k involved in this methodology depends, almost unavoidably, on some tuning parameter λ, typically a scale factor. Therefore, we actually have a *family* of kernels, k_λ, for $\lambda \in \Lambda$, where Λ is usually a subset of $\mathbb{R}^d$ $(d \geq 1)$. For instance, the popular family of *Gaussian kernels* with parameter $\lambda \in (0, \infty)$ is defined by

$$k_\lambda(x,y) = \exp\left(-\lambda \left\|x-y\right\|^2\right), \quad \text{for } x, y \in \mathcal{X}, \tag{4.2}$$

where $\|\cdot\|$ is a norm in $\mathcal{X}$. Unfortunately, there is no general rule to know *a priori* which kernel works best with the available data. In other words, the choice of λ is, to some extent, arbitrary but not irrelevant, as it could remarkably affect the final output. For example, very small or very large choices of λ in (4.2) result in null discrepancies, which have no ability to distinguish distributions. The selection of λ is hence a delicate problem that has not been satisfactorily solved so far. This is what we call the *kernel trap*: a bad choice of the parameter leading to poor results. Although this problem was not explicitly considered in Gretton et al. (2007), the authors are aware of this relevant question and uses a heuristic choice of λ for the finite-dimensional Gaussian kernel.

Further, a parameter-dependent method might be an obstacle for practitioners who are often reluctant to use procedures depending on auxiliary, hard-to-interpret parameters. We thus find here a particular instance of the trade-off between power and applicability: as stated in Tukey (1959), the *practical power* of a statistical procedure is defined as "the product of the mathematical power by the probability that the procedure will be used" (Tukey credits to Churchill Eisenhart for this idea). From this perspective, our proposal can be viewed as an attempt to make kernel-based homogeneity tests more usable by getting rid of the tuning parameter(s). Roughly speaking, the idea that we propose to avoid selecting a specific value of λ within the family $\{k_\lambda : \lambda \in \Lambda\}$ is to take the supremum over the set of parameters Λ of the resulting family of kernel-distances. We call this approach the *uniform kernel trick*, as we map the data into many functional spaces at the same time and use, as test statistic, the supremum of the corresponding kernel distances. We believe that this methodology could be successfully applied as well in supervised classification, though this topic is not considered in this work.

The topic under study: two-sample problems

Two-sample tests, also called homogeneity tests, aim to decide whether or not it can be accepted that two random elements have the same distribution, using the information provided by two independent samples. This problem is omnipresent in practice on account of their applicability to a great variety of situations, ranging from biomedicine to quality control. Since the classical Student's t-tests or rank-based (Mann-Whitney, Wilcoxon, ...) procedures, the subject has received an almost permanent attention from the statistical community. In this work we focus on two-sample tests valid, under broad assumptions, for general settings in which the data are drawn from two random elements X and Y taking values in a general space $\mathcal{X}$. The set $\mathcal{X}$ is the *sample space*, or *feature space* in the Machine Learning language. In the important particular case $\mathcal{X} = \mathrm{L}^2([0,1])$, X and Y are stochastic processes and the two-sample problem lies within the framework of Functional Data Analysis (FDA).

Many important statistical methods, including goodness of fit and homogeneity tests, are based on an appropriate metric (or discrepancy measure) that allows groups or distributions to be distinguished. Probability distances or semi-distances reveal to the practitioner the dissimilarity between two random quantities. Therefore, the estimation of a suitable distance helps detect significant differences between two populations. Some well-known, classic examples of such metrics are the Kolmogorov distance, that leads to the popular Kolmogorov-Smirnov statistic, and L^2-based discrepancy measures, leading to Cramér-von Mises or Anderson-Darling statistics. These methods, based on cumulative distribution functions, are no longer useful with high-dimensional or non-Euclidean data, as in FDA problems. For this reason we follow a different strategy based on more adaptable metrics between general probability measures.

The *energy distance* (see the review by Székely and Rizzo (2017)) and the associated *distance covariance*, as well as *kernel distance*, represent a step forward in this direction since they can be calculated with relative ease for high-dimensional distributions. In Sejdinovic et al. (2013) the relationships among these metrics in the context of hypothesis testing are discussed. In this chapter we consider an extension, as well as an alternative mathematical approach, for the two-sample test in Gretton et al. (2007). These authors show that kernel-based procedures perform better than other more classical approaches when dimension grows, although they are strongly dependent on the choice of the kernel parameter.

Kernel distances

To present the contributions of this chapter, we briefly refer to some important, mutually related, technical notions. As emphasized in Berlinet and Thomas-Agnan (2011), *Reproducing Kernel Hilbert Spaces* (RKHS in short) provide an excellent environment to construct helpful transformations in several statistical problems. Given a topological space $\mathcal{X}$ (in many applications $\mathcal{X}$ is a subset of a Hilbert space), remind that a *kernel* k is a real non-negative semidefinite symmetric function on $\mathcal{X} \times \mathcal{X}$. See Section 1.3 for additional details.

Let $\mathcal{M}_{\mathrm{p}}(\mathcal{X})$ be the set of (Borel) probability measures on $\mathcal{X}$. Under any of the conditions of Proposition 16, the functions in $\mathcal{H}_k$ are measurable and P-integrable, for each $\mathrm{P} \in \mathcal{M}_{\mathrm{p}}(\mathcal{X})$. Moreover, the function

$$\mu_{\mathrm{P}}(\cdot) = \int_{\mathcal{X}} k(\cdot, y)\, \mathrm{dP}(y),$$

(also in (4.1)) belongs to $\mathcal{H}_k$. The transformation $\mathrm{P} \mapsto \mu_{\mathrm{P}}$ from $\mathcal{M}_{\mathrm{p}}(\mathcal{X})$ to $\mathcal{H}_k$ is called the *(kernel) mean embedding* (see Sejdinovic et al. (2013) and Berlinet and Thomas-Agnan (2011, Chapter 4), Definition 13).

The *kernel distance* between P and Q in $\mathcal{M}_{\mathrm{p}}(\mathcal{X})$ is

$$d_k(\mathrm{P}, \mathrm{Q}) = \|\mu_{\mathrm{P}} - \mu_{\mathrm{Q}}\|_{\mathcal{H}_k} = \left(\int_{\mathcal{X}} \int_{\mathcal{X}} k(x, y)\, \mathrm{d}(\mathrm{P}-\mathrm{Q})(y)\, \mathrm{d}(\mathrm{P}-\mathrm{Q})(x) \right)^{1/2}, \tag{4.3}$$

where $\| \cdot \|_{\mathcal{H}_k}$ stands for the norm in $\mathcal{H}_k$ and $\mathrm{P} - \mathrm{Q}$ denotes the (signed) measure on $\mathcal{X}$. Therefore, $d_k(\mathrm{P}, \mathrm{Q})$ is the RKHS distance between the mean embeddings of the corresponding probability measures. Kernel distances were popularized in machine learning as tools to tackle several relevant statistical problems, such as homogeneity tests Gretton et al. (2006), independence Gretton et al. (2007), test of conditional independence Fukumizu et al. (2007) and density estimation Sriperumbudur (2011). The key idea behind this methodology can be seen as a particular case of the fruitful kernel trick paradigm.

Contributions: the uniform kernel trick

We consider a family of kernels $\{k_\lambda : \lambda \in \Lambda\}$, where Λ is certain parametric space. For the Gaussian kernel in (4.2), $\Lambda = (0, \infty)$, but in general λ could be a multidimensional parameter, as in the case of Matérn kernels or inverse quadratic kernels; see Sriperumbudur (2016, p. 1846). Each k_λ has an associated RKHS, $\mathcal{H}_{k,\lambda}$ (endowed with its intrinsic norm $\| \cdot \|_{\mathcal{H}_{k,\lambda}}$), and the corresponding probability distance $d_{k,\lambda}$. For $\mathrm{P}, \mathrm{Q} \in \mathcal{M}_{\mathrm{p}}(\mathcal{X})$, we want to test $\mathrm{H}_0 : \mathrm{P} = \mathrm{Q}$ using the distances within the collection $\{d_{k,\lambda} : \lambda \in \Lambda\}$. The current theoretical framework does not support the automatic (data-driven) choice of $\lambda \in \Lambda$, since the asymptotic theory is mainly developed for a fixed kernel, corresponding to a specific value of λ. However, the choice of λ is a non-trivial and sensitive issue with no obvious best solution, and which might affect the test performance.

There are various interesting proposals to deal with this problem in practice: the median heuristic of Gretton et al. (2007); sample-splitting and optimization methods in

Gretton, Sejdinovic, et al. (2012) and Liu et al. (2020); and aggregation methods such as Gretton, Sejdinovic, et al. (2012). In this chapter we explore an alternative to avoid making a parametric decision or splitting the data set. Our proposal can be included within the aggregative methods: we combine the information provided by different kernels by taking the supremum over the induced kernel metrics. Specifically, we use the quantity that "best separates" P and Q, that is, the supremum of all kernel distances given by

$$d_{k,\Lambda}(\mathrm{P},\mathrm{Q}) = \sup_{\lambda\in\Lambda}\left(d_{k,\lambda}(\mathrm{P},\mathrm{Q})\right) = \sup_{\lambda\in\Lambda}\left(\left\|\mu_{\mathrm{P}}^{\lambda} - \mu_{\mathrm{Q}}^{\lambda}\right\|_{\mathcal{H}_{k,\lambda}}\right), \quad \mathrm{P},\mathrm{Q}\in\mathcal{M}_{\mathrm{p}}(\mathcal{X}), \tag{4.4}$$

where, for $\lambda\in\Lambda$, $\mu_{\mathrm{P}}^{\lambda}$ and $\mu_{\mathrm{Q}}^{\lambda}$ are the mean embeddings of P and Q, respectively, in $\mathcal{H}_{k,\lambda}$. We call the quantity in (4.4) the *supremum (or uniform) kernel distance* of $\{k_\lambda : \lambda\in\Lambda\}$. Also, the *uniform kernel trick* refers to the overall idea of using (4.4) to eliminate the parameter in kernel-based statistics. Observe that d_k (4.3) is a particular case of $d_{k,\Lambda}$ in (4.4) when Λ has one element. Therefore, all the results in this chapter can be applied for usual kernel distances. In addition, in the family $\{k_\lambda : \lambda\in\Lambda\}$ we can include kernels from different parametric families, which would generate more robust test statistics that might work well under many types of alternatives.

The supremum kernel distance (4.4) entails several advantages and some mathematical challenges: First, the kernel selection problem is considerably simplified and solved in a natural way. Additionally, the approach is general enough to be applied in infinite-dimensional settings as FDA. This is interesting since in FDA there are only a few homogeneity tests in the literature. Some of them have been developed in the setting of ANOVA models (involving several samples) under homoscedasticity (equal covariance operators of the involved processes) and Gaussian assumptions. Hence, the current methodologies amount to testing the null hypothesis of equal means in all the populations; see, e.g., Cuevas et al. (2004) for an early contribution and J.-T. Zhang (2013) for a broader perspective. Our proposal is therefore quite related to more general approaches, not requiring any homoscedasticity assumption and still valid for a FDA framework. Examples of such similar tests are Hall and Van Keilegom (2007) and Pomann et al. (2016), as well as the random projections-based methodology in Cuesta-Albertos et al. (2007).

The inclusion of the supremum in (4.4) represents an additional difficulty. The asymptotic properties of the test statistic based on (4.4) are derived by following a different strategy from that of Gretton et al. (2007) and later works. The methodology proposed here allows us to cope with the supremum and applies directly to the case of unequal sample sizes. In short, our approach can be described as follows: First, we consider plug-in estimators of the kernel distances, obtained by replacing the unknown distributions by their empirical counterparts (see Subsection 1.2.2 for further details). Then, we use the powerful theory of empirical processes together with some recent results on the differentiability of the supremum (see Chapter 2) and functional Karhunen-Loève expansions of the underlying processes. These developments entail several technical difficulties from the mathematical point of view. However, they are worthwhile since they allow us to analyze the asymptotic behavior, under both the null and the alternative hypothesis, of the two-sample test based on (4.4).

The structure of this chapter

In Section 4.2 we provide some preliminaries regarding mean embedding basics and empirical processes (see also Subsections 1.3.1 and 1.2.2, respectively). While most of this background is well-known or can be found in the literature, it is included here to introduce the necessary notation and make the thesis as self-contained as possible. Section 4.3 contains the main theoretical contributions. First, we obtain a Donsker property for (unions of) unit balls in RKHS that could be of independent interest. We establish the asymptotic validity under the null hypothesis of the two-sample test based on the distance (4.4). The asymptotic statistical power (i.e., the behaviour under the alternative hypothesis of non-homogeneity) is also analysed. An empirical study, comparing the uniform kernel test with some other competitors is presented in Section 4.4. In the considered scenarios, SKD is competitive with other kernel-based methods, especially in the case of heteroscedastic populations. However, given the limited nature of the study, we cannot conclude that our proposal unequivocally outperforms existing approaches. Some concluding remarks are included in Section 4.5. Finally, Section 4.6 collects the proofs of the main theoretical results.

4.2 Preliminaries

In this section we describe various tools that we use throughout this chapter.

Kernel distances as integral probability metrics

Each $\mathrm{P} \in \mathcal{M}_{\mathrm{p}}(\mathcal{X})$ (Borel probability measure on $\mathcal{X}$), can be seen as a linear functional on $\mathcal{H}_k$ via the mapping

$$f \in \mathcal{H}_k \mapsto \mathrm{P}(f) = \int_{\mathcal{X}} f \, \mathrm{dP}, \tag{4.5}$$

whenever $\mathcal{H}_k \subset \mathrm{L}^1(\mathrm{P})$. This condition is also equivalent to saying that the function $x \mapsto k(x,\cdot)$ is Pettis integrable (with respect to P) and to the existence of the mean embedding μ_{P} in (4.1) as an element of $\mathcal{H}_k$ satisfying, by Riesz's representation theorem, that

$$\mathrm{P}(f) = \langle f, \mu_{\mathrm{P}} \rangle_{\mathcal{H}_k}, \quad \text{for } f \in \mathcal{H}_k. \tag{4.6}$$

Sufficient conditions guaranteeing the injectivity of the mean embedding transformation can be found in Sriperumbudur et al. (2011). Note that in (4.5) (and what follows) we use the standard notation in empirical processes theory: $\mathrm{P}(f)$ (see Section 1.2).

The existence of the mean embedding implies that the kernel distance in (4.3), as well as the supremum kernel distance in (4.4), are well-defined. Indeed, they are *integral probability metrics*; see Section 3.4.1 (see also Müller (1997) for further details). To see this, let us consider the unit ball of $\mathcal{H}_k$, that is, $\mathcal{F}_{\mathcal{H}_k}$. We have that

$$\begin{aligned} \|\mu_{\mathrm{P}} - \mu_{\mathrm{Q}}\|_{\mathcal{H}_k} &= \sup_{f \in \mathcal{F}_{\mathcal{H}_k}} \left(\langle f, \mu_{\mathrm{P}} - \mu_{\mathrm{Q}} \rangle_{\mathcal{H}_k} \right) \overset{(a)}{=} \sup_{f \in \mathcal{F}_{\mathcal{H}_k}} \left(\left\langle f, \int_{\mathcal{X}} k(\cdot, x) \, \mathrm{d}(\mathrm{P}-\mathrm{Q})(x) \right\rangle_{\mathcal{H}_k} \right) \\ &\overset{(b)}{=} \sup_{f \in \mathcal{F}_{\mathcal{H}_k}} \left(\int_{\mathcal{X}} \langle f, k(\cdot, x) \rangle_{\mathcal{H}_k} \, \mathrm{d}(\mathrm{P}-\mathrm{Q})(x) \right) \overset{(c)}{=} \sup_{f \in \mathcal{F}_{\mathcal{H}_k}} (\mathrm{P}(f) - \mathrm{Q}(f)), \end{aligned} \tag{4.7}$$

where (a) follows from the definition of mean embedding (4.1), (b) from Pettis integrability, and (c) from the reproducing property (see Definition 12); see also Gretton, Borgwardt, et al. (2012, Lemma 4). Thus, the kernel distance (4.3) is the integral probability metric generated by the class $\mathcal{F}_{\mathcal{H}_k}$. Therefore, the supremum kernel distance (4.4) admits the alternative representation

$$d_{k,\Lambda}(\mathrm{P},\mathrm{Q}) = \sup_{f\in\mathcal{F}_{\mathcal{H}_{k,\Lambda}}} (\mathrm{P}(f) - \mathrm{Q}(f)) \quad \text{with} \quad \mathcal{F}_{\mathcal{H}_{k,\Lambda}} = \bigcup_{\lambda\in\Lambda} \mathcal{F}_{\mathcal{H}_{k,\lambda}}, \tag{4.8}$$

where $\mathcal{F}_{\mathcal{H}_{k,\lambda}}$ is the unit ball in the RKHS space associated with k_λ. In other words, $d_{k,\Lambda}$ is the integral probability metric defined through the union of unit balls of the whole family of RKHS constructed with $\{k_\lambda : \lambda \in \Lambda\}$.

From the characterizations as integral probability metrics in (4.7) and (4.8), we conclude that d_k and $d_{k,\Lambda}$ satisfy the properties of a pseudo-metric (non-negativeness, symmetry, triangular property). However, to ensure the *identifiability property* of a metric d (i.e., $d(\mathrm{P},\mathrm{Q}) = 0$ if and only if $\mathrm{P} = \mathrm{Q}$) additional conditions are needed. It can be checked that when $\mathcal{X} = \mathbb{R}^d$, identifiability is satisfied for the usual kernels (such as the Gaussian kernel in (4.2)). However, when $\mathcal{X}$ is infinite-dimensional this type of results are more complicated; see Wynne and Duncan (2022) for a deep study of this topic for the Gaussian kernel (4.2). More details can also be found in Sriperumbudur et al. (2010) and Sriperumbudur et al. (2011).

Plug-in estimators and empirical processes

A simple and natural estimator of the supremum kernel distance (4.4) can be obtained by applying *the plug-in principle* in (4.8). Given two independent samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ from P and Q, respectively, we replace the unknown underlying probability measures P and Q with the observed empirical counterparts, $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\mathbb{Q}_m = \frac{1}{m}\sum_{i=1}^m \delta_{Y_i}$, (see Section 1.2). This leads to the estimator of $d_{k,\Lambda}(\mathrm{P},\mathrm{Q})$ in (4.8) given by

$$d_{k,\Lambda}(\mathbb{P}_n, \mathbb{Q}_m) = \sup_{f\in\mathcal{F}_{\mathcal{H}_{k,\Lambda}}} (\mathbb{P}_n(f) - \mathbb{Q}_m(f)) = \sup_{f\in\mathcal{F}_{\mathcal{H}_{k,\Lambda}}} \left(\frac{1}{n}\sum_{i=1}^n f(X_i) - \frac{1}{m}\sum_{j=1}^m f(Y_j) \right). \tag{4.9}$$

As a supremum over a class of functions is involved in (4.9), the theory of empirical processes comes into play naturally.

4.3 Main results

In this section we first show that (unions of) unit balls of RKHS are universal Donsker under mild conditions. This is an important technical result of independent interest that is the starting point in the proofs of the asymptotic results. We analyze the asymptotic behaviour of the plug-in estimator (4.9) of the supremum kernel distance in (4.4) and (4.8). The results are quite general as P and Q are assumed to be Borel probability measures on a separable metric space. The proofs are based on empirical processes theory together with the (extended) Delta method (Shapiro (1991, Theorem 2.1)) and some recent differentiability results for the supremum given in Chapter 2. This *differential approach* differs from previous methods (as those in Gretton, Sejdinovic, et al. (2012) or Gretton, Borgwardt, et al. (2012)) in which the theory of U-statistics is used to derive the asymptotic results. Our approach has some advantages: it is applicable to variables taking values in general spaces, including functional spaces, and the equal sample size constraint of previous works is removed. Furthermore, the results are applicable in other contexts (such as tests for equality between two copulas) by just changing the underlying stochastic process in the spirit of Chapter 3.

Another essential difference between our methodology and other approaches is the way in which the tuning parameter λ is treated. The asymptotic theory in Gretton, Borgwardt, et al. (2012) (and other related works) is derived for a fixed kernel, while the experiments incorporate the Gaussian kernel in (4.2) with a data-driven choice of λ. As pointed out by the authors, an automatic method for selecting λ is an interesting area of research with some theoretical implications: setting the kernel using the sample being tested might change the asymptotic distribution. Regarding this, we note that our procedure to deal with the tuning parameter λ is fully incorporated in the asymptotic analysis thanks to the use of the supremum kernel distance (4.4).

The hypotheses

We list some assumptions for later reference. We briefly explain the meaning and implications of each of them. Remind that k is a positive definite kernel and $\{k_\lambda : \lambda \in \Lambda\}$ is a family of positive definite kernels which might come from different parametric families.

(Reg) *Regularity assumption.* $\mathcal{X}$ is a separable metric space and each kernel is continuous as a real function of one variable (with the other kept fixed).

(Dom) *Dominance assumption.* There exists a constant $c > 0$ such that $k_\lambda \ll c\,k$, for all $\lambda \in \Lambda$. Further, k is bounded on the diagonal, that is, $\sup_{x\in\mathcal{X}}(k(x,x)) < \infty$.

(Ide) *Identifiability assumption.* If $\mathrm{P} \neq \mathrm{Q}$, there exists $\lambda \in \Lambda$ such that $\mu_{\mathrm{P}}^{\lambda} \neq \mu_{\mathrm{Q}}^{\lambda}$.

(Par) *Continuous parametrization.* Λ is a compact subset of $\mathbb{R}^k$ (with $k \in \mathbb{N}$) and, for a fixed $(x, y) \in \mathcal{X} \times \mathcal{X}$, the function $\lambda \mapsto k_\lambda(x, y)$ is continuous from Λ to $\mathbb{R}$.

(Sam) *Sampling scheme.* The sampling scheme is balanced, that is, $\frac{n}{(n+m)} \to \xi$, with $\xi \in [0, 1]$, as $n, m \to \infty$.

Assumptions (Dom), (Ide), and (Par) were already stated and commented in Section 2.7, in the context of Hadamard directional differentiability. Assumptions (Reg) and (Dom) together have important consequences. Firstly, they imply that $\mathcal{H}_{k,\lambda}$ is constituted by continuous and bounded functions, therefore measurable and integrable. Moreover, under these two conditions the mean embedding $\mu_{\mathrm{P}}^{\lambda}$ exits (for each P and λ). In particular, the supremum kernel distance (4.4) is well-defined. (Reg) and (Dom) are also essential to show that the class $\mathcal{F}_{\mathcal{H}_{k,\Lambda}}$ in (4.8) is universal Donsker, which is a key point in the proofs of the following theorems. Furthermore, we also observe that (Ide) is not specifically required to obtain the asymptotic distribution under H_0 in Theorem 52. Finally, (Sam) is necessary for the combination of the associated empirical processes to converge.

Examples of families of kernels

The hypotheses above can be verified for most families of kernels that are used in practice by properly choosing the parameter space. The most demanding assumption about the kernel family is perhaps (Dom). This condition is always satisfied (in any dimension) for finite families of kernels (i.e., when Λ is a finite set) that are bounded on the diagonal. In this case, each element of the collection of positive definite kernels is dominated by the sum of them. In particular, this always ensures the applicability of the results, both in high and infinite dimensions, since one practical implementation of the procedure is carried out by choosing a grid of points in the parameter space; see Section 4.4.

When $\mathcal{X} = \mathbb{R}^d$, a finite-dimensional space, the usual parametric families of kernels often generate a nested collection of RKHS; see H. Zhang and Zhao (2013). This means that for $\lambda_1, \lambda_2 \in \Lambda$, there exists a constant $c = c(\lambda_1, \lambda_2, d)$ such that $k_{\lambda_1} \ll c\, k_{\lambda_2}$ (or the other way around). In such cases, (Dom) is valid for a compact subset Λ of the whole parametric space by using one of the kernels of the family as the bounding kernel k in (Dom). Some important examples included in this setting are the families of Gaussian and Laplacian kernels, inverse multiquadrics kernels, B-spline kernels, Matérn kernels, among others; see H. Zhang and Zhao (2013, Theorems 3.5, 3.6, and 3.7) and Sriperumbudur (2016).

Nevertheless, the problem is more delicate in infinite dimension. If $\mathcal{X} = \mathbb{R}^d$ and for the usual parametric families of kernels, the best constant c in "inequalities" of the form $k_{\lambda_1} \ll c\, k_{\lambda_2}$ depends on the dimension d and blows up when d goes to infinity; see H. Zhang and Zhao (2013, Theorems 3.5 and 3.6). Therefore, when the domain is functional (for instance, if $\mathcal{X} = \mathrm{L}^2([0,1])$), the task of finding a dominating kernel is more involved. An example can be built when the parameter space

$$\Lambda = \left\{ m \in \mathrm{L}^2([0,1]) : m \text{ absolutely continuous with } \int_0^1 |m'|^2 < 1 \right\},$$

is the unit ball of the Cameron-Martin space associated to the Wiener measure in $\mathcal{C}([0,1])$, the space of continuous functions on $[0,1]$. A family of kernels $k_m(x,y)$ fulfilling (Dom) and (Ide) can be constructed using Minlos-Sazanov Theorem (see e.g., Wynne and Duncan (2022, Th. 24)) and relying on ideas by H. Zhang and Zhao (2013, Prop. 3.1) and Wendland (2004, Chapter 10).

Additionally, we observe that (Dom) is fulfilled for families of positive linear (or convex) combinations of a finite family of kernels. In this example, the set of parameter Λ is given by the weights of the considered combinations; see Gretton, Sejdinovic, et al. (2012). We finally refer to Berlinet and Thomas-Agnan (2011, Chapter 7) and Paulsen and Raghupathi (2016, Chapter 4) for a wider catalog of families of kernels within this context.

A Donsker property for units balls in RKHS

Establishing that a class of functions is (uniform) Donsker has important consequences. This property is equivalent to having an empirical Central Limit Theorem, which is at the heart of most asymptotic results in statistics. Therefore, this kind of Donsker-type results are relevant by themselves and of independent interest. For example, in Sriperumbudur (2016, Theorem 4.3) (see also Giné and Nickl (2008), Giné and Nickl (2021)) it is shown that $\mathcal{F}_{\mathcal{H}_{k,\Lambda}}$ in (4.8) is Donsker for some specific finite-dimensional parametric families and for a suitable subset of Λ. Then, this result is applied to derive asymptotic distributions of kernel density estimators. In Sriperumbudur (2016), the proofs of the Donsker property for RKHS unit balls are obtained when $\mathcal{X} = \mathbb{R}^d$ by direct covering (entropy-based) arguments. The underlying bounds in these references depend on the dimension d. Therefore, it seems difficult to extend these Donsker-type statements to the infinite-dimensional case. However, Theorem 51 below is suitable for the general framework where $\mathcal{X}$ might be an infinite-dimensional space, and thus useful in statistical problems with functional data.

The following theorem establishes that unit balls (and even the union of units balls) of RKHS are universal Donsker. In the first part of the proof (in Section 4.6) we use Marcus (1985, Theorem 1.1), while in the second one we show that the union of unit balls is included in a ball of the space $\mathcal{H}_k$ by using Aronszajn's inclusion theorem (Aronszajn (1950, Theorem I)).

Theorem 51. *Let $\mathcal{X}$ be a separable metric space. Assume that the kernel k is bounded on the diagonal, that is, $\sup_{x\in\mathcal{X}}(k(x,x)) < \infty$, and $k(x,\cdot)$ is continuous, for each $x \in \mathcal{X}$. Then, the class $\mathcal{F}_{\mathcal{H}_k}$ is universal Donsker. Moreover, if $\{k_\lambda : \lambda \in \Lambda\}$ satisfies (Dom), then the union $\mathcal{F}_{\mathcal{H}_{k,\Lambda}}$ in* (4.8) *is universal Donsker as well.*

This theorem extends Sriperumbudur (2016, Theorem 4.3), where the Donsker property was shown under more demanding analytical conditions, to any family of kernels satisfying (Dom).

Asymptotic behavior under the null hypothesis, $\mathrm{P} = \mathrm{Q}$

The next theorem provides the asymptotic distribution of the (normalized) estimator of the supremum kernel distance (4.4) when the two samples come from the same distribution. In the statement of the following results, $\mathbb{G}_{\mathrm{P}}$ and $\mathbb{G}_{\mathrm{Q}}$ are $\mathcal{F}_{\mathcal{H}_{k,\Lambda}}$-indexed P and Q Brownian bridges, respectively (see Section 1.2), "$\rightsquigarrow$" stands for the usual convergence in distribution of (real) random variables, and $\mathcal{H}_{k,\lambda}^*$ denotes the dual space of $\mathcal{H}_{k,\lambda}$.

Theorem 52. *Let us assume that (Reg), (Dom) and (Sam) hold. If* $\mathrm{P} = \mathrm{Q}$, *the statistic* (4.9) *satisfies that*

$$\sqrt{\frac{n\,m}{n+m}}\, d_{k,\Lambda}(\mathbb{P}_n, \mathbb{Q}_m) \rightsquigarrow \sup_{\lambda\in\Lambda}\left(\left(\sum_{j\in\mathbb{N}} Z_{j,\lambda}^2\right)^{1/2}\right), \quad n, m \to \infty, \tag{4.10}$$

where $d_{k,\Lambda}$ is defined in (4.8), *$Z_{j,\lambda} = \langle \mathbb{G}_{\mathrm{P}}, \varphi_{j,\lambda}\rangle_{\mathcal{H}_{k,\lambda}^*}$ (for each $\lambda \in \Lambda$ and $j \in \mathbb{N}$) and $\varphi_{j,\lambda}$ is the j-th eigenfunction of the covariance operator of $\mathbb{G}_{\mathrm{P}}$ on $\mathcal{H}_{k,\lambda}^*$.*

Moreover, $\{Z_{j,\lambda}\}_{j\in\mathbb{N},\lambda\in\Lambda}$ are jointly Gaussian and for a fixed $\lambda \in \Lambda$, $\{Z_{j,\lambda}\}_{j\in\mathbb{N}}$ are independent with $Z_{j,\lambda} \sim \mathcal{N}(0, \beta_{j,\lambda})$, where $\beta_{j,\lambda}$ is the eigenvalue associated to $\varphi_{j,\lambda}$.

In the first step of the proof of this theorem we use Theorem 51 to derive the weak convergence of the underlying process. The rest of the proof is rather technical. The basic ideas are as follows: we use of the Continuous Mapping Theorem to obtain the convergence of the statistic; subsequently, we apply a functional Karhunen-Loève-type theorem in the dual space $\mathcal{H}^*_{k,\lambda}$ (Lemma 57 in Section 4.6) to the resulting limiting process to achieve (4.10). Note that in the family $\{k_\lambda : \lambda \in \Lambda\}$ we can include kernels from different parametric families or mixtures of kernels from distinct families in order to *robustify* the test statistic.

Theorem 52 complements in several directions other previous works on this topic, starting from Gretton et al. (2007, Th. 8). See also J.-T. Zhang et al. (2022), J.-T. Zhang and Smaga (2022) for more recent references.

Asymptotic behavior under the alternative, $\mathrm{P} \neq \mathrm{Q}$

The following theorem establishes the asymptotic distribution of (the normalized version) of (4.9) under the alternative hypothesis of the homogeneity test. Therefore, it provides the consistency of the testing procedure based on the supremum kernel distance. Additionally, this result might be potentially useful in order to develop tests of *almost homogeneity*, that is, problems in which we are interested in testing $\mathrm{H}_0 : d_{k,\Lambda}(\mathrm{P},\mathrm{Q}) \leq \varepsilon$ versus $\mathrm{H}_1 : d_{k,\Lambda}(\mathrm{P},\mathrm{Q}) > \varepsilon$, for some $\varepsilon > 0$. Analogously, this idea is also applicable to provide statistical evidence in favor of almost homogeneity when H_0 and H_1 above are interchanged. Related ideas can be found in del Barrio et al. (2020) and Dette and Kokot (2022).

Theorem 53. *Let us assume that (Reg), (Dom), (Par), (Ide) and (Sam) hold. If $\mathrm{P} \neq \mathrm{Q}$, we have that*

$$\sqrt{\frac{nm}{n+m}}\left(d_{k,\Lambda}\left(\mathbb{P}_n, \mathbb{Q}_m\right) - d_{k,\Lambda}(\mathrm{P},\mathrm{Q})\right) \rightsquigarrow \sup_{\lambda\in\Lambda_0}\left(\mathbb{G}\left(h^{+,\lambda}\right)\right) = \sup_{L}(\mathbb{G}), \tag{4.11}$$

where

$$\mathbb{G} = \sqrt{1-\xi}\,\mathbb{G}_{\mathrm{P}} - \sqrt{\xi}\,\mathbb{G}_{\mathrm{Q}},$$

where $\mathbb{G}$ was defined in Subsection 1.2.2

$$h^{+,\lambda} = \frac{\mu^\lambda_{\mathrm{P}} - \mu^\lambda_{\mathrm{Q}}}{\left\|\mu^\lambda_{\mathrm{P}} - \mu^\lambda_{\mathrm{Q}}\right\|_{\mathcal{H}_{k,\lambda}}},$$

$$\Lambda_0 = \left\{\lambda \in \Lambda : \left\|\mu^\lambda_{\mathrm{P}} - \mu^\lambda_{\mathrm{Q}}\right\|_{\mathcal{H}_{k,\lambda}} = d_{k,\Lambda}(\mathrm{P},\mathrm{Q})\right\} \quad \text{and} \quad L = \left\{h^{+,\lambda} : \lambda \in \Lambda_0\right\}.$$

Theorem 53 directly provides the consistency of the homogeneity test based on the supremum kernel distance $d_{k,\Lambda}$ in (4.4). We also observe that $\mathbb{G}$ is a zero mean Gaussian process indexed by $\mathcal{F}_{\mathcal{H}_{k,\Lambda}}$. Further, $h^{+,\lambda}$ is called *witness function* in Gretton et al. (2007) as the maximum mean discrepancy over $\mathcal{F}_{\mathcal{H}_{k,\lambda}}$ is attained at this element, that is, $\mathrm{P}\left(h^{+,\lambda}\right) - \mathrm{Q}\left(h^{+,\lambda}\right) = \left\|\mu_{\mathrm{P}}^{\lambda} - \mu_{\mathrm{Q}}^{\lambda}\right\|_{\mathcal{H}_{k,\lambda}}$. Therefore, the limit in (4.11) corresponds to the supremum of $\mathbb{G}$ over the set of witness functions for which the value of the uniform kernel distance is achieved. Regarding the proof of Theorem 53, we mention that the extended Delta method (see Shapiro (1991, Theorem 2.1)) plays a key role. First, we use Theorem 51 to show that $\mathbb{G}$ is the limit of the underlying process. Later, Corollary 37 to derive (4.11).

The following result is a direct consequence of Theorem 53 when the family of kernels has a single element, k.

Corollary 54. *Let us assume that (Reg), (Dom) and (Sam) hold. Further, we assume that k is characteristic. If $\mathrm{P} \neq \mathrm{Q}$, we have that*

$$\sqrt{\frac{nm}{n+m}}\left(d_k\left(\mathbb{P}_n, \mathbb{Q}_m\right) - d_k(\mathrm{P},\mathrm{Q})\right) \rightsquigarrow \mathbb{G}\left(h^{+}\right), \tag{4.12}$$

where $\mathbb{G}$ is in theorem 53 and

$$h^{+} = \frac{\mu_{\mathrm{P}} - \mu_{\mathrm{Q}}}{\left\|\mu_{\mathrm{P}} - \mu_{\mathrm{Q}}\right\|_{\mathcal{H}_k}}.$$

In particular, the distribution of $\mathbb{G}\left(h^{+}\right)$ is normal with mean zero and variance $\mathbb{V}\mathrm{ar}\left(\mathbb{G}\left(h^{+}\right)\right) = (1-\xi)\,\mathbb{V}\mathrm{ar}_{\mathrm{P}}\left(h^{+}\right) + \xi\,\mathbb{V}\mathrm{ar}_{\mathrm{Q}}\left(h^{+}\right)$.

Corollary 54 extends some previous results in which it is assumed that $n = m$; see Borgwardt et al. (2006, Th. 2.5), Gretton et al. (2007, Th. 8), and Wynne and Duncan (2022, Th. 16).

4.4 Empirical results

The aim of this section is to provide some insight about the performance of the two-sample test based on the SKD in (4.4), both from simulations and real world data sets.

The purpose of these experiments and the methods under study

In the same spirit as Gretton et al. (2006, Section 8.1) or Gretton, Borgwardt, et al. (2012), we emphasize the interest of a new homogeneity test (based on kernel distances), suitable for high-dimensional data and not suffering from the degradation of classical two-sample tests when the dimension increases. Additionally, we show the advantages of avoiding the choice of the parameters in kernel distances, via our SKD proposal. The general idea is to check the SKD methodology as an attempt to robustify the test statistic against bad choices of the kernel or its parameter(s). In this empirical study we compare the following methods:

- SKD: test based on the SKD in (4.4) with a Gaussian kernel in (4.2).
- GKD: the kernel distance-based test of Gretton et al. (2006) with a data-driven choice of λ in (4.2) as the median distance between points in the aggregate sample.

- GKDSplit: test based on a Gaussian kernel distance where the estimation of the parameters is done by a splitting method which firstly appears in Fukumizu et al. (2009). The sample is divided into training and test subsamples to avoid data influence on the parameter.

- GKDSplitOpt: test based on a kernel distance where the parameter estimation is detailed in Gretton, Sejdinovic, et al. (2012). The data is divided into training and test sets. The target parameters are the coefficients of a convex combination of a finite family of kernels. The weights of the combination are selected to maximize the ratio between the empirical distance and the standard deviation of the discrepancy.

- ET: the energy test, a popular choice in this type of problems; see Székely and Rizzo (2017), Rizzo and Szekely (2022).

In our view, the "splitting-based" proposals are based on natural ideas that deserve attention. Still, there are some open issues to clarify, especially regarding the optimal splitting of the sample and the asymptotic behavior of the resulting data-driven tests. We hope that this study could encourage further research along these lines. As for the energy test ET (Székely and Rizzo (2017)), we have included it in the study because it is based in a successful statistical methodology, ultimately grounded on the underlying "distance covariance" association measure; in fact this method has become quite popular in high-dimensional two sample problems, via the `energy` package in `R`. We have considered as well a variant of this method, which is based on a different distance between the sample points. The standard statistic in ET is calculated in terms of the Euclidean distance, which can be seen, by a duality reasoning explained in Székely and Rizzo (2017), as a distance associated with the Brownian covariance. This suggests the possibility of broadening the choice of the distance to the whole range of the fractional Brownian motion with Hurst parameter $H \in (0,1)$; recall that $H = \frac{1}{2}$ for the standard Brownian motion. In our case, the choice $H = \frac{3}{4}$ led to results almost identical to those of $H = \frac{1}{2}$. Perhaps if heavy-tailed distributions were involved, this new alternative could make a real difference.

The present empirical study is intended as an illustration of our proposal. Therefore, it is far from exhaustive. A much more detailed experiment, including additional models and competitors might be worthwhile, but this is beyond the scope of this section.

The models

We include the models in Gretton et al. (2006), based on Gaussian distributions in high dimension with different means and diagonal covariance matrices. In addition, we also consider a new scenario with functional data corresponding to trajectories of Gaussian processes in $\mathrm{L}^2([0,1])$. We note that all the considered tests can be applied in the functional setting: GKD, SKD, GKDSplit and GKDSplitOpt are based on the aggregated matrix $(k_\lambda(Z_i, Z_j))_{i,j=1}^{n+m}$, where $Z_l = X_l$ for $l = 1, \ldots, n$ and $Z_{n+l} = Y_l$ for $l = 1, \ldots, m$. ET uses cross-distances between the data in the sample space.

More specifically, our simulation experiments are grouped in three blocks, respectively corresponding to different versions of homoscedasticity (Experiment 1) and heteroscedasticity (Experiment 2), plus a functional real data example.

Experiment 1. *Different means, homoscedastic case*

Model 1.1 *White noise.* We consider $\mathrm{P} \sim \mathcal{N}(0, \mathbb{I})$ and $\mathrm{Q} \sim \mathcal{N}(\mu \mathbf{1}, \mathbb{I})$, where $\mathbf{1} = \left(1, \overset{d)}{\ldots}, 1\right)^{\mathrm{T}}$ (the superindex denotes the transpose) and $\mathbb{I}$ is the $d \times d$ identity matrix. In this model we deal with two multivariate Gaussian distributions with identity covariance in large dimension: P is standard and Q has mean $\mu \mathbf{1}$. Hence, Q is a shifted version of P translated $\sqrt{d}\mu$ units in the direction given by the vector $\mathbf{1}$. The parameter μ takes the values 0 (null hypothesis), 0.01, 0.02 and 0.05 (alternative hypothesis).

Model 1.2 *Functional data.* In this case $\mathrm{P} \sim \mathcal{G}(0, \gamma)$ and $\mathrm{Q} \sim \mathcal{G}(\mu \mathbf{1}, \gamma)$, where $\mathcal{G}$ stands for a Gaussian process in $\mathrm{L}^2([0,1])$. The first parameter is the mean function and the second the covariance function. Here, $\mathbf{1}$ is the function identically equal to 1 and $\gamma(t_1, t_2) = \exp(-0.5\,|t_1 - t_2|)$. In this model, the "dimension" refers to the size of the grid used to approximate the process. The parameter μ takes the values 0 (null hypothesis), 0.01, 0.05 and 0.2 (alternative hypothesis).

Experiment 2. *Equal means, heteroscedastic cases*

Model 2.1 *Spread white noise.* We consider $\mathrm{P} \sim \mathcal{N}(0, \mathbb{I})$ and $\mathrm{Q} \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$. The measure P corresponds to a standard multidimensional Gaussian distribution and Q to σ times P. The parameter σ^2 takes the values $10^{0.01}$ and $10^{0.02}$. This scenario introduces different alternative hypotheses from those in Model 1.1. In this example, P is more concentrated around the mean than Q.

Model 2.2 *Equicorrelated marginals.* Here, $\mathrm{P} \sim \mathcal{N}(0, \mathbb{I})$ and $\mathrm{Q} \sim \mathcal{N}(0, \Sigma)$, where $\Sigma = \rho\,(\mathbf{1}\,\mathbf{1}^{\mathrm{T}} - \mathbb{I}) + \mathbb{I}$, with $\rho \in \{0.005, 0.01, 0.02, 0.05\}$. This scenario includes another different alternative from the ones in Model 1.1. In this case, the difference between P and Q lies on the (linear) dependence structure of the marginals.

Some technical aspects

Throughout this study we restrict ourselves to the family of Gaussian kernels in (4.2), where $\mathcal{X} = \mathbb{R}^d$ or $\mathrm{L}^2([0,1])$. In this case, it is easy to show that the kernel distance $d_{k,\lambda}(\mathrm{P}, \mathrm{Q}) \to 0$, when $\lambda \to 0$ or $\lambda \to \infty$ (and discrete part of P and Q is null). Given two random samples $X_1, \ldots, X_n \sim \mathrm{P}$ and $Y_1, \ldots, Y_m \sim \mathrm{Q}$, due to the fact that the empirical measures $\mathbb{P}_n$ and $\mathbb{Q}_m$ are discrete, the kernel distance

$$d_{k,\lambda}(\mathbb{P}_n, \mathbb{Q}_m) = O\left(\sqrt{\frac{1}{n} + \frac{1}{m}}\right), \quad \text{when } \lambda \to \infty.$$

This means that for small sample sizes, the plug-in estimator of the distance does not properly approximate its population counterpart. In particular, the maximum of the empirical distance is usually attained “at the tail”, i.e., on the extremes of the target interval for λ. This drawback is inherent to the classical kernel distance although it has not been explicitly mentioned in the literature. Therefore, it is convenient to slightly modify the kernel to make the empirical distance behave the same as in the continuous case for small sample sizes. We propose to use a smoothed Gaussian kernel for this experiments given by

$$k_\lambda(x,y) = \exp\left(-\lambda\left(\|x-y\|^2 + 0.1\left(\|x\|^2 + \|y\|^2\right)\right)\right).$$

This regularization is common in harmonic analysis to approximate the Dirac delta in spaces of distributions via smooth functions, called *mollifiers*. It could be seen as an ad-hoc correction to improve the approximation of the maximum of the estimated kernel distance to the corresponding “true” population maximum. The smoothing process can be eliminated when sample sizes are sufficiently large.

As shown in the literature, a data-driven choice of λ seems to have a good practical behavior. Specifically, in Gretton et al. (2006) (and in subsequent works), the value of λ is the median distance between points in the aggregate sample. A theoretical consequence of this choice is that the asymptotic theory, derived in Gretton et al. (2006) under the assumption that λ is fixed, does not longer apply to the data-driven case. Still, we include in our experiments, for comparison purposes, this data-driven choice as it is a common practice in the earlier literature. Since, to the best of our knowledge, the asymptotic distribution of the data-driven statistic is not known, we use a permutation test based on this statistic to obtain rejection regions rather than the other methods (Pearson curves, Gamma curves and bootstrap for U-statistics) explained in Gretton et al. (2006).

It is worth mentioning that in both tests SKD and ET a permutation procedure has been used to approximate the corresponding distributions. This is the methodology used in the energy `R`-package for the ET test and we have followed here the the same strategy for the SKD test. Let us recall that our theoretical results provide the asymptotic distribution of this test, as well as its consistency (see Theorems 52 and 53). However, the estimation of the quantiles of the limit distribution in (4.10) is far from trivial. As an additional complication, standard bootstrap approximation fails, as a consequence of the results in Fang and Santos (2019). This is why the permutation method appears as a natural choice.

Let us recall that the idea behind the SKD test is to dodge the parameter selection problem by considering “the whole parameter space”. Ideally, for the Gaussian kernel, an interval of the form $(0,\infty)$ could be considered in the SKD. As mentioned before, the extremes (0 and ∞) are not useful since the distance tends to zero when the parameter approaches to these values. Therefore, we use a parameter space of the form $\Lambda = [a,b]$, with $0 < a < b < \infty$. It is important to note that here the values a and b cannot be properly considered as tuning parameters, since the test is not particularly sensitive to their choice, provided that the interval is large enough. As we have experimentally verified, $\Lambda = [10^{-4}, 0.1]$ is adequate to carry out the test. In practice, we employ a grid of 11 points logarithmically separated between 10^{-4} and 0.1. The goal is to approximate the value of the supremum by the maximum over finite subsets. The simulation outputs below are based on averages over 200 replications. The permutation tests for GKD and SKD correspond to $B = 5000$ permutations. As for ET, we use the function `eqdist.etest` of the `R`-package Rizzo and Szekely (2022). Sample sizes are $n = m = 250$ in all experiments. The effect of increasing the dimension d is checked in the rank $d \in \{205, 405, 603, 803, 1003, 1203, 1401, 1601, 1801, 2001\}$. In all cases, the significance level of the test is set at $\alpha = 0.05$.

Outputs

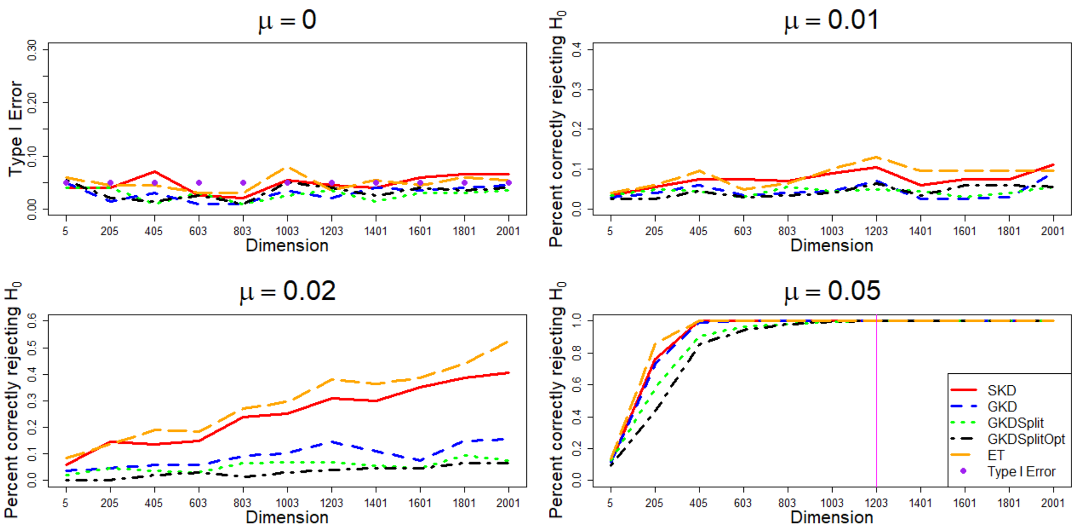


Figure 4.1: Performance of the tests under Model 1.1 with $\alpha = 0.05$. Four values of μ are shown: 0 (null hypothesis), 0.01, 0.02, and 0.05 (alternative hypothesis).

Outputs from Model 1.1 are displayed in Figure 4.1. Tests calibration, i.e., the behavior of the different tests under H_0, corresponds to the case $\mu = 0$. We observe that the size of the test is reasonably well controlled by the five tests. Under the alternative hypothesis $\mu = 0.01$ power curves oscillate slightly and are low. This is due to the proximity of the alternative hypothesis to the null. When $\mu = 0.02$, we observe that the power of all methods increases with dimension, with a slightly more pronounced growth for the ET. This is a kind of *dimension blessing* observed before for kernel distances in Gretton et al. (2007, Section 8.1). Note also that the SKD power is a little above the rest of kernel-based tests (GKD, GKDSplit and GKDSplitOpt). The vertical line in magenta in the last graph ($\mu = 0.05$) marks the point where the power of all tests is 1. The last tests in reaching this point when dimension gets larger are GKDSplit and GKDSplitOpt. This could be an indicator that the splitting methods do not work as well in these examples where the sample sizes ($n = m = 250$) are relatively small compared to the dimension we are dealing with.

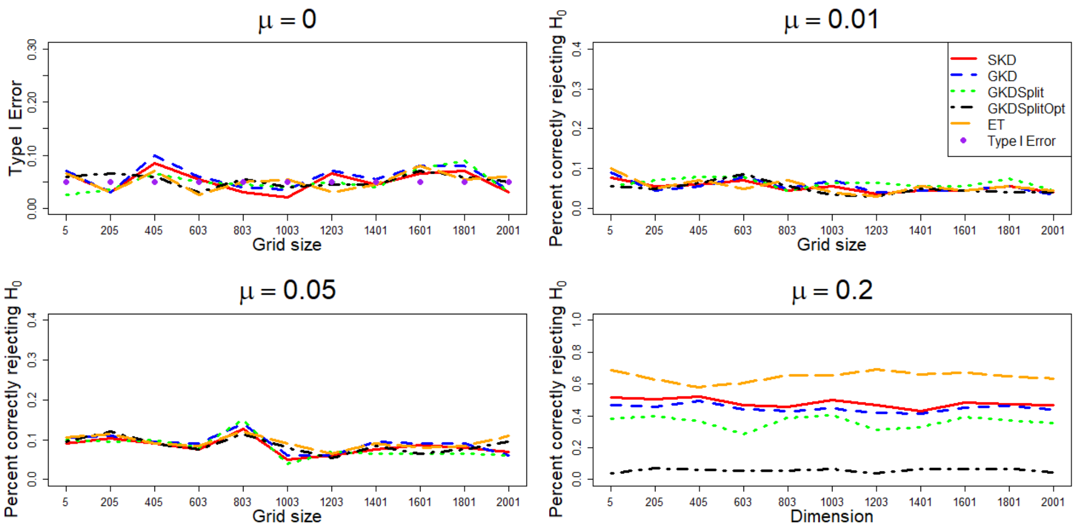


Figure 4.2: Performance of the tests under Model 1.2 with $\alpha = 0.05$. Four values of μ are shown: 0 (null hypothesis), 0.01, 0.05, and 0.2 (alternative hypothesis).

Results of Model 1.2 are summarized in Figure 4.2. Test calibration outputs are depicted for $\mu = 0$. As in the previous example, the second graph (case $\mu = 0.01$) shows a relatively small power in all cases, since both distributions are very close to each other. A gain in power is observed for $\mu = 0.05, 0.2$. The functional nature of the data is apparent in the fact that there is no clear pattern of "dimensionality blessing" associated with the increase of grid size. Indeed, unlike the other examples we are considering, the use of higher dimensional observations (a denser grid) does not entail a true gain in information, as grid observations are highly correlated, due to the continuity of the trajectories. ET obtains the best results under this scenario and SKD is competitive with the other kernel-based tests. Finally, it is surprising the absence of power of GKDSplitOpt in this model.

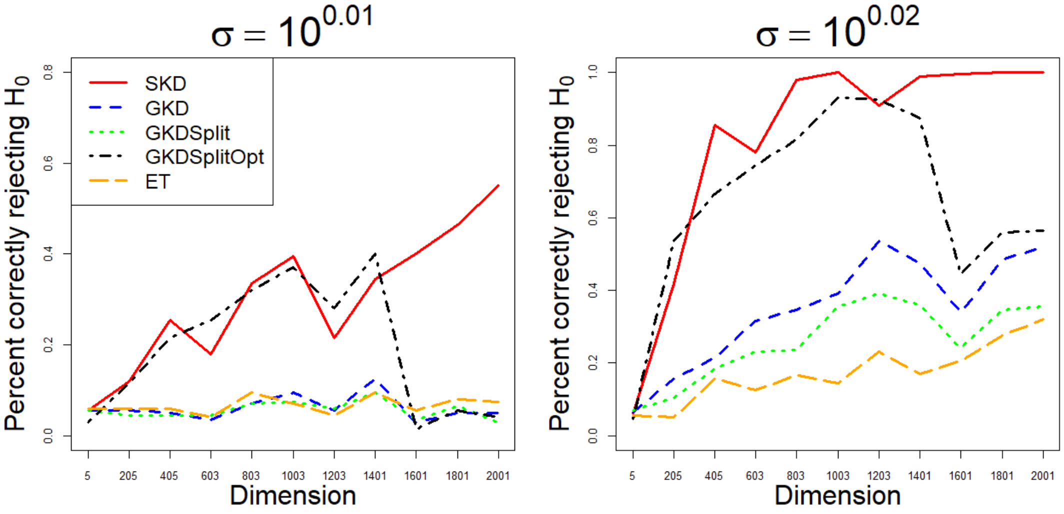


Figure 4.3: Performance of tests under Model 2.1 with $\alpha = 0.05$. Two values of σ^2 are shown: $10^{0.01}$ and $10^{0.02}$ (alternative hypothesis).

The outputs from the heteroscedastic Model 2.1 are placed in Figure 4.3. Here GKDSplitOpt and SKD behave very well and clearly outperform the other methods. A plausible explanation for this difference is that the (data-driven) median-based selection of λ of GKD is not a good choice for the heteroscedastic case when the value of the location parameter is the same in both populations. This heteroscedastic, same-location, scenario is also not the most favorable for the ET. Finally, it is noteworthy the loss of power of GKDSplitOpt from dimension 1601 onward. Again, this might be due to the small sample sizes in relation to the dimension of the problem.

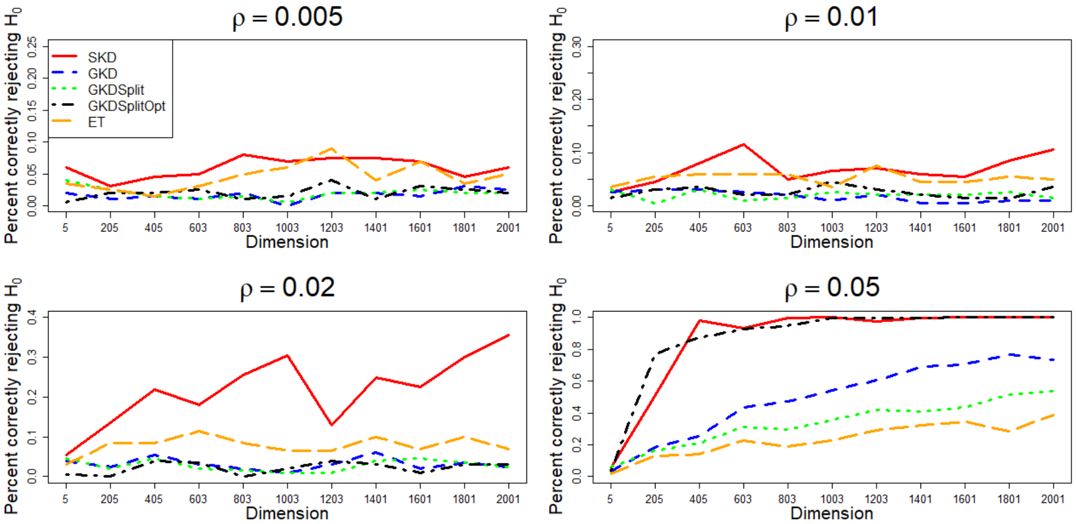


Figure 4.4: Performance of tests under Model 2.2 with $\alpha = 0.05$. Four values of ρ are shown: 0.005, 0.01, 0.02 and 0.05 (alternative hypothesis).

Results of Model 2.2 are shown in Figure 4.4. SKD seems to be particularly sensitive to dependence since correlations of $\rho = 0.05$ quickly lead to a power of almost 1. SKD obtains the best results in this scenario.

A real data example: Barcelona temperatures (1944-2019)

Daily values of maximum temperatures registered at Barcelona airport (El Prat) from years 1944 to 2019 are considered. The data set consists of 76 vectors of dimension 365, each of which corresponds to a year in that time period. The daily observations have been treated as discretization points to include the problem within the framework of functional data, every year providing a function in the sample. Those observations corresponding to the 29th of February in leap years are omitted and missing observations are interpolated. These data are available at https://www.ncei.noaa.gov, the web page of the National Centers for Environmental Information.

The purpose is to test the null hypothesis that the sample of temperatures from 1944 to 1981 comes from the same (functional) distribution to that of the period 1982-2019. The rejection of this null hypothesis could be interpreted as a hint of possible warming in the area. Indeed, we observe that, in absence of any significant climate

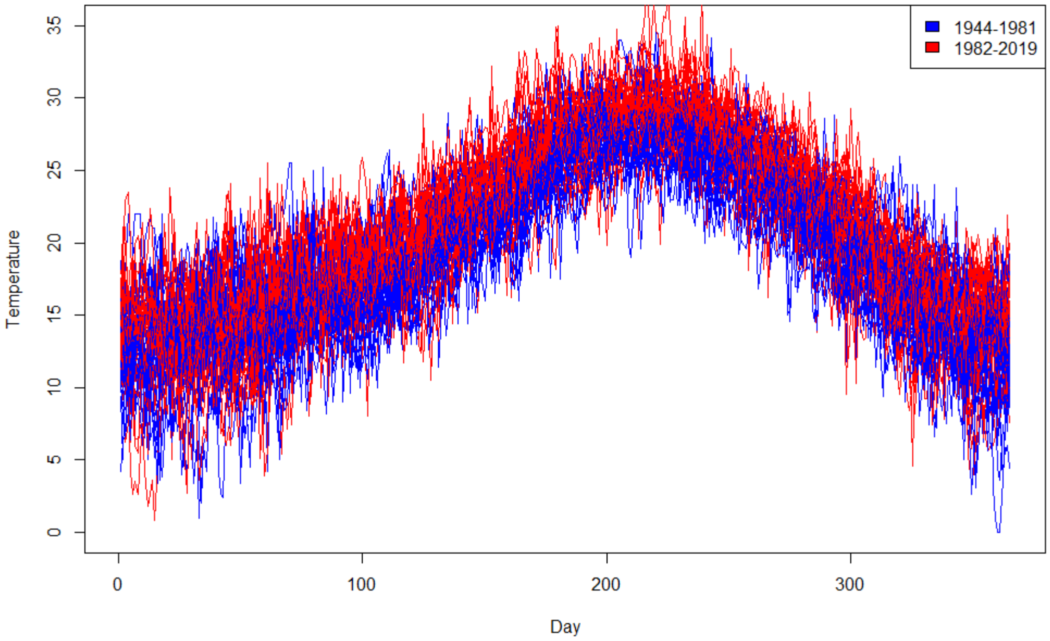


Figure 4.5: Maximum daily land surface temperature measured at El Prat Airport (Barcelona, Spain) between 1944 and 2019.

change, one would expect that both samples are made of independent trajectories from the same underlying process.

All the considered tests give a nearly null p-value. This is hardly surprising, in view of Figure 4.5, where the temperature curves are displayed (the blue curves correspond to the earlier period). While this is just a small experiment, presented here for illustration purposes, the results are consistent with those of many other deeper analysis published in recent years.

Conclusions of the empirical study

In the light of the results, we can conclude that, globally, the supremum kernel distance test (SKD) performs similarly to the GKD tests in the homoscedastic case, though the ET test appears to be the winner in this situation. In the heteroscedastic case, SKD obtains almost the best power results. A more complete study (including the derivation of the asymptotic distribution for the case of a data-driven selection of λ, the use of Pearson curves and/or modified bootstrap schemes, . . .) might be worthwhile in the future. On the other hand, according to Sejdinovic et al. (2013), energy tests can be expressed in terms of kernel distances. This idea might deserve further attention as well, in order to incorporate these “equivalent” kernels to the SKD paradigm. In any case, it is clear that the present study does not allow us to conclude any obvious superiority (or inferiority) of none of the considered methods. In fact, the aim of our limited empirical study is to show that the SKD method can be implemented and it is competitive. This goal has been hopefully achieved. More definitive conclusions should be reached via subsequent empirical experiments and, especially, with the use of these tests by practitioners in the coming years. Software to run the SKD-based test will soon be available as an `R`-package called `SKD2`.

4.5 Final remarks

In this chapter we develop the idea of combining different kernels by using the supremum of the corresponding distances. In particular, the classical two-sample problem, focused on high-dimensional and functional data, is considered. Despite the large amount of relevant literature on this topic, there is still room for improvements, as those provided here, in the line of obtaining more general results with a different technology of proofs. The use of differentiation techniques plus empirical processes methods allows us to address the asymptotic behavior of the test statistics under the null and alternative hypothesis, including the case of unbalanced samples; see Theorems 52 and 53. A key element in the proofs is the Donsker property established in Theorem 51, which extends previous similar results for finite-dimensional situations and could be potentially useful in other statistical procedures within the RKHS framework. This theorem extends previous similar results for finite-dimensional situations and could be potentially useful in other statistical procedures within the RKHS framework. The approach established in this chapter can also be potentially useful to analyze the asymptotic behavior of data-driven estimators of the kernel parameters. However, this interesting problem is beyond the scope of this thesis.

Note that Theorems 52 and 53 are meaningful from a conceptual point of view, even if the limit distributions are not particularly simple. The mere existence of the limit (non-degenerate) distributions is a primary guarantee that kernel-based statistics can be used to derive procedures achieving, at least asymptotically, a prescribed significance level under the null and providing consistency under the alternative.

In our empirical study, we deal with the complicated structure of the limit distribution under the null hypothesis by using permutation tests. This is also the approximation method used in other popular methodologies, such as the energy test (see Székely and Rizzo (2017) the implementation of this test in the `R`-package `energy`). Other alternative approximation techniques are also conceivable, including truncation in the limit expression in Theorems 52 plus estimation of the involved parameters.

4.6 Proofs of the main results

We need two auxiliar lemmata to prove Theorem 51. The first result corresponds to Marcus (1985, Theorem 1.1).

Lemma 55. *Let H be a real and separable Hilbert space. Let us consider a linear and continuous operator $T : H \to \mathcal{C}_{\mathrm{b}}(\mathcal{X})$, where $\mathcal{C}_{\mathrm{b}}(\mathcal{X})$ is the space of real bounded continuous* *functions on $\mathcal{X}$ endowed with the supremum norm. If B_H is the unit ball in H, then the class $B = T(B_H)$ is universal Donsker.*

We also require Aronszajn's inclusion theorem; see Aronszajn (1950, Theorem I).

Lemma 56. *Let k_1 and k_2 be two kernels on $\mathcal{X}$. Then, $\mathcal{H}_{k_1} \subset \mathcal{H}_{k_2}$ if and only if there exists a constant $c > 0$ such that $ck_2 - k_1$ is a positive definite kernel (i.e., $k_1 \ll ck_2$). In such a case, we also have that $\|f\|_{\mathcal{H}_{k_2}} \leq \sqrt{c}\,\|f\|_{\mathcal{H}_{k_1}}$, for all $f \in \mathcal{H}_{k_1}$.*

Proof of Theorem 51. The first part can be seen as a consequence of Lemma 55. First, by Berlinet and Thomas-Agnan (2011, Theorem 17), the functions in $\mathcal{H}_k$ are continuous. In particular, using Berlinet and Thomas-Agnan (2011, Corollary 3), we conclude that $\mathcal{H}_k$ is a separable Hilbert space. On the other hand, for $x \in \mathcal{X}$, by the reproducing property (see Definition 12) of k (twice) and Cauchy–Schwarz inequality, we have that

$$|f(x) - g(x)| = \left|\langle f - g, k(x,\cdot)\rangle_{\mathcal{H}_k}\right| \leq \|f - g\|_{\mathcal{H}_k} \|k(x,\cdot)\|_{\mathcal{H}_k} = \|f - g\|_{\mathcal{H}_k} \sqrt{k(x,x)}. \tag{4.13}$$

Therefore, as k is bounded on the diagonal, convergence in the RKHS norm entails uniform convergence. Further, from (4.13) we also see that the functions in $\mathcal{H}_k$ are bounded and hence $\mathcal{H}_k \subset \mathcal{C}_{\mathrm{b}}(\mathcal{X})$. Now, we can apply Lemma 55 to $H = \mathcal{H}_k$ and $T = I$, the inclusion map given by $I(f) = f$. According to (4.13), this linear transformation is continuous. As $B_H = \mathcal{F}_{\mathcal{H}_k}$, by Lemma 55, we thus conclude that $\mathcal{F}_{\mathcal{H}_k} = T(\mathcal{F}_{\mathcal{H}_k})$ is universal Donsker.

The second part is a by-product of the first one together with Aronszajn's inclusion theorem. According to Lemma 56, we have that $\|f\|_{\mathcal{H}_k} \leq \sqrt{c}\,\|f\|_{\mathcal{H}_{k,\lambda}}$, for all $f \in \mathcal{H}_{k,\lambda}$ and for all $\lambda \in \Lambda$. Therefore, $\mathcal{F}_{\mathcal{H}_{k,\Lambda}} \subset \sqrt{c}\mathcal{F}_{\mathcal{H}_k}$. Finally, from the first part of the theorem, the set $\sqrt{c}\mathcal{F}_{\mathcal{H}_k}$ is universal Donsker as it is the unit ball of the RKHS generated by the kernel ck. Therefore, $\mathcal{F}_{\mathcal{H}_{k,\Lambda}}$ is also universal Donsker (see A. van der Vaart and Wellner (1996, Theorem 2.10.1)) and the proof is complete. □

To prove Theorem 52, we need the following Karhunen-Loève-type result for the $\mathcal{F}_{\mathcal{H}_{k,\lambda}}$-indexed Brownian bridge.

Lemma 57. *Under the assumptions of Theorem 52, we have that:*

(a) For each $\lambda \in \Lambda$, the $\mathcal{F}_{\mathcal{H}_{k,\lambda}}$-indexed Brownian bridge $\mathbb{G}_{\mathrm{P}}$ can be extended almost surely to a continuous and linear map on $\mathcal{H}_{k,\lambda}$. Therefore, $\mathbb{G}_{\mathrm{P}}$ can be seen as a random element of the dual space $\mathcal{H}^*_{k,\lambda}$. For simplicity we also denote this extension in $\mathcal{H}^*_{k,\lambda}$ as $\mathbb{G}_{\mathrm{P}}$.

(b) As an element of $\mathcal{H}^*_{k,\lambda}$, $\mathbb{G}_{\mathrm{P}}$ admits the following representation:

$$\mathbb{G}_{\mathrm{P}} =_{\text{a.s.}} \sum_{j\in\mathbb{N}} Z_{j,\lambda}\,\varphi_{j,\lambda}, \quad \text{in } \mathcal{H}^*_{k,\lambda}. \tag{4.14}$$

In particular, we have that

$$\|\mathbb{G}_{\mathrm{P}}\|^2_{\mathcal{H}^*_{k,\lambda}} =_{\text{a.s.}} \sum_{j\in\mathbb{N}} Z^2_{j,\lambda}. \tag{4.15}$$

Proof. To show part *(a)*, we note that, from Theorem 51, $\mathcal{F}_{\mathcal{H}_{k,\lambda}}$ is a P-Donsker class and hence P-pre-Gaussian. Therefore, by Giné and Nickl (2021, Theorem 3.7.28), for almost all ω, the function $f \mapsto \mathbb{G}_{\mathrm{P}}(\omega)f$ $(f \in \mathcal{F}_{\mathcal{H}_{k,\lambda}})$ is prelinear and can be uniquely extended to a linear map on $\operatorname{span}\left(\mathcal{F}_{\mathcal{H}_{k,\lambda}}\right) = \mathcal{H}_{k,\lambda}$. Moreover, this extension is bounded and uniformly $\rho_{\mathrm{L}^2(\mathrm{P})}$-continuous in $\mathcal{H}_{k,\lambda}$. Finally, we observe that, thanks to (4.13),

$$\rho^2_{\mathrm{L}^2(\mathrm{P})}(f,g) \le \|f-g\|^2_{\mathcal{H}_{k,\lambda}} \int_{\mathcal{X}} k(x,x)\, \mathrm{dP}(x). \tag{4.16}$$

As by hypothesis k is bounded on the diagonal, we have that uniformly ρ_{P}-continuous functions on $\mathcal{H}_{k,\lambda}$ are also uniformly continuous functions with respect of the norm in $\mathcal{H}_{k,\lambda}$. In particular, $\mathbb{G}_{\mathrm{P}}$ is almost surely a continuous and linear functional on $\mathcal{H}_{k,\lambda}$, and thus an element of $\mathcal{H}^*_{k,\lambda}$. This finishes the proof of part *(a)*.

To prove part *(b)* we first note that, by *(a)*, $\mathbb{G}_{\mathrm{P}}$ is a Gaussian process in the Hilbert space $\mathcal{H}^*_{k,\lambda}$. The covariance operator $\mathcal{K}_{\mathbb{G}_{\mathrm{P}}}$ of $\mathbb{G}_{\mathrm{P}}$ is self-adjoint and compact. By the Fernique's theorem Bogachev (1998, p. 74), $\mathbb{G}_{\mathrm{P}}$ is Bochner square-integrable, $\mathcal{K}_{\mathbb{G}_{\mathrm{P}}}$ is a trace-class operator and

$$\operatorname{trace}\left(\mathcal{K}_{\mathbb{G}_{\mathrm{P}}}\right) = \int_{\mathcal{H}^*_{k,\lambda}} \|z\|^2_{\mathcal{H}^*_{k,\lambda}} \, \mathrm{d}\,\nu_{\mathbb{G}_{\mathrm{P}}}(z) = \mathbb{E}\left(\|\mathbb{G}_{\mathrm{P}}\|^2_{\mathcal{H}^*_{k,\lambda}}\right), \tag{4.17}$$

where $\nu_{\mathbb{G}_{\mathrm{P}}}$ is the measure induced by the process $\mathbb{G}_{\mathrm{P}}$ in $\mathcal{H}^*_{k,\lambda}$. The proof of (4.17) can be found in Bogachev (1998, p. 48).

Now, by the spectral theorem, there exists $\{(\beta_{j,\lambda}, \varphi_{j,\lambda})\}_{j\in\mathbb{N}} \in \left([0,\infty) \times \mathcal{H}^*_{k,\lambda}\right)^{\mathbb{N}}$ such that $\beta_{1,\lambda} \ge \beta_{2,\lambda} \ge \ldots$; $\mathcal{K}_{\mathbb{G}_P}\, \varphi_{j,\lambda} = \beta_{j,\lambda}\, \varphi_{j,\lambda}$, for $j \in \mathbb{N}$; and $\langle \varphi_{j_1,\lambda}, \varphi_{j_2,\lambda}\rangle_{\mathcal{H}^*_{k,\lambda}} = \delta_{j_1 j_2}$, for $j_1, j_2 \in \mathbb{N}$ with δ_{ij} the Kronecker's delta. As $\mathcal{K}_{\mathbb{G}_{\mathrm{P}}}$ is trace-class, we also have that $\operatorname{trace}\left(\mathcal{K}_{\mathbb{G}_P}\right) = \sum_{j\in\mathbb{N}} \beta_{j,\lambda}$. Additionally,

$$\mathbb{E}\left(\langle \mathbb{G}_{\mathrm{P}}, \varphi_{j,\lambda}\rangle_{\mathcal{H}^*_{k,\lambda}}\right) = 0 \quad \text{and} \quad \mathbb{E}\left(\langle \mathbb{G}_{\mathrm{P}}, \varphi_{j,\lambda}\rangle^2_{\mathcal{H}^*_{k,\lambda}}\right) = \langle \mathcal{K}_{\mathbb{G}_{\mathrm{P}}}\left(\varphi_{j,\lambda}\right), \varphi_{j,\lambda}\rangle_{\mathcal{H}^*_{k,\lambda}} = \beta_{j,\lambda}. \tag{4.18}$$

From (4.18), we have that $Z_{j,\lambda} = \langle \mathbb{G}_{\mathrm{P}}, \varphi_{j,\lambda}\rangle_{\mathcal{H}^*_{k,\lambda}} \sim \mathcal{N}\left(0, \beta_{j,\lambda}\right)$ $(j \in \mathbb{N})$ are jointly Gaussian and independent.

To finish this proof of (4.14), by Ledoux and Talagrand (1991, Theorem 6.1), it is enough to show absolute mean convergence, which is a necessary and sufficient condition. First, by orthogonality, we observe that for every $J \subset \mathbb{N}$ finite, we have that

$$0 \le \left\| \mathbb{G}_{\mathrm{P}} - \sum_{j\in J} Z_{j,\lambda}\, \varphi_{j,\lambda} \right\|^2_{\mathcal{H}^*_{k,\lambda}} = \|\mathbb{G}_{\mathrm{P}}\|^2_{\mathcal{H}^*_{k,\lambda}} - \sum_{j\in J} Z^2_{j,\lambda}.$$

Then by (4.17),

$$\mathbb{E}\left(\left| \|\mathbb{G}_{\mathrm{P}}\|^2_{\mathcal{H}^*_{k,\lambda}} - \sum_{j\in J} Z^2_{j,\lambda} \right|\right) = \operatorname{trace}\left(\mathcal{K}_{\mathbb{G}_{\mathrm{P}}}\right) - \sum_{j\in J} \beta_{j,\lambda} = \sum_{j\in\mathbb{N}\setminus J} \beta_{j,\lambda}, \tag{4.19}$$

which is the remainder of a convergent series. Hence, (4.14) holds. As (4.15) follows from (4.14), the proof is complete. □

The proof of part *(a)* in Lemma 57 essentially follows from Theorem 51. However, part *(b)*, where the series representation is obtained, must be discussed. Equation (4.14) shows the convergence of a series of functional random variables. This result looks like a standard Karhunen-Loève theorem, but some remarks should be done. The convergence of this series is in the dual space $\mathcal{H}_{k,\lambda}^*$, while Karhunen-Loève decomposition is stated classically on L^2-type spaces. In fact, our decomposition in (4.14) can be seen as a particular case of the results in Bay and Croix (2017). In Giné and Nickl (2021, Theorem 2.6.10) a similar decomposition is shown where the coordinates are deterministic while the basis is random, which is not useful for our purposes.

Proof of Theorem 52. From Theorem 51, the class $\mathcal{F}_{\mathcal{H}_{k,\Lambda}}$ is Donsker and hence we have that

$$\mathbb{G}_{n,m} = \sqrt{\frac{nm}{n+m}} \left(\mathbb{P}_n - \mathbb{Q}_m\right) \rightsquigarrow \mathbb{G}_{\mathrm{P}}, \quad \text{in } \ell^\infty\left(\mathcal{F}_{\mathcal{H}_{k,\Lambda}}\right). \tag{4.20}$$

Note that $d_{k,\Lambda}$ is the metric induced by the supremum norm in $\ell^\infty\left(\mathcal{F}_{\mathcal{H}_{k,\Lambda}}\right)$, hence $d_{k,\Lambda}$ is a continuous functional. From (4.20) and by the Continuous Mapping Theorem (see, for instance A. van der Vaart and Wellner (1996, Theorem 1.9.5)), we obtain that

$$\sqrt{\frac{nm}{n+m}}\, d_{k,\Lambda}\left(\mathbb{P}_n, \mathbb{Q}_m\right) \rightsquigarrow \sup_{\mathcal{F}_{\mathcal{H}_{k,\Lambda}}} \left(\mathbb{G}_{\mathrm{P}}\right). \tag{4.21}$$

From Lemma 57, the limit in (4.21) can be rewritten as

$$\sup_{\mathcal{F}_{\mathcal{H}_{k,\Lambda}}} \left(\mathbb{G}_{\mathrm{P}}\right) = \sup_{\lambda\in\Lambda}\left(\sup_{\mathcal{F}_{\mathcal{H}_{k,\lambda}}} \left(\mathbb{G}_{\mathrm{P}}\right)\right) = \sup_{\lambda\in\Lambda}\left(\|\mathbb{G}_{\mathrm{P}}\|_{\mathcal{H}_{k,\lambda}^*}\right). \tag{4.22}$$

Finally, from (4.22) and (4.15) we obtain the representation of the limit as in (4.10) and the proof of the theorem is complete. □

Proof of Corollary 54. From Theorem 51, we have that

$$\mathbb{G}_{n,m} = \sqrt{\frac{nm}{n+m}} \left(\mathbb{P}_n - \mathbb{Q}_m - (\mathrm{P}-\mathrm{Q})\right) \rightsquigarrow \mathbb{G} = \sqrt{1-\xi}\,\mathbb{G}_{\mathrm{P}} - \sqrt{\xi}\,\mathbb{G}_{\mathrm{Q}}, \quad \text{in } \ell^\infty\left(\mathcal{F}_{\mathcal{H}_k}\right). \tag{4.23}$$

From (4.7), the statistic in the right-hand side of equation (4.12) is precisely

$$\sqrt{\frac{nm}{n+m}} \left(\sigma\left(\mathbb{P}_n - \mathbb{Q}_m\right) - \sigma(\mathrm{P}-\mathrm{Q})\right), \tag{4.24}$$

where σ is denotes the supremum over the class $\mathcal{F}_{\mathcal{H}_k}$ (see also (2.1)).

Using Lemma 36, it can be checked that the paths of $\mathbb{G}$ in (4.23) are a.s. in $\mathcal{C}_{\mathrm{u}}\left(\mathcal{F}_{\mathcal{H}_k}, \rho\right)$, where

$$\rho = \max\left(\rho_{\mathrm{L}^2(\mathrm{P})}, \rho_{\mathrm{L}^2(\mathrm{Q})}\right), \tag{4.25}$$

is the natural L^2-metric of $\mathbb{G}$. From (4.16), it can be readily checked that $\mathcal{C}_{\mathrm{u}}\left(\mathcal{F}_{\mathcal{H}_k}, \rho\right) \subset \mathcal{C}_{\mathrm{u}}\left(\mathcal{F}_{\mathcal{H}_k}, d_{\mathcal{H}_k}\right)$ and hence $\mathbb{G} \in \mathcal{C}\left(\mathcal{F}_{\mathcal{H}_k}, d_{\mathcal{H}_k}\right)$ a.s. To finish the proof it is enough to apply Corollary 35 together with the functional Delta method (Proposition 5, see also A. van der Vaart and Wellner (1996, Section 3.9)). □

Proof of Theorem 53. The proof of this theorem is analogous to that of Corollary 54 using Corollary 37 instead of Corollary 35. Details are ommitted. □

Chapter 5

On uniqueness of the set of k-means

The notion of k-means appears, in a natural way, as the solution of an optimization problem in clustering and location models. It is not difficult to see that this problem does not have in general a unique solution. In this chapter we address this non-uniqueness issue. After stating the main definitions, we show some simple examples of non-uniqueness in order to gain some intuition on this phenomenon. Then, we establish (in terms of the Gromov-Hausdorff metric) a result of consistency adapted to the non-uniqueness scenario. We also provide a general characterization for non-uniqueness in terms of the asymptotic Gaussianity of the empirical risk; this result is, to the best of our knowledge, the first characterization of uniqueness available in the literature. It relies on some results of empirical processes theory, combined with Theorem 17 in Chapter 2. A test for the null hypothesis of uniqueness is derived as a consequence. The final part of the chapter is devoted to some numerical illustrations, including both Monte Carlo experiments and simulations aimed at checking the performance of the proposed test

5.1 Introduction

The k-means procedure is one of the most commonly used techniques for finding a given number of groups in a data set. The notion of k-means is a natural, almost elementary, idea with a clear interpretation and a great number of relevant applications. However, despite its simplicity, the underlying methodology still has some extraordinarily complex challenges associated with it (such as the choice of the parameter k) and many intriguing theoretical and computational aspects. In this chapter we focus on the problem of uniqueness of the set of k-means, both from a theoretical and practical point of view. As we discuss later, determining whether the set of k-means is unique or not could potentially shed light on possible reasonable choices of the parameter k, or, at least, to avoid bad choices of k.

In a stochastic context, given a random element X taking values in a separable Banach space $\mathcal{B}$ and a (prefixed) natural number k, the goal of k-means is to find the optimal set of centers of k groups, say $\mu_1, \ldots, \mu_k$, such that the expected square distance from X to its nearest center is minimal. More formally, if X induces the probability measure P on $\mathcal{B}$ (with norm $\|\cdot\|_{\mathcal{B}}$) and $k \in \mathbb{N}$, the *principal points* or *k-means set* (of P) is any subset (of cardinal k) of $\mathcal{B}$ minimizing (over all possible sets $\{a_1, \dots, a_k\} \subset \mathcal{B}$) the quantity

$$\Phi\left(\mathrm{P}; a_1, \dots, a_k\right) = \int_{\mathcal{B}} \min_{i=1,\dots,k} \left(\|x - a_i\|_{\mathcal{B}}^2\right) \mathrm{dP}(x). \tag{5.1}$$

The associated empirical version, based on n independent observations $X_1, \dots, X_n$ drawn from $\mathbb{P}_n$, corresponds to minimize the expression in (5.1) for $\mathbb{P}_n$, the *empirical measure* of the sample given by

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}, \tag{5.2}$$

where δ_a stands for the unit point mass at a. In other words, we have to minimize the function

$$\Phi_n\left(a_1, \dots, a_k\right) = \Phi\left(\mathbb{P}_n; a_1, \dots, a_k\right) = \frac{1}{n} \sum_{j=1}^{n} \min_{i=1,\dots,k} \left(\|X_j - a_i\|_{\mathcal{B}}^2\right). \tag{5.3}$$

The resulting optimal values in (5.3) are called the *empirical (or sample) k-means.*

The k-means procedure plays also a central role in localization problems in operations research. It can be motivated in terms of the so-called *facility location problem*: determine the optimal placement, $\mu_1, \dots, \mu_k$, for k facilities in such a way that the average (square) distance from a random individual to the closest facility is minimal. In statistics and machine learning, k-means techniques are generally used in clustering (see, e.g., Jain (2008) for a survey), where the aim is partitioning the space in a *Voronoi tesellation* of cells associated with the k-means. Hence, we want to divide the sampling space (or the sample) into a partition of k clusters. The i-th cluster is constituted by all elements whose closest center is μ_i ($i = 1, \dots, k$).

In addition to the usual and multiple applications of this technique (unsupervised classification, taxonomy, image analysis, information retrieval, market segmentation, computer vision, astronomy, etc.), the k-means clustering procedure is also used in *quantization*: constrain a (possibly continuous) large set to a small collection of values, the k-centers. In particular, the k-means method is generally used by practitioners as a powerful tool for summarizing data in cluster prototypes. The group centers provide a basic representation of the data that is usually very informative. The information contained in the principal points is extremely useful as a descriptive tool as it often allows understanding the underlying structure of the data and identifying prominent features.

The k-means methodology has important advantages: it is completely general, applicable to data in normed or metric spaces; easy to interpret and understand; with a clear population/sampling counterparts; several efficient heuristic algorithms are available; and with a sound asymptotic theory behind it. In spite of these appealing and positive aspects, this popular technique also entails some difficulties and challenges. First, the effective calculation of the sample k-means is a formidable computational (NP-hard) problem: algorithms have to cope with a non-convex optimization problem in a possibly high or even infinite-dimensional space. Moreover, the usual algorithms do not guarantee to reach a global optimum, but rather a local one; see Morissette and Chartier (2013) for an overview of various relevant clustering algorithms. Choosing of a good value for k

still receives considerable attention in this field and constitutes an area of active research: it is difficult to give a clear solution on the choice of the “best” number of groups, k, whatever method is used. Finally, as we discuss throughout this chapter, the k-means problem does not necessarily have a single solution in the population version. The lack of uniqueness might lead to important stability problems of the algorithms, as well as more subtle issues that can affect the understanding and interpretation of the results.

For many theoretical works involving the k-means methodology, the uniqueness of the k-means set is imposed as a requirement or a necessary assumption to obtain the results. Still, it is not difficult to find simple examples where this uniqueness prerequisite is violated. In Section 5.2 we present some of these examples (which we will use later for numerical simulations) to illustrate that there are in fact at least two distinct cases of k-means non-uniqueness that should be differentiated. We also introduce the Hausdorff metric which is suitable in this context for quantifying distances between sets. Section 5.3 brings together the main theoretical results. We show a general consistency theorem that does not assume uniqueness using the Hausdorff metric and the Gromov-Hausdorff distance. We also prove that the uniqueness of the population k-means set is equivalent to the asymptotic normality of the minimal empirical risk sequence. As an application of this result, we derive a hypothesis test for the uniqueness of the set of k-means. The proof relies on a general result on the asymptotic distribution of the empirical risk minimization over Donsker classes of functions. For this reason, at the end of Section 5.3 we provide several results that guarantee that the class of functions defining the minimization problem associated with k-means is Donsker. Some empirical results are included in Section 5.4.

5.2 Non-uniqueness of k-means

In this section we describe the consequences and practical implications of non-uniqueness in k-means. We also give some simple but illustrative examples of the different cases of non-uniqueness that might arise. These examples allow us to understand under what conditions k-means multiplicity usually appears.

We start with some definitions that we use throughout this work. We denote by $\mathcal{S}_{\mathrm{P}}(k)$ the collection of all the k-means sets of the probability measure P, that is,

$$\mathcal{S}_{\mathrm{P}}(k) = \{K = \{\mu_1, \dots, \mu_k\} \subset \mathcal{B} : K \text{ is a set of } k\text{-means for P}\}. \tag{5.4}$$

We say that P satisfies the *k-means uniqueness property*, UP(k), or $\mathrm{P} \in \mathrm{UP}(k)$ in short, if for that value k the set $\mathcal{S}_{\mathrm{P}}(k)$ in (5.4) has cardinal one. Therefore, if $\mathrm{P} \in \mathrm{UP}(k)$, there is a single set of k-means minimizing the functional in (5.1). Historically, this property has been an initial (and essential) assumption in all significant results that provide theoretical support for this methodology; see the seminal paper by Pollard (1981) on the consistency of the procedure and Pollard (1982) where the asymptotic normality of the empirical k-means is established. This hypothesis has been maintained in all subsequent works on this subject; see for example Cuesta and Matrán (1988) and Lember (2003).

However, as pointed out by García-Escudero et al. (1999), verifying the UP(k) is a difficult task in practice. There are only a few references related to the *uniqueness of the principal points*; see Li and Flury (1995), Tarpey (1994), Trushkin (1982) and Zoppe (1997). These contributions usually deal with very particular cases or univariate distributions which are not too relevant in clustering analysis. In general, it is considerably difficult to determine analytically whether a given multidimensional distribution verifies the UP(k) or not.

In the machine learning literature, it is commonly accepted that the lack of uniqueness is equivalent to k-means algorithms having instability problems. In short, it is accepted that the existence of a unique minimizer amounts to the stability of the k-means clustering; see Ben-David et al. (2006), Ben-David et al. (2007), Rakhlin and Caponnetto (2006) and the overview by Von Luxburg et al. (2010). This is of some practical significance since stability might be used for choosing the number of clusters as k can be selected as the value that provides the most stable results; see Caponnetto and Rakhlin (2006). However, as we point out below, this equivalence between uniqueness and stability is not entirely accurate because there are situations of non-uniqueness in which the algorithms are shown to be stable. This occurs when the different sets of k-means are "separated" from each other. This idea is elaborated in what follows.

First, we divide the possible cases in which there is no uniqueness in k-means into two different groups. We use the Hausdorff metric, suitable for quantifying differences between pairs of sets. We recall that, given two non-empty compact sets $A, C \subset \mathcal{B}$, the *Hausdorff distance* between A and C is defined by

$$d_H(A, C) = \inf \{\varepsilon > 0 : A \subset C^\varepsilon \text{ and } C \subset A^\varepsilon\}, \tag{5.5}$$

where, for any set $B \subset \mathcal{B}$, $B^\varepsilon = \bigcup_{x \in B} \{z : \|x - z\|_{\mathcal{B}} \leq \varepsilon\}$ is the ε-dilation of B. It is well-known that if $\mathcal{B}$ is a separable Banach space, then

$$\mathcal{H}(\mathcal{B}) = \{B \subset \mathcal{B} : B \neq \varnothing \text{ and } B \text{ is compact}\} \tag{5.6}$$

is a complete separable metric space, when endowed with the Hausdorff metric d_H.

We distinguish two different types of *multiplicity patterns*, that is, different situations in which there is no uniqueness in the k-means problem.

- *Continuous non-uniqueness*, CNU(k): For each $\varepsilon > 0$ and each set of k-means, $K_1 \in \mathcal{S}_{\mathrm{P}}(k)$, there exists $K_2 \in \mathcal{S}_{\mathrm{P}}(k)$ such that $K_1 \neq K_2$ and $d_H(K_1, K_2) < \varepsilon$. Here, every k-means set is an accumulation point (with respect to the Hausdorff metric) of the set $\mathcal{S}_{\mathrm{P}}(k)$.

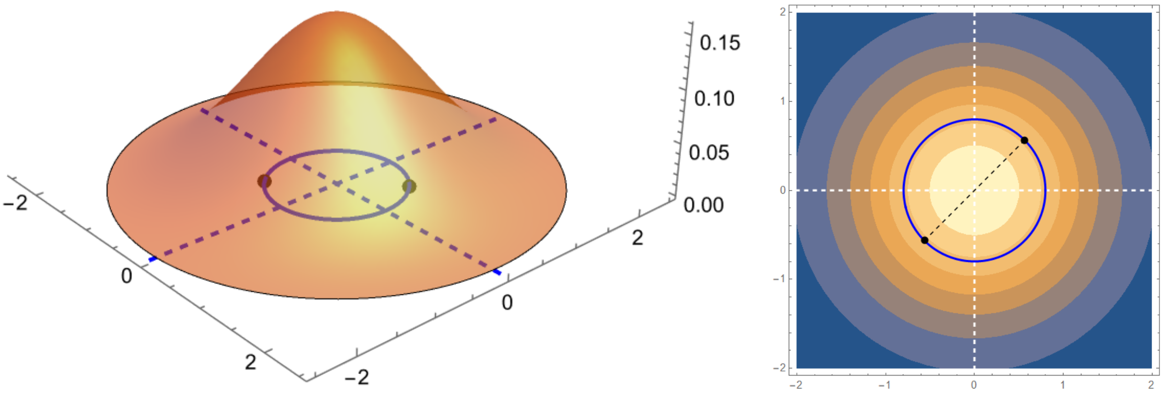
- *Discrete non-uniqueness*, DNU(k): The set of k-means $\mathcal{S}_{\mathrm{P}}(k)$ has cardinal greater than one and there exists $\varepsilon_0 > 0$ such that $d_H(K_1, K_2) \geq \varepsilon_0$, for each pair (K_1, K_2) of different sets in $\mathcal{S}_{\mathrm{P}}(k)$.

Whenever $\mathrm{P} \in$ CNU(k), there are sets of k-means arbitrarily close to each other. In practice, this is the worst possible situation because it effectively leads to k-means algorithms needing many iterations (and a large amount of time and computational cost) to stop. This is the case that is often identified by the machine learning community with the instability of the k-means algorithms. In contrast, in the DNU(k) case, there are multiple but isolated k-means. In practice, when taking a sample from a distribution $\mathrm{P} \in \mathrm{DNU}(k)$, the algorithms usually approach one of the k-means and converge fast without instability problems.

We illustrate these two possible situations of non-uniqueness with various examples. Three of them deal with distributions in $\mathbb{R}^2$ to facilitate the visualization of the probability densities and the associated k-means sets. These examples are also considered later in Section 5.4 to evaluate the proposed test for uniqueness and our theoretical results. Simple models not satisfying the UP(k) can be found in any dimension. In fact, k-means multiplicity could occur more easily in high dimension where geometric intuition is of little help. In fact, the last example below is infinite-dimensional.

Example 58 (Model C1k2)**.** One of the simplest examples of the CNU case is obtained by choosing a value of k greater than 1 in a population distributed as a standard bivariate normal. We consider $\mathrm{P} \sim \mathcal{N}((0,0), \mathbb{I}_2)$, where $\mathbb{I}_2$ is the identity matrix of dimension 2. When we select $k = 2$, due to the circular symmetry of the normal density, infinite sets of k-means appear. It is not difficult to check that these sets are formed by any two diametrically opposite points on a circumference centered at the origin of radius $\sqrt{\frac{2}{\pi}}$; see Figure 5.1.

There is an obvious symmetry pattern in $\mathcal{S}_{\mathrm{P}}(2)$ (the class of 2-means sets of this example): any set of 2-means can be brought to another one by a rotation around the origin. This pattern is repeated when we choose any value of k greater than 1.



(a) (b)

Figure 5.1: Model C1k2. (a) Density plot and (b) contour plot. The two points in black are one of the sets of 2-means located on a circumference centered at the origin of radius $\sqrt{\frac{2}{\pi}}$. This corresponds to the case CNU(2).

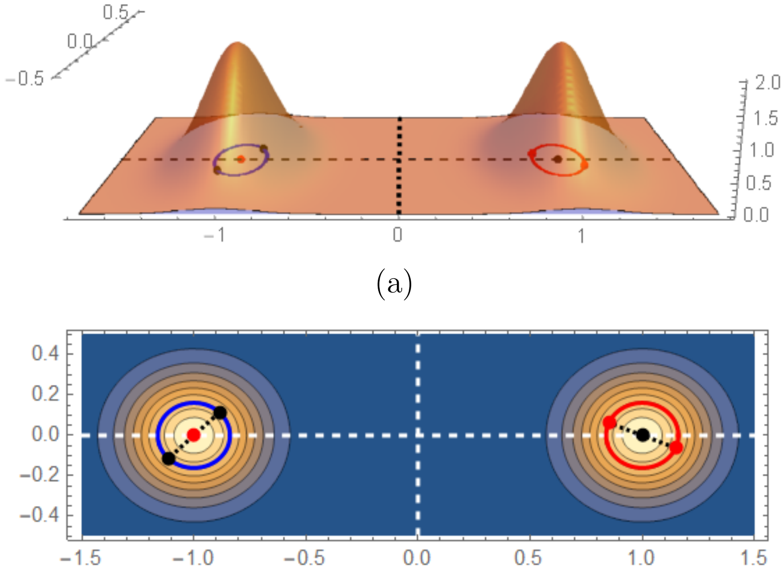
Example 59 (Model C2k3)**.** Another example of the CNU case is obtained when selecting $k = 3$ with a suitable mixture of two bivariate normal distributions. Specifically, we consider

$$\mathrm{P} \sim \frac{1}{2}\mathcal{N}\left((-1,0), \frac{\mathbb{I}_2}{25}\right) + \frac{1}{2}\mathcal{N}\left((1,0), \frac{\mathbb{I}_2}{25}\right),$$

where $\mathbb{I}_2$ is the identity matrix of dimension 2.

Here, we obtain infinite sets of 3-means within the case CNU(3). It can be seen that the sets are formed by two diametrically opposite points on a circumference centered on the mean of one of the normal distributions of the mixture of radius 0.16 together with the center of the other normal. The graphical representation of this situation is presented in Figure 5.2.

Observe that there is an isometric transformation (a rotation plus a reflection), taking one of these k-means set into another one.



(a)

(b)

Figure 5.2: Model C2k3. (a) Density plot and (b) contour plot. The two sets of three points in black and red are two of the infinite sets of 3-means of the case CNU(3).

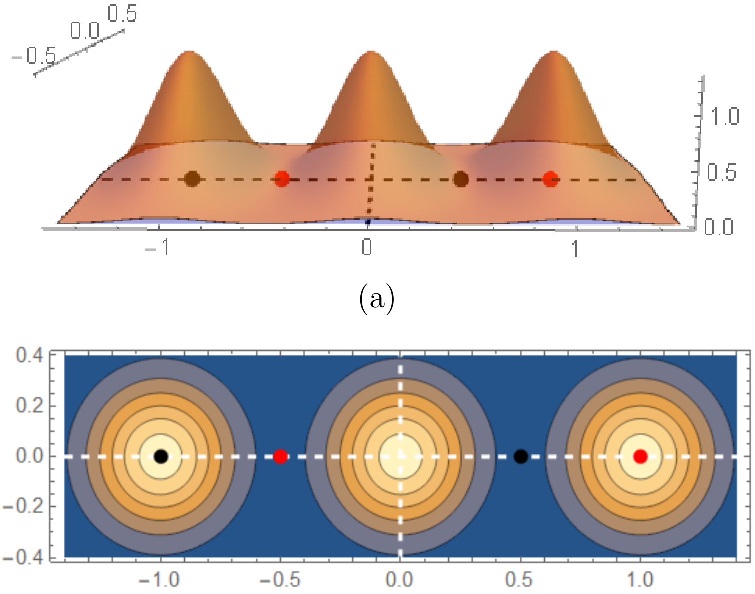
Example 60 (Model C3k2)**.** We consider an example of discrete non-uniqueness. We select $k = 2$ in a mixture of three bivariate normal distributions (with equal weights). Let

$$\mathrm{P} \sim \frac{1}{3}\mathcal{N}\left((-1,0), \frac{\mathbb{I}_2}{25}\right) + \frac{1}{3}\mathcal{N}\left((0,0), \frac{\mathbb{I}_2}{25}\right) + \frac{1}{3}\mathcal{N}\left((1,0), \frac{\mathbb{I}_2}{25}\right),$$

where $\mathbb{I}_2$ is the identity matrix of dimension 2.

We obtain two sets of 2-means included in DNU(2) given by $\left\{(-1,0), \left(\frac{1}{2},0\right)\right\}$ and $\left\{\left(-\frac{1}{2},0\right), (1,0)\right\}$; see Figure 5.3.

As it can be seen in Figure 5.3 (b), there are two isolated sets of 2-means and one of them can be transformed into the other one by means of a reflection with respect to the ordinate axis.



(a)

(b)

Figure 5.3: Model C3k2. (a) Density plot and (b) contour plot. The two sets of two points in black and red are the two sets of 2-means of the case DNU(2).

The uniqueness assumption can easily fail in high dimensions. In fact, it seems easier for the symmetries observed in the above examples to occur in higher dimensions as the following example shows.

Example 61 (Infinite-dimensional example)**.** Let us consider $(B_1(t), B_2(t))$ a vector of independent and standard Brownian motions on $[0, 1]$, and let us fix $k \geq 2$. Consider the random trajectory $(t, B_1(t), B_2(t))$ in $\mathbb{R}^3$. In this case, there is a symmetry of rotation around the time-axis (x-axis) and and infinite k-means appear, in further instance of the continuous no-uniqueness paradigm CNU(k).

It is important to note that in the above examples the multiplicity in the k-means appears when: (1) the distribution of the population has a certain symmetry; and (2) the value of k has not been conveniently chosen. In practice, it is very relevant to detect either of these two circumstances. Symmetries occur frequently in nature and science, and their understanding has led to significant advances in various fields: symmetries are fundamental in the standard model of particle physics, quantum mechanics, crystallography, the structure of biological molecules such as DNA and proteins, computer vision, etc. Symmetry also plays an important role in many statistical and data analysis problems.

5.3 Main results

In this section we collect the main results. We show the consistency of the empirical k-means in terms of the Gromov-Hausdorff metric even in the cases of non-uniqueness. We also give a characterization of the uniqueness of the set of k-means that we use to construct a uniqueness test. This characterization follows from a general asymptotic result for the empirical risk. Finally, we provide various necessary conditions for the class of functions defining the k-means risk minimization problem to be Donsker. This last result is important from a theoretical point of view because it is the basic assumption to obtain the characterization of the uniqueness property.

5.3.1 Strong consistency without uniqueness

We have shown in the previous section that some simple models might not satisfy the uniqueness property UP(k). Particularly in the CNU(k) case, the standard k-means algorithms could show a remarkable instability, providing different outputs, depending on the initial conditions or failing to fulfill the standard stopping criteria. However, if the sample size n grows to infinity and we obtain a set of empirical k-means for each n, one might ask about the "limit behavior" of such sequence of sets. The following Theorem 62 provides an answer. It is an elaboration from some previous results by Cuesta and Matrán (1988) and Lember (2003, Th. 3.1). Essentially, the result establishes that the set of empirical k-means always approaches some population k-mean set and, reciprocally, every population k-mean set is a limit of a sequence of empirical k-means sets. Here, the results of convergence for sets of k-means is stated in terms of the Hausdorff metric d_H in (5.5), and the Gromov-Hausdroff metric, both defined on the space $\mathcal{H}(\mathcal{B})$ in (5.6).

The *Gromov-Hausdorff metric* (see Burago et al. (2022) for details) is defined by

$$d_{GH}(A, C) = \inf\left(\{d_H(T(A), S(C))\}\right), \quad A, C \in \mathcal{H}(\mathcal{B}), \tag{5.7}$$

where d_H is a the Hausdorff metric in (5.5) and the infimum ranges over all possible choices of T, S and $\mathcal{M}$, with $T : A \to \mathcal{M}$ and $S : C \to \mathcal{M}$ being isometric embeddings and $\mathcal{M}$ is a compact metric space.

We use the following assumptions:

(Geo) *Geometric assumption.* $\mathcal{B}$ is a separable and uniformly convex Banach space.

(Int) *Integrability assumption.* The measure P satisfies $\int_{\mathcal{B}} \|x\|_{\mathcal{B}}^2 \, \mathrm{dP}(x) < \infty$ (strong second moment).

(Sym) *Symmetry assumption.* For any $K_1, K_2 \in \mathcal{S}_{\mathrm{P}}(k)$ there is an isometry $T : \mathcal{B} \to \mathcal{B}$ such that $T(K_2) = K_1$.

Assumption (Geo) requires that $\mathcal{B}$ is uniformly convex or uniformly rotund; see Clarkson (1936). This is fulfilled for Hilbert spaces as well as for L^p spaces, with $1 < p < \infty$. This condition is imposed by Cuesta and Matrán (1988) to derive their consistency results. This assumption entails the less restrictive, but more technical, hypothesis in Lember (2003, Th. 3.1). The requirement (Int) is necessary for the functional (5.1) to be well-defined. Assumption (Sym) might seem too restrictive. However, it holds in all examples of k-means non-uniqueness that we are aware of. It is fulfilled in the examples in Section 5.2 where the sets of k-means present self-similarity patterns easy to formalize in terms of isometries (translations, rotations, reflections, ...).

Theorem 62. *Let X be a random element taking values in $\mathcal{B}$ with probability distribution P fulfilling (Geo) and (Int). We consider the sequence $(K_n)_{n\in\mathbb{N}}$, with $K_n \in \mathcal{S}_{\mathbb{P}_n}(k)$ an empirical k-mean with $\mathbb{P}_n$ in* (5.2)*. We have that:*

(i) Any d_H-adherent point of $(K_n)_{n\in\mathbb{N}}$ belongs to $\mathcal{S}_{\mathrm{P}}(k)$ a.s.

(ii) If additionally (Sym) holds, for any $K_0 \in \mathcal{S}_{\mathrm{P}}(k)$, there is a subsequence of $(K_n)_{n\in\mathbb{N}}$, say $\left(K_{n_j}\right)_{j\in\mathbb{N}}$, such that $d_{GH}\left(K_{n_j}, K_0\right) \to 0$, almost surely, as $j \to \infty$.

Proof. Part (i) essentially follows from Lember (2003, Th. 3.1). Such result is very general but quite technical. Therefore, we check all its requirements below. To begin with, we consider the ordinary weak topology in $\mathcal{B}$ as τ in Lember (2003). Then, the key assumption Lember (2003, Assumption B, p. 29) is fulfilled. This condition imposes that "every closed ball of $\mathcal{B}$ is sequentially τ-compact". This is guaranteed by (Geo) since, by Milman-Pettis's theorem, every uniformly convex space is reflexive and hence the weak and the weak* topology coincide. In particular, the following Radon-Riesz property holds: if x_n converges weakly to x in $\mathcal{B}$ and $\|x_n\|_{\mathcal{B}} \to \|x\|_{\mathcal{B}}$, then $\|x_n - x\|_{\mathcal{B}} \to 0$. As pointed out in Lember (2003, p. 29), this is another name for the *Kadec-Klee property* imposed on Lember (2003, Assumption (2), Th. 3.1). Thus, we can apply Lember (2003, Th. 3.1) to conclude that every subsequence of the empirical k-means $(K_n)_{n\in\mathbb{N}}$ has a further subsequence converging (almost surely) in the Hausdorff metric to some set $K_0 \in \mathcal{S}_{\mathrm{P}}(k)$. Observe that if $\left(K_{n_j}\right)_{j\in\mathbb{N}}$ is a subsequence of the empirical k-means converging almost surely to $K_0 \notin \mathcal{S}_{\mathrm{P}}(k)$, all subsequences of $\left(K_{n_j}\right)_{j\in\mathbb{N}}$ should converge almost surely to K_0, which contradicts Lember (2003, Th. 3.1).

To prove (ii), we consider $K_0 \in \mathcal{S}_{\mathrm{P}}(k)$. Let $(K_n)_{n\in\mathbb{N}}$ be a sequence with $K_n \in \mathcal{S}_{\mathbb{P}_n}(k)$. Using (i), there exists a subsequence, say $\left(K_{n_j}\right)_{j\in\mathbb{N}}$, and $K \in \mathcal{S}_{\mathrm{P}}(k)$ such that $d_H\left(K_{n_j}, K\right) \to 0$, almost surely, as $j \to \infty$. Now, using (Sym), let T be an isometry on $\mathcal{B}$ such that $T(K) = K_0$. Let us denote by $\mathcal{X}_{n_j}$ the subsequence of data sets corresponding to K_{n_j}. Since T is an isometry, a sequence of k-means sets corresponding to the sequence $T\left(\mathcal{X}_{n_j}\right)$ is $T\left(K_{n_j}\right)$. As $d_H\left(K_{n_j}, K\right) \to 0$ a.s., we also obtain that

$$d_H\left(T\left(K_{n_j}\right), K_0\right) = d_H\left(T\left(K_{n_j}\right), T(K)\right) \overset{(*)}{=} d_H\left(K_{n_j}, K\right) \to 0, \text{ a.s.} \tag{5.8}$$

To see $(*)$, recall that, given $A, C \in \mathcal{H}(\mathcal{B})$, the Hausdorff distance in (5.5) between them can be alternatively expressed as

$$d_H(A, C) = \max\left(\max_{a\in A}(d(a, C)), \max_{c\in C}(d(c, A))\right), \tag{5.9}$$

where $d(a, C) = \inf_{c\in C}\left(\|a - c\|_{\mathcal{B}}\right)$ and analogously for $d(c, A)$. As a consequence of (5.9), Hausdorff metric remains invariant when the same isometry is applied to both sets. So, equality $(*)$ in (5.8) holds.

Finally, from the definition in (5.7) of d_{GH} and (5.8) we obtain that

$$d_{GH}\left(K_{n_j}, K_0\right) \le d_H\left(T\left(K_{n_j}\right), K_0\right) \to 0, \quad \text{a.s. as } j \to \infty.$$

This conclude the proof. □

An heuristic statement of Theorem 62 (ii) could be as follows: a set belongs to $\mathcal{S}_{\mathrm{P}}(k)$ if and only if it is the limit of some subsequence of empirical k-means, but not necessarily a subsequence of the actual sequence that we have. However, if we could ensure the uniqueness of the sequence of empirical k-means $(K_n)_{n\in\mathbb{N}}$, a.s., our result would establish that, the set of all d_H-accumulation points of such sequence is precisely $\mathcal{S}_{\mathrm{P}}(k)$, the set of population k-means. This situation has been corroborated in the numerical simulations of the empirical examples in Section 5.4. Intuitively, the uniqueness of the empirical k-means seems quite natural: by taking a sample, the possible population symmetry is broken with probability 1 and there is a single set of k-means. However, to the best of our knowledge, this "empirical uniqueness" does not seem simple to prove. It is essentially a matter of establishing the uniqueness of a sample sequence defined in terms of the '$\arg\min$' of a suitable functional. The interesting paper by Cox (2020) deals with the uniqueness of $\arg\min$-type statistics with a especial focus of M-statistics and maximum likelihood estimators. The methodology in that paper includes differentiability assumptions which make little sense in the k-means framework. This suggests us that the full study of affordable sufficient conditions for the uniqueness of the empirical k-means is far beyond the scope of this thesis.

5.3.2 Risk minimization for k-means over Donsker classes

The k-means problem, as stated in (5.1) and (5.3), can be viewed as a risk minimization problem over an appropriate class of functions. We follow here this approach, in the framework of empirical processes theory, to address the question of the uniqueness in k-means.

As it is common in clustering, we assume that the k-means live in a certain subset of the space $\mathcal{B}$. Hence, for a fixed $B \subset \mathcal{B}$, we consider the collection of functions

$$\mathcal{F}_{V_k(B)} = \{f_a : a = (a_1, \dots, a_k) \in V_k(B)\}, \tag{5.10}$$

where $f_a : \mathcal{B} \to \mathbb{R}$ is defined by

$$f_a(z) = \min_{i=1,\dots,k} \left(\|z - a_i\|_{\mathcal{B}}^2\right), \quad z \in \mathcal{B}, \tag{5.11}$$

and

$$V_k(B) = \left\{(a_1, \dots, a_k) \in B^k : a_i \neq a_j \text{ for } i \neq j\right\}. \tag{5.12}$$

The *expected risk* of a function $f \in \mathcal{F}_{V_k(B)}$ is given by

$$\mathrm{P}(f) = \int_{\mathcal{B}} f \, \mathrm{dP}.$$

The *expected risk* of the class $\mathcal{F}_{V_k(B)}$, $\mathrm{R}\left(\mathcal{F}_{V_k(B)}\right)$, is the minimum risk that we can achieve with functions in $\mathcal{F}_{V_k(B)}$. In other words,

$$\mathrm{R}\left(\mathcal{F}_{V_k(B)}\right) = \inf_{f\in\mathcal{F}_{V_k(B)}} (\mathrm{P}(f)). \tag{5.13}$$

When the k-means are restricted to live in $B \subset \mathcal{B}$, the k-means problem in (5.1) corresponds to finding the functions $f_\mu \in \mathcal{F}_{V_k(B)}$ (or, equivalently, the elements $\mu = (\mu_1, \ldots, \mu_k) \in V_k(B)$) such that $\mathrm{P}(f_\mu) = \mathrm{R}\left(\mathcal{F}_{V_k(B)}\right)$. In other words, $\{\mu_1, \ldots, \mu_k\} \in \mathcal{S}_\mathrm{P}(k)$ (an element of the set of all k-means) with $\mu_i \in B$ if and only if the function f_μ in (5.11) with $\mu = (\mu_1, \ldots, \mu_k)$ is a risk-minimizer of P in $\mathcal{F}_{V_k(B)}$. Note that the parametrization of $\mathcal{F}_{V_k(B)}$ in terms of the set $V_k(B)$ in (5.12) is not injective: different permutations of $a \in V_k(B)$ lead to the same function $f_a \in \mathcal{F}_{V_k(B)}$.

As the underlying probability P is usually unknown, instead of using $\mathrm{P}(f)$ and $\mathrm{R}\left(\mathcal{F}_{V_k(B)}\right)$ in (5.13), the *empirical risk* of a function $f \in \mathcal{F}_{V_k(B)}$ and the class $\mathcal{F}_{V_k(B)}$ are usually employed in practice. That is,

$$\mathbb{P}_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i) \quad \text{and} \quad \mathrm{R}_n\left(\mathcal{F}_{V_k(B)}\right) = \inf_{f \in \mathcal{F}_{V_k(B)}} \left(\mathbb{P}_n(f)\right), \tag{5.14}$$

where $X_1, \ldots, X_n$ is a set of independent variables distributed as P (a training sample from X) and $\mathbb{P}_n$ is the empirical measure in (5.2). Observe that $\mathrm{R}_n\left(\mathcal{F}_{V_k(B)}\right)$ is nothing but the mean of the *within cluster sum of squares* in k-means.

In the following theorem we calculate the asymptotic distribution of the empirical risk $\mathrm{R}_n\left(\mathcal{F}_{V_k(B)}\right)$ in (5.14). This result is key to characterize the uniqueness of the set of k-means. When dealing with risk minimization, it is often required that the class of functions in use is (uniform) Glivenko-Cantelli (see, e.g., A. van der Vaart and Wellner (1996)) to ensure that $\mathbb{P}_n(f)$ and $\mathrm{P}(f)$ are close to each other uniformly over the class of functions (for all probability measures P on $\mathcal{B}$). However, to derive the asymptotic distribution of $\mathrm{R}_n\left(\mathcal{F}_{V_k(B)}\right)$ it is essential that $\mathcal{F}_{V_k(B)}$ satisfies the Central Limit Theorem –i.e., $\mathcal{F}_{V_k(B)}$ has to be a *Donsker class* (see Definition 11)–, which is obviously a more demanding condition. Donsker classes are rather general and they have been already considered in this setting; see Caponnetto and Rakhlin (2006).

The following result provides the asymptotic distribution of the empirical risk $\mathrm{R}_n\left(\mathcal{F}_{V_k(B)}\right)$ in (5.14). It is a consequence of Corollary 26.

Theorem 63. *Let us assume (Int) and that the following two conditions are satisfied.*

(Bnd) Boundedness assumption. *The set B is bounded in $\mathcal{B}$. In other words, we restrict the search for the k-means to a bounded set of the space.*

(Dnk) Donsker assumption. *The class $\mathcal{F}_{V_k(B)}$ in* (5.10) *is* P*-Donsker.*

Then we have that,

$$T_n(k) = \sqrt{n}\left(\mathrm{R}_n\left(\mathcal{F}_{V_k(B)}\right) - \mathrm{R}\left(\mathcal{F}_{V_k(B)}\right)\right) \rightsquigarrow T(k) = \inf_{f \in \overline{S}_\mathrm{P}(k,B)} \left(\mathbb{G}_\mathrm{P}(f)\right), \tag{5.15}$$

where

$$\overline{S}_\mathrm{P}(k,B) = \left\{f \in \overline{\mathcal{F}}_{V_k(B)} : \mathrm{P}(f) = \mathrm{R}\left(\mathcal{F}_{V_k(B)}\right)\right\}, \tag{5.16}$$

is the set of all minimizers of P *over the class $\overline{\mathcal{F}}_{V_k(B)} \equiv$ the completion of $\mathcal{F}_{V_k(B)}$ with respect to the metric $\rho_{\mathrm{L}^2(\mathrm{P})}$.*

Proof. First, by (Bnd), $M = \sup_{b \in B} (\|b\|_{\mathcal{B}}^2) < \infty$. For any $f_a \in \mathcal{F}_{V_k(B)}$ with $a = (a_1, \dots, a_k) \in V_k(B)$, by (Int), we have that

$$\begin{aligned} \mathrm{P}(f_a) &= \int_{\mathcal{B}} \min_{i=1,\dots,k} \left(\|z - a_i\|_{\mathcal{B}}^2 \right) \mathrm{dP}(z) \\ &= 2^2 \int_{\mathcal{B}} \min_{i=1,\dots,k} \left(\left\| \frac{z - a_i}{2} \right\|_{\mathcal{B}}^2 \right) \mathrm{dP}(z) \\ &\overset{(1)}{\leq} 2^2 \int_{\mathcal{B}} \min_{i=1,\dots,k} \left(\frac{1}{2} \left(\|z\|_{\mathcal{B}}^2 + \|a_i\|_{\mathcal{B}}^2 \right) \right) \mathrm{dP}(z) \\ &\leq 2 \left(\int_{\mathcal{B}} \|z\|_{\mathcal{B}}^2 \, \mathrm{dP}(z) + M \right) < \infty, \end{aligned} \tag{5.17}$$

where (1) follows by the convexity of the norm and the square. Therefore, the functional $\mathrm{P} : \mathcal{F}_{V_k(B)} \to \mathbb{R}$ defined by $\mathrm{P}(f) = \int_{\mathcal{B}} f \, \mathrm{dP}$ belongs to $\ell^\infty \left(\mathcal{F}_{V_k(B)} \right)$.

Now, we use similar ideas as those in the proof of Theorem 50. As $\mathcal{F}_{V_k(B)}$ is P-Donsker, the space $\left(\mathcal{F}_{V_k(B)}, \rho_{\mathrm{P}} \right)$ is totally bounded, where ρ_{P} is the intrinsic pseudometric (see Definition 9 or Giné and Nickl (2021, Remark 3.7.27)). Also, the P-Brownian bridge $\mathbb{G}_{\mathrm{P}} \in \mathcal{C}_{\mathrm{u}} \left(\mathcal{F}_{V_k(B)}, \rho_{\mathrm{P}} \right)$ a.s. By (5.17), the class $\mathcal{F}_{V_k(B)}$ is bounded in $\mathrm{L}^1(\mathrm{P})$, i.e., $\sup_{f \in \mathcal{F}_{V_k(B)}} (|\mathrm{P}(f)|) < \infty$. This condition, joint to the fact that $\left(\mathcal{F}_{V_k(B)}, \rho_{\mathrm{P}} \right)$ is totally bounded, implies that $\left(\mathcal{F}_{V_k(B)}, \rho_{\mathrm{L}^2(\mathrm{P})} \right)$ is totally bounded; as it follows from the same ideas as in the proof of Giné and Nickl (2021, Theorem 3.7.40, p. 262). Note that the trajectories of $\mathbb{G}_{\mathrm{P}}$ also belong to $\mathcal{C}_{\mathrm{u}} \left(\mathcal{F}_{V_k(B)}, \rho_{\mathrm{L}^2(\mathrm{P})} \right)$ with probability 1 because $\rho_{\mathrm{P}} \leq \rho_{\mathrm{L}^2(\mathrm{P})}$. Finally, we can apply the extended Delta method for the infimum together with Corollary 26 to derive the asymptotic result in (5.15). □

5.3.3 A characterization and test of k-means uniqueness

We establish here necessary and sufficient conditions for the uniqueness of the set of k-means within a set $B \subset \mathcal{B}$. First, we observe that the set of functions $\overline{\mathcal{F}}_{V_k(B)}$ includes the set of all possible limits (in the completion of $\mathcal{F}_{V_k(B)}$ in (5.10) with respect to $\rho_{\mathrm{L}^2(\mathrm{P})}$) of minimizing sequences of the risk. Further, $\overline{S}_{\mathrm{P}}(k, B)$ in (5.16) is just the set of minimizers of the risk in the $\rho_{\mathrm{L}^2(\mathrm{P})}$-completion of $\mathcal{F}_{V_k(B)}$. Note that we need to complete the class $\mathcal{F}_{V_k(B)}$ to ensure the existence of minimizers of the risk.

We say that P satisfies the *k-means uniqueness property in B*, $\mathrm{UP}(k, B)$, or $\mathrm{P} \in \mathrm{UP}(k, B)$, if the set $\overline{S}_{\mathrm{P}}(k, B)$ in (5.16) has cardinal one. The following result, a consequence of Theorem 63, characterizes this situation in terms of the properties of the limit variable $T(k)$ in (5.15).

Corollary 64. *Under assumptions (Bnd) and (Dnk), the following three assertions are equivalent.*

(i) $\mathrm{P} \in \mathrm{UP}(k, B)$ *(uniqueness of k-means on B).*

(ii) The variable $T(k)$ in (5.15) *is normally distributed with mean zero.*

(iii) The variable $T(k)$ *in* (5.15) *has zero mean.*

Proof. Assume that (i) holds. We then have that there exists a unique minimizer $f^- \in \overline{\mathcal{F}}_{V_k(B)}$ such that $\overline{S}_{\mathrm{P}}(k,B) = \{f^-\}$. From Theorem 63, we obtain that $T(k) = \mathbb{G}_{\mathrm{P}}(f^-)$, which has normal distribution with mean zero. Therefore, (ii) is satisfied. The implication (ii) ⇒ (iii) is direct. Finally, assume that (iii) holds. Observe that $T(k) \leq_{\mathrm{st}} \mathbb{G}_{\mathrm{P}}(f)$, for each $f \in \overline{S}_{\mathrm{P}}(k,B)$, where '$\leq_{\mathrm{st}}$' stands for the usual stochastic order. Since $\mathbb{E}(T(k)) = 0 = \mathbb{E}(\mathbb{G}_{\mathrm{P}}(f))$, we conclude that $T(k) =_{\mathrm{st}} \mathbb{G}_{\mathrm{P}}(f)$; see Shaked and Shanthikumar (2007, Theorem 1.A.8.). Hence, we obtain that $T(k)$ is normally distributed. Finally, as $T(k) = \inf_{f \in \overline{S}_{\mathrm{P}}(k,B)} (\mathbb{G}_{\mathrm{P}}(f))$ is the infimum of normal variables, its distribution only can be normal when the infimum is taken over a set of cardinal one; if $f_1, f_2 \in \overline{S}_{\mathrm{P}}(k,B)$ with $f_1 \neq f_2$, then $\min(\mathbb{G}_{\mathrm{P}}(f_1), \mathbb{G}_{\mathrm{P}}(f_2))$ is not normally distributed. Hence, we conclude that the cardinal of $\overline{S}_{\mathrm{P}}(k,B)$ is necessarily one and (i) holds. □

Corollary 64 makes it possible to develop various discrepancy measures to perform the following hypothesis test for uniqueness of k-means:

$$\begin{aligned} &\mathrm{H}_0 : \mathrm{P} \in \mathrm{UP}(k) && (\text{uniqueness of } k\text{-means}), \\ &\mathrm{H}_1 : \mathrm{P} \notin \mathrm{UP}(k) && (\text{non-uniqueness of } k\text{-means}). \end{aligned} \tag{5.18}$$

We note that H_0 is equivalent to the asymptotic distribution of $T_n(k)$ in (5.15) being normal with zero mean. Therefore, we can use any normality test available in the literature if we have a sufficient number of data. On the other hand, under H_0 the mean of $T(k)$ is zero and under H_1, $\mathbb{E}(T(k)) < 0$ (strictly less than zero). Therefore, the test in (5.18) is equivalent to either of the following two tests:

$$\text{(a)} \quad \begin{aligned} &\mathrm{H}_0 : \ T(k) \sim \mathcal{N}(0,\sigma^2), \\ &\mathrm{H}_1 : \ \text{no } \mathrm{H}_0, \end{aligned} \qquad \text{(b)} \quad \begin{aligned} &\mathrm{H}_0 : \ \mathbb{E}(T(k)) = 0, \\ &\mathrm{H}_1 : \ \mathbb{E}(T(k)) < 0, \end{aligned} \tag{5.19}$$

where $T(k)$ is the limiting distribution of $T_n(k)$ in equation (5.15).

We apply these ideas in Section 5.4 to evaluate the performance of the uniqueness test in (5.18) (and (5.19)) with simulated data.

5.3.4 Donsker's theorems for k-means

In this section we give conditions guaranteeing that the class of functions $\mathcal{F}_{V_k(B)}$ in (5.10) is P-Donsker, and hence assumption (Dnk) in Theorem 63 is fulfilled.

In the classical, finite dimensional setting, of the k-means problem; when X takes values in a ball of $\mathbb{R}^d$ (for the Euclidean distance), then the class $\mathcal{F}_{V_k(B)}$ in (5.10) is universal Donsker; see Caponnetto and Rakhlin (2006, Lemma 3.1). This boundedness assumption is usually imposed in many works related to k-means. In Telgarsky and Dasgupta (2013) some additional results are provided for distributions with $p \geq 4$ finite moments.

A subset $B \subset \mathcal{B}$ is said to be *bounded* respect to the metric d if there exist $a \in \mathcal{B}$ and $r > 0$ such that $B \subset B_d(a,r)$, the ball of center a and radius r respect to the metric d. The sub-index d is omitted when the metric is understood by the context. The metric induced by the norm $\|\cdot\|_{\mathcal{B}}$ is denoted by $d_{\mathcal{B}}$. In the finite dimensional context, it is reasonable to ask that the k-means be restricted to a bounded set $B \subset \mathcal{B}$. However, in the infinite dimensional (functional) setting more demanding conditions are needed by the loss of the Heine-Borel's property.

Sufficient conditions to ensure that a class of functions $\mathcal{F}$ satisfies the Donsker property are related to *the size of the covering.* One way to cover a subset of a metric space is using balls. Given a subset of $B \subset \mathcal{B}$ we say that B is *totally bounded* if for every $\varepsilon > 0$ there exists $N \in \mathbb{N}$ and $x_1, \ldots, x_N \in \mathcal{B}$ such that $B \subseteq \bigcup_{i=1}^{N} B_d(x_i, \varepsilon)$. Then, it is said that B is *covered* by $\{B_d(x_i, \varepsilon)\}_{i=1}^{N}$. Given $\varepsilon > 0$, $N(\varepsilon, B, d)$ is the minimal number of balls of radius ε required to cover B. The quantity $N(\varepsilon, B, d)$ is known as *ε-covering number.* Obviously, every totally bounded set is bounded and $N(\varepsilon, B, d) \to 1$ when $\varepsilon \to \infty$.

Alternatively, a class of real functions $\mathcal{F}$ endowed with a metric d can be covered with brackets. Given $l, u \in \mathcal{F}$, the *bracket*

$$[l, u] = \{f \in \mathcal{F} : l \leq f \leq u\}.$$

An ε-bracket (respect to the metric d) is a bracket where $d(l, u) < \varepsilon$. The *ε-bracketing number* $N_{[\,]}(\varepsilon, \mathcal{F}, d)$ is the minimal number of ε-brackets necessary to cover B. The Donsker property is typically ensured for a class $\mathcal{F}$ through conditions that are associated with the covering and/or bracketing numbers; see A. van der Vaart and Wellner (1996, Section 2.5).

The following lemma links bracketing numbers of $\mathcal{F}_{V_k(B)}$ in (5.10) to covering numbers of B.

Lemma 65. *Let $B \subset \mathcal{B}$ be a totally bounded set and assume that (Int) is satisfied. For the class $\mathcal{F}_{V_k(B)}$ in* (5.10)*, we have that*

$$N_{[\,]}\left(\varepsilon, \mathcal{F}_{V_k(B)}, \|\cdot\|_{\mathrm{L}^2(\mathrm{P})}\right) \leq N\left(\frac{\varepsilon}{C}, B, d_{\mathcal{B}}\right)^k, \tag{5.20}$$

where $C = 2\left(\int_{\mathcal{B}} \|z\|_{\mathcal{B}}^2 \,\mathrm{dP}(z) + \sup_{b \in B} \left(\|b\|_{\mathcal{B}}\right)\right)$.

Proof. Let us first prove the following Lipschitz property for the functions in $\mathcal{F}_{V_k(B)}$:

$$|f_a(z) - f_b(z)| \leq 2\left(\|z\|_{\mathcal{B}} + \operatorname{diam}(B)\right) d_\infty(a, b), \tag{5.21}$$

where $a = (a_1, \ldots, a_k)$ and $b = (b_1, \ldots, b_k) \in V_k(B)$ and

$$d_\infty(a, b) = \max_{i=1,\ldots,k} \left(\|a_i - b_i\|_{\mathcal{B}}\right).$$

Indeed, we have that

$$\begin{aligned}
|f_a(z) - f_b(z)| &= \left| \min_{i=1,\dots,k} \left(\|z - a_i\|_{\mathcal{B}}^2 \right) - \min_{i=1,\dots,k} \left(\|z - b_i\|_{\mathcal{B}}^2 \right) \right| \\
&= \left| \max_{i=1,\dots,k} \left(- \|z - a_i\|_{\mathcal{B}}^2 \right) - \max_{i=1,\dots,k} \left(- \|z - b_i\|_{\mathcal{B}}^2 \right) \right| \\
&\leq \max_{i=1,\dots,k} \left(\left| \|z - a_i\|_{\mathcal{B}}^2 - \|z - b_i\|_{\mathcal{B}}^2 \right| \right) \\
&= \max_{i=1,\dots,k} \left(\left| \|z - a_i\|_{\mathcal{B}} + \|z - b_i\|_{\mathcal{B}} \right| \left| \|z - a_i\|_{\mathcal{B}} - \|z - b_i\|_{\mathcal{B}} \right| \right) \\
&\leq \left(2 \|z\|_{\mathcal{B}} + \max_{i=1,\dots,k} \left(\|a_i\|_{\mathcal{B}} + \|b_i\|_{\mathcal{B}} \right) \right) \max_{i=1,\dots,k} \left(\|a_i - b_i\|_{\mathcal{B}} \right).
\end{aligned}$$

The first summand above can be bounded by $2 \left(\|z\|_{\mathcal{B}} + \sup_{u \in B} \left(\|u\|_{\mathcal{B}} \right) \right)$ and the second one is, by definition, $d_\infty(a, b)$, so the Lipschitz property (5.21) is obtained.

Now, we apply A. van der Vaart and Wellner (1996, Theorem 2.7.11) to derive (5.20). □

The sufficient conditions for a class to be P-Donsker in A. W. van der Vaart (2000, Theorem 19.5) are quite standard in empirical processes literature. The following theorem allows us to express this condition in terms of the, more geometrically motivated, covering numbers of B.

Theorem 66. *Assume that (Int) is satisfied and*

$$\int_0^\infty \sqrt{\log \left(N \left(\varepsilon, B, d_{\mathcal{B}} \right) \right)} \, \mathrm{d}\varepsilon < \infty. \tag{5.22}$$

Then, $\mathcal{F}_{V_k(B)}$ *in* (5.10) *is a* P*-Donsker class.*

Proof. By Lemma 65 we have that

$$\int_0^\infty \sqrt{\log \left(N_{[\,]} \left(\varepsilon, \mathcal{F}_{V_k(B)}, \| \cdot \|_{\mathrm{L}^2(\mathrm{P})} \right) \right)} \, \mathrm{d}\varepsilon < \sqrt{k} \int_0^\infty \sqrt{\log \left(N \left(\frac{\varepsilon}{C}, B, d_{\mathcal{B}} \right) \right)} \, \mathrm{d}\varepsilon.$$

Hence, if right hand side is finite, by A. W. van der Vaart (2000, Theorem 19.5) the class $\mathcal{F}_{V_k(B)}$ is Donsker. □

The previous theorem extends the findings in Pollard (1982), where the asymptotic Gaussianity of the empirical k-means is established. Let us now comment on the crucial condition (5.22) concerning the set B. This requirement might appear restrictive at first glance. However, it is satisfied by a broad range of examples. To begin with, it holds for every finite-dimensional bounded set. Moreover, every Vapnik-Červonenkis (VC) class satisfies this condition, with its entropy numbers growing polynomially; for this and related results, we refer to A. van der Vaart and Wellner (1996, Section 2.6).

In the realm of functional data analysis, there are also several examples in which $\mathcal{F}_{V_k(B)}$ is Donsker. For instance, unit balls in α-Hölder continuous, Sobolev, and Besov function spaces satisfy this condition under mild restrictions on the parameters; see A. van der Vaart and Wellner (2023, Section 2.7). It is worth noting that these sets are dense in the unit balls of the space of continuous and bounded functions $\mathcal{C}_{\mathrm{b}}$ and L^p spaces, respectively. In essence, the preceding lines imply that if we confine ourselves to dense subspaces of "smooth functions", a common practice in functional data analysis, there is no loss of generality for the uniqueness of k-means problem.

Other commonly used spaces satisfying this condition are monotone and convex functions on subsets of the Euclidean space. The following corollary summarize this discussion.

Corollary 67. *Assume one of the following conditions is satisfied:*

(a) B is a bounded subset of a finite dimensional Banach space $\mathcal{B}$.

(b) B is a Vapnik-Červonenkis (VC) class of functions endowed with the $\mathrm{L}^p(\mathrm{P})$-metric.

(c) B is the unit ball of the space α-Hölder continuous functions $\mathcal{C}^\alpha(\mathcal{X})$, where $\mathcal{X} \subset \mathbb{R}^l$ bounded, convex and with no empty interior.

(d) B is the unit ball of the Sobolev space $W^{\alpha,p}(\mathcal{X})$, where $\mathcal{X}$ is a Lipschitz domain of $\mathbb{R}^l$, endowed with L^r-norm and $\alpha > \max\left(0, d\left(\frac{1}{p} - \frac{1}{r}\right)\right)$. More generally, B is a subset of the unit ball of the Besov space $\mathcal{B}^\alpha_{p,q}(\mathcal{X})$ endowed with the L^r-norm.

(e) B is the class of monotone functions on $[0,1]$ endowed with an L^p-norm.

(f) B is the class of all convex functions on a compact convex subset $C \subset \mathbb{R}^l$ with values on $[0,1]$ endowed with L^p-norm.

Then $\mathcal{F}_{V_k(B)}$ in (5.10) is a P-Donsker class.

Proof. The proof of this corollary is based on checking condition (5.22) and a direct application of Theorem 66 for each of the stated scenarios. To see (a), we denote the dimension of this space by r. Then, B is totally bounded and the covering numbers $N(\varepsilon, B, d_{\mathcal{B}})$ are proportional to $\frac{1}{\varepsilon^r}$. Consequently, (66) holds. In part (b), the decreasing of the covering numbers is polynomial in $\frac{1}{\varepsilon}$; see A. van der Vaart and Wellner (2023, Theorem 2.6.4). In part (c), to check that the decreasing of the logarithm of covering numbers is polynomial in $\frac{1}{\varepsilon}$ we refer to A. van der Vaart and Wellner (2023, Theorem 2.7.1). To see (d), use A. van der Vaart and Wellner (2023, Theorem 2.7.4). For (e), A. van der Vaart and Wellner (2023, Theorem 2.7.9). Part (f) follows from A. van der Vaart and Wellner (2023, Theorem 2.7.14). □

Finally, let us introduce the concept of Riesz's property: given a class of real functions $\mathcal{F}$ endowed with the norm $\|\cdot\|$, it is said that $\|\cdot\|$ satisfies the *Riesz's property* if for every pair $f, g \in \mathcal{F}$ such that $|f| \leq |g|$, then $\|f\| \leq \|g\|$. Norms with this property on $\mathcal{F}$ fulfil $N\left(\varepsilon, \mathcal{F}, d_{\|\cdot\|}\right) \leq N_{[\,]}\left(2\varepsilon, \mathcal{F}, d_{\|\cdot\|}\right)$. In other words, Riesz's property gives a relation between covering and bracketing numbers that can be also exploited in this context. L^p-norms and supremum norm satisfy Riesz's property.

5.4 Empirical results

In this section we provide some insight about the validity of the theoretical results established in Section 5.3. We analyze some examples to illustrate the different scenarios discussed in this chapter: UP(k), uniqueness of the k-means set; CNU(k) continuous non-uniqueness; and DNU(k) discrete non-uniqueness. The CNU(k) case could be seen as "more pathological" since it entails the existence of different sets of k-means arbitrarily close (in Hausdorff distance), which usually leads to algorithm instability problems.

The numerical experiments include:

(a) Graphical illustrations of the behavior of the empirical k-means sets in different scenarios of non-uniqueness.

(b) Some simulations designed to evaluate the performance of the uniqueness test in (5.18).

(c) A few Monte Carlo experiments, where many samples are drawn from a known underlying distribution to check the empirical level of the k-means uniqueness test ($\alpha = 0.05$ being the theoretical significance level in all examples).

The models under study

We follow here the notation introduced in Section 5.2. Thus, Cikj stands for a model in which the data are drawn for a mixture of i distributions and we decide to look for $k = j$ centers. Of course, ideally, if we had prior information on the underlying distribution (which seldom happens in practice), the k-means parameter should be typically chosen to match the numbers of mixture components, that is, $i = j$. But, as we have commented in Section 5.2, non-uniqueness might arise from a "wrong" choice of k when compared with the true numbers of mixture components.

We explore five different situations, as described below. Here $\mathcal{U}(A)$ denotes a uniform distribution with support A, and $\mathcal{N}(\mu, \Sigma)$ represents a Gaussian distribution with mean μ and covariance matrix Σ.

C1k2 Two different population distributions are considered: $\mathrm{P} \sim \mathcal{N}\left((0,0), \mathbb{I}_2\right)$, where $\mathbb{I}_2$ is the identity matrix of dimension 2, and $\mathrm{P} \sim \mathcal{U}\left(B\left((0,0), \frac{1}{5}\right)\right)$, where $B\left(0, \frac{1}{5}\right)$ is the open ball of center $(0,0)$ and radius $\frac{1}{5}$. In both models there is a single population center and we chose (wrongly) $k = 2$ so that infinite sets of 2-means appear as in Figure 5.4.

C2k2, C2k3 Here

$$\mathrm{P} \sim \frac{1}{2}\mathcal{N}\left((-1,0), \frac{\mathbb{I}_2}{25}\right) + \frac{1}{2}\mathcal{N}\left((1,0), \frac{\mathbb{I}_2}{25}\right),$$

and

$$\mathrm{P} \sim \frac{1}{2}\mathcal{U}\left(B\left((-1,0), \frac{1}{5}\right)\right) + \frac{1}{2}\mathcal{U}\left(B\left((1,0), \frac{1}{5}\right)\right).$$

In other words, we consider a mixture of two well-separated distributions.

C3k2, C3k3 We consider

$$\mathrm{P} \sim \frac{1}{3}\mathcal{N}\left((-1,0),\frac{\mathbb{I}_2}{25}\right)+\frac{1}{3}\mathcal{N}\left((0,0),\frac{\mathbb{I}_2}{25}\right)+\frac{1}{3}\mathcal{N}\left((1,0),\frac{\mathbb{I}_2}{25}\right),$$

and

$$\mathrm{P} \sim \frac{1}{3}\mathcal{U}\left(B\left((-1,0),\frac{1}{5}\right)\right)+\frac{1}{3}\mathcal{U}\left(B\left((0,0),\frac{1}{5}\right)\right)+\frac{1}{3}\mathcal{U}\left(B\left((1,0),\frac{1}{5}\right)\right).$$

We deal with a mixture of three distributions well-separated from each other.

Some graphical illustrations

The colored points, and lines, in the before figures correspond to the empirical estimators of the k-means sets we have obtained from very large samples of the respective models (so we do not have analytically calculated the population k-means). The light gray points are a few observations from these distributions, just to give a more complete idea of the whole situation.

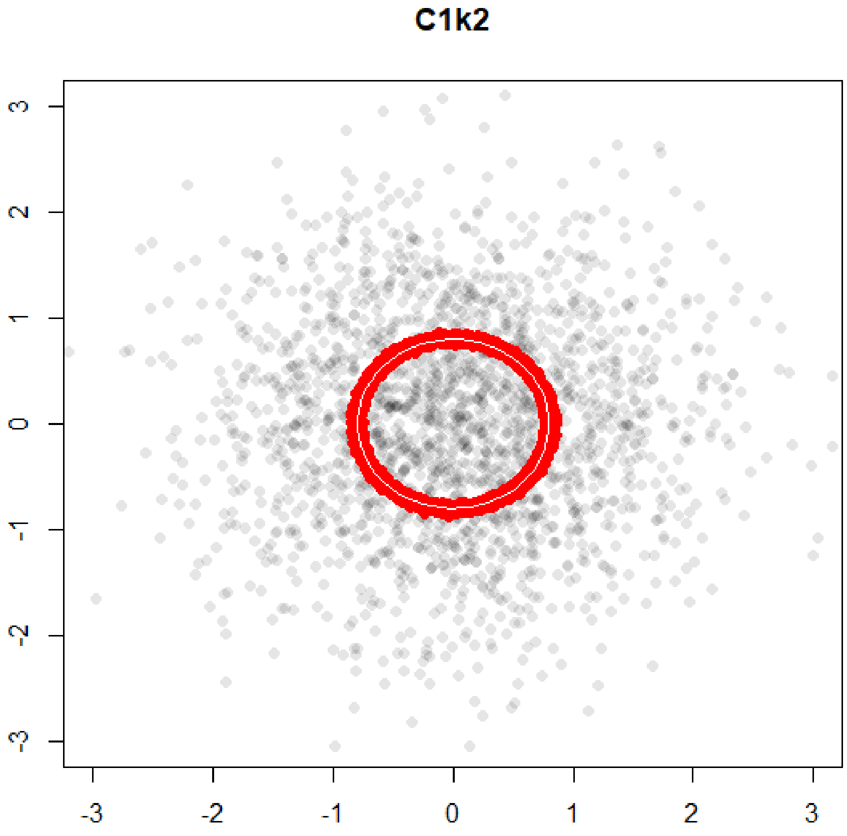


Figure 5.4: Model C1k2. In light black, a sample of a standard Gaussian distribution of size 2000. In red, centers resulting from running the Hartigan-Wong k-means algorithm over 1000 samples of size $n = 2 \times 10^5$ from a standard Gaussian distribution with $k = 2$.

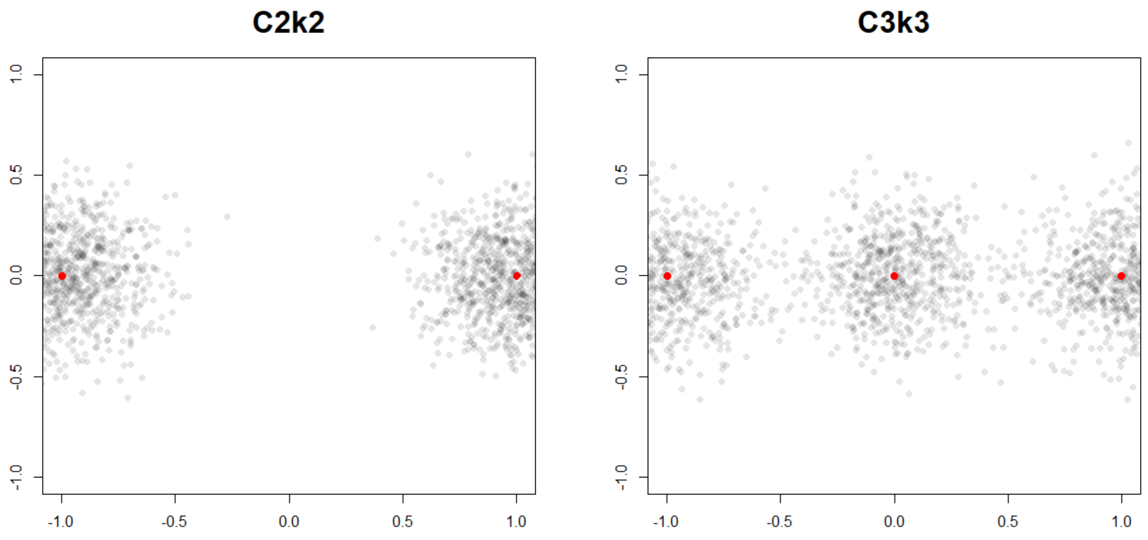


Figure 5.5: Model C2k2 in the left. In light black, a sample of the mixture $\frac{1}{2}\mathcal{N}\left((-1,0),\frac{\mathbb{I}_2}{25}\right)+\frac{1}{2}\mathcal{N}\left((1,0),\frac{\mathbb{I}_2}{25}\right)$ of size 2000. In red, centers resulting from running the Hartigan-Wong k-means algorithm over 1000 samples of size $n = 2\times10^5$ from a standard Gaussian distribution with $k = 2$. Model C3k3 on the right. Analogous graph taking the mixture $\frac{1}{3}\mathcal{N}\left((-1,0),\frac{\mathbb{I}_2}{25}\right)+\frac{1}{3}\mathcal{N}\left((0,0),\frac{\mathbb{I}_2}{25}\right)+\frac{1}{3}\mathcal{N}\left((1,0),\frac{\mathbb{I}_2}{25}\right)$ and $k = 3$.

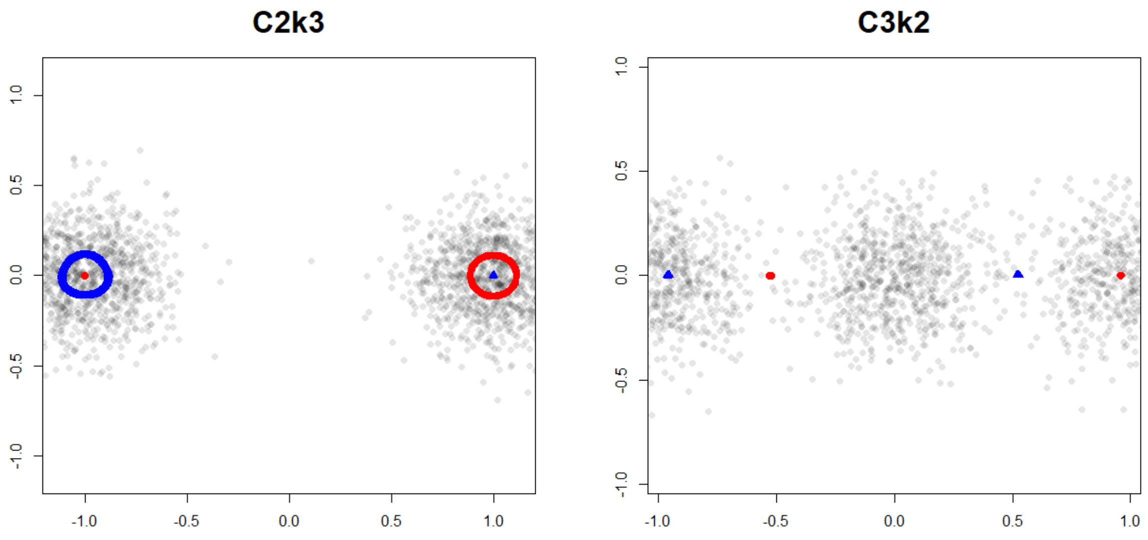


Figure 5.6: Model C2k3 in the left. In light black, a sample of the mixture $\frac{1}{2}\mathcal{N}\left((-1,0),\frac{\mathbb{I}_2}{25}\right)+\frac{1}{2}\mathcal{N}\left((1,0),\frac{\mathbb{I}_2}{25}\right)$ of size 2000. In red-bullets and blue-triangles, centers resulting from running the Hartigan-Wong k-means algorithm over 1000 samples of size $n = 2\times10^5$ from a standard Gaussian distribution with $k = 2$. Infinite sets of 3-means. Model C3k3 on the right. Analogous graph taking the mixture $\frac{1}{3}\mathcal{N}\left((-1,0),\frac{\mathbb{I}_2}{25}\right)+\frac{1}{3}\mathcal{N}\left((0,0),\frac{\mathbb{I}_2}{25}\right)+\frac{1}{3}\mathcal{N}\left((1,0),\frac{\mathbb{I}_2}{25}\right)$ and $k = 2$. Two sets of 2-means.

Test experiments

Before presenting the results of testing the hypothesis of uniqueness of the set of k-means (see (5.18)), let us describe first the different testing procedures. According to Corollary 64, the test in (5.18) is equivalent to test that the quantity $T_n(k)$ in (5.15) is asymptotically normal.

Given a sample $X_1, \dots, X_n \sim \mathrm{P}$, we proceed as follows:

1. Draw $B = 1000$ bootstrap samples $X_1^*, \dots, X_n^* \sim \mathbb{P}_n$ (of size n).

2. Compute the quantities $T_n^*(k)$, where

$$T_n^*(k) = \sqrt{n}\left(\mathrm{R}_n^*\left(\mathcal{F}_{V_k(B)}\right) - \mathrm{R}_n\left(\mathcal{F}_{V_k(B)}\right)\right),$$

   $\mathrm{R}_n^*\left(\mathcal{F}_{V_k(B)}\right)$ being the risk computed with the bootstrap sample.

3. Implement a test of Gaussianity over the bootstrap sample of the previous step in (5.19) (a) or a test for the mean in (5.19) (b). Here, we have used this last approach as well as two different normality tests available in literature:

   Normal mean (NM): We conduct a test for the mean based on (5.19) (b) over the bootstrap sample of the estimator of the risk.

   Jarque-Bera (JB): We carry out a normality test based on third and fourth moments; see Jarque and Bera (1987).

   Anderson-Darling (AD): We run a normality test based on a weighted L^2-norm of the distribution functions; see Anderson and Darling (1954).

   Regarding the asymptotic validity of the bootstrap methodology, let us recall that the functional given by the infimum is fully Hadamard directional differentiable if and only if there is an unique set of k-means; see Remark 27. Hence, under the null hypothesis the naive bootstrap is a.s. consistent; see Fang and Santos (2019, Theorem 3.1).

The results of the simulations are summarized in Table 5.1. All the considered mixtures are made of Gaussian distributions. Note that the Type I error is not controlled until $n = 2 \times 10^5$. The best results are obtained by the test of means NM. However, this test tends to over-reject under the null hypothesis systematically with lower sample sizes. It is also remarkable that JB and AD are not able to discriminate between non-uniqueness continuous cases (C1k2 and C2k3) and uniqueness (null-hypothesis: C2k2 and C3k3).

Monte Carlo experiments

As commented above, based on the the result of the testing experiment, which is far from being exhaustive, the best option for testing uniqueness of the set of k-means is NM; see Table 5.1. However, large sample sizes are required. Table 5.2 shows the results obtained from the indicated models, when the mixtures are made of uniform distributions

Test		C1k2	C2k2	C2k3	C3k3	C3k2
NM	$n = 10^5$	**1.00**	0.100	**0.990**	0.110	**1.00**
	$n = 2 \times 10^5$	**1.00**	0.060	**0.990**	0.050	**0.890**

		C1k2	C2k2	C2k3	C3k3	C3k2
JB	$n = 10^5$	0.055	0.050	0.040	0.095	0.785
	$n = 2 \times 10^5$	0.030	0.050	0.100	0.030	0.730

		C1k2	C2k2	C2k3	C3k3	C3k2
AD	$n = 10^5$	0.040	0.070	0.060	0.070	0.660
	$n = 2 \times 10^5$	0.025	0.055	0.065	0.035	0.625

Table 5.1: Rejection proportion with level $\alpha = 0.05$ in the five different proposed models under different testing procedures. The sample sizes used are $n = 10^5, 2 \times 10^5$.

(as pointed out above), where the k-means problem should be easier to solve. Further, all the involved quantities can be computed explicitly. Hence, it is the best possible scenario for the k-means problem. In this experiment (where, in all cases, we are under the null hypothesis of uniqueness) the following steps have been followed:

1. Take $B = 1000$ samples $X_1, \ldots, X_n \sim \mathrm{P}$ (of size n).
2. Compute the quantities
$$T_n(k) = \frac{\sqrt{n}}{\sigma}\left(\mathrm{R}_n\left(\mathcal{F}_{V_k(B)}\right) - \mathrm{R}\left(\mathcal{F}_{V_k(B)}\right)\right),$$
where $\sigma^2 = \mathbb{V}\mathrm{ar}_{\mathrm{P}}\left(f^-\right)$ and f^- is the unique minimizer of the risk. The population constants σ and $\mathrm{R}\left(\mathcal{F}_{V_k(B)}\right)$ are explicitly computed.
3. Perform the test for the mean in (5.19) (b).

Note that there are two main differences with respect to the previous experiments. First, the samples are taken from the “true” population distribution P (since it is assumed to be known). Second, standardization of the risk is carried out with the true population parameters.

Some conclusions

Though our empirical study is far from exhaustive, we may draw some provisional conclusions. First, the overall results are consistent with our theory in the sense that the asymptotic test seems to work, both in the sense of controlling the type I error and in the sense of detecting departures from the null hypothesis. Still, the convergence is slow, as we observe that large sample sizes are needed in order to control the Type I error.

Second, the outputs of the Monte Carlo experiments show some counter-intuitive aspects, when compared with those of the bootstrap simulations. For some reason (not

n	C2k2	C3k3
10^4	0.315	0.340
2×10^4	0.200	0.245
5×10^4	0.135	0.200
10^5	0.075	0.175
2×10^5	0.095	0.100
5×10^5	0.045	0.100
10^6	0.050	0.055

Table 5.2: Type I error with $\alpha = 0.05$ of the NM test under Monte Carlo data obtained from mixtures of uniform distributions under models C2k2 and C3k3.

completely clear to us), larger sample sizes are required to properly control the significance level. Recall that simulations are based on normal mixtures which, in principle, should lead to harder situations that those of the uniform counterparts considered in the Monte Carlo experiments. Clearly, more research is needed in this regard.

Chapter 6

Conclusions and future work

The goal of this section is to provide a perspective on the main contributions of this thesis. The supremum norm is a widely applicable mathematical tool in Statistics. This is mainly due to its simplicity and versatility. The supremum norm is *simple* in that, intuitively, it extends pointwise distance. Furthermore, it characterizes uniform convergence, so that this intuition materializes in results about continuity. So much so that, for example, the Kolmogorov-Smirnov test has been the foundation of many non-parametric statistical works for over 70 years. As mentioned, in addition to being simple, it is *versatile* since there is a vast number of duality theorems in function spaces that characterize norms and distances such as the supremum over the appropriate set (Hahn-Banach, Kantorovich-Rubinstein, Riesz for measures, Riesz in Hilbert spaces, Riesz siblings, ...). From a computational perspective, it is also noteworthy because it is naturally approximated through the maximum over an increasing sequence of finite sets that eventually fill the domain.

Regarding the results obtained in this thesis, we can say that they are developed around two main axes that complement each other: theoretical and computational results. The theoretical contributions, as evident in Chapters 3, 4, and 5, are based on the application of the extended Delta method (see Proposition 5, also Shapiro (1991, Theorem 2.1)). The key result enabling this is found in Theorem 17, where the Hadamard directional differentiability of the supremum (and other related functionals) in the space of bounded functions equipped with the supremum norm is proven. Thus, thanks to the consequences of this theorem obtained throughout Chapter 2, the following statistical results are established:

- In Section 3.1 of Chapter 3, we generalize the results obtained in Raghavachari (1973) on the Kolmogorov-Smirnov and Kuiper statistics under the alternative hypothesis. Briefly, these results were obtained through a careful analysis of the statistics, for which the author imposed an unnecessary continuity restriction. This constraint has been replicated in subsequent works. With the approach proposed in this thesis, this restriction is eliminated by explicitly calculating the distribution of these statistics in a very general setting which allows for multivariate extensions.

  Alongside this classical result, other applications in non-parametric statistics are also presented, namely, the asymptotic distribution of the estimator of the distance between copulas, the asymptotic distribution of Berk-Jones type statistics, and the asymptotic (general) distribution of integral probability distances, also known as *maximum mean discrepancies* (MMD).

- In Chapter 4, a test for two-sample problems in high dimensions and functional data based on kernel-type distances is derived. Specifically, a new test is proposed based on the distance resulting from calculating the supremum of a family of distances. Intuitively, under the alternative hypothesis, the discrepancy between different probability measures is maximized, thereby increasing power and robustness. Unlike previous works on this topic, heuristics to select a parameter (usually data-driven, dependent on the data) that escapes the theoretically proven results to date are avoided in this approach. Finally, the test based on this new proposal also allows for considering an entire family of kernels (associated with the relevant distance family), making it more robust against a poor choice that might not appropriately capture the data differences in the true populations.

  Section 4.4 of this chapter also includes, an empirical analysis where the power of the test is compared with other popular proposals. Although these results are not exhaustive, the performance of our proposal is better than that of other kernel-based distance tests. The software related to these experiments will be publicly available in an `R` package named `skd2`.

- In Chapter 5, we address the issue of uniqueness in sets of k-means. Specifically, using the same theoretical framework as in the previous chapters, it is proven that the normalized sum of squares follows a Gaussian distribution if and only if there is a unique set of k-means in the considered population (for a fixed k). Additionally, a consistency result on the estimation of the population k-means through a sequence of sample k-means is provided.

  The proposed test (based on a normality test or a test for the mean) is derived from these theoretical results, and a small empirical study is conducted. Although more computational work is needed, the theoretical results can be visualized in Monte Carlo experiments as well as in the rejection proportion experiments with simulated data.

Up to this point, this thesis has resulted in two scientific articles: Cárcamo et al. (2020), which encompasses Chapters 2 and 3; and Cárcamo et al. (2022), which covers Chapter 4.

Future work

We briefly comment on some lines of future work that constitute the natural continuation of this thesis. As it is always the case in science, no work is ever truly finished because new questions continually arise. This thesis is no exception.

- The works derived from this thesis fit into the theoretical framework provided by the differentiability of the supremum. Since the publication Cárcamo et al. (2020), others have emerged, such as Dette and Kokot (2022), Hundrieser et al. (2022), and del Barrio et al. (2024). We expect to add two more to the list soon, focusing on the empirical process with estimated parameters and the application of functional regression models with functional response.

- Following the theoretical and empirical results presented in Chapter 4, it is necessary to conduct a depth empirical study. The SKD appears to be a reasonable alternative to classical kernel distances and energy tests. Additionally, there are proposals for homogeneity tests that have been excluded from this study, along with others related to random projections or dimensionality reduction. From a theoretical standpoint, it is known that the popular *energy test* can be expressed as a kernel distance (see Sejdinovic et al. (2013)). Therefore, it is reasonable to inquire whether including this family of distances in the SKD would lead to increased power in models where the *energy tests* performs better. In summary, a comprehensive experiment with various scenarios and real data is needed to shed light on when to use one test over another in two sample problems.

- The work presented in Chapter 5 is still in progress. The k-means problem is inherently complex. So far, non-uniqueness scenarios have been proposed with simulated data. Intuition suggests that non-uniqueness scenarios are related to models that exhibit a certain degree of symmetry. Therefore, two main lines of work open up. On one hand, the necessary and sufficient condition for uniqueness is purely analytical and probabilistic. It would be interesting to find sufficient conditions regarding the geometric nature. On the other hand, conducting a broader study with real data to demonstrate the usefulness of the uniqueness test for the set of k-means.

In summary, Statistics is a branch of Applied Mathematics where both abstract theoretical results and concrete applications with intensive computation continue to be of interest even today. This manuscript is a proof of it. On the theoretical front, all key results find their foundation in Theorem 17. This result, at a considerable level of abstraction, allows for a profound understanding and a unified mathematical framework for classical results from 50 years ago, others not so old, and some yet to come. On the computational side, particularly as these lines are written with Chapter 5 in mind, they are those experiments—easy to comprehend yet challenging to execute—that defy intuition and pave the way for the theoretical development.

Chapter 7

Conclusiones y trabajo futuro

El principal objetivo de esta sección es poner en perspectiva las aportaciones principales de esta tesis. La norma del supremo es una herramienta matemática de amplia aplicación en Estadística. Esto se debe principalmente a su simplicidad y versatilidad. La norma del supremo es *simple* en tanto a que, intuitivamente, extiende la distancia punto a punto. Caracteriza, además, la convergencia uniforme, por lo que aparece de manera natural en los resultados sobre funciones continuas. Tanto es así que, por ejemplo, el test de Kolmogorov-Smirnov es la base de muchos trabajos de estadística no paramétrica desde hace más de 70 años. Además de simple es *versátil* pues existe una vasta cantidad de teoremas de dualidad en espacios de funciones que caracterizan normas y distancias como el supremo sobre el conjunto apropiado (Hahn-Banach, Kantorovich-Rubinstein, Riesz para medidas, Riesz en espacios de Hilbert, hermanos Riesz, . . .). Desde el punto de vista computacional es manejable pues en definitiva, el máximo sobre una sucesión creciente de conjuntos finitos que, eventualmente, llenan el dominio supone una aproximación natural del supremo de una función.

En cuanto a los resultados obtenidos en esta tesis, podemos decir que se dan en dos ejes principales que se complementan mutuamente: teórico y computacional. Las aportaciones teóricas, como puede verse en los apítulos 3, 4 y 5, se basan en la aplicación del Método delta extendido (ver Proposición 5, también Shapiro (1991, Theorem 2.1)). El resultado principal que hace eso posible en el Teorema 17, donde se prueba la diferenciabilidad Hadamard direccional del supremo (y otros funcionales relacionados) en el espacio de las funciones acotadas equipado con la norma del supremo. Así, gracias a corolarios de este teorema obtenidos a lo largo del Capítulo 2, se prueban los siguientes resultados estadísticos:

- En la Sección 3.1 del Capítulo 3, se generalizan los resultados obtenidos en Raghavachari (1973) sobre el estadístico de Kolmogorov-Smirnov y de Kuiper bajo la hipótesis alternativa. Brevemente, los resultados de aquel artículo fueron obtenidos mediante un cuidadoso análisis de los estadísticos, para lo que el autor impone una restricción de continuidad que no es necesaria. Dicha limitación ha sido impuesta en los trabajos subsiguientes. Con el enfoque propuesto en esta tesis, esta es eliminada, calculando explícitamente la distribución asintótica de los estimadores bajo la alternativa un contexto más general. Asimismo, permite una extensión multivariante.

Junto a este resultado clásico se proporcionan también otras aplicaciones en Estadística no paramétrica: la distribución asintótica del estimador de la distancia entre cópulas, distribución asintótica de estadísticos del tipo Berk-Jones y distribución asintótica (general) de distancias integrales de probabilidad, también conocidas como *discrepancias máximas medias* (MMD).

- En el Capítulo 4 se obtiene un contraste de hipótesis para los problemas de dos muestras en alta dimensión y datos funcionales basado en las distancias de tipo kernel. Concretamente, se propone un nuevo test basado en el supremo de una familia de distancias. Intuitivamente, bajo la hipótesis alternativa se maximiza la discrepancia entre las medidas de probabilidad, aumentando así la potencia. A diferencia de los trabajos sobre esta temática, evitamos heurísticas para seleccionar los parámetros (métodos *data-driven*, dependiente de los datos, generalmente) que escapan a los resultados teóricos probados hasta la fecha. Finalmente, el test basado en esta nueva propuesta permite, además, tener en cuenta toda una familia de núcleos (asociados a las distancias pertinentes), por lo que es más robusto ante una mala elección que no capture las propiedades de los datos apropiadamente.

  En la Sección 4.4 un análisis empírico donde se ha comparado la potencia del test con otras propuestas. Como puede verse, aunque estos resultados no son exhaustivos, el desempeño de nuestra propuesta es mejor que el de otros tests basados en distancias kernel. El software relativo a estos experimentos estará disponible en breve públicamente en un paquete de `R` bajo el nombre `skd2`.

- En el Capítulo 5 se aborda la cuestión de unicidad en los conjuntos de k-medias. Concretamente, con la misma metodología que en los capítulos anteriores, se prueba que la suma de cuadrados dentro de los grupos normalizada sigue una distribución gaussiana si y solo si hay un único conjunto de k-medias en la población considerada (fijado k). Asimismo, se da también un resultado de consistencia para los estimadores (sucesión de k-medias muestrales) de las k-medias (poblacionales).

  De estos resultados teóricos se deriva el test de unicidad (basado en test de normalidad o medias) y se hace un pequeño estudio empírico. Aunque hace falta más trabajo computacional, pueden visualizarse los resultados teóricos demostrados, tanto en experimentos Montecarlo como en experimentos de proporción de rechazo con datos simulados.

Hasta el momento, esta tesis ha dado lugar a dos artículos científicos: Cárcamo et al. (2020), que engloba los Capítulos 2 y 3; y Cárcamo et al. (2022), que recoge el Capítulo 4.

Trabajo futuro

Comentamos aquí brevemente algunas líneas de trabajo futuro que son la continuidad natural de esta tesis. Como siempre sucede en Ciencia, ningún trabajo está totalmente terminado porque siempre surgen nuevos interrogantes. Esta tesis no es una excepción:

- Esta tesis y los trabajos derivados de ella encajan en el marco teórico que proporciona la diferenciabilidad del supremo. Desde la publicación de Cárcamo et al. (2020), han aparecido otros trabajos con similares ideas como Dette and Kokot (2022), Hundrieser et al. (2022) y del Barrio et al. (2024). Esperamos, próximamente, añadir dos más a la lista relacionados con el proceso empírico con parámetros estimados y la aplicación de los modelos de regresión funcional con respuesta funcional a la inferencia sobre ecuaciones diferenciales.

- Tras los resultados, tanto teóricos como empíricos, expuestos en el Capítulo 4, es necesario hacer un estudio empírico más profundo. Las SKD parece una alternativa razonable a las distancias kernel clásicas y a los energy test. Asimismo, hay propuestas de tests de homogeneidad que han quedado fuera de este estudio y otros relacionados, como aquellos basados en proyecciones aleatorias o reducción de dimensión. Desde el punto de vista teórico, es sabido que los populares *energy test* se pueden expresar como una distancia kernel (ver Sejdinovic et al. (2013)). Por tanto, es razonable preguntarse si incluir esta familia de distancias en la SKD llevaría a un aumento de la potencia en aquellos modelos donde los *energy* tenían mejores resultados. En síntesis, hace falta realizar un experimento amplio con situaciones más o menos frecuentes y datos reales que arroje algo de luz sobre cuándo se debe usar un test u otro en problemas de dos muestras.

- El trabajo presentado en el Capítulo 5 se encuentra aún en vías de desarrollo. El problema de k-medias es, en sí mismo, una cuestión compleja. Hasta ahora se han propuesto escenarios de no unicidad limitados con datos simulados. La intuición nos dice que los ejemplos de no unicidad están relacionados con modelos que poseen cierto grado de simetría. Por tanto, se nos abren dos líneas principales de trabajo. Por un lado, la unicidad es una condición puramente analítica y probabilística. Sería interesante dar, además, una condición suficiente de tipo geométrico. Por otro, sería deseable hacer un estudio más amplio con datos reales para así mostrar la utilidad del test de unicidad propuesto.

En síntesis, la Estadística es una rama de la Matemática aplicada donde aún siguen teniendo interés tanto los resultados teóricos abstractos como las aplicaciones concretas de intensa computación. Este manuscrito es una prueba de ello. En lo que al plano teórico respecta, todos los resultados clave tienen su punto de apoyo en el Teorema 17. Este resultado, desde una nivel de abstracción considerable, permite comprender de manera profunda y bajo el mismo marco matemático resultados clásicos de hace 50 años, otros no tan antiguos y algunos que están por llegar. Por otro lado, desde la perspectiva computacional, con el Capítulo 5 en mente al redactar estas líneas; son esos experimentos, fáciles de comprender y complejos de ejecutar los que desafían la intuición sugiriendo el camino que debe seguir el desarrollo teórico.

Glossary

$\rightsquigarrow$ Weak convergence of probability measures in the sense of Hoffamnn-Jørgensen (see A. van der Vaart and Wellner, 1996). 17, 18, 26, 68

$\ll$ usual (partial) order in the set of kernels. $k_1 \ll k_2$ if and only if $k_2 - k_1$ is a positive definite kernel. 43, 66

$\mathbb{G}_{\mathrm{P}}$ P-Brownian bridge. 18, 41, 68

$\mathbb{P}_n$ empirical probability measure associated with a sample of size n coming from P. 17, 65, 84

$\mathbf{1}_A$ indicator function of the set A (if $x \in A$, $\mathbf{1}_A(x) = 1$ and it takes the value 0 otherwise). 20, 31, 48

$\mathcal{C}$ space of continuous functions. Usually, the domain and its topology are also specified. If a topology is not specified with the domain, standard topology is assumed. Additionally, $\mathcal{C}_{\mathrm{b}}$ is the space of *bounded* continuous function, $\mathcal{C}_{\mathrm{u}}$ is the space of *uniformly* continuous functions, and $\mathcal{C}_{\mathrm{pl}}$ is the space of continuous *pre-linear* functions. Indices can be combined. 32, 34, 36, 40, 41, 43, 54, 94

$\mathcal{F}_{V_k(B)}$ $= \left\{ f(z) = \min_{j=1,\dots,k} \left(\|z - a_j\|^2_{\mathcal{B}} \right) : (a_1, \dots, a_k) \in V_k(B) \right\}$, with B totally bounded in the corresponding space and $V_k(B) = \{(a_1, \dots, a_k) \in B^k : a_i \neq a_j \text{ for } i \neq j\}$. $\mathcal{F}_{V_k(B)}$ is the associated class of functions for empirical risk minimization in k-means problem. 48, 92

$\mathcal{F}_{\mathcal{H}_{k,\Lambda}}$ $= \bigcup_{\lambda \in \Lambda} \mathcal{F}_{\mathcal{H}_{k,\lambda}}$, where $\{k_\lambda : \lambda \in \Lambda\}$ is a family of reproducing kernels indexed by Λ and $\mathcal{F}_{\mathcal{H}_{k,\lambda}}$ is the respective unit ball in the corresponding RKHS. 42, 56

$\mathcal{F}_{\mathcal{H}_k}$ $= \left\{ h \in \mathcal{H}_k : \|h\|_{\mathcal{H}_k} \leq 1 \right\}$, unit ball of the reproducing kernel Hilbert space $\mathcal{H}_k$. 23, 40, 48, 56, 64, 65, 68, 81

$\mathcal{M}$ space of Borel measures. Usually, the domain and its topology are also specified. If a topology is not specified with the domain, standard topology is assumed. Additionally, $\mathcal{M}_{\mathrm{p}}$ denoted the subset of probability measures, a the subset of $\mathcal{M}$. 62

L with exponent $p \in [1, \infty]$, the space of classes of functions f such that $\int |f|^p < \infty$. The measure and the domain can be specified as $\mathrm{L}^p(\mathfrak{X}, \nu)$. L^1 is the space of integrable functions. 23, 55, 64

i. i. d. independent and identically distributed. 20

$\rho_{\mathrm{L}^2(\mathrm{P})}$ Intrinsic pseudometric of square integrable functions respect to the measure P. 18, 57, 80, 93

ρ_{P} Intrinsic pseudometric associated with the measure P. 18, 41, 94

o_{p} "Little" o for convergence in probability (in measure respect to the probability measure). 17

$=_{\mathrm{st}}$ equality in distribution. 54

Bibliography

Álvarez-Esteban, P. C., Del Barrio, E., Cuesta-Albertos, J. A., & Matrán, C. (2012). Similarity of samples and trimming. *Bernoulli*, *18*(2), 606–634.

Álvarez-Esteban, P. C., del Barrio, E., Cuesta-Albertos, J., Matrán, C., et al. (2016). A contamination model for the stochastic order. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, *25*(4), 751–774.

Anderson, T. W., & Darling, D. A. (1954). A test of goodness of fit. *Journal of the American statistical association*, *49*(268), 765–769.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, *68*(3), 337–404.

Baíllo, A., Cárcamo, J., & Getman, K. (2019). New distance measures for classifying x-ray astronomy data into stellar classes. *Advances in Data Analysis and Classification*, *13*(2), 531–557.

Banach, S. (1932). *Théorie des opérations linéaires*.

Bay, X., & Croix, J.-C. (2017). Karhunen-loève decomposition of gaussian measures on banach spaces. *arXiv preprint arXiv:1704.01448*.

Beare, B. K., & Fang, Z. (2017). Weak convergence of the least concave majorant of estimators for a concave distribution function. *Electronic Journal of Statistics*, *11*, 3841–3870.

Ben-David, S., Pál, D., & Simon, H. U. (2007). Stability of k-means clustering. *Learning Theory: 20th Annual Conference on Learning Theory, COLT 2007, San Diego, CA, USA; June 13-15, 2007. Proceedings 20*, 20–34.

Ben-David, S., Von Luxburg, U., & Pál, D. (2006). A sober look at clustering stability. *Learning Theory: 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 22-25, 2006. Proceedings 19*, 5–19.

Berk, R. H., & Jones, D. H. (1979). Goodness-of-fit test statistics that dominate the kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, *47*(1), 47–59.

Berlinet, A., & Thomas-Agnan, C. (2011). *Reproducing kernel hilbert spaces in probability and statistics*. Springer Science & Business Media.

Bickel, P. J., & Wichura, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *The Annals of Mathematical Statistics*, *42*(5), 1656–1670.

Bogachev, V. I. (1998). *Gaussian measures*. American Mathematical Soc.

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., & Smola, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, *22*(14), e49–e57.

Brezis, H., & Brézis, H. (2011). *Functional analysis, sobolev spaces and partial differential equations* (Vol. 2). Springer.

Burago, D., Burago, Y., & Ivanov, S. (2022). *A course in metric geometry* (Vol. 33). American Mathematical Society.

Caponnetto, A., & Rakhlin, A. (2006). Stability properties of empirical risk minimization over Donsker classes. *Journal of Machine Learning Research*, *7*(12).

Cárcamo, J. (2017). Integrated empirical processes in L^p with applications to estimate probability metrics. *Convergence of sequential quasi-Monte Carlo smoothing algorithms*, *23*(4B), 3412–3436.

Cárcamo, J., Cuevas, A., & Rodríguez, L.-A. (2020). Directional differentiability for supremum-type functionals: Statistical applications. *Bernoulli*, *26*(3), 2143–2175.

Cárcamo, J., Cuevas, A., & Rodríguez, L.-A. (2022). A uniform kernel trick for high-dimensional two-sample problems. *arXiv preprint arXiv:2210.02171*.

Clarkson, J. A. (1936). Uniformly convex spaces. *Transactions of the American Mathematical Society*, *40*(3), 396–414.

Conway, J. B. (2019). *A course in functional analysis* (Vol. 96). Springer.

Cox, G. (2020). Almost sure uniqueness of a global minimum without convexity. *Annals of statistics*, *48*(1), 584–606.

Cuesta, J., & Matrán, C. (1988). The strong law of large numbers for k-means and best possible nets of banach valued random variables. *Probability theory and related fields*, *78*(4), 523–534.

Cuesta-Albertos, J. A., Fraiman, R., & Ransford, T. (2007). A sharp form of the cramér–wold theorem. *Journal of Theoretical Probability*, *20*(2), 201–209.

Cuevas, A., Febrero, M., & Fraiman, R. (2004). An anova test for functional data. *Computational statistics & data analysis*, *47*(1), 111–122.

DasGupta, A. (2008). *Asymptotic theory of statistics and probability* (Vol. 180). Springer.

del Barrio, E., González Sanz, A., & Loubes, J.-M. (2024). Central limit theorems for semi-discrete wasserstein distances. *Bernoulli*, *30*(1), 554–580.

del Barrio, E., Inouzhe, H., & Matrán, C. (2020). On approximate validation of models: A kolmogorov–smirnov-based approach. *TEST*, *29*(4), 938–965.

Denuit, M., Dhaene, J., Goovaerts, M., & Kaas, R. (2006). *Actuarial theory for dependent risks: Measures, orders and models*. John Wiley & Sons.

Dette, H., & Kokot, K. (2022). Detecting relevant differences in the covariance operators of functional time series: A sup-norm approach. *Annals of the Institute of Statistical Mathematics*, *74*(2), 195–231.

Dette, H., Möllenhoff, K., Volgushev, S., & Bretz, F. (2018). Equivalence of regression curves. *Journal of the American Statistical Association*, *113*(522), 711–729.

Dudley, R. M. (2002). *Real analysis and probability*. Cambridge University Press.

Dudley, R. M. (2014). *Uniform central limit theorems* (Vol. 142). Cambridge university press.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*, *95*, 125–140.

El-Fallah, O., Kellay, K., Mashreghi, J., & Ransford, T. (2014). *A primer on the dirichlet space* (Vol. 203). Cambridge University Press.

Fang, Z., & Santos, A. (2019). Inference on directionally differentiable functions. *The Review of Economic Studies*, *86*(1), 377–412.

Fermanian, J.-D. (2013). An overview of the goodness-of-fit test problem for copulas. *Copulae in Mathematical and Quantitative Finance: Proceedings of the Workshop Held in Cracow, 10-11 July 2012*, 61–89.

Fernholz, L. T. (1983). *Von mises calculus for statistical functionals* (Vol. 19). Springer Science & Business Media.

Filippova, A. (1962). Mises' theorem on the asymptotic behavior of functionals of empirical distribution functions and its statistical applications. *Theory of Probability & Its Applications*, *7*(1), 24–57.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, *222*(594-604), 309–368.

Fortet, R., & Mourier, E. (1953). Convergence de la répartition empirique vers la répartition théorique. *Annales scientifiques de l'École Normale Supérieure*, *70*(3), 267–285.

Freitag, G., Lange, S., & Munk, A. (2006). Non-parametric assessment of non-inferiority with censored data. *Statistics in medicine*, *25*(7), 1201–1217.

Fukumizu, K., Gretton, A., Lanckriet, G., Schölkopf, B., & Sriperumbudur, B. K. (2009). Kernel choice and classifiability for rkhs embeddings of probability distributions. *Advances in neural information processing systems*, *22*.

Fukumizu, K., Gretton, A., Sun, X., & Schölkopf, B. (2007). Kernel measures of conditional dependence. *Advances in neural information processing systems*, *20*.

García-Escudero, L. A., Gordaliza, A., & Matrán, C. (1999). A central limit theorem for multivariate generalized trimmed k-means. *Annals of statistics*, 1061–1079.

Genest, C., & Nešlehová, J. G. (2014). On tests of radial symmetry for bivariate copulas. *Statistical Papers*, *55*, 1107–1119.

Giné, E., & Nickl, R. (2008). Uniform central limit theorems for kernel density estimators. *Probability Theory and Related Fields*, *141*(3-4), 333–387.

Giné, E., & Nickl, R. (2021). *Mathematical foundations of infinite-dimensional statistical models* (2nd ed.). Cambridge university press.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., & Smola, A. (2006). A kernel method for the two-sample-problem. *Advances in neural information processing systems*, *19*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, *13*(1), 723–773.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., & Smola, A. (2007). A kernel statistical test of independence. *Advances in neural information processing systems*, *20*.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., & Sriperumbudur, B. K. (2012). Optimal kernel choice for large-scale two-sample tests. *Advances in neural information processing systems*, *25*.

Hall, P., & Van Keilegom, I. (2007). Two-sample tests in functional data analysis starting from discrete data. *Statistica Sinica*, 1511–1531.

Hampel, F., Ronchetti, E., Rousseeuw, P., & Stahel, W. (2011). *Robust statistics: The approach based on influence functions*. Wiley.

Hille, E., & Phillips, R. (1957). Functional analysis and semigroups, ams colloq. *Public. XXXI. The Russian translation:(1962), Funktsionalnyi analiz i polugruppy. M.*

Hjort, N. L. (1990). Goodness of fit tests in models for life history data based on cumulative hazard rates. *The Annals of Statistics*, 1221–1258.

Hsing, T., & Eubank, R. (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators* (Vol. 997). John Wiley & Sons.

Huber, P. J. (2004). *Robust statistics* (2nd ed.). Wiley.

Huber, P. J. (2011). Robust statistics. In *International encyclopedia of statistical science* (pp. 1248–1251). Springer.

Hundrieser, S., Klatt, M., Staudt, T., & Munk, A. (2022). A unifying approach to distributional limits for empirical optimal transport. *arXiv preprint arXiv:2202.12790.*

Jager, L., & Wellner, J. A. (2004). On the "Poisson boundaries" of the family of weighted kolmogorov statistics. *Lecture Notes-Monograph Series*, 319–331.

Jager, L., & Wellner, J. A. (2007). Goodness-of-fit tests via phi-divergences. *Annals of Statistics*, *35*(5), 2018–2053.

Jain, A. K. (2008). Data clustering: 50 years beyond k-means. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 3–4.

Janson, S. (1997). *Gaussian Hilbert spaces*. Cambridge university press.

Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, 163–172.

Kallianpur, G., & Rao, C. R. (1955). On fisher's lower bound to asymptotic variance of a consistent estimate. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, *15*(4), 331–342.

Ledoux, M., & Talagrand, M. (1991). *Probability in banach spaces: Isoperimetry and processes* (Vol. 23). Springer Science & Business Media.

Lember, J. (2003). On minimizing sequences for k-centres. *Journal of Approximation Theory*, *120*(1), 20–35.

Leonard, I., & Taylor, K. (1983). Supremum norm differentiability. *International Journal of Mathematics and Mathematical Sciences*, *6*(4), 705–713.

Leonard, I., & Taylor, K. (1985). Essential supremum norm differentiability. *International Journal of Mathematics and Mathematical Sciences*, *8*(3), 433–439.

Li, L., & Flury, B. (1995). Uniqueness of principal points for univariate distributions. *Statistics & probability letters*, *25*(4), 323–327.

Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., & Sutherland, D. J. (2020). Learning deep kernels for non-parametric two-sample tests. *International conference on machine learning*, 6316–6326.

Marcus, D. J. (1985). Relationships between donsker classes and sobolev spaces. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *69*(3), 323–330.

Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in mathematica. *Tutorials in Quantitative Methods for Psychology*, *9*(1), 15–24.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in applied probability*, *29*(2), 429–443.

Neininger, R., & Rüschendorf, L. (2004a). A general limit theorem for recursive algorithms and combinatorial structures. *The Annals of Applied Probability*, *14*(1), 378–418.

Neininger, R., & Rüschendorf, L. (2004b). On the contraction method with degenerate limit equation. *Annals of probability*, *32*(3B), 2838–2856.

Neuhaus, G. (1971). On weak convergence of stochastic processes with multidimensional time parameter. *The Annals of Mathematical Statistics*, *42*(4), 1285–1295.

Nickl, R., & Pötscher, B. M. (2007). Bracketing metric entropy rates and empirical central limit theorems for function classes of besov-and sobolev-type. *Journal of Theoretical Probability*, *20*, 177–199.

Parr, W. C. (1985). The bootstrap: Some large sample theory and connections with robustness. *Statistics & probability letters*, *3*(2), 97–100.

Paulsen, V., & Raghupathi, M. (2016). *An introduction to the theory of reproducing kernel hilbert spaces*. Cambridge University Press.

Pettis, B. (1938). On integration in vector spaces. *Transactions of the American Mathematical Society*, *44*(2), 277–304.

Pollard, D. (1981). Strong consistency of k-means clustering. *The Annals of Statistics*, 135–140.

Pollard, D. (1982). A central limit theorem for k-means clustering. *The Annals of Probability*, *10*(4), 919–926.

Pomann, G.-M., Staicu, A.-M., & Ghosh, S. (2016). A two sample distribution-free test for functional data with application to a diffusion tensor imaging study of multiple sclerosis. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, *65*(3), 395.

Rachev, S. T., Klebanov, L. B., Stoyanov, S. V., & Fabozzi, F. (2013). *The methods of distances in the theory of probability and statistics* (Vol. 10). Springer.

Raghavachari, M. (1973). Limiting distributions of kolmogorov-smirnov type statistics under the alternative. *The Annals of Statistics*, *1*(1), 67–73.

Rakhlin, A., & Caponnetto, A. (2006). Stability of k-means clustering. *Advances in neural information processing systems*, *19*.

Rao, M. M. (1997). *Real and stochastic analysisrecent advances* (Vol. 8). CRC Press.

Rice, J. (1995). *Mathematical statistics and data analysis*. Duxbury Press.

Rizzo, M., & Szekely, G. (2022). *Energy: E-statistics: Multivariate inference via the energy of data* [R package version 1.7-10]. https://CRAN.R-project.org/package=energy

Römisch, W. (2004). Delta method, infinite dimensional. *Encyclopedia of Statistical Sciences*, *3*.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, *40*(2), 99.

Schmoyer, R. L. (1988). Linear interpolation with a nonparametric accelerated failure-time model. *Journal of the American Statistical Association*, *83*(402), 441–449.

Scholkopf, B., & Smola, A. J. (2018). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press.

Segers, J. (2012). Asymptotics of empirical copula processes under non-restrictive smoothness assumptions. *Bernoulli*, *18*(3), 764–782.

Seijo, E., & Sen, B. (2011). A continuous mapping theorem for the smallest argmax functional. *Electronic Journal of Statistics*, *5*, 421–439.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., & Fukumizu, K. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The annals of statistics*, 2263–2291.

Serfling, R. (2009). *Approximation theorems of mathematical statistics*. Wiley.

Shaked, M., & Shanthikumar, J. G. (2007). *Stochastic orders*. Springer.

Shapiro, A. (1990). On concepts of directional differentiability. *Journal of optimization theory and applications*, *66*, 477–487.

Shapiro, A. (1991). Asymptotic analysis of stochastic programs. *Annals of Operations Research*, *30*, 169–186.

Shorack, G. R., & Wellner, J. A. (2009). *Empirical processes with applications to statistics*. SIAM.

Sommerfeld, M., & Munk, A. (2018). Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *80*(1), 219–238.

Sriperumbudur, B. K. (2011). Mixture density estimation via hilbert space embedding of measures. *2011 IEEE International Symposium on Information Theory Proceedings*, 1027–1030.

Sriperumbudur, B. K. (2016). On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, *22*(3), 1839–1893.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., & Lanckriet, G. R. (2012). On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, *6*, 1550–1599.

Sriperumbudur, B. K., Fukumizu, K., & Lanckriet, G. R. (2011). Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, *12*(7).

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., & Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, *11*, 1517–1561.

Székely, G. J., & Rizzo, M. L. (2017). The energy of data. *Annual Review of Statistics and Its Application*, *4*, 447–479.

Tarpey, T. (1994). Two principal points of symmetric, strongly unimodal distributions. *Statistics & Probability Letters*, *20*(4), 253–257.

Telgarsky, M. J., & Dasgupta, S. (2013). Moment-based uniform deviation bounds for k-means and friends. *Advances in Neural Information Processing Systems*, *26*.

Trushkin, A. (1982). Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting functions. *IEEE Transactions on Information Theory*, *28*(2), 187–198.

Tukey, J. W. (1959). A quick compact two sample test to duckworth's specifications. *Technometrics*, *1*(1), 31–48.

van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge university press.

van der Vaart, A., & Wellner, J. (1996). *Weak convergence and empirical processes: With applications to statistics*. Springer.

van der Vaart, A., & Wellner, J. (2023). *Weak convergence and empirical processes: With applications to statistics* (2nd ed.). Springer International Publishing.

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer science & business media.

Villani, C. (2008). *Optimal transport: Old and new*. Springer Berlin Heidelberg.

Von Luxburg, U., et al. (2010). Clustering stability: An overview. *Foundations and Trends® in Machine Learning*, *2*(3), 235–274.

von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *The annals of mathematical statistics*, *18*(3), 309–348.

von Mises, R. (1964). Mathematical theory of probability and statistics. *Mathematical Theory of Probability and Statistics*.

Wellner, J. A., & Koltchinskii, V. (2003). A note on the asymptotic distribution of berk—jones type statistics under the null hypothesis. *High Dimensional Probability III*, 321–332.

Wendland, H. (2004). *Scattered data approximation* (Vol. 17). Cambridge university press.

Wynne, G., & Duncan, A. B. (2022). A kernel two-sample test for functional data. *Journal of Machine Learning Research*, *23*(73), 1–51.

Zhang, H., & Zhao, L. (2013). On the inclusion relation of reproducing kernel hilbert spaces. *Analysis and Applications*, *11*(02), 1350014.

Zhang, J.-T. (2013). *Analysis of variance for functional data*. CRC press.

Zhang, J.-T., Guo, J., & Zhou, B. (2022). Testing equality of several distributions in separable metric spaces: A maximum mean discrepancy based approach. *Journal of Econometrics*.

Zhang, J.-T., & Smaga, Ł. (2022). Two-sample test for equal distributions in separate metric space: New maximum mean discrepancy based approaches. *Electronic Journal of Statistics*, *16*(2), 4090–4132.

Zolotarev, V. M. (1983). Probability metrics. *Teoriya Veroyatnostei i ee Primeneniya*, *28*(2), 264–287.

Zoppe, A. (1997). On uniqueness and symmetry of self-consistent points of univariate continuous distributions. *Journal of classification*, *14*(1), 147–158.