



INTERNATIONAL DOCTORAL
SCHOOL OF THE USC

Marcos
Matabuena

PhD Thesis

Contributions on metric spaces
with applications in
personalized medicine

Santiago de Compostela, 2022

Doctoral Programme in Information Technology Research



TESE DE DOUTORAMENTO

**CONTRIBUTIONS ON METRIC
SPACES WITH APPLICATIONS
IN PERSONALIZED MEDICINE**

Marcos Matabuena Rodríguez

ESCOLA DE DOUTORAMENTO INTERNACIONAL DA UNIVERSIDADE DE SANTIAGO DE
COMPOSTELA

PROGRAMA DE DOUTORAMENTO EN INVESTIGACIÓN EN TECNOLOXÍAS DA INFORMACIÓN

SANTIAGO DE COMPOSTELA
2022





Declaración do autor da tese

Don Marcos Matabuena Rodríguez

Título da tese: Contributions on metric spaces with applications in personalized medicine.

Presento a miña tese, seguindo o procedemento adecuado ao Regulamento, e declaro que:

- 1. A tese abarca os resultados da elaboración do meu traballo.*
- 2. De ser o caso, na tese faise referencia ás colaboracións que tivo este traballo.*
- 3. Confirmo que a tese non incorre en ningún tipo de plaxio doutros autores nin de traballos presentados por min para a obtención doutros títulos.*
- 4. A tese é a versión definitiva presentada para a súa defensa e coincide a versión impresa coa presentada en formato electrónico.*

E comprométome a presentar o Compromiso Documental de Supervisión no caso de que o orixinal non estea na Escola.

En Santiago de Compostela, 28 de Septiembre de 2022

Asdo. Marcos Matabuena Rodríguez





Autorización do director da tese

Contributions on metric spaces with applications in personalized medicine

Don Paulo Félix Lamas, Profesor Titular da Área de Ciencia da Computación e Intelixencia Artificial da Universidade de Santiago de Compostela

Dona Balbina Casas Méndez, Profesora Titular da Área de Estatística e Investigación Operativa da Universidade de Santiago de Compostela

INFORMAN:

*Que a presente tese correspóndese co traballo realizado por **Don Marcos Matabuena Rodríguez**, baixo a nosa dirección, e autorizamos a súa presentación, considerando que reúne os requisitos esixidos no Regulamento de Estudos de Doutoramento da USC, e que como directores desta non incorre nas causas de abstención establecidas na Lei 40/2015.*

De acordo co indicado no Regulamento de Estudos de Doutoramento, declara tamén que a presente tese de doutoramento é idónea para ser defendida en base á modalidade de Monográfica con reprodución de publicacións, nos que a participación do/a doutorando/a foi decisiva para a súa elaboración e as publicacións se axustan ao Plan de Investigación.

En Santiago de Compostela, 28 de Septiembre de 2022

Asdo. Paulo Félix Lamas
Director/a tese

Asdo. Balbina Casas Méndez
Director/a tese



He never explained anything, just posed problems, one of them, V. I. Arnold, once said. He didn't chew them over. He gave the student complete independence...

(about Andrei Kolmogorov)

Who reads much and walks much sees much and knows much

Miguel de Cervantes

It seems to me that a reasonable explanation was given by I. G. Petrovskii, who taught me in 1966: genuine mathematicians do not gang up, but the weak need gangs in order to survive. They can unite on various grounds (it could be super-abstractness, anti-Semitism or "applied and industrial" problems), but the essence is always a solution of the social problem - survival in conditions of more literate surroundings. By the way, I shall remind you of a warning of L. Pasteur: there never have been and never will be any "applied sciences", there are only applications of sciences (quite useful ones!).

Vladimir Arnold



Freedom is learned in school: freedom or slavery.

Emilio Lledó

It is impossible to be a mathematician without being a poet in soul.

Sofia Kovalevskaya

Acknowledgments

A doctoral thesis is a long road full of adventures, misfortunes, and sacrifices, where the student acts like a child searching for new ideas to explore and play with to improve his field of study or, more importantly, to understand how the world works and build a better society.

In this particular enterprise, driven mainly by a utopian impetus, I should first like to acknowledge the work done by my high school teacher José Rodríguez for being the best teacher I've had in my career so far. José Rodríguez provided me with the formal basis to be autonomous, have critical thinking, and ultimately be free, which is the ultimate goal of education.

I would like to thank my thesis advisors, Paulo Félix and Balbina Casas Méndez, for the independence provided, their many comments and corrections, and all the support they gave me. They have been listening to me since our first meeting in which we discussed ambitious projects, and nobody believed that a young student would be able to develop them.

What would life be without diversity?

Apart from talent, creativity, and luck, the success of a scientist depends to a large extent on possessing high levels of resilience. If I own this quality, it is thanks to my long experience in the world of sports that was crucial in building me as a tenacious, ambitious, honest, courageous, and dreamer. In this respect, meeting Juan Manuel Díaz García, my athletics coach, was vital in my life. He taught me to think big, believe in myself, and not give up in the face of adversity. For all this, thank you very much, Juan. I would also like to thank Dr. Juan José García Cota (the doctor that discovered my injury problems) for demonstrating what it means to be a true professional with integrity and the importance of honesty and values in life.

An essential point of my doctoral thesis and my short scientific career is extensive communication with researchers from different areas. These communications allowed me to see the world from another point of view, to understand which topics we should investigate in order to be beneficial to society, and most importantly, to understand the true genius and humility of the great masters as well as the affection and enthusiasm with which ambitious

young researchers should be treated. I will never forget a spontaneous e-mail from Arnold Jansen sending me additional references for my extension of the energy distance in survival analysis. To Prof. Winfried Stute, for his teachings on what questions we should address in research, the importance of proper education, and the pursuit of happiness in what one does. To Dr. Fernando Huelin Trillo for all the facilities and for believing in me to develop high-level research with data from the high-performance center of Pontevedra. To Prof. Phil Hayes for believing in my unusual ideas for changing physical activity analysis and for trusting me to guide them in the statistical part to his doctorates like Sherveen Riazziati. To the great mathematicians Prof. Russell Lyons and Prof. Gabor Szekely for answering all my questions and the former for providing feedback on one of my works, despite my limited English at the beginning of my doctoral thesis. To Prof. Oscar Hernan Madrid Padilla and Prof. Alex Petersen, my initial collaborators, for believing in an inexperienced person on the other side of the world and helping me to pursue my goals. To Prof. Borja Cruz, my faithful collaborator in physical activity, for thinking we could do something big using my distributional representations of accelerometer data. To Prof. Dino Sejdinovic for helping me, collaborating on kernel methods research, and proving to be a true master in this area. To Prof. Robert Wagner and Prof. Thomas Lumley for helping me with my diabetes projects and broadening my biomedical and biostatistical research horizon. To my Spanish colleagues in Germany, Prof. Enrique Zuazua and Prof. Celia Martín Vicario, the first one for inviting me to give a seminar at his university, being a reference for young researchers, and for answering so kindly my doubts; and the second one for helping me to develop my idea of a diabetes risk score. To Prof. Jorge Mota and Prof. Pedro Abdalla, my collaborators during my stay in Oporto, for their friendship and kindness, which made me spend one of the happiest moments of my life in Portugal.

To Prof. Gábor Lugosi for being my first great mathematical teacher in person and for his humility in building a wonderful project on the inference of uncertainty quantification in metric spaces. To Jean Pauphilet for introducing me to mathematical optimization. To Jesús Cortés for discussing multiple questions on neuroimaging problems.

To Prof. Ciprian Crainiceanu and Marta Karas for their kindness; they believed in me and increased my vision and knowledge about multilevel models.

I could cite many more international colleagues, but I would like to thank them all for the sake of brevity.

At the local level, someone who played an essential role in my life was Dr. Francisco Gude. He hired me and trusted me to carry out the exciting challenges of AEGIS diabetes study. Moreover, he never doubted my judgment, providing me with complete independence.

I would also like to thank Carlos Meixide, my first student who always believed in my criteria and in the idea that young people must be encouraged and we should be brave with our beliefs.

Finally, thanks to my family for all their support, especially my parents and grandmother, for loving me as I am and encouraging me to never give up, despite the many adversities of the past few years, both academic and personal.

To everyone who helped me, thank you, Yes, we can!. My adventure is only the beginning of a long way: The Ithaka Travel.

As you set out for Ithaka hope your road is a long one, full of adventure, full of discovery. Laistrygonians, Cyclops, angry Poseidon—don't be afraid of them: you'll never find things like that on your way as long as you keep your thoughts raised high, as long as a rare excitement stirs your spirit and your body.

Laistrygonians, Cyclops, wild Poseidon—you won't encounter them unless you bring them along inside your soul, unless your soul sets them up in front of you.

Hope your road is a long one. May there be many summer mornings when, with what pleasure, what joy, you enter harbors you're seeing for the first time; may you stop at Phoenician trading stations to buy fine things, mother of pearl and coral, amber and ebony, sensual perfume of every kind—as many sensual perfumes as you can; and may you visit many Egyptian cities to learn and go on learning from their scholars. Keep Ithaka always in your mind. Arriving there is what you're destined for. But don't hurry the journey at all. Better if it lasts for years, so you're old by the time you reach the island, wealthy with all you've gained on the way, not expecting Ithaka to make you rich.

Ithaka gave you the marvelous journey. Without her you wouldn't have set out. She has nothing left to give you now.

And if you find her poor, Ithaka won't have fooled you. Wise as you will have become, so full of experience, you'll have understood by then what these Ithakas mean.

Konstantinos Kavafis

Contents

I Preliminaries	1
Notation	3
Resumen	5
Resumo	13
1 Introduction and aims	21
2 Statistical analysis in metric spaces and reproducing kernel Hilbert spaces	27
2.1 Complex data in medicine	28
2.2 Statistical analysis in metric spaces	30
2.3 Statistical learning in reproducing kernel Hilbert spaces	35
II Distributional representations from wearable and biosensor technology	43
3 Distributional representations of biosensor data for continuous stochastic process with application in CGM technology	45
3.1 Sample and procedures	48
3.2 Definition and estimation of the glucodensity	49
3.3 Regression models with glucodensities	52
3.4 Clinical validation of the glucodensity	54
3.5 Hypothesis testing and clustering analysis with glucodensities	58
3.6 Discussion	62
4 Distributional representations of biosensor data for mixed stochastic processes with application in physical activity analysis	69
4.1 NHANES 2003-2006 dataset	71
4.2 Functional representation of accelerometer data and regression models	74

4.3	Results	82
4.4	Discussion	85
4.5	Application in NHANES 2011-2014: Discovering clinical physical activity phenotypes in the U.S. population	87
III Missing data modeling from complex statistical objects		97
5	Statistical independence, variable selection and conformal inference with missing responses in long-term glucose modeling using distributional representations	99
5.1	Data analysis outline	100
5.2	Methods	101
5.3	An application in modelling long-term changes in glucose levels	109
5.4	Results	111
5.5	Discussion	115
6	Hypothesis testing in the presence of complex paired missing data by maximum mean discrepancy: An application to continuous glucose monitoring	119
6.1	Motivation from glucodensity representation	120
6.2	Hypotheses and statistics	121
6.3	Illustrative data analysis	127
6.4	Discussion	130
IV Conclusions		133
7	Conclusions	135
7.1	New opportunities	138
7.2	Future work	139
V Appendices		141
A	Theory of U and V-statistics	143
A.1	The notions of U-and V-statistics to derive limit distributions	143
B	Proofs: Chapters 5 and 6	147
B.1	Chapter 5	147
B.2	Chapter 6	153

C R package: Biosensors.usc	155
C.1 Installation	155
Bibliography	165
Publications	191
List of Figures	195
List of Tables	199

Part I

Preliminaries

Notation

\mathcal{X}, \mathcal{Y}	set of covariates, set of responses
(\mathcal{Y}, d)	metric space with set \mathcal{Y} equipped with distance d
f, g	density functions f, g
F, G	distribution functions F, G
F^{-1}, G^{-1}	quantile functions from F, G
$d_{W_2}(f, g)$	2-Wasserstein metric
$P_X, P_Y, P_{X,Y}$	probability distributions over X, Y and (X, Y)
E	expectation operator
$P(\mathcal{X})$	set of distribution functions over \mathcal{X}
$\mathcal{W}_p(\mathcal{X})$	p -Wasserstein space
$m(\cdot)$	regression function
\mathcal{F}	class of functions
\mathcal{H}	reproducing kernel Hilbert space (RKHS)
$k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$	kernel function over \mathcal{X}
$\rho(\cdot, \cdot)$	semi-metric negative type
$F(\mathcal{X}) := \{f : \mathcal{X} \rightarrow \mathbb{R}\}$	set of real functions whose domain is \mathcal{X}
$\langle \cdot, \cdot \rangle$	dot product
$\mathcal{M}_f(\mathcal{X})$	set of finite Borel measures on \mathcal{X}
$C(\mathcal{X}) := \{f : \mathcal{X} \rightarrow \mathbb{R}; f \text{ is continuous}\}$	set of continuous real functions whose domain is \mathcal{X}
$\phi : \mathcal{M}_f(\mathcal{X}) \rightarrow \mathcal{H}, \phi(F)$	Evaluation of kernel mean embedding in an arbitrary distribution F
\mathbb{K}	Kernel of Nadaraya-Watson estimator
h	smoothing parameter Nadaraya-Watson estimator
σ	smoothing parameter Gaussian and Laplacian kernel
δ	missing data indicator
$\pi(\cdot) = P(\delta = 1 X = \cdot)$	missing data mechanism under missing completely at random (MCAR)
w_i^*	standardized inverse probability weighting estimator
$\tilde{\cdot}$	empirical estimator

Resumen

En la última década se han producido avances reseñables en la transducción de señales biológicas y el posterior reconocimiento de eventos de interés fisiopatológico [153]. Los progresos en esta área han estimulado la aparición de una prometedora y robusta tecnología de biosensores a partir de enfoques no invasivos y de bajo coste [274]. La adopción de esta tecnología en la práctica clínica ha de conducir necesariamente al desarrollo de nuevos métodos para obtener información más sofisticada sobre los cambios en la salud de los pacientes en tiempo real [153, 274]. Un ejemplo paradigmático lo encontramos en los sistemas de lazo cerrado para la dispensación de insulina mediante una bomba que se ajusta dinámicamente según la lectura continua de la glucosa, que actualmente está revolucionando la gestión de pacientes con diabetes tipo 1 [135, 229].

Con la ganancia de la información registrada sobre los pacientes y una mayor disponibilidad de estos sensores entre la población general, el desarrollo de modelos estadísticos y de aprendizaje automático es el siguiente gran paso para establecer las bases metodológicas de los nuevos paradigmas clínicos de la medicina digital y de precisión [93, 138, 139, 153, 162, 298].

La medicina de precisión pretende optimizar la calidad de la atención sanitaria individualizando el proceso asistencial de acuerdo a los cambios de la condición clínica del paciente a lo largo del tiempo [138]. En este nuevo paradigma las decisiones se apoyan fundamentalmente en enfoques guiados por datos. Matemáticamente, la medicina de precisión se formaliza como un problema de optimización dinámica. Las acciones potenciales podrían ser: la selección del fármaco a utilizar, la selección de la dosis, el momento de la administración, la recomendación de una dieta o un ejercicio específico [138].

Desafortunadamente, la práctica generalizada en la actualidad es que los tratamientos no se prescriben de forma personalizada y se centran en optimizar la salud del individuo medio de la población de estudio. En consecuencia, el efecto de los tratamientos resulta subóptimo a lo largo de una amplia lista de enfermedades y fenotipos de pacientes [27, 28, 240, 243, 266].

Indudablemente, la prescripción del tratamiento óptimo es el objetivo crucial de la medicina personalizada [138], pero para alcanzarlo se requiere de la modelización de una amplia lista de problemas previos [139]. De acuerdo con las contribuciones presentadas aquí, algunas de las nuevas direcciones de investigación tienen como objeto el diseñar nuevas representaciones

de los datos proporcionados por biosensores y dispositivos vestibles que nos permitan registrar los cambios en la salud de los pacientes con mayor precisión y capturar más información para alimentar diferentes algoritmos predictivos con más exactitud. En el contexto de la medicina digital, centrada en la explotación de la información proporcionada por dispositivos electrónicos de medida, cada vez resulta más común evaluar el impacto del tratamiento en el estado del paciente mediante biomarcadores digitales (definidos a partir de métricas derivadas de la monitorización de distintas variables ambientales, biomecánicas o fisiológicas [153]), presentando a menudo una naturaleza compleja, ya sea en forma de perfiles funcionales para representar la variabilidad de la frecuencia cardíaca [196], el gasto energético [177], o la concentración de glucosa [178], o bien como grafos de conectividad cerebral, que contienen información fundamental acerca de la estructura y topología de los distintos patrones de actividad neuronal [289]. En este contexto, resulta fundamental diseñar nuevas pruebas de hipótesis como medida confirmatoria de la eficacia de un fármaco, por ejemplo, en el marco de un ensayo clínico aleatorizado [49, 51, 57, 125, 139, 140].

Motivado por el progreso en el campo de la medicina de precisión con los elementos de la medicina digital, el principal propósito de esta tesis es desarrollar nuevas metodologías estadísticas y de aprendizaje automático para analizar objetos estadísticos complejos [62] como son los datos funcionales euclidianos, o construcciones más generales en espacios métricos, como las distribuciones de probabilidad, que viven en espacios no lineales sin estructura de espacio vectorial [178, 215].

En primer lugar abordamos el reto de analizar series temporales biológicas en pacientes que se encuentran monitorizados fuera del entorno hospitalario, durante aquellas actividades que forman parte de su rutina diaria. Este reto presenta una dificultad de partida al abordar un análisis global comparativo de las series de tiempo registradas, debido a que los pacientes son monitorizados en periodos de distinta duración, siguen diferentes estilos de vida, o existen importantes diferencias cronobiológicas entre ellos. Para solucionar dicho problema, encontramos en la bibliografía distintas representaciones que extraen algunas características resumen de la serie, en forma de valores promedio o de indicadores de la variabilidad [141, 187]. Así todo, a día de hoy las métricas vectoriales de naturaleza composicional constituyen el estándar de facto utilizado para resumir la información de los datos de los biosensores en múltiples dominios, como en la diabetes a partir del uso de monitores continuos de la glucosa, o en la actividad física mediante dispositivos de acelerometría [31, 64, 177, 178]. Estas métricas cuantifican la proporción de tiempo que el paciente se encuentra en cada una de las zonas objetivo o en categorías definidas de antemano. Precisamente, su gran simplicidad e interpretabilidad explican su éxito entre los usuarios [187].

No obstante, uno de los mayores inconvenientes del uso de las métricas vectoriales de naturaleza composicional en la práctica es que no podemos cuantificar el impacto en la salud

del tiempo que permanece el paciente sobre un continuo de valores registrados por el dispositivo sensor. En su lugar, estas métricas proporcionan un resumen de la información de las series temporales en un conjunto reducido de intervalos, con la consiguiente pérdida de información. Aparte de esta importante desventaja, resumir la información en intervalos presenta otros importantes inconvenientes. Necesitamos cierto conocimiento experto para definir los puntos de corte, que pueden depender de la tarea de modelización específica y de la población de estudio a analizar. Precisamente, este aspecto se viene cuestionando insistentemente en las últimas investigaciones sobre la actividad física con acelerómetros [187].

En esta tesis, y con el fin de superar estos inconvenientes proponemos usar la extensión natural a nivel funcional de las métricas vectoriales de naturaleza composicional, que permite capturar más información. La idea de partida es considerar la distribución de probabilidad o la función de densidad de los diferentes segmentos de la serie temporal biológica registrada [178]. Una de las ventajas de esta nueva representación es que no requiere la categorización de la información en intervalos definidos a partir de conocimiento experto.

Desde un punto de vista formal, proponemos una representación distribucional cuyo análisis deriva del análisis de datos composicionales, pero según una perspectiva funcional [119, 265]. A este respecto, el análisis de datos con esta representación presenta algunos retos importantes. Por ejemplo, dado que los datos composicionales viven en un simplex de probabilidad, no existe una estructura de espacio vectorial (a menos que nos restrinjamos a las combinaciones convexas), lo que implica la necesidad de aplicar técnicas alternativas de análisis de datos para transformar el espacio de entrada no lineal en un espacio lineal más apropiado [215]. El proceso se puede resumir como sigue: 1) transformación composicional a un espacio vectorial euclidiano; y 2) aplicación de modelos de regresión euclidianos. Sin embargo, esto presenta limitaciones sustanciales para realizar tareas inferenciales precisas y para construir métodos matemáticos con garantías teóricas con muestras finitas. Además, las transformaciones composicionales clásicas u otros enfoques más recientes no están bien definidos cuando hay ceros en alguna de las partes de la representación. En la extensión funcional este problema se agrava debido a que el soporte de la distribución de probabilidad varía entre los distintos individuos que son objeto de monitorización.

Para superar esta limitación, proporcionamos un nuevo marco de análisis de datos funcionales composicionales basado en el enfoque de la teoría del transporte óptimo, junto al análisis estadístico en espacios métricos, concretamente en espacios de Hilbert de núcleo reproductor (*reproducing kernel Hilbert spaces - RKHS*). De forma más específica, proponemos un conjunto de núcleos con geometría de la distancia 2–Wasserstein, que es fácilmente computable en el caso particular de distribuciones de probabilidad unidimensionales. Este marco de análisis permite construir modelos de regresión, métodos de selección de variables, métodos de cuantificación de la incertidumbre, y test de hipótesis. Abordamos el contexto usual de datos

independientes e idénticamente distribuidos, así como otros más complejos como el de los diseños de encuesta, o de datos perdidos, habituales estos últimos en los estudios longitudinales cuando el paciente renuncia a continuar su participación.

La eficacia real de las representaciones distribucionales propuestas se evaluarán en los ámbitos de la diabetes y la actividad física. Para ello, utilizaremos los datos de un estudio longitudinal sobre la diabetes: el Estudio de Glicación e Inflamación de A Estrada (AEGIS) [101], en el que se dispone para cada paciente de información sobre la monitorización continua de la glucosa (*continuous glucose monitoring - CGM*) durante aproximadamente una semana. Con esta información de alta resolución se pretende responder a algunas preguntas de interés clínico sobre el impacto que determinadas representaciones distribucionales tienen sobre la capacidad de predicción de los valores de glucosa a largo plazo. En el caso de la actividad física, utilizamos datos de sucesivas encuestas NHANES (*National Health and Nutrition Examination Survey*) en EE.UU., que disponen de información de los hábitos de actividad física de la población americana medidos de forma objetiva mediante dispositivos de acelerometría. En este caso, ilustraremos el uso y las ventajas de las representaciones distribucionales en este contexto para analizar la relación entre la actividad física, la mortalidad y supervivencia de los participantes en todo el espectro de intensidades de actividad física registradas por el acelerómetro. En los distintos análisis realizados proporcionaremos evidencias sólidas sobre los beneficios de usar estas representaciones en términos de capacidad de predicción y sensibilidad clínica en los dominios mencionados.

Aquí conviene recordar que la diabetes es probablemente la pandemia más importante de este siglo [120, 256] con unas tasas de prevalencia alarmantes que se espera que aumenten en los próximos años, debido principalmente al empeoramiento de los estilos de vida, motivado tanto por un incremento de los niveles de inactividad como por el empobrecimiento en la calidad de las dietas en todo el mundo. Diversas organizaciones y especialistas reclaman nuevas políticas de salud pública para mejorar el control y la propagación de la enfermedad. Los enfoques de telemedicina, medicina digital y de precisión [45, 145] pueden ser un paso adelante e inspirar nuevas soluciones de gran efectividad para revertir este problema. En esta dirección, algunos trabajos recientes han planteado la hipótesis de que la inclusión de biomarcadores digitales procedentes de la motorización continua de la glucosa puede mejorar la detección temprana de la enfermedad [105]. Hace diez años la motorización continua de la glucosa ya supuso un gran avance en la transformación de la gestión de la diabetes [135, 229]. Hoy en día, la adopción de estas tecnologías encuentra más aplicaciones, tanto en poblaciones sanas como enfermas. Por ejemplo, los dispositivos de monitorización continua de la glucosa se utilizan en la nutrición de precisión para personalizar las dietas en función de cómo la ingesta de alimentos específicos modifica nuestros valores de glucosa [23].

En cuanto a la importancia de analizar los datos de actividad física, destacaremos que hoy

en día la actividad física se considera la intervención no farmacológica de bajo coste más eficaz para combatir un amplio espectro de enfermedades, como la diabetes o el síndrome metabólico [208]. La actividad física es también una intervención efectiva para retrasar el deterioro cognitivo y funcional asociado al envejecimiento [231]. Por todo ello, en la actualidad, hay un importante consenso para promover el deporte como un elemento esencial en la construcción de sociedades modernas saludables y la digitalización de la práctica deportiva puede ser un punto esencial para alcanzar dicho objetivo. La adopción de enfoques de medicina personalizada en esta área de conocimiento es una tarea pendiente [233], y esta tesis introduce nuevas aportaciones en esta dirección.

A continuación delimitamos con mayor precisión los objetivos y contenidos de cada uno de los capítulos que componen esta tesis doctoral.

El capítulo 2 tiene como objetivo introducir las herramientas necesarias para entender el formalismo del análisis estadístico de objetos complejos que toman valores en espacios métricos. Asimismo, motivaremos el uso de estas herramientas de análisis mediante algunos ejemplos de interés por sus múltiples aplicaciones en la medicina contemporánea. Primero comenzaremos introduciendo la noción de media y varianza de Fréchet que generaliza la noción de centro y dispersión usual en el contexto de los espacios métricos [82]. Introduciremos algunas propiedades básicas del problema de estimación, que puede ser escrito en forma de un problema de M-estimación [277] y, en consecuencia, permite usar la teoría existente para derivar la teoría límite en algunos casos usando las herramientas clásicas de la teoría de procesos empíricos [220]. De acuerdo a estas ideas introducimos la noción de modelo de regresión lineal en espacios métricos [214], de nuevo expresando el problema como un problema de M-estimación, y haciendo uso de la teoría asociada a esta familia general de estimadores. A continuación, puesto que el objetivo de esta tesis es ilustrar los métodos en el caso particular de las representaciones distribucionales con la geometría inducida por la distancia 2-Wasserstein, y esta construcción está motivada por el problema del transporte óptimo de distribuciones de probabilidad [205], introduciremos las métricas de Wasserstein. A continuación, en tanto que planteamos realizar las distintas tareas de modelado bajo el paraguas de los espacios de Hilbert con núcleo reproductor (RKHS) haremos una breve introducción a los elementos centrales de este paradigma, que permite analizar e integrar datos de naturaleza compleja. Además, introduciremos la noción de *kernel mean embedding* [191], que sirve como base formal para definir distancias estadísticas entre las representaciones distribucionales y otros objetos estadísticos complejos, y que nos servirá en capítulos posteriores para definir test estadísticos de igualdad de distribución, test de independencia estadística o realizar análisis clúster. El capítulo 3 tiene como objetivo introducir desde un punto de vista formal las nuevas representaciones distribucionales para aquellas series temporales que son generadas por un proceso estocástico en tiempo continuo, como es el caso de un medidor de CGM, que mide los valores de la glucosa

intersticial de forma dinámica y cada pocos minutos. Asimismo, proporcionaremos evidencias de las ventajas de la nueva representación frente a las métricas existentes de resumen de la información de los datos de CGM, como las métricas composicionales del tiempo en rango, medidas de variabilidad glucémica u otros biomarcadores clínicos existentes. Finalmente, como las nuevas representaciones distribucionales generan datos funcionales composicionales que no se pueden analizar con las técnicas usuales de datos funcionales, proporcionaremos una serie de técnicas de análisis estadístico en espacios métricos para analizar distintos problemas que pueden aparecer en la clínica, como tests de hipótesis para evaluar la efectividad de un tratamiento, análisis de regresión para predecir la evolución del paciente, o análisis clúster para descubrir nuevos fenotipos clínicos.

De forma similar, el capítulo 4 tiene como objeto introducir formalmente y validar las nuevas representaciones distribucionales, pero en el caso en que el proceso estocástico es mixto, como ocurre con los datos de acelerómetros. Aquí la validación se realiza utilizando datos del estudio NHANES, que provienen de un diseño experimental de encuesta. Extendemos algunos métodos predictivos previos a datos complejos, para poder analizar las ventajas potenciales de las nuevas representaciones, como el método *kernel ridge* para la regresión, y utilizamos el estimador Nadaraya-Watson para la clasificación en espacios métricos. Al final de este capítulo proponemos como aplicación en la población mayor la identificación de nuevos fenotipos clínicos de la actividad física, con importantes implicaciones en la práctica para monitorizar la actividad física y como variables de seguimiento de la supervivencia y pronóstico de los pacientes.

En el capítulo 5 proponemos nuevos métodos de aprendizaje estadístico supervisado para problemas con datos perdidos en la variable respuesta, basados en métodos previos del paradigma RKHS. Los nuevos métodos incluyen algoritmos de cuantificación de la incertidumbre, medidas de dependencia estadística o algoritmos de selección de variables. La tarea de modelización consiste en predecir cómo cambian los valores de glucosa a cinco años en términos del biomarcador A1c, estándar *de facto* actual para el control y diagnóstico de la enfermedad. Los resultados obtenidos en este capítulo son una prueba de que la inclusión de la monitorización continua de la glucosa conduce a una mejora en la predicción de la condición glucémica del paciente a largo plazo.

En el capítulo 6, y motivados por la necesidad de comparar las posibles variaciones en las representaciones distribucionales en dos instantes de tiempo, por ejemplo, antes y después de administrar un determinado tratamiento, proponemos nuevas pruebas de hipótesis para datos emparejados y perdidos con objetos estadísticos complejos, mediante métodos núcleo. Validaremos estos métodos, diseñados con carácter general, mediante un análisis de la evolución de la glucosa a cinco años en distintos subgrupos de pacientes con los datos del estudio AEGIS.

Por último, en el capítulo 7, de Conclusiones, discutimos los avances metodológicos de

la tesis en el modelado de objetos estadísticos complejos y, más importante, su potencial clínico en la nueva era de la medicina digital y de precisión. También discutimos los diversos trabajos activos motivados por esta tesis doctoral. Al final de este capítulo presentamos nuevos problemas abiertos que hemos identificado y que consideramos abordar en el futuro, para proporcionar a los usuarios nuevas herramientas de análisis de datos con aplicaciones en la ciencia médica.

Se acompaña la tesis de un conjunto de Apéndices en los que escribimos formalmente las pruebas matemáticas que, por espacio y claridad, se han separado del resto del material, como es el caso de la consistencia de la estrategia *bootstrap* introducida en el capítulo 5 y el estadístico de prueba introducido en el capítulo 6. También proporcionamos una pequeña guía de uso del paquete software desarrollado para el entorno R, *biosensor.usc*, con el objetivo de ilustrar cómo se pueden utilizar las técnicas aquí propuestas. Por último, como las distintas pruebas introducidas en esta tesis dependen de la teoría de U -estadísticos introducimos algunos resultados básicos de esta teoría [137, 244].

Resumo

Na última década producíronse importantes avances na transdución de sinais biolóxicos e o posterior recoñecemento de eventos de interese fisiopatolóxico [153]. Os progresos nesta área estimularon a aparición dunha prometedora e robusta tecnoloxía de biosensores a partires de enfoques non invasivos e de baixo custo [274]. A adopción desta tecnoloxía na práctica clínica impulsará o desenvolvemento de novos métodos para obter información máis sofisticada sobre os cambios na saúde dos pacientes en tempo real [153, 274]. Un exemplo paradigmático atopámolo nos sistemas de lazo pechado para a dispensación de insulina mediante unha bomba que se axusta dinámicamente segundo a medición continua dos valores de glicosa, que está a revolucionar a xestión de pacientes con diabetes tipo 1 [135, 229].

Coa ganancia da información rexistrada sobre os doentes e unha maior dispoñibilidade destes sensores entre a poboación xeral, o desenvolvemento de modelos estatísticos e de aprendizaxe automática é o seguinte gran paso para establecer as bases metodolóxicas dos novos paradigmas clínicos da medicina dixital e de precisión [93, 138, 139, 153, 162, 298].

A medicina de precisión pretende optimizar a calidade da atención sanitaria individualizando o proceso asistencial de acordo aos cambios da condición clínica do doente ao longo do tempo [138]. Neste novo paradigma clínico as decisións apóianse fundamentalmente en enfoques guiados por datos. Matematicamente, a medicina de precisión formalízase como un problema de optimización dinámica. As accións potenciais poderían ser: a selección do fármaco a empregar, a selección da dose, o momento da administración, a recomendación dunha dieta ou a realización dun exercicio físico específico [138].

Desafortunadamente, a práctica xeneralizada na actualidade é que os tratamentos non se prescriben de forma personalizada e céntranse en optimizar a saúde do individuo medio da poboación de estudo. En consecuencia, o efecto dos tratamentos resulta subóptimo ao longo dunha ampla lista de enfermidades e fenotipos de doentes [27, 28, 240, 243, 266].

Indubidablemente, a prescrición do tratamento óptimo é o obxectivo crucial da medicina personalizada [138], pero para alcanzalo requírese da modelización dunha ampla lista de problemas previos [139]. De acordo coas contribucións presentadas aquí, algunhas das novas direccións de investigación teñen como obxecto deseñar novas representacións dos datos proporcionados por biosensores e dispositivos vestibles que nos permitan rexistrar os cam-

bios na saúde dos doentes con maior precisión e capturar máis información para alimentar diferentes algoritmos predictivos con máis exactitude. No contexto da medicina dixital, centrada na explotación da información proporcionada por dispositivos electrónicos de medición, cada vez resulta máis común avaliar o impacto do tratamento no estado do doente mediante biomarcadores dixitais (definidos a partir de métricas derivadas da monitorización de distintas variables ambientais, biomecánicas ou fisiolóxicas [153]), presentando a miúdo unha natureza complexa, xa sexa en forma de perfís funcionais para representar a variabilidade da frecuencia cardíaca [196], o gasto enerxético [177], ou a concentración de glicosa [178], ou ben como grafos de conectividade cerebral, que conteñen información fundamental acerca da estrutura e topoloxía dos distintos patróns de actividade neuronal [289]. Neste contexto, resulta fundamental deseñar novas probas de hipóteses como medida confirmatoria da eficacia dun fármaco, por exemplo, no marco dun ensaio clínico aleatorizado [49, 51, 57, 125, 139, 140].

Motivado polo progreso no campo da medicina de precisión cos elementos da medicina dixital, o principal propósito desta tese é desenvolver novas metodoloxías estatísticas e de aprendizaxe automática para analizar obxectos estatísticos complexos [62], como son os datos funcionais euclidianos, ou construcións máis xerais en espazos métricos, como as distribucións de probabilidade, que viven en espazos non lineais sen estrutura de espazo vectorial [178, 215].

Nesta tese, abordamos en primeiro lugar o reto de analizar series temporais biolóxicas en doentes que se atopan monitorizados fóra da contorna hospitalaria, durante aquelas actividades que forman parte da súa rutina diaria. Este reto presenta dificultades para levarse a cabo cunha análise global comparativa das series de tempo rexistradas, debido a que os doentes son monitorizados en períodos de distinta duración, seguen diferentes estilos de vida, ou existen importantes diferenzas cronobiolóxicas entre eles. Para solucionar o devandito problema, atopamos na bibliografía distintas representacións que extraen características resumo da serie temporal, en forma de valores promedio ou de indicadores da variabilidade [141, 187]. Así todo, a día de hoxe as métricas vectoriais de natureza composicional constitúen o estándar *de facto* para resumir a información dos datos biosensores en múltiples dominios, como na diabetes mellitus, a partir do uso de monitores continuos da glicosa, ou na actividade física mediante dispositivos de acelerometría [31, 64, 177, 178]. Estas métricas cuantifican a proporción de tempo que o doente se atopa en cada unha das zonas obxectivo, ou noutras categorías definidas de antemán. Precisamente, a súa gran simplicidade e interpretabilidade explican o seu éxito entre os usuarios [187].

Porén, un dos maiores inconvintes no uso das métricas vectoriais de natureza composicional é que non podemos cuantificar o impacto na saúde do tempo que permanece o doente nun continuo de valores rexistrados polo dispositivo de monitorización. No seu lugar, estas métricas proporcionan un resumo da información das series temporais nun conxunto reducido de intervalos, coa conseguinte perda de información. Á parte desta importante desvantaxe,

resumir a información en intervalos presenta outros importantes inconvenientes. Necesitamos certo coñecemento experto para definir os puntos de corte, que poden depender da tarefa de modelización específica e da poboación de estudo a analizar. Precisamente, este aspecto vén sido cuestionado insistentemente nas últimas investigacións sobre a actividade física con acelerómetros [187].

Nesta tese, e co fin de superar estes inconvenientes propoñemos usar a extensión natural a nivel funcional das métricas vectoriais de natureza composicional, que permiten capturar máis información. A idea de partida é considerar a distribución de probabilidade ou a función de densidade dos diferentes segmentos da serie temporal biolóxica rexistrada [178]. Unha das vantaxes desta nova representación é que non require a categorización da información en intervalos definidos a partir do coñecemento experto.

Desde un punto de vista formal, propoñemos unha representación distribucional cuxa análise deriva da análise de datos composicionais, pero de acordo a unha perspectiva funcional [119, 265]. A este respecto, a análise de datos con esta representación presenta algúns retos importantes. Por exemplo, dado que os datos composicionais viven nun simplex de probabilidade, non existe unha estrutura de espazo vectorial (a menos que nos restrinxamos ás combinacións convexas), o que implica a necesidade de aplicar técnicas alternativas de análise de datos para transformar o espazo de entrada non lineal a un espazo lineal máis apropiado [215]. O proceso pódese resumir como segue: 1) transformación composicional a un espazo vectorial euclidiano; e 2) aplicación de modelos de regresión euclidianos. Con todo, esta metodoloxía presenta limitacións substanciais para realizar tarefas inferenciais precisas e para construír métodos matemáticos con garantías teóricas con mostras finitas. Ademais, as transformacións composicionais clásicas ou outros enfoques máis recentes non están ben definidos cando hai ceros nalgunha das partes da representación. Na extensión funcional este problema agrávase debido a que o soporte da distribución de probabilidade varía entre os distintos individuos que son obxecto de monitorización.

Para superar esta limitación, proporcionamos un novo marco de análise de datos funcionais composicionais baseado no enfoque da teoría do transporte óptimo, xunto á análise estatística en espazos métricos, concretamente en espazos de Hilbert de núcleo reprodutor (*reproducing kernel Hilbert spaces - RKHS*). De forma máis específica, propoñemos un conxunto de núcleos equipados coa xeometría da distancia 2–Wasserstein, que é facilmente computable no caso particular de distribucións de probabilidade unidimensionais. Este marco de análise permite construír modelos de regresión, métodos de selección de variables, métodos de cuantificación da incerteza, e test de hipóteses. Abordaremos o contexto usual de datos independentes e idénticamente distribuídos, así como outros máis complexos como o dos deseños de enquisa, ou de datos perdidos, habituais estes últimos nos estudos lonxitudinais cando o doente renuncia a continuar a súa participación.

A eficacia real das representacións distribucionais avaliarase nos eidos da diabetes e a actividade física. Para iso, utilizaremos os datos dun estudo lonxitudinal sobre a diabetes: o Estudo de Glicación e Inflamación da Estrada (AEGIS) [101], no que se dispón para cada doente de información sobre a monitorización continua da glicosa (*continuous glucose monitoring- CGM*) durante aproximadamente unha semana. Con esta información de alta resolución preténdese responder a algunhas preguntas de interese clínico sobre o impacto que determinadas representacións distribucionais teñen sobre a capacidade de predición dos valores de glicosa a longo prazo. No caso da actividade física, empregamos datos das enquisas NHANES (*National Health and Nutrition Examination Survey*) realizadas en EE.UU., que dispoñen de información dos hábitos de actividade física da poboación americana medidos de forma obxectiva mediante dispositivos de acelerometría. Neste caso, ilustraremos o uso e as vantaxes das representacións distribucionais neste contexto para analizar a relación entre a actividade física, a mortalidade e supervivencia dos participantes en todo o espectro de intensidades de actividade física rexistradas polo acelerómetro. Nas distintas análises realizadas proporcionaremos evidencias sólidas sobre os beneficios de usar estas representacións en termos de capacidade de predición e sensibilidade clínica nos dominios mencionados.

Cumpre salientar aquí que a diabetes é probablemente a pandemia máis importante deste século [120, 256] cunhas taxas de prevalencia alarmantes que se espera que aumenten nos vindeiros anos, debido principalmente ao empeoramento dos estilos de vida, motivado tanto por un incremento dos niveis de inactividade como polo empobrecemento na calidade das dietas en todo o mundo. Diversas organizacións e especialistas reclaman novas políticas de saúde pública para mellorar o control e a propagación da enfermidade. Os enfoques de telemedicina, medicina dixital e de precisión [45, 145] poden ser un paso adiante e inspirar novas solucións de gran efectividade para revertir este problema. Nesta dirección, algúns traballos recentes expuxeron a hipótese de que a inclusión de biomarcadores dixitais procedentes da motorización continua da glicosa pode mellorar a detección temperá da enfermidade [105]. Fai dez anos a monitorización continua da glicosa xa supuxo un gran avance na transformación da xestión da diabetes [135, 229]. Hoxe en día, a adopción destas tecnoloxías atopa máis aplicacións, tanto en poboacións sas como enfermas. Por exemplo, os dispositivos de monitorización continua da glicosa utilízanse na nutrición de precisión para personalizar as dietas en función de como a inxesta de alimentos específicos modifica os nosos valores de glicosa [23].

En canto á importancia de analizar os datos de actividade física, destacamos que hoxe en día a actividade física considérase a intervención non farmacolóxica de baixo custo máis eficaz para combater un amplo espectro de enfermidades, como a diabetes ou a síndrome metabólico [208]. A actividade física é tamén unha intervención efectiva para atrasar a deterioración cognitiva e funcional asociada ao envellecemento [231]. Por todo iso, na actualidade, hai un importante consenso para promover o deporte como un elemento esencial na construción de sociedades

modernas saudables e a dixitalización da práctica deportiva pode ser un punto esencial para alcanzar o devandito obxectivo. A adopción de enfoques de medicina personalizada nesta área de coñecemento é unha tarefa pendente [233], e esta tese introduce novas achegas nesta dirección.

A continuación delimitamos con maior precisión os obxectivos e contidos de cada un dos capítulos que compón esta tese de doutoramento.

O capítulo 2 ten como obxectivo introducir as ferramentas necesarias para entender o formalismo da análise estatística de obxectos complexos que toman valores en espazos métricos. Así mesmo, motivaremos o uso destas ferramentas de análise mediante algúns exemplos de interese polas súas múltiples aplicacións na medicina contemporánea. Primeiro comezaremos introducindo a noción de media e varianza de Fréchet que xeneraliza a noción de centro e dispersión usual no contexto dos espazos métricos [82]. Introduciremos algunhas propiedades básicas do problema de estimación, que pode ser escrito en forma de problema de M-estimación [277] e, en consecuencia, permite empregar a teoría existente para derivar a teoría límite nalgúns casos mediante as ferramentas clásicas da teoría de procesos empíricos [220]. De acordo a estas ideas introducimos a noción de modelo de regresión lineal en espazos métricos [214], de novo expresando o problema como un problema de M-estimación, e facendo uso da teoría asociada a esta familia xeral de estimadores. A continuación, posto que o obxectivo desta tese é ilustrar os métodos no caso particular das representacións distribucionais coa xeometría inducida pola distancia 2-Wasserstein, e esta construción está motivada polo problema do transporte óptimo de distribucións de probabilidade [205], introduciremos as métricas de Wasserstein. A continuación, e dado que realizamos as distintas tarefas de modelado baixo o paraugas dos espazos de Hilbert con núcleo reprodutor (RKHS) faremos unha breve introdución aos elementos centrais deste paradigma, que permite analizar e integrar datos de natureza complexa. Ademais, introduciremos a noción de *kernel mean embedding* [191], que serve como base formal para definir distancias estatísticas entre as representacións distribucionais e outros obxectos estatísticos complexos, e que nos servirá en capítulos posteriores para definir test estatísticos de igualdade de distribución, test de independencia estatística ou realizar análise de conglomerados.

O capítulo 3 ten como obxectivo introducir desde un punto de vista formal as novas representacións distribucionais para aquelas series temporais que son xeradas por un proceso estocástico en tempo continuo, como é o caso dun medidor de CGM, que mide os valores da glicosa intersticial de forma dinámica e cada poucos minutos. Así mesmo, proporcionamos evidencias das vantaxes da nova representación fronte ás métricas existentes de resumo da información dos datos de CGM, como as métricas composiciónais do tempo en rango, medidas de variabilidade glicémica ou outros biomarcadores clínicos existentes. Finalmente, como as novas representacións distribucionais xeran datos funcionais composiciónais que non se poden analizar coas técnicas usuais de datos funcionais, proporcionamos unha serie de técnicas

de análise estatística en espazos métricos para analizar distintos problemas que poden aparecer na clínica, como tests de hipóteses para avaliar a efectividade dun tratamento, análise de regresión para predicir a evolución do doente, ou análises de conglomerados para descubrir novos fenotipos clínicos.

De forma semellante, o capítulo 4 ten como obxecto introducir formalmente e validar as novas representacións distribucionais, pero no caso en que o proceso estocástico é mixto, como ocorre cos datos de acelerómetros. Aquí a validación realízase empregando datos do estudo NHANES, que proveñen dun deseño experimental de enquisa. Estendemos algúns métodos predictivos previos a datos complexos, para poder analizar as vantaxes potenciais das novas representacións, como o método *kernel ridge* para a regresión, e utilizamos o estimador Nadaraya-Watson para a clasificación en espazos métricos. Ao final deste capítulo propoñemos como aplicación na poboación maior a identificación de novos fenotipos clínicos da actividade física, con importantes implicacións na práctica para monitorizar a actividade física e como variables de seguimento da supervivencia e prognóstico dos pacientes.

No capítulo 5 propoñemos novos métodos de aprendizaxe estatística supervisada para problemas con datos perdidos na variable resposta, baseados na aprendizaxe con métodos núcleo en espazos RKHS. Os novos métodos inclúen algoritmos de cuantificación da incerteza, medidas de dependencia estatística ou algoritmos de selección de variables. A tarefa de modelización consiste en predicir como cambian os valores de glicosa a cinco anos en termos do biomarcador A1c, estándar actual para o control e diagnóstico da enfermidade. Os resultados obtidos neste capítulo son unha proba de que a inclusión da monitorización continua da glicosa conduce a unha mellora na predición da condición glicémica do paciente a longo prazo.

No capítulo 6, e motivados pola necesidade de comparar as posibles variacións nas representacións distribucionais en dous instantes de tempo, por exemplo, antes e despois de administrar un determinado tratamento, propoñemos novas probas de hipóteses para datos emparellados e perdidos con obxectos estatísticos complexos, mediante métodos núcleo. Validamos estes métodos, deseñados con carácter xeral, mediante unha análise da evolución da glicosa a cinco anos en distintos subgrupos de pacientes cos datos do estudo AEGIS.

Por último, no capítulo 7, de Conclusións, discutimos os avances metodolóxicos da tese no modelado de obxectos estatísticos complexos e, máis importante, o seu potencial clínico na nova era da medicina dixital e de precisión. Tamén discutimos os diversos traballos activos que foron motivados por esta tese doutoral. Ao final deste capítulo presentamos novos problemas abertos que identificamos e que consideramos abordar no futuro, para proporcionar aos usuarios novas ferramentas de análise de datos con aplicacións na ciencia médica.

Se acompaña a tese dun conxunto de Apéndice nos que escribimos formalmente as probas matemáticas que, por espazo e claridade, foron separadas do resto do material, como a consistencia da estratexia bootstrap introducida no capítulo 5 e o test estatístico introducido no

capítulo 6. Tamén proporcionamos unha pequena guía de uso do paquete software desenvolvido para o entorno R, `biosensor.usc`, co fin de ilustrar como se poden empregar as técnicas aquí propostas. Por último, como as distintas probas introducidas nesta tese dependen da teoría de U -estatísticos introducimos algúns resultados básicos desta teoría [137, 244].

1 Introduction and aims

The last decade has seen remarkable advances in biological signal transduction and subsequent recognition of events of pathophysiological interest [153]. In addition, progress in this area has stimulated the emergence of promising and robust biosensor technology through noninvasive and low-cost approaches [153]. The adoption of this technology in clinical practice will surely lead to the development of new methods to obtain more sophisticated information about changes in patients' health in real-time [153, 274]. A paradigmatic example is closed-loop insulin pump systems that dynamically adjust based on continuous glucose reading, revolutionizing the management of patients with type 1 diabetes [135, 229].

With the increase of recorded patient information and greater availability of these sensors among the general population, the development of statistical and machine learning models is the next big step in establishing the methodology foundations for the new clinical paradigms of digital and precision medicine [93, 138, 139, 153, 162, 298].

Precision medicine aims to optimize healthcare quality by individualizing the care process according to changes in the patient's clinical condition over time [138]. In this new clinical paradigm, decisions are fundamentally supported by data-driven approaches. Mathematically, precision medicine can be formalized as a dynamic optimization problem. For example, potential actions could be: selecting the appropriate drug, selecting the appropriate dose, determining the time of administration, or recommending a specific diet or exercise [138].

Unfortunately, the widespread practice today is that treatments are not prescribed on a personalized basis and focus is on optimizing the health of the average individual in the study population. Consequently, the effect of treatments is suboptimal across a broad list of diseases and patient phenotypes [27, 28, 240, 243, 266].

Undoubtedly, prescribing the optimal treatment is the key goal of personalized medicine [138], but achieving it requires modeling an extensive list of upstream challenges [139]. According to the contributions presented here, some new research directions aim to design new representations of the data provided by biosensors and wearable devices. These new methods will allow us to record patient health changes more accurately and feed different predictive algorithms with more information. In the context of digital medicine, which focuses on exploiting the information provided by electronic measurement devices, it is becoming increasingly

common to assess the impact of the treatment on the patient's condition using digital biomarkers (environmental, biomechanical, and physiological metrics measured with biosensors [153]). Notwithstanding, it is very common that data is complex in nature, e.g. functional nature, such as heart rate variability metrics [196], energy expenditure [177], glucose concentration [178], or as brain connectivity graphs, which contain essential information about the structure and topology of different patterns of neuronal activity [289]. In this new research area, it is essential to design new hypothesis tests as a confirmatory measure of a drug's efficacy, for example, in the context of a randomized clinical trial [49, 51, 57, 125, 139, 140].

Motivated by the progress in the field of precision medicine with the core elements of digital medicine, the primary purpose of this dissertation is to develop new statistical and machine learning methodologies to analyze complex statistical objects [62] such as Euclidean functional data, or more general constructions in metric spaces, such as probability distributions, that live in non-linear spaces without vector space structure [178, 215].

We first address the challenge of analyzing biological time series in patients who are monitored in free-living environments. This challenge presents initial difficulties when approaching a global comparative analysis of the recorded time series because patients are monitored for periods of different duration, follow different lifestyles, or have critical chronological differences. To solve this problem, we find in the literature different representations that extract some summary characteristics of the time series, by means of average values or other metrics of variability [141, 187]. Still, to date the compositional vector metrics constitute the gold-standard method to summarize information from biosensor data in multiple domains, such as in diabetes from the use of continuous glucose monitors, or in physical activity using accelerometry devices [31, 64, 177, 178]. Furthermore, these metrics quantify the proportion of time that the patient is in each of the target zones or predefined categories. Precisely, their great simplicity and interpretability explain the success of these metrics among users [187].

One of the significant drawbacks of using a compositional nature vector metrics in practice is that we cannot quantify the health impact of the time spent by the patient on a continuum of values recorded by the sensor device. Instead, these metrics provide a summary of time-series information over a reduced set of intervals, with a consequent loss of information. Apart from this critical disadvantage, summarizing the information in intervals has other essential drawbacks. We need expert knowledge to define the cut-off points, which may depend on the specific modeling task and the study population to be analyzed. This aspect has been strongly questioned in recent research on physical activity with accelerometers [187].

In this dissertation, to overcome these drawbacks, we propose to use the natural extension at the functional level of vector metrics of compositional nature, which allow for capturing more information. The underlying idea is to consider the probability distribution or density function of the different segments of the recorded biological time series [178]. One of the

advantages of this new representation is that it does not require categorizing the information into a set of intervals previously defined from expert knowledge.

From a formal point of view, we propose a distributional representation whose analysis is derived from compositional data analysis, but according to a functional perspective, [119, 265]. In this regard, analyzing data with this representation presents some critical challenges. For example, since compositional data live in a probability simplex, there is no vector space structure (unless we restrict ourselves to convex combinations), which implies the need to apply alternative data analysis techniques to transform the nonlinear input space to a more appropriate linear space [215]. The process can be summarized as follows: 1) compositional transformation of original input data to a Euclidean vector space, and 2) the application of Euclidean regression models. However, this presents substantial limitations for performing accurate inferential tasks and constructing mathematical methods with theoretical guarantees with finite samples. Moreover, classical compositional transformations or other more recent approaches are not well defined when there are zeros in any part of the representation. In the functional extension, this problem is compounded by the fact that the support of the probability distribution varies among the individuals being monitored.

To overcome this limitation, we provide a new framework for compositional functional data analysis based on the optimal transport theory, coupled with statistical analysis in metric spaces, namely Reproducing kernel Hilbert spaces (RKHS). More specifically, we propose a set of kernels with 2–Wasserstein distance geometry, which is easily computable in the particular case of one-dimensional probability distributions. This analysis framework allows us to build regression models, variable selection methods, uncertainty quantification methods, and hypothesis testing. In addition, we will address the usual context of independent and identically distributed data, as well as more complex designs such as survey designs, or missing data, the latter being common in longitudinal studies when the patient is lost to follow-up.

The effectiveness of the proposed distributional representations will be evaluated in the areas of diabetes and physical activity. For this purpose, we will use data from a longitudinal study on diabetes: the A Estrada Glycation and Inflammation Study (AEGIS) [101], in which information on continuous glucose monitoring (CGM) is available for each patient for approximately one week. This high-resolution information is intended to answer some clinically exciting questions about the impact that certain distributional representations have on the predictive ability of long-term glucose values. In the case of physical activity, we use data from the US NHANES (National Health and Nutrition Examination Survey) survey, which provides information on the physical activity habits of the American population measured objectively using accelerometry devices. Here, we will illustrate the use and advantages of distributional representations in this context to analyze the relationship between physical activity, mortality, and survival of participants across the spectrum of physical activity intensities recorded by

the accelerometer. In the various analyses performed, we will provide strong evidence on the benefits of using these representations to improve predictive ability and clinical sensitivity in the domains mentioned above.

It should be taken into account that diabetes is probably the essential pandemic of this century with alarming prevalence rates that are expected to increase in the coming years, mainly due to worsening lifestyles, motivated both by an increase in inactivity levels and by the impoverishment in the quality of diets worldwide [120, 256]. Therefore, several organizations and specialists are calling for new public health policies to improve the control and spread of the disease. Telemedicine, digital medicine, and precision medicine approach [45, 145] can be a step forward and inspire new, highly effective solutions to reverse this problem. According to these principles, some recent work has hypothesized that the inclusion of digital biomarkers from continuous glucose monitorization may improve the early detection of the disease [105]. Ten years ago, continuous glucose monitorization was already a breakthrough in transforming diabetes management [135, 229]. Nowadays, adopting these technologies is finding more applications in healthy and diseased populations. For example, continuous glucose monitoring devices are used in precision nutrition to personalize diets based on how the intake of specific foods modifies our glucose values [23].

Regarding the importance of analyzing physical activity data, we will highlight that physical activity is currently considered the most effective low-cost non-pharmacological intervention to combat a broad spectrum of diseases, such as diabetes or metabolic syndrome [208]. Physical activity is also an effective intervention to delay cognitive and functional decline associated with aging [231]. For all these reasons, there is currently a solid consensus to promote sport as an essential element in building healthy modern societies. In this respect, the digitization of sports practice can be essential to achieving that goal. The adoption of personalized medicine approaches in this area of knowledge is a pending task [233], and this dissertation introduces new contributions in this direction.

In the following, we delimit more precisely the objectives and contents of each of the chapters that compose this dissertation.

Chapter 2 aims to introduce the necessary tools to understand the formalism of the statistical analysis of complex objects that take values in metric spaces. Likewise, we will illustrate the use of these analysis tools by providing some examples of interest due to their multiple applications in contemporary medicine. First, we will begin by introducing the notion of Fréchet mean and variance, which generalizes the usual notion of center and dispersion in the context of metric spaces [82]. We will introduce some basic properties of the estimation problem, which can be written in the form of an M-estimation problem [277] and, consequently, allows us to use existing knowledge to derive the limit theory in some cases using the classical tools of empirical process theory [220]. According to these ideas, we introduce the notion of linear

regression model in metric spaces [214], again expressing the problem as an M-estimation problem and using the theory associated with this general family of estimators. Next, this dissertation aims to illustrate the methods in the particular case of distributional representations with the geometry induced by the 2-Wasserstein distance. This construction is motivated by the problem of optimal transport of probability distributions [205]; we will make a brief introduction to the Monge problem and to the properties of considering Wasserstein metrics from the empirical point of view. Next, we perform the various modeling tasks under the umbrella of reproducing kernel Hilbert spaces (RKHS); we will briefly introduce the central elements of this theory, which allows us to analyze and integrate data of complex nature. In addition, we will introduce the notion of kernel mean embedding [191], which serves as a formal basis for defining statistical distances between distributional representations and other complex statistical objects, and which will serve us in later chapters to define statistical tests of equality of distribution, statistical independence tests or to perform cluster analysis. Finally, we will show how to construct new kernels from the 2-Wasserstein distance, which can be of potential interest together with classical kernels such as the Gaussian and the Laplacian.

Chapter 3 aims to introduce from a formal point of view the new distributional representations for those time series generated by a stochastic process in continuous time, as is the case of a CGM monitor, which measures interstitial glucose values dynamically over time. We will also provide evidence of the advantages of the new representation over existing metrics for summarizing CGM data information, such as time-in-range compositional metrics, measures of glycemic variability, or other existing clinical biomarkers. Finally, as the new distributional representations generate functional compositional data that cannot be analyzed with standard functional data techniques, we will provide a set of techniques on metric spaces to analyze different clinical problems, such as hypothesis tests to evaluate the effectiveness of a treatment, regression analysis to predict patient evolution, or cluster analysis to discover new clinical phenotypes.

Similarly, chapter 4 aims to introduce the new distributional representations when the stochastic process is mixed, as is the case with accelerometer data. Here the validation is performed using data from the NHANES study, derived from an experimental survey design. We extend some previous predictive methods to complex data in this set-up, such as the kernel ridge method for regression, in order to analyze the potential advantages of these new representations, and we propose a Nadaraya-Watson estimator for classification in metric spaces. At the end of this chapter, we propose as a clinical application in the elderly population the identification of new clinical phenotypes of physical activity, with important implications in practice for monitoring physical activity and as variables for follow-up of patient survival and prognosis.

In chapter 5, we propose new supervised statistical learning methods for problems with

missing data in the response variable, by extending some previous methods of the RKHS paradigm. The new methods include uncertainty quantification algorithms, statistical dependence measures, or variable selection algorithms. The modeling task is to predict how five-year glucose values change in the A1c biomarker, the current standard for the control and diagnosis of the diabetes disease. The results obtained in this chapter are evidence that the inclusion of continuous glucose monitoring improves the prediction of the patient's long-term glycemic condition.

In chapter 6, motivated by the need to compare possible variations in distributional representations at two points in time, for example, before and after treatment, we propose new hypothesis tests for paired and missing data with complex statistical objects, using kernel methods. We will validate these generally designed methods by analyzing the 5-year glucose evolution in different subgroups of patients with data from the AEGIS study.

Finally, in the Conclusions chapter, we discuss the methodological advances of the dissertation in modeling complex statistical objects and, more importantly, their clinical potential in the new era of digital and precision medicine. We also discuss the various active works motivated by this dissertation. At the end of this chapter, we present new open problems that we have identified and consider addressing in the future to provide users with new data analysis tools with applications in medical science.

In the Appendices, we write the mathematical proofs that, for space and clarity, have been moved from the corresponding chapters, such as the consistency of the bootstrap strategy described in chapter 5 and the test statistic introduced in chapter 6. We also provide a short guide to our R package, `biosensor.usc`, to illustrate how practitioners can use the proposed techniques. Finally, since the various tests introduced in this dissertation depend on the U -statistic theory, we introduce the standard results of this theory [137, 244].

2 Statistical analysis in metric spaces and reproducing kernel Hilbert spaces

Object-oriented data statistics [172] or data analysis on nonstandard spaces [122] born with regression modeling by Gauss and his contemporaries, for the purpose of answering different physical questions involving angular and spherical data that arise from the astronomy field [217]. Nowadays, the specific use of these methods allows us to model the raw data recollected by the many measuring instruments and sensors, exploiting their geometrical and intrinsic characteristics. According to the mentioned example of angular and spherical data, several works have shown the potential use of specific data analysis strategies in order to provide new findings in a broad spectrum of applications [61, 171, 204, 206, 217].

From a general perspective of metric spaces, many data analysis techniques have been proposed in the last decades, mainly motivated by practical problems. Examples include the notion of centroid in metric space (Fréchet mean) [82] with applications to phylogenetical tree analysis [201], or the extension of conditional linear regression models to metric spaces with numerous applications in neuroimage and other medical problems [178, 214]. We can also reference less general mathematical constructions by Riemann [155], and hyperbolic manifolds [270], with interesting local geometrical properties and applications, for example, in shape analysis [121].

Despite the recent scientific progress it can be said that the statistical analysis of complex objects is still in its infancy. On the one hand, the number of well-established methods and theoretical results from the multivariate Euclidean spaces with no equivalent in metric spaces is large. Some examples of problems that require more attention can include the selection of relevant variables in regression modeling [275], the assessment of uncertainty in model outputs, the hypothesis testing, or the achievement of optimal theoretical guarantees in the non-asymptotic regimes to understand the empirical model's behavior with finite samples [47, 285]. On the other hand, in metric spaces only a distance function is available, and the underlying optimization problems can be complex, impairing the computational efficiency in high-dimensional and large problems.

This chapter briefly introduces the core elements of statistical analysis on metric spaces.

First, we motivate the interest of medical science in this sort of analysis with some relevant applications. Subsequently, we introduce the notion of Fréchet's mean and variance, together with the global Fréchet model that generalizes the standard linear regression model to those responses that take values in metric spaces. Then, we briefly introduce the framework of reproducing kernel Hilbert spaces (RKHS). RKHS play a vital role in analyzing complex statistical objects since they preserve the properties and advantages of Euclidean geometries. First, we introduce the notion of positive definite kernel functions, the core elements of this framework. Then, we present some statistical distances between random variables that we will use in subsequent chapters, such as the energy distance and the maximum mean discrepancy.

2.1 Complex data in medicine

Bellow, we provide two important examples of complex data that arise in medical applications.

2.1.1 Graph and matrix structures

One of the fascinating research areas that has benefit from the emergence of new statistical techniques for complex objects is neuroimaging [29, 41, 50, 62]. Surely, the analysis of functional magnetic resonance imaging (fMRI) supplies the primary gold standard test to evaluate brain structures. The use of fMRI data and specific statistical test has driven substantial scientific progress about how the brain works and draws new insights into the brain behavior [41, 211, 289].

In practice, the standard routine to summarize the rich and unique signature about brain structures provided by fMRI data is to construct an individual profile from fMRI data estimating the intra- and inter-connections between cerebral regions utilizing continuous (correlation matrix) and discrete (graph structures) representations [289].

In the continuous case, we can consider the metric space (\mathcal{Y}, d) where \mathcal{Y} is the set of Pearson correlation matrices of a fixed dimension r , and d denotes the Frobenius metric $d_{\text{FRO}}(A, B) = \sqrt{\sum_{i,j=1}^r (A_{ij} - B_{ij})^2}$. Similarly, we can consider the space of networks with a fixed number, say r , of nodes and the same metric in Laplacian matrices. In order to identify the space of the graphs with the space of Laplacian matrices, we must introduce some technical assumptions.

Let $G_m = (V, E)$ be an arbitrary network with a set of nodes $V = \{v_1, \dots, v_r\}$ and a set of edge weights $E = \{w_{ij} : w_{ij} \geq 0; i, j = 1, \dots, r\}$, where $w_{ij} = 0$ indicates that v_i, v_j are unconnected. We assume: i) G_m is simple, i.e., there are no self-loops or multi-edges; ii) G_m is weighted, undirected and labeled; iii) the edge weights w_{ij} are bounded, i.e, there exists $w \geq 0$ such that $0 \leq w_{ij} \leq w$. The first assumption is required for the one-to-one correspondence between a network G_m and its Laplacian matrix, which is the central tool

to represent networks. Assumption two guarantees that the adjacency matrix $A = (w_{ij})$ is symmetric, i.e $w_{ij} = w_{ji}$, $\forall i, j$. Assumption three limits the maximum strength of connections between two nodes and prevents extremes. Any network satisfying these assumptions can be uniquely associated with its Laplacian graph $L = (l_{ij})$

$$l_{ij} = \begin{cases} -w_{ij} & i \neq j \\ \sum_{k \neq i} w_{ik} & i = j \end{cases},$$

for $i, j = 1, \dots, r$, which motivates to characterize the corresponding space of networks by

$$\mathcal{Y} = \{L = (l_{ij}) : L = L^T; L1_r = 0_r; -w \leq l_{ij} \leq 0 \text{ for } i \leq j\},$$

where 1_r and 0_r are the r -vectors of ones and zeros, respectively. A precise characterization of the properties of the space of Laplacian graphs can be found in [92].

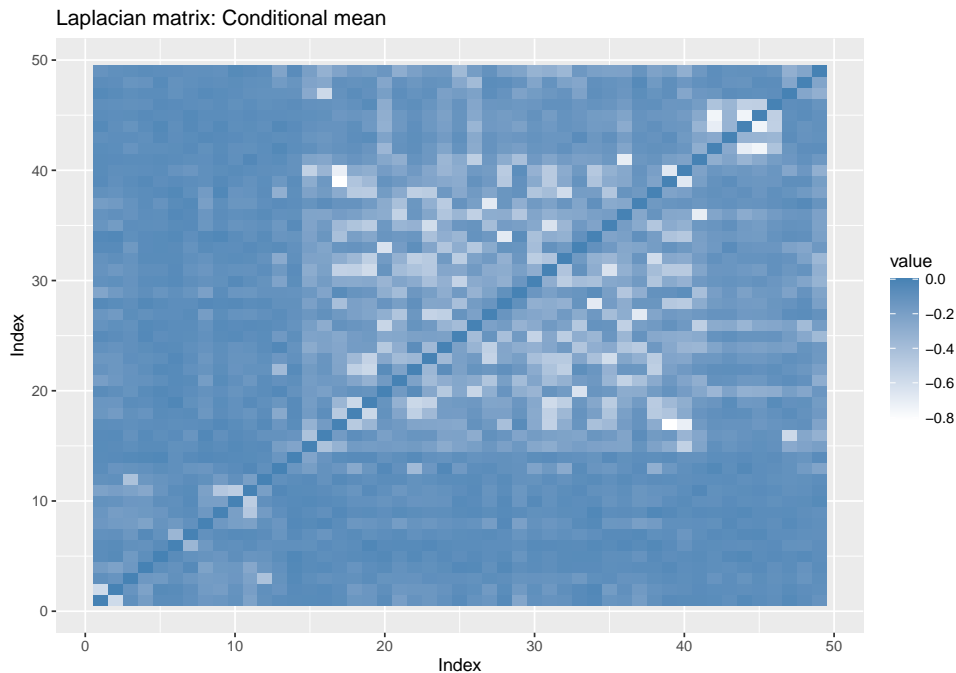


Figure 2.1: Laplacian conditional matrix mean estimation from a person with schizophrenia.

In order to illustrate a Laplacian matrix structure, Figure 2.1 shows the Laplacian conditional matrix mean estimation from a person with schizophrenia. We can see that along the 48 cerebral regions examined, the cerebral connectivity is limited to specific areas, and in general, it is low (value 0 in the contour plot). The Laplacian matrix structures allow us to analyze the spatial interconnections between different regions in integrated statistical objects, unlike the primitive data analysis that only quantifies the specific cerebral activity intensity marginally. As we can appreciate, this type of structures shows sparsity, and specific statistical methods are being developed for this sort of data [227].

2.1.2 Tree structures

Tree structures are increasingly common for registering medical information. The first application of tree structures in medicine emerges in phylogenetical tree analysis [58, 114, 115] where the models exploit the inheritance tree structure between ancestors in the analysis of evolution. In this domain, new algorithms develop notions such as principal component analysis (PCA) for non-Euclidean spaces, and new regression models in abstract spaces, e.g., tropical spaces [293], have been proposed to handle general dependence relationships. More recently, tree structures have been exploited in the analysis of pulmonary function and other human body internal structures [287]. As an example, Figure 2.2 shows the three component blood vessel trees from one person, in which the information captured by the device poses a hierarchical tree structure.

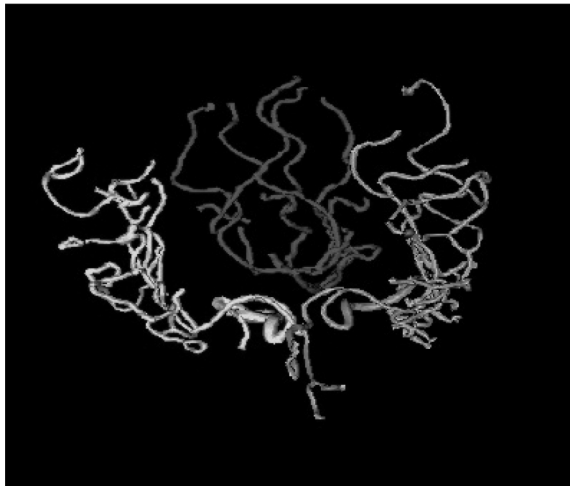


Figure 2.2: The three component blood vessel trees from a given patient.

2.2 Statistical analysis in metric spaces

Definition 1. A metric space is a pair (\mathcal{Y}, d) where \mathcal{Y} is a set and d is a metric on \mathcal{Y} :

$$d : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+, \quad (2.1)$$

such that for any $\omega_1, \omega_2, \omega_3 \in \mathcal{Y}$, the following holds:

1. $d(\omega_1, \omega_2) = 0$ iff $\omega_1 = \omega_2$,
2. $d(\omega_1, \omega_2) = d(\omega_2, \omega_1)$,
3. $d(\omega_1, \omega_3) \leq d(\omega_1, \omega_2) + d(\omega_2, \omega_3)$.

The theory of metric spaces has a large spectre of applications in measuring the similarity between two arbitrary elements of the space using only a distance function. Importantly, in

1948, Fréchet generalized the notion of centroid to metric spaces through the definition of the so called Fréchet mean [82]. Thus, in the particular case of Euclidean data, the arithmetic mean, median, and geometric mean can be considered as different Fréchet means under different choices of distance functions. The Fréchet variance is the corresponding generalized measure of dispersion around the Fréchet mean. The extension of these concepts to this general set-up allows us to operate with general structures that, as we have just seen, appear in many modern medical applications.

Formally, let (\mathcal{Y}, d) be a separable and bounded metric space. Consider a random object $Y \sim P_Y$, where Y takes values in \mathcal{Y} and P_Y denotes the distribution function from Y . The Fréchet mean and variance, denoted as μ_Y and σ_Y^2 respectively, can be defined as [82]:

$$\mu_Y = \arg \min_{\omega \in \mathcal{Y}} E(d^2(Y, \omega)), \quad \sigma_Y^2 = E(d^2(Y, \mu_Y)). \quad (2.2)$$

Let $\{Y_i\}_{i=1}^n$ be a collection of random objects i.i.d. according to P_Y . The empirical Fréchet mean and variance are defined as follows:

$$\tilde{\mu}_Y = \arg \min_{\omega \in \mathcal{Y}} \frac{1}{n} \sum_{i=1}^n d^2(Y_i, \omega), \quad \tilde{\sigma}_Y^2 = \frac{1}{n} \sum_{i=1}^n d^2(Y_i, \tilde{\mu}_Y). \quad (2.3)$$

In order to guarantee the convergence of the empirical versions of Fréchet mean and variance to population counterparts μ_Y and σ_Y^2 , and to provide a central limit theorem to be used in statistical inference, we must introduce the following technical assumptions from the theory of M-estimators.

Assumption 1. *The objects μ_Y and σ_Y^2 exist and are unique, and for any $\epsilon > 0$ $\inf_{d(\mu_Y, \omega) > \epsilon} E(d^2(Y, \omega)) > E(d^2(Y, \mu_Y))$.*

This assumption is instrumental to establish the weak convergence of the empirical process $H_n(\omega) = \frac{1}{n} \sum_{i=1}^n d^2(Y_i, \omega)$ to the population process $H(\omega) = E(d^2(Y, \omega))$, which implies the consistency of $\tilde{\mu}_Y$,

$$d(\tilde{\mu}_Y, \mu_Y) = o_P(1). \quad (2.4)$$

By observing that:

$$|\tilde{\sigma}_Y^2 - \sigma_Y^2| \leq \left| \frac{1}{n} \sum_{i=1}^n [d^2(Y_i, \tilde{\mu}_Y) - d^2(Y_i, \mu_Y)] \right| + \left| \frac{1}{n} \sum_{i=1}^n d^2(Y_i, \mu_Y) - \sigma_Y^2 \right|, \quad (2.5)$$

and due to the consistency of $\tilde{\mu}_Y$ the consistency of $\tilde{\sigma}_Y^2$ can be easily derived:

$$d(\tilde{\sigma}_Y^2, \sigma_Y^2) = o_P(1). \quad (2.6)$$

For a central limit theorem to hold for the empirical Fréchet variance we need an assumption on the complexity of the metric space \mathcal{Y} , which can be quantified by a bound on the entropy

integral for metric δ -balls $B_\delta(\omega)$ of \mathcal{Y} , given by

$$J(\delta, \omega) = \int_0^1 [1 + \log[\eta(\epsilon\delta/2, B_\delta(\omega), d)]]^{1/2} d\epsilon, \quad (2.7)$$

where $B_\delta(\omega)$ is a δ -ball in the metric d , centered at ω , and $\eta(\epsilon\delta/2, B_\delta(\omega), d)$ is the covering number for $B_\delta(\omega)$ using open balls of radius $\epsilon\delta/2$.

Importantly, in the setting of metrics spaces, we do not have a guarantee that the central limit theorem holds; that is why we must introduce specific entropy assumptions.

Assumption 2. For any $\omega \in \mathcal{Y}$, it holds that $J(\delta, \omega) \rightarrow 0$ as $\delta \rightarrow 0$.

Assumption 3. The entropy integral of \mathcal{Y} is finite, that is $\int_0^1 [1 + \log(\eta(\epsilon, \mathcal{Y}, d))]^{1/2} d\epsilon < \infty$.

The central limit theorem for the Fréchet variance is as follows.

Theorem 1. [63] Suppose Assumptions 1-3 hold. Then

$$\sqrt{n}(\tilde{\sigma}_Y^2 - \sigma_Y^2) \rightarrow N(0, \sigma_F^2),$$

in distribution, where $\sigma_F^2 = \text{Var}(d^2(Y, \mu_Y))$.

2.2.1 Anova test

Anova test is one of the core statistical tests to detect differences in localization (mean) and scale (variation) between two or more populations. In this thesis, we are interested in using this test to compare differences, for example, between women and men in their glucose values in Chapter 3. Theorem 1 provides the theoretical foundations to develop an Anova test in metric spaces and derive the asymptotic limit distribution under the null hypothesis. Below, we introduce the formal details of the Anova test proposed in [63].

Let $\{Y_i\}_{i=1}^n$ denote an i.i.d random sample belonging to k different disjoint groups whose distribution functions are F_1, F_2, \dots, F_k , each of size n_j ($j = 1, \dots, k$), so that $\sum_{j=1}^k n_j = n$.

Authors define $\tilde{\mu}_j = \arg \min_{g \in \mathcal{Y}} \frac{1}{n_j} \sum_{i \in G_j} d^2(Y_i, g)$ and $\tilde{\sigma}_j^2 = \frac{1}{n_j} \sum_{i \in G_j} d^2(Y_i, \tilde{\mu}_j)$, the pooled sample Fréchet mean and the pooled sample Fréchet variance:

$$\tilde{\mu}_p = \arg \min_{g \in \mathcal{Y}} \sum_{j=1}^k \sum_{i \in G_j} d^2(Y_i, g), \quad \tilde{\sigma}_p^2 = \frac{1}{n} \sum_{j=1}^k \sum_{i \in G_j} d^2(Y_i, \tilde{\mu}_p),$$

and finally the empirical variance for $\tilde{\sigma}_j^2$ for each group:

$$\widehat{\text{Var}}(\tilde{\sigma}_j^2) = \frac{1}{n_j} \sum_{i \in G_j} d^4(Y_i, \tilde{\mu}_j) - \left\{ \frac{1}{n_j} \sum_{i \in G_j} d^2(Y_i, \tilde{\mu}_j) \right\}^2$$

where $j = 1, \dots, k$. Then, with

$$F_n = \tilde{\sigma}_p^2 - \sum_{j=1}^k \lambda_{j,n} \tilde{\sigma}_j^2, \quad R_n = \sum_{j < l} \frac{\lambda_{j,n} \lambda_{l,n}}{\tilde{\sigma}_l^2 \tilde{\sigma}_j^2} (\tilde{\sigma}_j^2 - \tilde{\sigma}_l^2),$$

under the null hypothesis $H_0 : F_1 = F_2 = \dots = F_k$, the proposed test statistic is

$$T_n = \frac{nR_n}{\sum_{j=1}^k \frac{\lambda_{j,n}}{\widetilde{Var}(\tilde{\sigma}_j^2)}} + \frac{nF_n^2}{\sum_{j=1}^k \lambda_{j,n}^2 \widetilde{Var}(\tilde{\sigma}_j^2)} \rightarrow \chi_{k-1}^2, \quad (2.8)$$

in distribution, where the weights $\lambda_{j,n} = n_j n^{-1}$ sum 1. The consistency of the Efrón naive calibration strategy of the test statistics can be obtained conditioned to the observed random sample using similar theoretical arguments as used in Theorem 1. Further details can be found in [63].

2.2.2 Empirical Fréchet mean: the Wasserstein metrics

A Wasserstein distance is a family of metrics to assess the similarity between density functions f and g (F and G denote their distribution functions on a ground space \mathcal{X}).

Let \mathcal{X} be a separable Banach space and $P(\mathcal{X})$ the set of distribution functions over \mathcal{X} . The p -Wasserstein space on \mathcal{X} is defined as

$$\mathcal{W}_p(\mathcal{X}) = \left\{ f : F \in P(\mathcal{X}) \text{ and } \int_{\mathcal{X}} \|x\|^p f(x) < \infty \right\}, \quad p \geq 1. \quad (2.9)$$

Recall that if $F, G \in P(\mathcal{X})$, then $\Pi(F, G)$ is defined to be the set of measures $\pi \in P(\mathcal{X}^2)$ having as distribution function F and G as marginals. The p -Wasserstein distance between f and g is defined as the minimal transportation cost between the distributions F and G :

$$d_{\mathcal{W}_p}(f, g) = \inf_{\pi \in \Pi(F, G)} (C_p(\pi))^{1/p} = \left(\inf_{\pi \in \Pi(F, G)} \int_{\mathcal{X} \times \mathcal{X}} \|x_1 - x_2\|^p d\pi(x_1, x_2) \right)^{1/p}, \quad (2.10)$$

that can be interpreted, from a physical point of the view, as the minimum of amount of work to change one distribution function into another. A proof that \mathcal{W}_p is a metric can be found in Villani [283].

In general, estimating the p -Wasserstein metric is challenging and requires using a specific numerical scheme to obtain the optimal solution. However, there are exceptions for the coefficients $p = 1$ and $p = 2$. For example, for $p = 2$, we can write the optimal solution in terms of the quadratic distance of quantile functions:

$$d_{\mathcal{W}_2}(f, g) = \sqrt{\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt}. \quad (2.11)$$

Still, this exception only holds for univariate density functions. Consider a random sample $\{f_i\}_{i=1}^n$, and let \mathcal{Y} be the space of univariate density functions with support in $T \subset \mathbb{R}$, equipped with the metric $d_{\mathcal{W}_2}$. In this case, we can find a closed-form expression to estimate the Fréchet centroid as the usual mean in Euclidean spaces but from the Quantile representations associated with the density functions:

$$\tilde{\mu}_Y = \arg \min_{f \in \mathcal{Y}} \frac{1}{n} \sum_{i=1}^n d^2(f_i, f) = \arg \min_{f \in \mathcal{Y}} \frac{1}{n} \sum_{i=1}^n \int_0^1 (F_i^{-1}(t) - F^{-1}(t))^2 dt. \quad (2.12)$$

For a different $p \in (1, 2)$, we must resort to specific gradient optimization approximations to derive the notion of centroid [218], and the posterior estimation of the Fréchet variance.

2.2.3 Global Fréchet regression

Linear regression methods are the most commonly used to estimate the conditional mean between a random response variable Y and a predictor random vector X . Their interpretability and closed-form expressions are the primary characteristics making them attractive to practitioners.

Authors in [214] generalize the notion of linear regression to the case of the response variable taking values in a metric space, but the predictors remaining Euclidean. We use later this regression method in Chapter 3 to estimate the conditional glucose profiles according to patient characteristics.

Let $(X, Y) \sim P_{X, Y}$ be a multivariate random variable, where $X \in \mathbb{R}^p$, and $Y \in \mathcal{Y}$, being (\mathcal{Y}, d) a properly bounded separable metric space. For each fixed $X = x \in \mathbb{R}^p$, the conditional mean estimation can be obtained by solving the following optimization problem in metric space:

$$m(x) = \arg \min_{\omega \in \mathcal{Y}} E \left[[1 + (X - x) \Sigma^{-1} (x - \mu)] d^2(Y, \omega) \right] = \arg \min_{\omega \in \mathcal{Y}} E \left[w(x, X) d^2(Y, \omega) \right],$$

where $\Sigma = Cov(X, X)$, $\mu = E(X)$, and $w(x, X) = [1 + (X - \mu) \Sigma^{-1} (x - \mu)]$ is a proper weight-function for enforcing the constraint that the conditional mean takes a linear structure.

Let us suppose that $\mathcal{Y} = \mathbb{R}$ and $d(x, y) = |x - y|$, we have

$$\begin{aligned} m(x) &= \arg \min_{\omega \in \mathcal{Y}} E \left[[1 + (X - \mu) \Sigma^{-1} (x - \mu)] (Y - \omega)^2 \right] \\ &= \arg \min_{\omega \in \mathcal{Y}} E \left[(Y - \omega)^2 \right] + \arg \min_{\omega \in \mathcal{Y}} E \left[(Y - \omega)^2 [(X - \mu) \Sigma^{-1}] \right] (x - \mu) \\ &= \mu_Y + \langle \beta, x - \mu \rangle, \end{aligned} \quad (2.13)$$

that is, we can rewrite the solutions in terms of an intercept, μ_Y , and a slope coefficient $\beta \in \mathbb{R}^p$, recovering the traditional multivariate linear shape, that can be interpreted in terms

of a linear projection $m(x) = \mu_Y + \langle \beta, x - \mu \rangle$, where $\langle x, y \rangle$, denotes a dot product between two elements $x, y \in \mathbb{R}^p$.

Importantly, this prior interpretation only holds when \mathcal{Y} is a separable Hilbert metric space, as established in [214] and not only when the response variable takes values in the real line.

Assume that an i.i.d. random sample $\{(X_i, Y_i)\}_{i=1}^n$ from $P_{X,Y}$ is available, we can obtain a estimation of conditional mean as follows:

$$\tilde{m}(x) = \arg \min_{\omega \in \mathcal{Y}} \frac{1}{n} \sum_{i=1}^n \left[1 + (X_i - \bar{X}) \tilde{\Sigma}^{-1} (x - \bar{X}) d^2(\omega, Y_i) \right], \quad (2.14)$$

where $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, and $\tilde{\Sigma} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T$.

A more general abstract regression approach was proposed recently, for the case of the response and predictors taking values in a separable metric space [47]. A random forest algorithm with these same settings was proposed in [222].

2.3 Statistical learning in reproducing kernel Hilbert spaces

The reproducing kernel Hilbert spaces (RKHS) framework constitutes one of the most influential families of statistical learning algorithms in metric spaces [26]. Methods in RKHS are a generalization from finite-dimensional Euclidean spaces to an infinite-dimensional context preserving the advantages of Euclidean geometries [26]. Furthermore, they can naturally integrate different sources of information, model non-linear relations of dependence, and analyze complex statistical objects such as graphs, strings, or curves in a straightforward way [34, 191]. Furthermore, they can perform statistical learning in different modeling problems with excellent ratios of convergence [43].

Suppose that we observe a random sample $\mathcal{D}_n = \{(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}\}_{i=1}^n$ from a distribution function $P_{X,Y}$. The structure of many statistical learning problems from a functional analysis perspective is as follows:

$$\tilde{m} = \arg \min_{m \in \mathcal{F}} \mathcal{R}_{L, \mathcal{D}_n}(m) + \lambda \Omega(m), \quad (2.15)$$

where $\mathcal{R}_{L, \mathcal{D}_n}(m)$ denotes the empirical risk functional that can be defined in terms of loss-function $L : (y, y') \in \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ as follows $\mathcal{R}_{L, \mathcal{D}_n}(m) = \frac{1}{n} \sum_{i=1}^n L(Y_i, m(X_i))$. \mathcal{F} is the class of functions over \mathcal{X} that we use to perform the statistical learning, $\lambda > 0$ denotes the regularization parameter of regularization function $\Omega(f) : \mathcal{F} \rightarrow \mathbb{R}^+$. We also assume that convexity property holds in the loss-function L . Suppose for example the standard mean regression problem as a particular case:

$$\tilde{m} = \arg \min_{m \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (Y_i - m(X_i))^2 + \lambda \|m\|_{\mathcal{F}}^2. \quad (2.16)$$

In practice, the election of a very general class \mathcal{F} can be counterproductive. Let us suppose that we fix the functional space $\mathcal{F} = L^2([0, 1]) = \{f : \int_0^1 f^2(s) ds < \infty\}$, where $f : [0, 1] \rightarrow \mathbb{R}$ is a measurable function. We can find examples in which the functional in Equation 2.15 may diverge to infinity. Therefore, we must restrict the functional space \mathcal{F} in which we perform the statistical learning for ensuring good properties, while we guarantee that it is rich enough for the specific modeling task.

A natural way to do this is by performing the statistical learning in a proper reproducing Kernel Hilbert space (RKHS) that we denote by \mathcal{H} . A key element of the RKHS theory is the existence of a kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$ that allows transforming the previous infinite-dimensional problem into a finite-dimensional one [26]. Importantly, the function minimizing the empirical risk functional in Equation 2.15 has a closed-form expression given by $\tilde{m}(x) = \sum_{i=1}^n \alpha_i k(x, X_i)$, according to the well-known representer theorem. Thus, many modeling statistical problems can be expressed as a linear combination of the kernel function evaluated at the training points [191].

Next, we formally review the RKHS framework [26]. First, consider a set \mathcal{X} and denote $\mathcal{F}(\mathcal{X}) = \{f : \mathcal{X} \rightarrow \mathbb{R}\}$ the space of real-valued functions on \mathcal{X} with their vectorial space-structure.

Definition 2. (*reproducing kernel Hilbert spaces - RKHS*) Let \mathcal{X} be a set, let $\mathcal{H} \subseteq \mathcal{F}(\mathcal{X})$ be a Hilbert space. Then \mathcal{H} is called a RKHS if there exist a kernel k on \mathcal{X} satisfying

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

Moreover, we call $k(\cdot, \cdot)$ a reproducing kernel of \mathcal{H} .

2.3.1 Positive-definite kernels

The key point for the success of statistical learning in RKHS spaces is that we can transform many modeling problems in terms of a linear operator by means of the positive-definite kernel k [26, 191]. Next, we introduce key properties of kernels.

Definition 3. (*positive semi-definite kernel [26]*) Let \mathcal{X} be a set, then a symmetric kernel in their arguments k on \mathcal{X} is called positive semi-definite if for all $m \in \mathbb{N}$ and for all $x_1, \dots, x_m \in \mathcal{X}$ the gram matrix K given by $K_{ij} = k(x_i, x_j), i, j = 1, \dots, m$ is positive semi-definite, and for all $z \in \mathbb{R}^m$ it is hold that $z^t K z > 0$.

Examples of positive semi-definite kernels on $\mathcal{X} = \mathbb{R}^n$ are introduced bellow:

- Gaussian kernel with bandwidth $\sigma > 0$,

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \quad (2.17)$$

- Laplacian kernel with bandwidth $\sigma > 0$,

$$k(x, y) = \exp\left(-\frac{\|x - y\|}{\sigma}\right) \quad (2.18)$$

- Polynomial kernel of degree $d \in \mathbb{N}$,

$$k(x, y) = \langle x, y \rangle^d \quad (2.19)$$

- Sigmoid kernel with κ , where κ is a real-function, and $\theta < 0$,

$$k(x, y) = \tanh(\kappa\langle x, y \rangle) + \theta. \quad (2.20)$$

We can introduce several operations between kernels that preserve semi-definite character.

1. if k_1 is a kernel, and $b \geq 0$, then $k_1 + b$ is a kernel.
2. If k_1 and k_2 are kernels, and $\alpha_1, \alpha_2 \geq 0$, then $\alpha_1 k_1 + \alpha_2 k_2$ is a kernel.
3. If k_1 and k_2 are kernels, then $k(x, y) := k_1(x, y) k_2(x, y)$ is a kernel.
4. If k_1, k_2, \dots are kernels, and $k(x, y) := \lim_{n \rightarrow \infty} k_n(x, y)$ exists for all x, y , then k is a kernel.

Based on these primitives, we can derive more complicated closure operations. For instance, a polynomial function $f: \mathbb{R} \rightarrow \mathbb{R}$ with positive coefficients

$$f(r) = \sum_{i=0}^d a_i r^i, \quad \forall d \in \mathbb{N}, a_i \geq 0. \quad (2.21)$$

We can also derive the polynomial kernel $k(x, y) = (\langle x, y \rangle + c)^d$ ($c \geq 0, d \in \mathbb{N}$). Another example is the exponential kernel $k(x, y) = \exp(\sigma\langle x, y \rangle)$ ($\sigma > 0$) since

$$\exp(\sigma r) = \sum_{i=0}^{\infty} \frac{\sigma^i}{i!} r^i. \quad (2.22)$$

A last example. The following function

$$k(x, y) := \frac{\exp(2\sigma\langle x, y \rangle)}{\sqrt{\exp(2\sigma\langle x, x \rangle)} \sqrt{\exp(2\sigma\langle y, y \rangle)}} \quad (2.23)$$

is a kernel.



Proposition 2. (uniqueness of the kernel) [26] Let \mathcal{X} be a set and let \mathcal{H} be a RKHS on \mathcal{X} . Assume both k and \tilde{k} are reproducing kernels on \mathcal{H} . Then $k = \tilde{k}$.

Theorem 3. (alternative characterization of RKHS) [26] Let \mathcal{X} be a set and for all $x \in \mathcal{X}$ let $\delta_x : \mathcal{F}(\mathcal{X}) \rightarrow \mathbb{R}$ be the function such that for all $f \in \mathcal{F}(\mathcal{X})$ it holds that $\delta_x(f) = f(x)$. Then, a Hilbert space $\mathcal{H} \subseteq \mathcal{F}(\mathcal{X})$ is a reproducing kernel Hilbert space if and only if for all $x \in \mathcal{X}$ the function δ_x is continuous on \mathcal{H} .

Theorem 4. (separability and continuity) [26] Let \mathcal{X} be a set, let k a continuous, bounded and positive semi-definite kernel on \mathcal{X} and let \mathcal{H} be the RKHS with reproducing kernel k . Then, \mathcal{H} is separable Hilbert space consisting only of continuous functions. Furthermore, given an orthonormal basis $(\phi_n)_{n \in \mathbb{N}}$ of \mathcal{H} it holds for all $x, y \in \mathcal{X}$ that

$$k(x, y) = \sum_{n=1}^{\infty} \phi_n(x) \phi_n(y). \quad (2.24)$$

The Gaussian kernel on \mathbb{R}^n holds all these conditions as well as all the kernels proposed in this thesis.

2.3.2 Negative-type metrics and semi-definite kernels

Let (\mathcal{X}, d) be a metric space. We say that (\mathcal{X}, d) has a negative type if for all $n \geq 1$ and all lists of n red points x_i and n blue points x'_i in \mathcal{X} , the sum $2 \sum_{i=1}^n \sum_{j=1}^n d(x_i, x'_j)$ of the distances between the $2n^2$ ordered pairs of points of opposite color is at least the sum $\sum_{i=1}^n \sum_{j=1}^n d(x_i, x_j) + d(x'_i, x'_j)$ of the distances between the $2n^2$ ordered pairs of points of the same color [168].

We can arrive at this formal property in metric spaces from a well-know and non-trivial property of Euclidean spaces. For all $n \geq 1$, $x_1, \dots, x_n \in \mathcal{X}$, and $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ with $\sum_{i=1}^n \alpha_i = 0$, we have [168],

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j d(x_i, x_j) \leq 0, \quad (2.25)$$

where α_i can be interpreted as the indicator that x_i is red minus the indicator that x_i is blue.

We say that (\mathcal{X}, d) has a strict negative type if, for every n and all n -tuples of distinct points x_1, \dots, x_n , equality holds in 2.25 iff $\alpha_i = 0$ for all i . Again, Euclidean spaces have strict negative type. A simple example of a metric space with non-strict negative type is l_1 distance on \mathbb{R}^2 on a 3-point space.

Given two (Borel) probability measure F and G on \mathcal{X} with finite first moment, we can generalize the notion of negative type for infinite spaces. We say that (\mathcal{X}, d) has a negative type if

$$\int d(x_1, x_2) d(F - G)^2(x_1, x_2) \leq 0. \quad (2.26)$$

We say that (\mathcal{X}, d) has a strong negative type if it has a negative type and equality holds only when $F = G$ [167, 168].

The following result shows that semi-metrics of negative type and symmetric positive semi-definite kernels are closely related [24]. Let

$$k(x, y) = \frac{1}{2} [d(x, x_0) + d(y, x_0) - d(x, y)], \quad (2.27)$$

where $x_0 \in \mathcal{X}$ is an arbitrary family of fixed points. Then, it can be show that k is positive semi-definite iff d is a semi-metric of negative type. We have a family of kernels, one for each choice of x_0 . Conversely, if d is a semi-metric of negative type and k is a kernel in this family, then

$$d(x, y) = k(x, x) + k(y, y) - 2k(x, y). \quad (2.28)$$

2.3.3 Kernel mean embeddings, characteristics and universal kernels

One of the main advantages of statistical learning in the *RKHS* framework is that we can embed complex statistical objects into them and exploit the inherent Hilbert structure to analyze these objects in an infinite-dimensional space that can preserve, for example, the distributional properties of the original input data [191]. Therefore, we can express different statistical problems as linear operations between inner products utilizing kernel function k that transform the original input space. Many times, the final estimators possess closed-form expressions. Along this thesis, we use this idea of embedding to transform a probability distribution in the original space to an element of a RKHS.

Let \mathcal{X} be a separable metric space. Let k be a continuous bounded positive semi-definite kernel and let \mathcal{H} be the RKHS with reproducing kernel k . We denote as $\mathcal{M}_f(\mathcal{X})$ the set of finite Borel measures on \mathcal{X} .

Definition 4. (*kernel mean embedding function*) [191] *The kernel mean embedding of probability measures in $\mathcal{M}_f(\mathcal{X})$ is defined by a mapping $\phi : \mathcal{M}_f(\mathcal{X}) \rightarrow \mathcal{H}$ with the property that*

$$\phi(F) = \int_{\mathcal{X}} k(x, \cdot) F(dx). \quad (2.29)$$

Definition 5. (*characteristic kernel*) [191] *The kernel k is said to be characteristic if ϕ is injective.*

Definition 6. (*universal kernel*) [191] *The kernel k is said to be universal if \mathcal{H} is dense on $\mathcal{C}(\mathcal{X})$, i.e. for every function $f \in \mathcal{C}(\mathcal{X})$, and $\epsilon > 0$ there exist a function g induced by k with $\|f - g\|_{\infty} \leq \epsilon$.*

Examples of universal kernels are provided bellow

- $k(x, y) = k(\langle x, y \rangle)$ defined on \mathbb{R}^d , with power series expansion

$$k(r) = \sum_{i=0}^{\infty} a_i r^i, \quad (2.30)$$

is a universal kernel iff for all i , we have a_i strictly positive.

- $k(x, y) = k(x - y)$ defined on \mathbb{R}^d , with a Fourier transformation

$$F[k](\omega) = (2\pi)^{-\frac{d}{2}} \int_{\mathbb{R}^d} e^{i\langle \omega, r \rangle} k(r) dr \quad (2.31)$$

is a universal kernel iff for all ω , we have the Fourier coefficients strictly positive.

2.3.4 Energy distance and distances between embeddings

Distances between kernel mean embeddings such as the maximum mean discrepancy (MMD) [191] and the energy distance (ED) [262] are two families of statistical distances between arbitrary random elements that take values in separable Hilbert spaces. With their increase in popularity at the beginning of this century, data analysis methods derived from them have become an essential tool in many statistical modeling tasks, e.g., hypothesis testing, clustering analysis, variable selection, and screening variables [263].

The equivalence between these two families of distances, at the population and finite sample levels, was established in a series of recent papers using the connections between negative-type semi-metrics and conditional symmetric positive definite kernels [180, 241]. Next, we provide the formal definition of energy distance in Euclidean spaces.

Definition 7. (*Euclidean energy distance*) Let be $X \sim F, Y \sim G$ be two \mathcal{X} -random variables satisfying $E(\|X\|^2) < \infty$ and $E(\|Y\|^2) < \infty$. The Euclidean energy distance of order $\alpha \in (0, 2]$ is defined as:

$$\epsilon_\alpha(X, Y) = 2E(\|X - Y\|^\alpha) - E(\|X - X'\|^\alpha) - E(\|Y - Y'\|^\alpha), \quad (2.32)$$

where X' and Y' are i.i.d. random copies of random variables X and Y , respectively.

Energy distance can be extended to obtain a more general family of statistical distances in separable Hilbert Spaces. For this purpose, it is enough to consider an arbitrary semi-metric of negative type $\rho(\cdot, \cdot)$ [24].

Definition 8. (*general formulation of energy distance*) Let $\rho : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$ be a metric of negative type, and $X \sim F, Y \sim G$ two random variables in \mathcal{X} satisfying $E(\rho^2(X, o)) < \infty$ and $E(\rho^2(Y, o)) < \infty$, where $o \in \mathcal{X}$ is an arbitrary fixed element of the space. We define the energy distance as

$$\epsilon_\rho(X, Y) = 2E(\rho(X, Y)) - E(\rho(X, X')) - E(\rho(Y, Y')), \quad (2.33)$$

which is equivalent to Eq. (2.32) when $\rho(x, y) = \|x - y\|^\alpha$.

Finally, we define the maximum mean discrepancy in the RKHS framework.

Definition 9. (*maximum mean discrepancy (MMD)*) Let $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$ be a symmetric and positive definite kernel, and $X \sim F, Y \sim G$ be two random variables in \mathcal{X} satisfying $E(k(X, X')) < \infty$ and $E(k(Y, Y')) < \infty$. We define the maximum mean discrepancy as

$$\begin{aligned} \text{MMD}(X, Y)_k^2 &= \|\phi(F) - \phi(G)\|^2 \\ &= \left\| \int k(x, \cdot) F(dx) - \int k(y, \cdot) G(dy) \right\|^2 \\ &= E(k(X, X')) + E(k(Y, Y')) - 2E(k(X, Y)). \end{aligned} \quad (2.34)$$

where X' and Y' are i.i.d. random copies of random variables X and Y , respectively.

Given two samples i.i.d $\{X_i\}_{i=1}^{n_1} \sim F$ and $\{Y_i\}_{i=1}^{n_2} \sim G$, $n_1 + n_2 = n$, we can straightforwardly estimate their empirical counterparts, that we denote as \hat{F} and \hat{G} , respectively. Using this estimates, we estimate the aforementioned statistics as

$$\tilde{\epsilon}_\rho(X, Y) = \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \rho(X_i, Y_j) - \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \rho(X_i, X_j) - \frac{1}{n_2^2} \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} \rho(Y_i, Y_j), \quad (2.35)$$

and

$$\widetilde{\text{MMD}}(X, Y)_k = \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} k(X_i, X_j) + \frac{1}{n_2^2} \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} k(Y_i, Y_j) - \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} k(X_i, Y_j) \quad (2.36)$$

The choice of the kernel function k , both in the energy distance and the maximum mean discrepancy, is critical for the performance of different modeling tasks. However, it is not easy to establish a general criterion since each choice of semi-metric characterizes distributional differences, giving more or less priority to a specific moment in the computation. This information may not be available in practice and may be challenging to obtain from expert knowledge or prior studies. Established principles for kernel choice on the basis of maximizing the two-sample test power can be found in [99, 128].

The asymptotic behaviour of the previous statistics under the null (equality in distribution) and alternative hypothesis can be derived applying the standard theory of U- and V-statistics (see Appendix). The exact derivations can be found in the original references [97, 261, 262].

Part II

**Distributional representations
from wearable and biosensor
technology**

3 Distributional representations of biosensor data for continuous stochastic process with application in CGM technology

The steadily increasing availability and prominence of biosensor data have given rise to new methodological challenges for their statistical analysis. A primary feature of these data is that the monitored individuals are in free-living conditions, making a direct analysis of the recorded time series between groups of patients problematic if not infeasible. A clear example of such data is found in the study of diabetes, where continuous glucose monitoring (CGM) is increasingly used. The elevation of glucose is distinct between individuals and is influenced by factors such as mealtimes, diet composition, or physical exercise [71]. Consequently, an exciting topic of debate is how to exploit the enormous wealth of information recorded by CGM to draw more reliable conclusions about glucose homeostasis rather than the cursory summary measures such as fasting plasma glucose (FPG) or glycated hemoglobin (A1c) [295].

Since 2010, the American Diabetes Association (ADA) has included measurement of A1c levels for both diagnosis and diabetes control [17]. A1c levels reflect underlying glucose levels over the preceding 3 months, and its within-patient reproducibility is superior to that of fasting plasma glucose and oral glucose tolerance tests (OGTTs) [242]. However, recent works have provided evidence for the need to go beyond A1c and use new measures for glycemic control [100, 110], in order to capture more diverse aspects of the temporally evolving glucose levels beyond the average, for example, glucose variability and time-in-range metrics. The metric time-in-range measures the proportion of time an individual's glucose levels are maintained in different target zones. In the case of diabetes, these can include ranges corresponding to hypoglycemia and hyperglycemia. In an innovative article [21], authors validated the time-in-range metric, showing that it is a good predictor of long-term microvascular complications despite just measuring glucose values seven times per day. [160] reached similar conclusions but using CGM technology only for 24 hours in each patient. At the same time, it is well-known that two patients may have the same glycosylated hemoglobin and a completely different glycemic profile [22]. These new findings have led clinical specialists to consider that continuous glucose measurement during long monitoring periods can lead to more accurate research and clinical

practice results than standard methods [111]. In fact, since 2012, the European Medicine Agency [79] recommends the use of CGM to validate the effect of drugs for treatment or prevention of diabetes mellitus.

Traditionally, real-time continuous glucose monitoring combined with insulin pump therapy has been shown to improve metabolic control and to reduce the rate of hypoglycemia in adults with type 1 diabetes [59, 74, 142]. Notwithstanding, more recent applications of CGM have been more general. For example, they involve screening patients, optimizing diet, epidemiological studies, assessing patient prognosis, supporting treatment prescriptions, and having even been used in healthy populations [83, 105, 161]. In addition to the increasing utility of CGM data, the technology is gradually becoming cheaper, and new devices capable of measuring glucose in a non-invasive way are quickly emerging [200]. All of these advances are facilitating the adoption of CGM in standard clinical practice.

In 2012, a panel of experts discussed how to represent CGM data in an “easy to view format” [25]. They also analyzed the convenience of using glycemic variability measures and other summary measures such as time-in-range to extract the CGM’s recorded information. In 2019, ADA launched an updated consensus guide for promoting the correct and standardized use of time-in-range metrics in standard clinical practice, defining several practitioners’ target zones. A more recent review about the CGM metric establishes time-in-range as a gold-standard measure [199].

Motivated by the problem of analyzing data gathered via CGM more precisely while still leveraging the advantages possessed by time-in-range metrics, we propose an approach based on the construction of a functional profile of glucose values for each subject. Conceptually, the approach is a natural extension of time-in-range metrics in which the intervals simultaneously shrink in size and increase in number so that the new profile effectively measures the proportion of time each patient spends at each specific glucose concentration rather than a coarsely defined range. As a result, the new functional profile, which we refer to as a glucodensity, automatically and simultaneously captures all parameters arising from individual glucose distributions. To illustrate our new glucose representation graphically, Figure 3.1 shows a set of constructed glucodensities that represent the data objects for which we will propose using a tailored set of statistical methods. The glucose profile patterns are clearly heterogeneous between individuals, both in mean, variability, or any other distributional characteristics including the hypo-hyper glucose range, where glucodensities have different support depending on patient condition. For example, in normoglycemic patients, glucose generally oscillates between $75 - 150\text{mg/dL}$ while in some patients with diabetes, glucose can reach concentrations of 400mg/dL in the range of severe hyperglycemia. Moreover, the shape of the glucodensities are entirely different. Moreover, the shape of the glucodensities is entirely different, with existing variability patterns along all glucose concentrations between normoglycemic and diabetes

patients.

Mathematically, glucodensities constitute functional-distributional data since each glucodensity represents a distribution of glucose concentrations. As such, these complex and constrained curves cannot be directly analyzed with the usual techniques. To overcome this, we introduce a framework for the analysis of distances between glucodensities by compiling suitable methods based on the calculation of glucodensities distances. We also reveal our representation's superior clinical capacity compared to classical measures of diabetes control and diagnostics. Finally, we demonstrate that our representation has a higher sensitivity than the standard time-in-range metric to explain the glyceimic differences between patients in various settings, including regression analysis. A new shiny interface to use the methods outlined in this chapter is available at <https://tec.citius.usc.es/diabetes>.

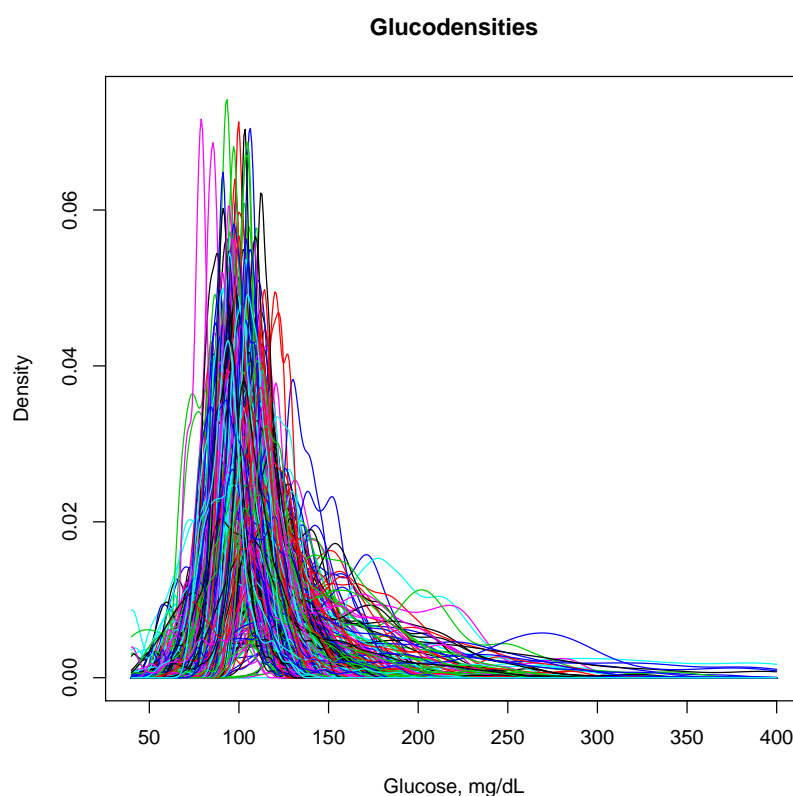


Figure 3.1: Glucodensities are estimated from a random sample of the AEGIS study with diabetic and normoglycemic patients. For each patient, this glucose representation estimates the proportion of time spent at each glucose concentration over a continuum, representing a more sophisticated approach to assess glucose metabolism.

The structure of this chapter is as follows. First, we briefly describe the AEGIS study. We then formally introduce the concept of glucodensity, the estimation methods, and some essential statistical background to understand the statistical procedures introduced in the chapter. Subsequently, we explain the regression models used in the validation of the representation. Afterwards, we show the results that demonstrate the superiority of glucodensity over glucose

	Men ($n = 220$)	Women ($n = 361$)
Age, years	47.8 ± 14.8	48.2 ± 14.5
A1c, %	5.6 ± 0.9	5.5 ± 0.7
FPG, mg/dL	97 ± 23	91 ± 21
HOMA-IR, mg/dL. μ IU/mL	3.97 ± 5.56	2.74 ± 2.47
BMI, kg/m ²	28.9 ± 4.7	27.7 ± 5.3
CONGA, mg/dL	0.88 ± 0.40	0.86 ± 0.36
MAGE, mg/dL	33.6 ± 22.3	31.2 ± 14.6
MODD	0.84 ± 0.58	0.77 ± 0.33

Table 3.1: Characteristics of AEGIS study participants with CGM monitoring by sex. Means and standard deviations are shown. A1c: glycated haemoglobin; FPG, fasting plasma glucose; HOMA-IR, homeostasis model assessment-insulin resistance; BMI, body mass index; CONGA, glycemic variability in terms of continuous overall net glyceemic action; MAGE, mean amplitude of glyceemic excursions; MODD, mean of daily difference.

representations in the state of the art. Then, we illustrate the use with real data of the glucodensities methodology in two-sample testing and cluster analysis. Finally, we discuss the new perspectives opened by the use of glucodensities.

3.1 Sample and procedures

3.1.1 Study design

A subset of the subjects in the A Estrada Glycation and Inflammation Study (AEGIS; trial NCT01796184 at www.clinicaltrials.gov) provided the sample for the present work. In the latter cross-sectional study, an age-stratified random sample of the population (aged ≥ 18) was drawn from Spain's National Health System Registry. A detailed description has been published elsewhere [101]. For a one-year period beginning in March, subjects were periodically examined at their primary care center where they: i) completed an interviewer-administered structured questionnaire; ii) provided a lifestyle description; iii) were subjected to biochemical measurements, and iv) were prepared for CGM (lasting 6 days). The subjects who made up the present sample were the 581 (361 women, 220 men) who completed at least 2 days of monitoring, out of an original 622 persons who consented to undergo a 6-day period of CGM. Another 41 original subjects were withdrawn from the study due to non-compliance with protocol demands ($n = 4$) or difficulties in handling the device ($n = 37$). The characteristics of the participants are shown in the Table 3.1.

3.1.2 Ethical approval and informed consent

The present study was reviewed and approved by the Clinical Research Ethics Committee from Galicia, Spain (CEIC2012-025). Written informed consent was obtained from each participant in the study, which conformed to the current Helsinki Declaration.

3.1.3 Laboratory determinations

Glucose was determined in plasma samples from fasting participants by the glucose oxidase peroxidase method. A1c was determined by high-performance liquid chromatography in a Menarini Diagnostics HA-8160 analyzer; all A1c values were converted to DCCT-aligned values [112]. Insulin resistance was estimated using the homeostasis model assessment method (HOMA-IR) according to [182].

3.1.3.1 Glycaemic variability

Glycaemic variability was measured in terms of continuous overall net glycemc action (CONGA) [183], the mean amplitude of glycaemic excursions (MAGE) [246], and the mean of the daily differences (MODD) [188] in glucose concentration.

3.1.3.2 CGM Procedures

At the start of each monitoring period, a research nurse inserted a sensor (Enlite™, Medtronic, Inc, Northridge, CA, USA) subcutaneously into the subject's abdomen and instructed him/her in the use of the iPro™ CGM device (Medtronic, Inc, Northridge, CA, USA). The sensor continuously measures the interstitial glucose level 40-400 (range mg/dL) of the subcutaneous tissue, recording values every 5 min. Participants were also provided with a conventional OneTouchR VerioR Pro glucometer (LifeScan, Milpitas, CA, USA) as well as compatible lancets and test strips for calibrating the CGM. All subjects were asked to make at least three capillary blood glucose measurements (usually before the main meals). These readings were taken without checking the current CGM reading. The sensor was removed on the seventh day, and the data downloaded and stored for further analysis. If the number of data-acquisition "skips" per day totaled more than 2 h, the entire day's data were discarded.

3.1.4 Time-in-range metric

The time-in-range metric was calculated with two different methods. First, we estimate the deciles of CGM records with normoglycemic patients and use the deciles as cut-offs (Table 3.2). Second, we use the cut-off points established by the ADA in the 2019 Medical guideline [20] (Table 3.3).

3.2 Definition and estimation of the glucodensity

Suppose that a random sample of n patients is available. For patient i , denote the gathered glucose monitoring data by pairs (t_{ij}, X_{ij}) , $j = 1, \dots, m_i$, where the t_{ij} represent recording times that are typically equally spaced across the observation interval, and X_{ij} is the glucose

Range 1	< 85
Range 2	85 – 90
Range 3	91 – 94
Range 4	95 – 98
Range 5	99 – 101
Range 6	102 – 105
Range 7	106 – 109
Range 8	110 – 115
Range 9	116 – 124
Range 10	> 125

Table 3.2: Cut-offs for time-in-range metrics using own estimations throught normoglycemic individuals of AEGIS study

Range 1	< 54
Range 2	54 – 69
Range 3	70 – 180
Range 4	181 – 250
Range 5	> 250

Table 3.3: Cut-offs for time-in-range metrics following ADA guidelines [20]

level at time $t_{ij} \in [0, T_i]$. Note that the number of records m_i , the spacing between them, and the overall observation length T_i can vary by patient. One can think of these data as discrete observations of a continuous latent processes $Y_i(t)$, with $X_{ij} = Y_i(t_{ij})$. The glucodensity for this patient is defined in terms of this latent process as $f_i(x) = F'_i(x)$, where

$$F_i(x) = \frac{1}{T_i} \int_0^{T_i} \mathbf{1}(Y_i(t) \leq x) dt, \tag{3.1}$$

$$\text{for } \inf_{t \in [0, T_i]} Y_i(t) \leq x \leq \sup_{t \in [0, T_i]} Y_i(t), \tag{3.2}$$

is the proportion of the observation interval in which the glucose levels remain below x . Since F_i are increasing from 0 to 1, the data to be modeled are a set of probability density functions f_i , $i = 1, \dots, n$.

Of course, neither F_i nor the glucodensity f_i is observed in practice, but one can construct an approximation through a density estimate $\tilde{f}_i(\cdot)$ obtained from the observed sample. In the case of CGM data, the glucodensities may have different support and shape. Therefore, we suggest using a non-parametric approach to estimate each density function. For example, using a kernel-type estimator, we have

$$\tilde{f}_i(x) = \frac{1}{m_i} \sum_{j=1}^{m_i} \mathbb{K}_{h_i}(x - X_{ij}),$$

where $h_i > 0$ is the smoothing parameter and $\mathbb{K}_{h_i}(s) = \frac{1}{h_i} \mathbb{K}(\frac{s}{h_i})$. The choice of K does not have a big impact on the efficiency of the estimator, but the value of h_i is crucial [193].

In the standard setting of independent random samples, a vast number of approaches for selecting the smoothing parameter are available in the literature. Common strategies include cross-validation, minimizing the estimated mean integrated squared error (MISE), or a “rule of thumb” derived from the assumption that the density is Gaussian. In this last case, the choice can be explicitly written as $\tilde{h}_i = 1.06\tilde{\sigma}_i m_i^{-1/5}$, where $\tilde{\sigma}_i$ is the sample standard deviation of the X_{ij} [249].

In our particular setup, we are estimating the density function of a stochastic process/time series, which is more difficult in theory. However, in a seminal work in this area [106], authors have shown that the rule of thumb and other traditional smoothing parameter selection strategies behave well. Additionally, the number of density function estimators that exists is considerable, and we can also employ other approaches as the use of orthogonal expansions (e.g., Fourier or Wavelet basis), splines, and histograms. For further details, the reader is referred to [14, 124, 193].

3.2.1 Distance-based Descriptive Statistics

Let $[a, b]$ be an interval of the real line, which may be unbounded, and suppose that each glucodensity f_i has support contained in $[a, b]$. From a statistical point of view, the sample f_1, \dots, f_n may be modeled and analyzed using methods of functional data analysis [226, 288]. However, since the f_i must be positive and satisfy $\int_a^b f_i(x)dx = 1$, classical methods have in recent years been adapted to account for the nonlinear, distributional structure of density samples [119, 215]. The general approach is to define a metric or distance between densities that, in turn, leads to descriptive statistics that respect the unique density properties. For example, define the data space of glucodensities as $A := \{f : [a, b] \rightarrow \mathbb{R}^+ : \int_a^b f(x)dx = 1 \text{ and } \int_a^b x^2 f(x)dx < \infty\}$. Given two arbitrary glucodensities $f, g \in A$, the 2-Wasserstein distance [283] between f and g is

$$d_{\mathcal{W}_2}(f, g) = \sqrt{\int_0^1 (F^{-1}(x) - G^{-1}(x))^2 dx}, \quad (3.3)$$

where F and G are the cumulative distribution functions (cdfs) of the density functions f and g .

The 2-Wasserstein distance is a natural distance to measure the similarity between density functions through its representation in the space of the quantile (inverse cdf) functions, and it has already been successfully applied in biological problems [218]. Furthermore, it has computational and modeling advantages compared to the usual $L^2[a, b]$ metric when glucodensities

have different support within $[a, b]$. Finally, it has a physical interpretation in the theory of optimal transport [9, 283].

As glucodensities are distributional data, the subsequent application of the usual techniques for functional data, such as estimation of mean, covariance, and regression models, may lead to misleading results. Hence, we have chosen to use models based on the 2-Wasserstein distance, although other choices are possible. As a starting point, based on the notion of distance we can generalise the mean and variance of a random variable that takes values in an abstract space with metric structure [82]. As we will see, similar adaptations can be developed for regression, hypothesis testing, or to perform cluster analysis. Given a distance $d : A \times A \rightarrow \mathbb{R}^+$ between density functions, of which d_{W_2} is one example, and a random variable f defined on A , the *Fréchet mean* of f is

$$\mu_f = \arg \min_{g \in A} E(d^2(f, g)).$$

The *Fréchet variance* of f is then

$$\sigma_f^2 = E(d^2(f, \mu_f)).$$

If the choice of distance is the Wasserstein metric d_{W_2} , these are given the names of Wasserstein mean and variance, respectively. In this particular case, (3.3) implies that μ_f is the density whose quantile function is the pointwise mean of the random quantile function F^{-1} . Moreover, σ_f^2 is interpreted as the integral of the pointwise variance of F^{-1} . In general, calculation of the Fréchet mean is not easy, and we must resource to computational approximations [205].

In the following subsections, we will extend this notions to statistical methods for regression, clustering, and hypothesis testing.

3.3 Regression models with glucodensities

3.3.1 Non-parametric regression with glucodensity as the predictor

Let f be a functional random variable taking values in (A, d_{W_2}) and Y a random variable that takes values in the real line. We assume the following regression relationship between f and Y , which represent the predictor and response variables, respectively:

$$Y = m(f) + \epsilon \tag{3.4}$$

where $m : A \rightarrow \mathbb{R}$ is an unknown smooth function, and the random error ϵ satisfies $E(\epsilon) = 0$.

Given a sample $\{(f_i, Y_i) \in A \times \mathbb{R}\}_{i=1}^n$, most non-parametric estimators $\tilde{m}(\cdot)$ have the form of a weighted average of the responses

$$\tilde{m}(x) = \sum_{i=1}^n w_{ni}(x) Y_i. \tag{3.5}$$

In general, the weights $w_{ni}(x)$ depend on the distance selected to measure the similarities between the density functions f_i and x , with larger distances receiving lower weights, and satisfy $\sum_{i=1}^n w_{ni}(x) = 1$ [76]. A typical choice would be the Nadaraya–Watson weights

$$w_{ni}(x) = \frac{\mathbb{K}\left(\frac{d(x, f_i)}{h}\right)}{\sum_{i=1}^n \left(\mathbb{K}\left(\frac{d(x, f_i)}{h}\right)\right)}, \quad (3.6)$$

where h is a smoothing parameter and $\mathbb{K} : \mathbb{R} \rightarrow \mathbb{R}$ is a known univariate probability density function called the kernel. For more details about this procedure see [76].

3.3.2 Regression with glucodensity as the response

In the case of the regression methods with a density function as response, the literature is not very extensive to the current date [38, 107, 198, 216, 265]. In this chapter, we use the method proposed in [216] which allows us to incorporate the desired metric d_{W_2} and is a direct generalization of classical linear regression. The primary rationale for our use of this approach is that, unlike the other approaches mentioned above, there is a methodology developed to perform inferential procedures such as confidence bands and hypothesis testing in order to establish the significance of the input variables in the model [213].

Let f be a random variable (e.g. a glucodensity) that takes values in the space of (A, d_{W_2}) defined above. Consider a random vector $U \subset \mathbb{R}^d$ that contains the set of predictors. Our interest is in the Fréchet regression function, or function of conditional Fréchet means (see chapter 2.2.3 for more details),

$$\tilde{m}(u) := \arg \min_{g \in A} E(d_{W_2}^2(f, g) | U = u), \quad u \in \mathbb{R}^d. \quad (3.7)$$

The approach in [216] imposes a particular model for \tilde{m} that, in direct analogy to classical linear regression, takes the form of a weighted Fréchet mean:

$$\tilde{m}(u) = \arg \min_{g \in A} E(s(U, u) d_{W_2}^2(f, g)), \quad u \in \mathbb{R}^d. \quad (3.8)$$

Here, the weight function is

$$s(U, u) = 1 + (U - \mu)^T \Sigma^{-1} (u - \mu), \quad \mu = E(U), \Sigma = \text{Cov}(U), \quad (3.9)$$

and Σ is assumed to be positive definite.

Given a sample (U_i, f_i) , $i = 1, \dots, n$, of independent pairs each distributed as (U, f) , one can proceed to estimate $\tilde{m}(u)$ for any desired input u . Due to the intimate connection between the Wasserstein metric and quantile functions as in (3.3), for most inferential procedures it is sufficient to estimate the conditional Wasserstein mean quantile function $\bar{Q}(u)$ corresponding

to $\bar{f}(u)$. Let D be the set of quantile functions, Q_i the quantile function corresponding to the random density f_i , and define empirical weights $s_{in}(u) = 1 + (U_i - \bar{U})^T \tilde{\Sigma}^{-1}(u - \bar{U})$, where \bar{U} and $\tilde{\Sigma}$ are the sample mean and variance of the U_i , respectively. The natural estimator under d_{W_2} is the weighted empirical mean quantile function

$$\tilde{Q}(u) = \arg \min_{Q \in D} \sum_{i=1}^n s_{in}(u) \|Q - Q_i\|^2, \quad (3.10)$$

where $\|\cdot\|$ denotes the $L^2[0,1]$ norm on D .

A straightforward algorithm for computing $\tilde{Q}(u)$ is shown in [213]. In addition, two algorithms are given to estimate the confidence bands at a given significance level α for both the quantile functional parameter $\bar{Q}(\cdot)$ and the density $m(\cdot)$.

3.3.3 Density estimation and software details

The density function of each individual was estimated with a non-parametric Nadaraya-Watson procedure. For this purpose, we used a Gaussian kernel and rule of thumb as a smoothing parameter. As some computations involving the 2-Wasserstein metric only require a quantile function as input, these were estimated using the empirical quantile function of the observations.

Concerning prediction, the two regression methods previously described were used in glucodensity validation: i) The non-parametric kernel functional regression model with the 2-Wasserstein distance having the glucodensity as predictor [76]; and ii) A global 2-Wasserstein regression model where the glucodensity is the response [216]. In addition, with standard vector-valued time-in-range metrics, k -nearest neighbor algorithms were employed with $k = 10$ neighbors. These time-in-range metrics we first transformed using the isometric log-ratio (ilr) transformation for compositional data prior to fitting the model. In order to avoid problems associated with zero values in any of these predefined ranges, a fixed positive constant was added to each range, which were then normalized to add to 1.

All analyses were carried out using R software. Functional data analysis was performed using the `fda.usc` package [72], which is freely available at <https://cran.r-project.org/>. In the case of the ANOVA test of [63] as well as Fréchet regression in [216] using the 2-Wasserstein distance, we use our own implementations. The glucodensities and their quantile representation were estimated using the R basis functions.

3.4 Clinical validation of the glucodensity

To validate the glucodensity representation, we use the AEGIS database [101]. This database contains the continuous glucose monitoring data between 2-6 days of 581 patients from a

Biomarker	Clinical significance
A1c	Gold standard marker in diabetes diagnosis and control
HOMA-IR	Measurements to quantify insulin resistance and β -cell function
CONGA MODD MAGE	Summary indices of glucose variability

Table 3.4: Clinical importance of biomarkers used in the statistical analysis

general population's random sample. To develop the validation task, we use two different regression models: i) a non-parametric regression model where the unique predictor is glucodensity, and ii) a linear regression model where the response is a glucodensity. The first model was used to predict glycated hemoglobin (A1c) [133], homeostatic model assessment (HOMA-IR) [19], and the following measures of glycemic variability [101, 189, 245]: continuous overall net glycemic action (CONGA), mean amplitude of glycemic excursions (MAGE) and mean of daily differences (MODD), through glucodensity representation. In contrast, the second was used to predict the glucodensity with the five variables above. Figure 3.1 gives a visualization of the sample of glucodensities used in these models. Biological significance in variables under consideration is described in Table 3.4.

3.4.1 Prediction of biomarkers using the glucodensity

The aim of the first set of regression analyses is to demonstrate that the glucodensity is sufficiently rich in its information content to recover the biomarkers mentioned above with high precision. To quantify this precision, we estimated the R^2 after fitting a non-parametric model for each biomarker as the outcome variable, using the glucodensity as the sole predictor (i.e., independent variable). The R^2 estimates for A1c, HOMA-IR, MAGE, MODD, CONGA were 0.79, 0.79, 0.92, 0.86, and 0.92 respectively. To supplement the results, Figure 3.2 shows the predicted values against the observed values, where the outstanding predictive capacity of the glucodensity can be seen independently of high or low response values.

3.4.2 Prediction of the glucodensity using biomarkers

In the second regression analysis with the glucodensity as the outcome variable, we aim to show that the previous measurements commonly used in the clinical practice cannot capture the glucodensity with high accuracy. This fact is not completely surprising because, as noted by some authors [295], the information provided by a CGM is more precise than that contained in summary measures. To accomplish this, we computed a suitable version of R^2 for this task

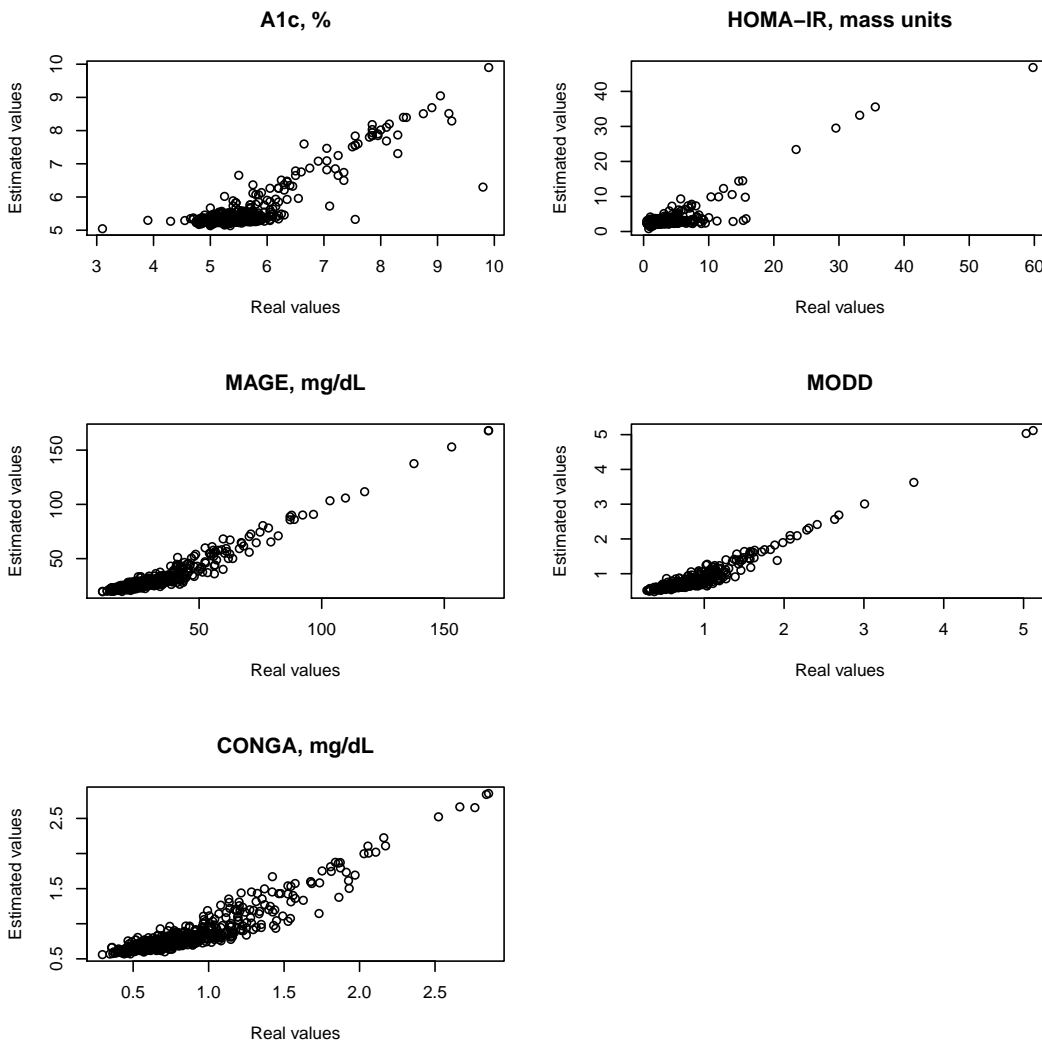


Figure 3.2: Real values vs estimated values when glucodensity is predictor.

after fitting a regression model where the response is a glucodensity, and the previous variables are the predictors. In this case, the R^2 estimated was 0.74. As predicted, compared to the previous section's results, we could not accurately capture the complex nature of glucodensities, even while using the combined predictive power of several commonly used summary measures. Moreover, in some cases, the prediction differences can be significant (see Figure 3.3).

3.4.3 Comparison of time-in-range metrics with glucodensities

To illustrate the higher clinical sensitivity of glucodensities compared to time-in-range metrics, we compared each representation's ability to predict A1c, HOMA-IR, and glycemic variability metrics MODD, MAGE, and CONGA, using the data from the AEGIS study. The predictive capacity of the glucodensity representation was illustrated above, and this section gives the corresponding results for time-in-range metrics, where these were calculated according to two

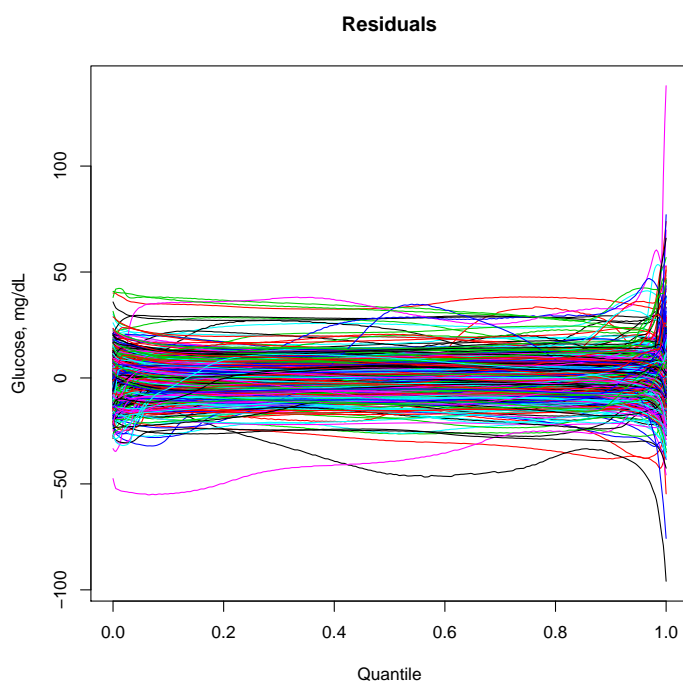


Figure 3.3: Residuals in quantile space.

	A1c	HOMA-IR	CONGA	MAGE	MODD
Normoglycemic cut-off	0.63	0.22	0.68	0.65	0.65
ADA cut-off	0.61	0.08	0.73	0.69	0.60

Table 3.5: R^2 estimated with time-in-range metrics under consideration

sets of cut-offs. In the first, the normoglycemic individuals' deciles from the AEGIS study were used, while those proposed by the ADA were used in the second. Tables 3.2 and 3.3 show the exact cutoff values for both cases. Since the time-in-range metrics constitute a sample of compositional data [207], the isometric log-ratio (ilr) transformation was employed in combination with a k -nearest neighbor algorithm as a regression model for predicting the scalar variables.

Figure 3.4 compares the real and estimated values of the previous five variables under the two time-in-range metrics under consideration. Table 3.5 provides the estimates of R^2 for each variable and metric.

The predictive capacity is significantly worse than that attained by the glucodensity methodology. The superiority of the glucodensity is particularly noteworthy in the case of the HOMA-IR variable, where the association is relatively weak for time-in-range metrics. Even for the other variables where the values of R^2 are moderate, the larger residuals seen in diabetes patients with more severe alterations of glucose metabolism indicate that time-in-range metrics are particularly poorly suited for such patients. Interestingly, we do not observe substantial or

consistent differences between the two time-in-range metrics used, as deciles perform better than ADA criteria for two of the variables, while the ordering was reversed in other instances.

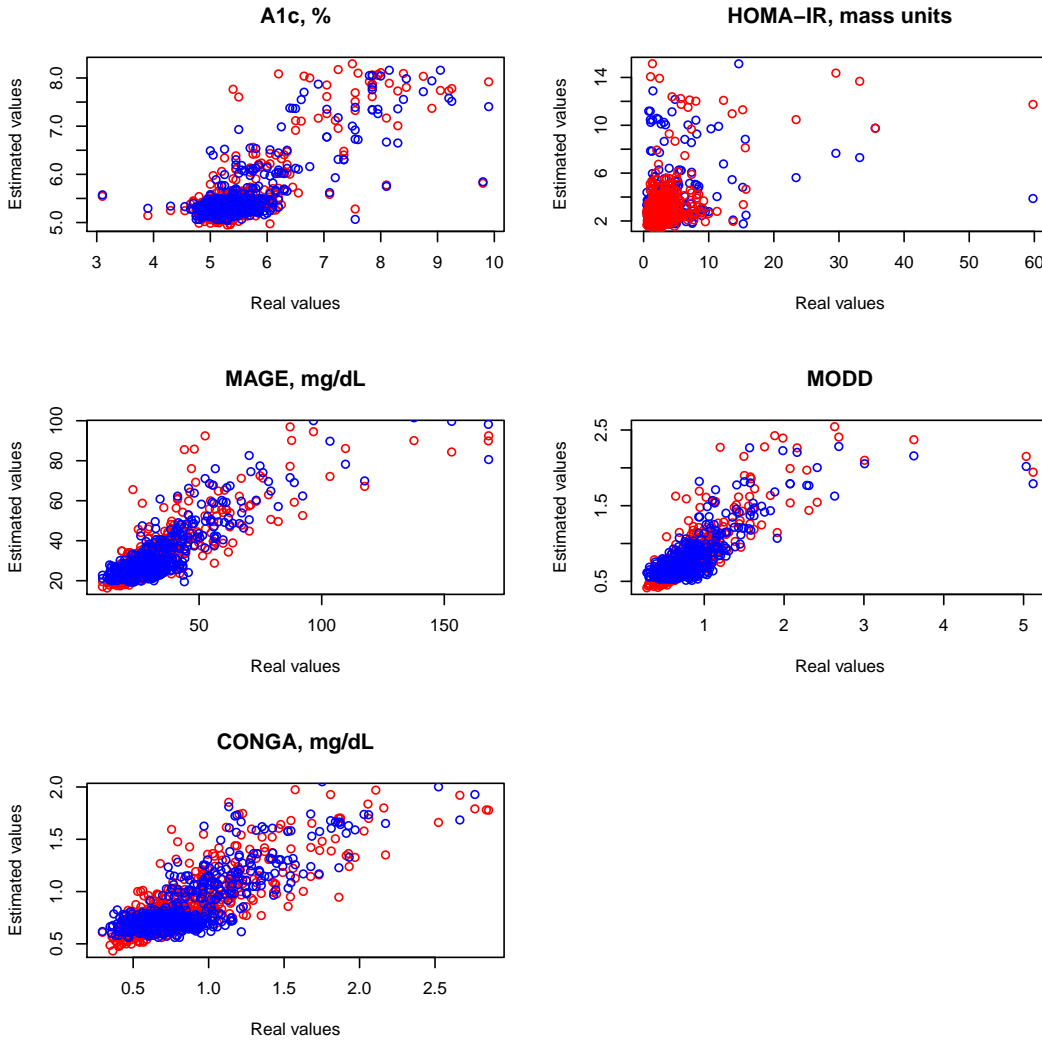


Figure 3.4: Real values vs. estimated values when time-in-range metric is the predictor. (Blue) time-in-range metric with cut-offs calculated with normoglycemic patients of AEGIS database. (Red) time-in-range metric using the cut-offs suggested by ADA.

3.5 Hypothesis testing and clustering analysis with glucodensities

3.5.1 Analysis of variance

As a special case of regression, suppose we have a sample f_1, \dots, f_n of glucodensities defined on (A, d_W) belonging to k different groups G_1, G_2, \dots, G_k that partition $\{1, \dots, n\}$ each of them of size n_j ($j = 1, \dots, k$), so that $\sum_{j=1}^k n_j = n$. If the goal is to simply test whether the Wasserstein means are equal for each group, [213] developed testing procedures based on model (3.8) for this purpose. An advantage of this model is its flexibility, which allows for

multiple factor layouts as well as tests for interactions. However, the theoretical properties of these tests require a type of equal variance assumption that may be restrictive for some data sets.

More generally, one may wish to test the null hypothesis that the population distributions of the k groups share common Wasserstein means and variances, against the alternative that at least one of the groups has a different population distribution compared to the others in terms of either its Wasserstein mean or variance. In this scenario, [63] investigated a test statistic based on the group proportions $\lambda_{j,n} = n_j n^{-1}$, the groupwise sample Wasserstein means $\tilde{\mu}_j = \arg \min_{g \in A} \sum_{i \in G_j} d_{W_2}^2(f_i, g)$ and variances $\tilde{\sigma}_j^2 = n_j^{-1} \sum_{i \in G_j} d_{W_2}^2(f_i, \tilde{\mu}_j)$, the pooled Wasserstein mean $\hat{\mu}_p = \arg \min_{g \in A} \sum_{j=1}^k \sum_{i \in G_j} d_{W_2}^2(f_i, g)$ and variance $\tilde{\sigma}_p^2 = n^{-1} \sum_{j=1}^k \sum_{i \in G_j} d_{W_2}^2(f_i, \hat{\mu}_p)$, and finally the quantities

$$\widetilde{Var}(\tilde{\sigma}_j^2) = \frac{1}{n_j} \sum_{i \in G_j} d_{W_2}^4(f_i, \tilde{\mu}_j) - \left\{ \frac{1}{n_j} \sum_{i \in G_j} d_{W_2}^2(f_i, \tilde{\mu}_j) \right\}^2$$

as estimates of the variance of $\tilde{\sigma}_j^2$.

Then, with

$$F_n = \tilde{\sigma}_p^2 - \sum_{j=1}^k \lambda_{j,n} \tilde{\sigma}_j^2, \quad R_n = \sum_{j < l} \frac{\lambda_{j,n} \lambda_{l,n}}{\tilde{\sigma}_l^2 \tilde{\sigma}_j^2} (\tilde{\sigma}_j^2 - \tilde{\sigma}_l^2),$$

the proposed test statistic is

$$T_n = \frac{nR_n}{\sum_{j=1}^k \frac{\lambda_{j,n}}{\widetilde{Var}(\tilde{\sigma}_j^2)}} + \frac{nF_n^2}{\sum_{j=1}^k \lambda_{j,n}^2 \widetilde{Var}(\tilde{\sigma}_j^2)}. \quad (3.11)$$

Dubey and Müller [63] demonstrated that the corresponding test is distribution-free, in that the limiting distribution of T_n does not depend on the underlying distribution under some assumptions. In practice, it was also demonstrated that it could be useful to calibrate the test under the null hypothesis via a simple empirical bootstrap over the preceding statistics. For formal details, we refer the reader to [63].

3.5.2 Energy distance methods with glucodensities

The energy distance is a statistical distance between two distribution functions proposed in 1984 by Gábor J. Székely [263]. This distance is inspired by the concept of gravitational energy between two bodies and has experienced a rise in appeal for modern statistical applications due to its applicability to data of a complex nature such as functions, graphs, or objects that live in negative type space.

Consider independent random variables $Y, Y' \sim F$ and $Z, Z' \sim G$ that are defined on a (semi)metric space (Ω, ρ) of negative type, where $\rho : V \times V \rightarrow \mathbb{R}$ is the semi-metric. Though the notation in this section is quite general, in particular we have in mind the case $(\Omega, \rho) = (A, d_{\mathcal{W}_2})$ corresponding to glucodensities. Let us recall that the energy distance associated with ρ between the distribution F and G is

$$\epsilon_\rho(F, G) = 2E(\rho(Y, Z)) - E(\rho(Y, Y')) - E(\rho(Z, Z')).$$

Given random samples $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} F$ and $Z_1, \dots, Z_m \stackrel{\text{iid}}{\sim} G$, the sample energy distance is

$$\begin{aligned} \tilde{\epsilon}_\rho(F, G) = & \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m \rho(Y_i, Z_j) - \frac{1}{n^2} \sum_{i=1}^n \sum_{i=1}^n \rho(Y_i, Y_j) \\ & - \frac{1}{m^2} \sum_{i=1}^m \sum_{i=1}^m \rho(Z_i, Z_j). \end{aligned}$$

The asymptotic distribution of the above statistics for a null hypothesis ($H_0 : F = G$) as well as for the alternative ($H_a : F \neq G$) is dependent on the chosen semi-metric ρ . Besides, its expression is difficult to calculate and to implement in practice. Hence, when using the energy distance based methods, the distribution under the null hypothesis is usually calibrated with a permutation method. Alternatives to calibrate the distribution under the null hypothesis include the wild or a weighted bootstrap, as described in [127, 151]. The energy distance can also be extended to handle samples from more than two populations. Given k independent samples $Y_{j1}, \dots, Y_{jn_j} \stackrel{\text{iid}}{\sim} F_j$, $j = 1, \dots, k$, the energy distance statistic is

$$\begin{aligned} \tilde{\epsilon}_\rho(F_1, \dots, F_k) & \sum_{1 \leq j < l \leq k} \frac{n_j n_l}{2n} [2g_{jl} - g_{jj} - g_{ll}], \\ g_{jl} & = \frac{1}{n_j n_l} \sum_{i=1}^{n_j} \sum_{i'=1}^{n_l} \rho(Y_{ji}, Y_{li'}), \end{aligned}$$

where $n = n_1 + \dots + n_k$.

We now explain how this statistic can be adapted to perform clustering. Consider random pairs (Y_i, I_i) , $i = 1, \dots, n$, where Y_i is observed and takes values in (Ω, ρ) , while $I_i \in \{1, \dots, k\}$ is an unobserved label of cluster membership. The task is to recover the true clusters $C_j^* = \{i : I_i = j\}$, $j = 1, \dots, k$. Let C_1, \dots, C_k be a generic partition of $\{1, \dots, n\}$, and denote the size of each cluster by $|C_j|$. Then a clustering may be chosen by optimizing the statistic

$$S_\rho(C_1, \dots, C_k) = \sum_{1 \leq j < l \leq k} \frac{n_j n_l}{2n} [2\tilde{g}_{jl} - \tilde{g}_{jj} - \tilde{g}_{ll}], \quad (3.12)$$

$$\tilde{g}_{jl} = \frac{1}{|C_j| |C_l|} \sum_{(i, i') \in C_j \times C_l} \rho(Y_i, Y_{i'}) \quad (3.13)$$

over all possible clusters C_j . At first view, this seems computationally intractable due to the appearance of distances between the elements of each cluster. However, defining

$$W_\rho(C_1, \dots, C_k) = \sum_{j=1}^k \frac{|C_j|}{2} \tilde{g}_{jj}, \quad (3.14)$$

it can be proven that $S_\rho + W_\rho$ is constant. This implies that maximizing S_ρ is equivalent to minimizing W_ρ .

More specifically, in [80], the authors show the equivalence between the previous energy distance optimization problem with the kernel k -means optimization problem. This equivalence allows us to solve them by applying popular heuristics algorithms, such as Hartigan's and Lloyd's ones.

3.5.3 Example of hypothesis testing and clustering analysis

Below, we illustrate the methodology of glucodensities in hypothesis testing and cluster analysis with the 2-Wasserstein distance. We use the ANOVA test [63] and the k -groups algorithm [80].

An interesting question to address in an epidemiological study is whether there are differences between men and women in the glycemic profile. The ANOVA test is an important instrument to establish whether there are statistically significant differences in mean and variance with glucodensities, where there are two or more patient groups. After applying this method with AEGIS data, the test yields a p-value equal to 0.10. Therefore, there is no statistically significant difference between men and women at the significance level of 5 percent.

Figure 3.5 shows the glucodensity samples for each gender using their quantile representations. The pointwise means of these quantile functions constitute the quantile function of the sample Wasserstein mean glucodensities. These, together with pointwise standard deviation curves, are also shown in Figure 3.5. On average, the groups are quite similar. However, certain discrepancies are observed between both groups in terms of their variance, although not large enough for the test to show statistically significant differences.

Cluster analysis is an essential tool for identifying subgroups of patients with similar characteristics. As an example, with the diabetes patients' data from the AEGIS study, we perform a cluster analysis using three clusters. To establish when a patient has diabetes, we use the doctor's previous diagnostic criteria, or if individuals currently have their glucose values measured with A1c and FPG in the ADA ranges to be classified in that category.

Figure 3.6 contains the results of applying the cluster analysis in diabetes patients. The algorithm has identified three differentiated groups of patients. The first group comprises patients with normal glucose values, probably because they are on medication and the diagnosis of diabetes was made in the past. The second group comprises patients with slightly altered

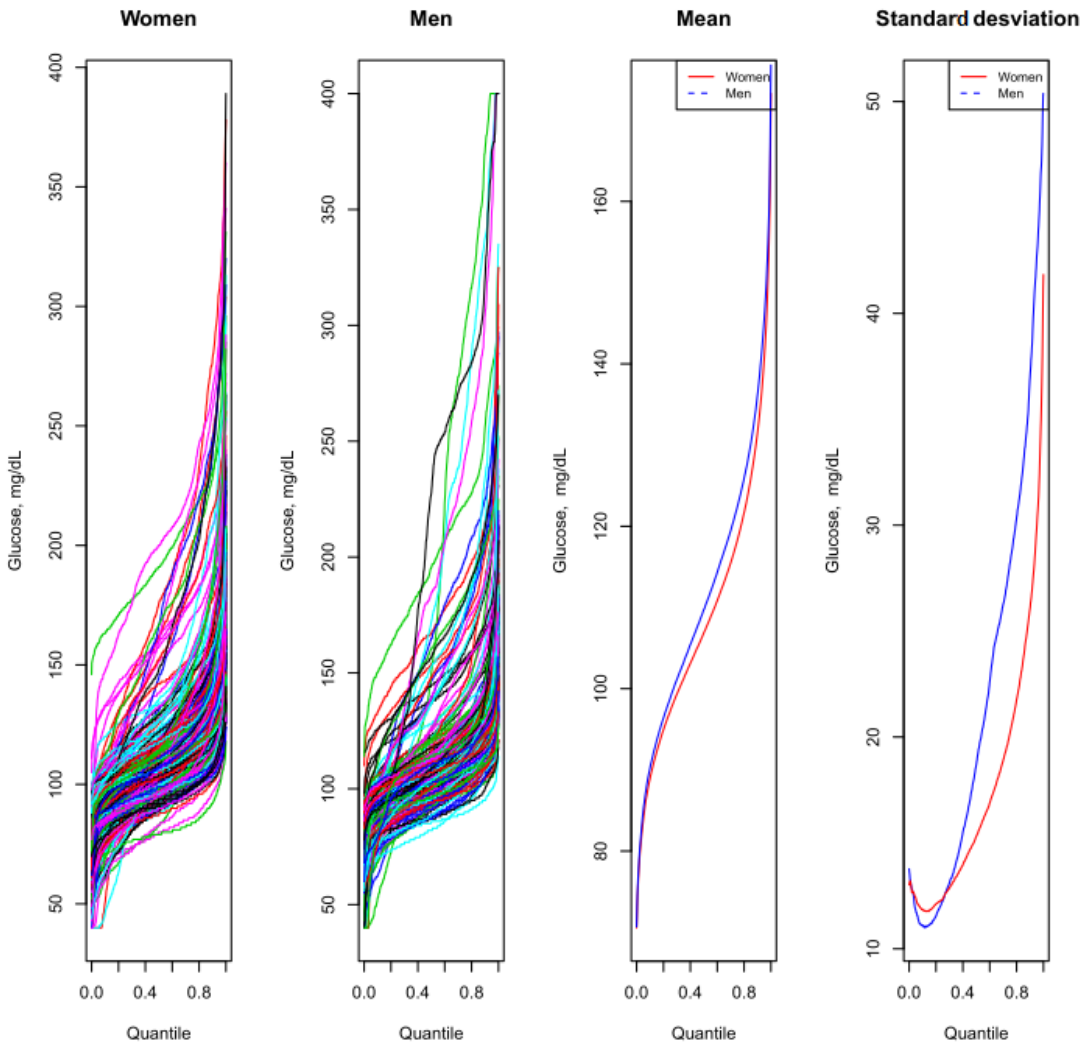


Figure 3.5: (Left two panels) Glucodensities for women and men of the AEGIS study, plotted as quantile functions; (Third panel) 2-Wasserstein mean quantile functions for each group; (Fourth Panel) Cross-sectional standard deviation curves for quantile functions in each group.

diabetes metabolism. Finally, the last group comprises patients with severely altered glucose values, and as can be seen in the glucodensities, their glucose is continuously fluctuating. The two-dimensional graphical representation of the density function of A1c and FPG helps to validate these findings.

3.6 Discussion

The primary contribution of this chapter is to propose a new representation of CGM data called glucodensity. We have validated this representation from a clinical point of view, proving that it is more accurate than the previous time-in-range metrics.

3.6.1 Diabetes etiology and biological components to capture

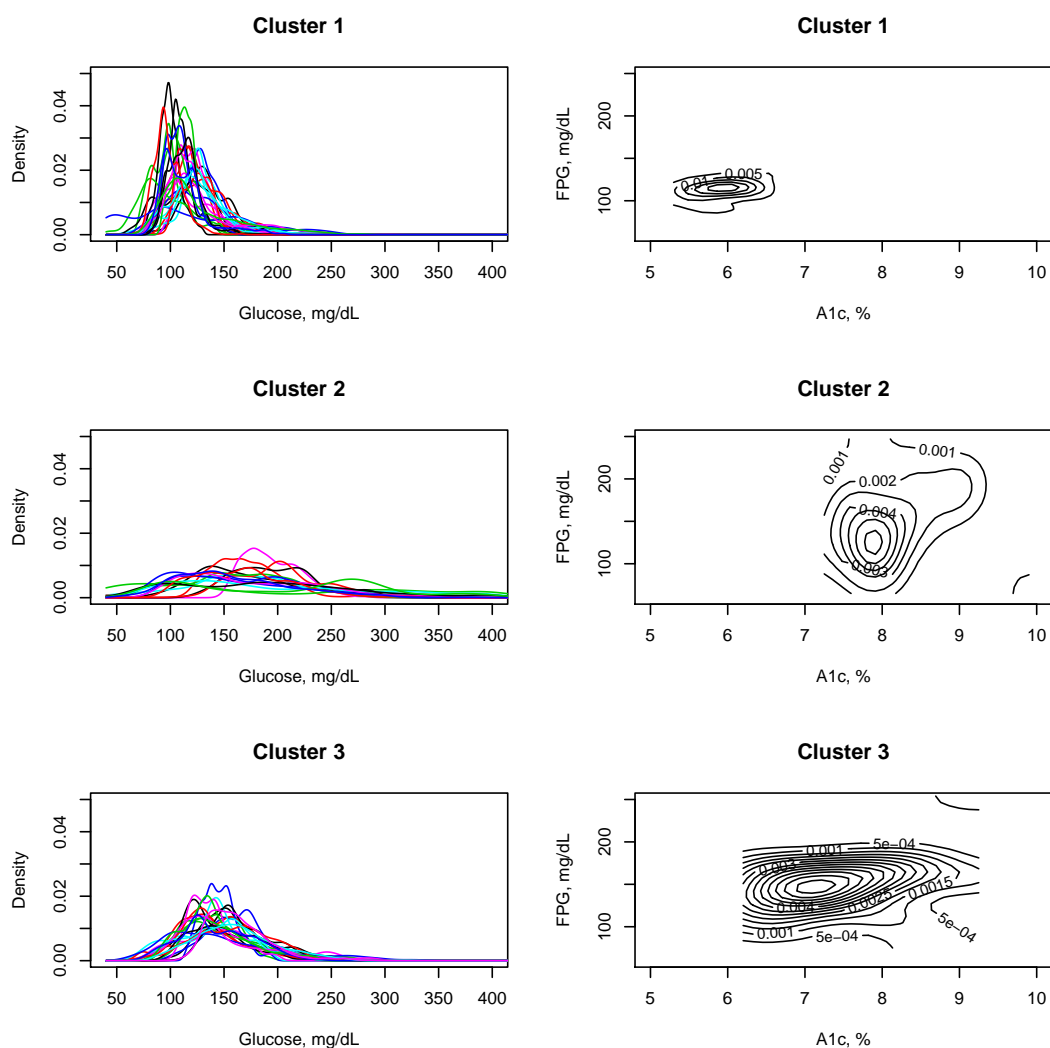
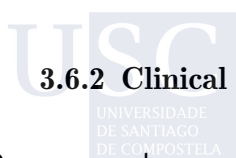


Figure 3.6: Clustering analysis of diabetes patients in AEGIS study

Diabetes encompasses a heterogeneous group of impaired glucose metabolism, such as the frequent presence of hyperglycemia or hypoglycemia [17]. Anomalous glucose fluctuations are another essential trait of dysglycemic regulation [189, 190]. Accordingly, the use of glycemic measures that capture other aspects beyond the average, particularly: i) the impact of time spent along the whole spectrum of glucose concentrations on the glucose deregulation process, and ii) the oscillations of glucose concentration associated with cellular damage [190], is crucial in the management of patients with diabetes, as well as in the assessment of glucose metabolism with a high degree of precision.



3.6.2 Clinical validation of glucodensity

Our proposal accurately captures the components of diabetes mentioned above. Using clinical data, we evaluated the clinical sensitivity against established biomarkers in diabetes. We found

a high association between A1c, HOMA-IR, CONGA, MODD, MAGE, and glucodensity. In the case of the HOMA-IR variable, the predictive ability does not seem excellent, although, to the best of our knowledge, no known marker shows a predictive ability against that variable. Still, our model can provide consistent values in moderate and large HOMA-IR values. While the fit for the variable A1c was not perfect, we must consider that the time scale for the A1c and the glucodensities were quite different. A1c is a measure that reflects the average glucose over 2-3 months while patients undergo CGM for less than 1 week. Our R^2 of 0.79 is better than the average glucose recorded by the monitoring period ($R^2=0.61$), which indicates that an individual's glucose distributional values may give extra information to the long-term glucose averages.

In the prediction of glucodensity from A1c, HOMA-IR, and glycemic variability measures, the estimated R^2 shows a moderate relationship between those variables. However, we are introducing the essential variables of the glucose deregulation process. A possible explanation of this is that the use of the summary measures commonly used in diabetes can hardly capture an individual's glycemic profile. Glucose metabolism is very complex and highly dependent on the patient's conditions. For example, the cellular mechanisms are different in type I and type II diabetes. In the former, there is an inhibition of β -cell function and consequent non-insulin production, while insulin secretion is reduced in the latter [268]. In this context, the introduction of the concept of glucodensity provides greater clinical accuracy to the possible decisions derived from such representation compared to traditional methods because we utilize the entire distribution of glucose concentrations of an individual over time.

3.6.3 Time-in-range metrics vs. glucodensity

While time-in-range metrics may also achieve the previous aim, they do so to a clearly lesser extent than the glucodensity. Our proposal can capture the differences between individuals in each glucose concentration. Notwithstanding, time-in-range only measures glucose differences along intervals with the subsequent loss of information. Also, time-in-range metrics are substantially limited since the target zones must be defined previously, and these may also depend on the study population or the aim of the analysis.

Empirical results demonstrate the advantages of our proposal out of the theoretical framework. The ability of glucodensity to predict A1c, HOMA-IR, and the CONGA, MAGE, and MODD variability measures is surprisingly high, much higher than that achieved with the range metric despite using two different target zones: the deciles of normoglycemic patients glucose values and the target zones prescribed by the ADA.

The estimated R^2 between glucodensities and A1c is similar to that reported by other authors between A1c and average glucose values [197]. However, in this study, patients are monitored only for 2-6 days and not for weeks. Two possible factors should be considered

in the analysis of the results. First, there are both diabetic and non-diabetic patients in our database, and, second, the glucodensity captures A1c better because it represents the entire distribution of glucose concentration values, while glycation rates are known to increase with glucose concentrations [251]. In particular, the estimated R^2 between A1c and the mean glucose in our database is only 0.61.

3.6.4 Statistical considerations

From a statistical standpoint, glucodensities are a special constrained type of functional data known as distributional data; therefore, we cannot use the usual statistical techniques directly. To alleviate this limitation, this paper proposes a framework for the analysis of these distributional data based on distances with existing techniques for hypothesis testing, cluster analysis, and regression models. However, it is essential to remark that alternative approaches are available, including functional transformations [215, 279] which embed the densities in an unconstrained Hilbert Space, after which standard functional analysis techniques can be applied. Nevertheless, these particular transformations cannot be applied directly in our setting due to differences in support of the glucodensities. Moreover, functional transformation has the significant disadvantage that does not allow some standard inferential tasks, such as building a theoretical confidence interval. However, we can address this issue, for example, through resampling techniques with distance methods proposed that exploit 2-Wasserstein geometry [213]. Also, the application of these transformations implies a loss of interpretation; what is the interpretation of an Anova test in a transformed space concerning the original mean function?

3.6.5 Limitations

A potential limitation of our representation is that it ignores the order of events. Instead, it analyzes only the distribution of glucose values. Nevertheless, following different animal models in diabetes, the event sequence may not be a critical component in diabetes modeling. The main factor of microvascular and macrovascular complications is chronic hyperglycemia [48, 253], and this is captured with high accuracy by our models. Moreover, an essential aspect of managing diabetes patients is hypoglycemia control, and our proposal also captures this. Finally, the third component of dysglycemia [189], glucose variability, can be accurately predicted by our representations, at least through metrics CONGA, MAGE, and MODD.

From another point of view, for other authors as Zaccardi and Khunti [295], it is expected that different glucose fluctuations on different time scales may provide extra information on glucose homeostasis. Two extensions of our models could potentially take into account this variability. The first one is to utilize functional multilevel models [56] applied to transformed

glucodensities, using the distributional transformations discussed above. A second approach would be to build similar densities of glucose speed and acceleration values, both marginally and as multivariate functions in the statistical models.

The sample size used may also be a limitation from a statistical point of view. Nevertheless, in the field of diabetes, the AEGIS study is one the world's largest database and, unlike other studies, is composed of randomly selected individuals from a general population [297]. Finally, for study validation, perhaps the most reliable way of validating the new representation is in terms of the patients' long-term prognosis. However, to the best of our knowledge, no study with a reasonable sample size has this information from CGM technology's intensive use. Moreover, the clinical validation carried on was performed from variables associated with the biological and molecular mechanisms of diabetes development, diabetes status, and future diabetes prognosis, as we can see in the literature.

3.6.6 Potential Applications

Adopting the concept of glucodensity in clinical practice and biomedical research could be very promising in the following ways.

- To have a simple and more accurate representation of the glycaemic profile of an individual. This representation is especially useful in managing diabetic patients and assessing the effects of an intervention.
- To establish if there are statistically significant differences between patients subjected to different interventions, for example, in a clinical trial.
- To identify different subtypes of patients based on their glycaemic condition and other variables. Cluster analysis of glucodensities can create new patient subtypes based on the risk of diabetes or other complications. Furthermore, it allows us to better describe the disease's etiology by creating groups of subjects whose glucose profiles and other clinical characteristics are similar.
- To establish the prognosis or risk of a patient or analyze the relationship of an individual's glycemic profile with different clinical variables in epidemiological studies.
- To predict changes in the glycemic profile based on the individuals' characteristics and the intervention performed. For example: how does the glucodensity vary according to the diet?
- To recommend the most advantageous treatments for a patient. Following the previous idea, a causal inference model could be fitted where the response is glucodensity, for

Chapter 3. Distributional representations of biosensor data for continuous stochastic process with application in CGM technology

example, to establish which diet is the most beneficial for the individual to achieve suitable glucose levels.

4 Distributional representations of biosensor data for mixed stochastic processes with application in physical activity analysis

Physical activity level is an influential causal factor associated with the development of chronic diseases, mortality, life-span, and increased medical costs [8, 33, 208]. At the same time, regular physical exercise is one of the most effective interventions to control glucose values in diabetic patients [185], reduce weight [81], minimize the effects of aging [237], and improve health in general [36, 84], often without introducing pharmacological treatment. Most medical guidelines recommend 150 minutes per week of aerobic exercise for the general population [185]. However, to ensure the intervention's success, a personalized training prescription and evaluation are required [35, 176].

Traditionally, in epidemiological studies, physical activity level in the general population has been measured using methods that introduce subjectivity, such as surveys, sleep-logs, and daily diaries [252]. Similarly, in professional sports, subjective assessment metrics such as the rate of perceived exertion (RPE) [70] have been widely used. With the boom of digital medicine [144] and the possibilities of monitoring patients in real-time through biosensors, the objective measurement of physical activity is becoming increasingly common [179]. The estimation of energy expenditure using accelerometers is probably the most general and reliable procedure for this purpose at the moment.

Accelerometer data provides a vast source of information that quantifies the intensity, volume, and direction of physical activity in real-time in the period in which the device is worn. For the last 15 years, multiple epidemiological studies have used these devices to infer physical activity patterns in various cohorts. For example, in [271], the authors describe the physical activity patterns in the American population using simple summary measures by age-groups; in [94], the authors analyze how physical activity patterns vary minute-to-minute through functional data analysis techniques with children from New York. Other studies use accelerometers to resolve complex questions such as the relationships between physical activity levels and short-term mortality or life-span [69, 166, 267]. Precisely in this domain, a remarkable recent study [254] showed that physical activity patterns may predict mortality more accurately than

well-established epidemiological variables such as age, smoking, and the presence of cancer. Answering these questions with precision is essential to guide public health policies and design physical activity routines that optimize the population's health [60, 225]. The National Health and Nutrition Examination Survey (NHANES) is a public database containing information on the American population's physical activity levels during the period 2003-2006, and is the best-known database containing accelerometer monitoring. Other cohorts with available accelerometer data include the Baltimore longitudinal study [195], and more recent studies with the UK Biobank [258] or the International Children's accelerometer Database [248], and have provided new clinical knowledge with different study populations and other or similar sampling designs.

In the current era of precision medicine [138], these devices are also beneficial for individualized prescription of physical exercise, given that the data obtained is vital for the control and measurement of exercise performed in general and sports populations. For these reasons, accelerometer technology has also been gradually used to evaluate interventions and more beneficial physical activity therapies in clinical trials [194].

From a statistical point of view, the analysis of this data is usually complicated, and summary measurements must be used to compress the information recorded by the curves obtained with these devices. One of the main methodological obstacles that must be overcome is that the curves can have different lengths, and the subjects are not in standardized conditions, so a direct time series or functional data analysis is not usually workable. Given the inherent difficulty for direct examination of this data, practitioners often define several target zones and quantify the proportion of time (or total time) that the individual spends in each target zone when the device is worn. In many domains, such as diabetes, these metrics are commonly referred to as time-in-range metrics [30, 65]. When the characteristic vector obtained is a ratio-vector, several authors have recently suggested to use specific compositional data analysis techniques [31, 64, 65]. Naturally, time-in-range metrics suffer from a loss of information, as the information is discrete in intervals. In addition, the cut-off points chosen may be arbitrary and dependent on the characteristics of the population under study.

In this chapter, a similar strategy to that of glucodensities is applied, exploiting the connection between the Wasserstein geometry and quantile functions, with the dynamic accelerometer data being represented as probability distributions. Here, the induced probability distribution differs from the glucodensity extracted from CGM data in that it is a mixed distribution containing an atom at zero representing the proportion of inactive time. With the intended purpose of drawing more representative conclusions of physical activity levels at the population level than can typically be achieved with observational studies, many of the main cohorts' physical activity studies are designed with a complex survey structure including demographical characteristics in sample selection, as is notably the case in the NHANES 2003-2006 dataset.

In order to handle these sampling characteristics in our analysis, we propose adaptations of well-known estimators, such as kernel smoothing [286] or machine learning approaches such as kernel ridge regression [284], that accommodate the survey design structure as well as complex objects like the proposed physical activity distribution representation of accelerometer data. Including these nonparametric regression models can be a valuable option for practitioners to utilize their data more effectively by considering the study design's specific nature. Although there does not exist a vast amount of literature on nonparametric regression using survey data [108, 165], such techniques have the potential to contribute to obtain new clinical findings by modeling complex data relations that are common in biology and related fields. Since survey data can lead to more reliable conclusions than observational data [2, 164], the development of these tools for more complex data objects such as physical activity distributions has a high potential impact. As it is increasingly frequent in standard clinical practice to use medical devices that monitor patient conditions with high temporal resolution, the techniques proposed in this chapter can potentially be used to handle the resulting complex statistical objects quite broadly.

4.1 NHANES 2003-2006 dataset

The National Health and Nutrition Examination Survey (NHANES) 2003-2006 is an extensive, stratified, multistage survey conducted by the Centers for Disease Control (CDC) that collects health and nutrition data on the US population. The NHANES 2003-2006 data are publicly available from the CDC (<https://www.cdc.gov/nchs/nhanes/index.htm>) and are broadly categorized into six areas: demographics, dietary, examination, laboratory, questionnaires, and limited access. The accelerometer data for a particular NHANES cohort can be downloaded from the "Physical Activity Monitor" subcategory under the "Examination data" tab. In this work, a subset of patients that are between 68 and 85 years old will be used. This subset is different than that employed by [149] that involve patients within a wider age range (50-85 years old). The decision to restrict to a narrower age range in our applications was made for two reasons. First, although the Area Under the Curve (AUC) metric was high in predicting five-year mortality, the predictive model fitted in [149] did not classify any individuals as dead, partially due to the large imbalance between classes in this data set. As a consequence, the classical sensitivity vs. specificity analysis that is captured by the ROC curve is not sensible, and AUC may not be the best metric to assess the usefulness and predictive capacity of a clinical diagnostic model in this type of supervised modelling. Second, we think it is more clear to constrain the analysis to a more specific target population that can show more realistically the impact of physical inactivity than a more general and heterogeneous sample that involves lower-risk patients. At the same time, one must interpret the impact of physical activity on mortality with caution as opposed to the blind use of standard model performance metrics. For example,

in this domain, if a model incorrectly predicts that a patient will die, any negative impact may be very minor, with potential positive impacts such as identifying a high-risk individual who may be able to transform their lifestyle in five years, reversing their medical condition. While we cannot hope to predict mortality using only physical activity levels, these tools can serve as instruments to identify highly inactive patients phenotypes with a more urgent need for physical activity programs defined according to their specific characteristics. In general, models used to predict mortality in five years have limited predictive capacity, as it is likely necessary to use longitudinal models that realistically capture dynamic health evolution in this type of predictive task.

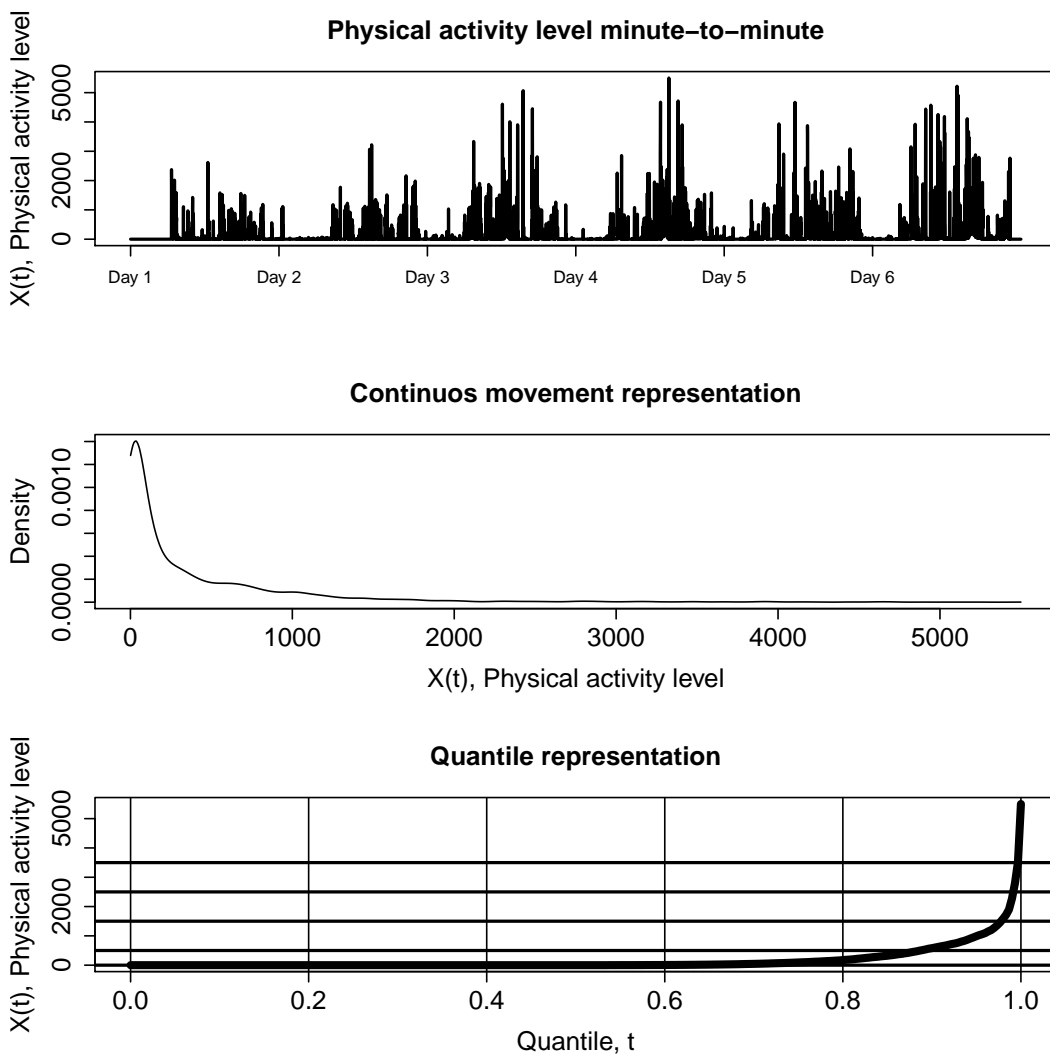


Figure 4.1: Example of transforming the raw accelerometer signal distributional profile for a randomly selected individual: (top) Physical activity recordings in real time; (middle) density function for active movement; and (bottom) quantile representation.

In order to explore the data, we estimate some basic characteristics of the patients examined using the survey weights for this target population, computed using the R package

Table 4.1: Variable summaries for the chosen cohort grouped as survivors/decedents. The reported values are mean (standard deviation) for continuous variables and counts (%) for categorical variables. Total activity count (TAC); total log-transformed activity count (TLAC); total minutes of moderate/vigorous physical activity (MVPA); active to sedentary/sleep/non-wear transition probability (ASTP); sedentary/sleep/non-wear to active transition probability (SATP); Coronary heart disease CHD (CHD); Congestive heart failure (CHF)

Variable	Survivors	Decedents
TAC	193735.4 (115239.8)	129456.5 (76978.72)
Age	72.3 (4.5)	75.5 (5.5)
MVPA	12.3 (4.5)	4.5 (9.6)
ASTP	0.3 (0.1)	0.37 (0.12)
Sedentary time	1126.7 (112.1)	1180 (110.5)
TLAC	2651.7 (764.3)	2328.7 (766.7)
Mobility problem	367 (37%)	129 (60%)
SATP	0.078 (0.02)	0.075 (0.02)
Education		
Less than high school	393 (39%)	81 (38%)
High school	262 (26%)	64 (30%)
More than high school	349 (35%)	71 (33%)
Drinking Status		
Moderate Drinker	442 (44%)	75 (35%)
Non-Drinker	496 (49%)	118 (55)
Heavy Drinker	40 (3%)	15 (7%)
Missing alcohol	26 (2%)	8 (4%)
Smoking Status		
Never	460 (46%)	68 (13%)
Former	466 (46%)	110 (51%)
Current	78 (8%)	38 (18%)
CHF	115 (11%)	37 (17%)
Gender		
Male	519 (52%)	145 (67%)
Female	485 (48%)	71 (34%)
Diabetes	163 (18%)	48 (22%)
Cancer	231 (23%)	57 (26%)
BMI		
Normal	283 (28%)	75 (35%)
Underweight	7 (1%)	7 (3%)
Overweight	414 (41%)	76 (35%)
Obese	300 (29%)	58 (27%)
CHD	115 (11%)	37 (17%)
Stroke	71 (7%)	31 (14%)
Race		
White	648 (65%)	161 (74 %)
Mexican American	172 (17%)	22 (18.8%)
Other Hispanic	17 (2%)	0 (0%)
Black	144 (14%)	28 (10%)
Other	23 (2%)	5 (2%)
Wear time	878.4 (161.0)	892.2 (164.98)

rnhanesdata [149]. Table 4.1 contains sample characteristics of individuals according to their mortality status after five years. Furthermore, in Figure 4.1, we show the raw data during several days of one participant selected randomly from the database. For each individual, while the device is worn, accelerometer devices collect an estimate of minute-by-minute energy expenditure. However, given that the device is not worn all day, the recorded signal cannot be continuous, and could possess intricate missing data patterns.

In addition to the reasons given above, one of the main benefits of restricting attention to aging populations in our analysis is that it mitigates against the effects of missing data. In aging populations, intra- and inter-day homogeneity of the raw physical activity profiles is much higher than in young populations, so that the impact of missing data is much lower. In order to further increase the reliability of the analysis, we use the following preprocessing strategy extracted from [254] in order to remove participants with poor quality in their accelerometry data. Those participants who i) had fewer than three days of data with at least 10 hours of estimated wear time or were deemed by NHANES to have poor quality data, or ii) had non-wear periods, identified as intervals with at least 60 consecutive minutes of zero activity counts and at most 2 minutes with counts between 0 and 100 were removed. These protocol instructions were extracted from state-of-the-art accelerometer research (see, for example, [271]).

4.2 Functional representation of accelerometer data and regression models

4.2.1 Functional representation of accelerometer data

First, we introduce the formal definition of the new representation. For the i -th patient, let T_i indicate the number of days (including partial days) for which accelerometer records are available and n_i be the number of observations recorded in form of pairs (t_{ij}, A_{ij}) , $j = 1, \dots, n_i$. Here, the t_{ij} are a sequence of time points in the interval $[0, T_i]$ in which the accelerometer records activity information, and A_{ij} is the measurement of the accelerometer at time t_{ij} . Unlike continuous glucose monitoring data, accelerometer readings of exactly zero are quite frequent, representing physical inactivity. Thus, in our distributional representation, we will assign positive probability mass at zero equal to the fraction of total time that the individual is physically inactive. In addition, the range of values measured by the accelerometer varies widely between individuals and groups, which can present difficulties when trying to apply common distributional data analysis methods, for example, functional transformations [119, 215, 279] that can be an alternative strategy to handle the representation that we specify below.

In order to handle accelerometer data gathered over different monitoring periods in free-living conditions, we propose to utilize a cumulative distribution function $F_i(x)$ for each individual. Formally, consider a latent process $V_i(t)$ such that the accelerometer measures

$A_{ij} = V_i(t_{ij})$ ($j = 1, \dots, n_i$), and define F_i as

$$F_i(x) = \frac{1}{T_i} \int_0^{T_i} \mathbf{1}(V_i(t) \leq x) dt, \quad \text{for } x \geq 0. \quad (4.1)$$

This definition corresponds to using $x = 0$ as a cutoff for inactivity; in the NHANES data set, it always holds that $F_i(0) > 0$. Thus, if U_i is a random variable uniformly distributed on $[0, T_i]$ that is independent of V_i , F_i is the distribution function of $V_i(U_i)$. In practice, one could use another reasonable cutoff for inactivity. For example, other studies have used accelerometer measures between 0-100 to quantify the inactive range. In this case, one would define F_i as the distribution of the censored random variable which takes the value 100 whenever $V_i(U_i) \leq 100$ and $V_i(U_i)$ otherwise. Analogously, we may be interested in truncating the latent process from above, for example to combine measurements representing high-intensity exercise, e.g., device observations greater than or equal to 3500. For instance, this idea can be exploited to establish high-intensity exercise benefits in the prediction of mortality or another relevant outcome. Practically speaking, an upper threshold of this type would lead to a simpler model that could be beneficial in the predictive task. Then F_i would be the distribution of the censored random variable taking values $V_i(U_i)$ whenever this is at most 3500, and 3500 otherwise. Combination of lower and upper cutoffs would be treated in a similarly straightforward manner.

In the remainder of the chapter, we define F_i as in (4.1), and denote $\mathbb{P}_{inactive}^i = F_i(0)$, $F_{active}^i(x) = F_i(x) - F_i(0)$ for $x > 0$, and $f_{active}^i(x) = [F_{active}^i]'(x)$. Hence, $F_i(x) = \mathbb{P}_{inactive}^i + \int_0^x f_{active}^i(s) ds$, which more clearly demonstrates the mixed nature of the distribution. In real world settings, $\mathbb{P}_{inactive}^i$ and $f_{active}^i(\cdot)$ are not observed, but must be estimated from the observed sample, which we carry out using the following two-step strategy. First, we estimate the proportion of inactivity-time, that is

$$\tilde{\mathbb{P}}_{inactive}^i = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{1}_{\{A_{ij}=0\}}.$$

Second, we estimate the continuous physical activity profile as conditional smooth density function. Letting \mathbb{K} denote a univariate probability density function and $h^i > 0$ the bandwidth parameter, define

$$\tilde{f}_{active}^i(x) = (1 - \tilde{\mathbb{P}}_{inactive}^i) \frac{1}{n_i^{active} h^i} \sum_{j:A_{ij}>0} \mathbb{K}\left(\frac{A_{ij} - x}{h^i}\right),$$

where $n_i^{active} = \sum_{j=1}^{n_i} \mathbf{1}_{\{A_{ij}>0\}}$. In our experiments, the Gaussian kernel was used for \mathbb{K} and the bandwidth parameter was selected through Silverman's rule of thumb [250]. More discussion about density estimation procedure and smoothing parameter selection with biosensor data can be found in [178]. Hence, given n samples of accelerometer measures belonging to n individuals $\{A_{ij}\}_{j=1}^{n_i}$, $i = 1, \dots, n$, we can use the above estimates to form empirical quantile functions $\tilde{Q}_i = \tilde{F}_i^{-1}$, where $\tilde{F}_i(x) = \tilde{\mathbb{P}}_{inactive}^i + \int_0^x \tilde{f}_{active}^i(s) ds$.

4.2.2 Statistical Framework for the Distributional Representation

While the representation of physical activity levels via the inactivity probability $\tilde{\mathbb{P}}^i_{inactive}$ and activity density \tilde{f}^i_{active} estimates provides a rich and fairly comprehensive representation of the accelerometer data, the mathematical constraints of these objects makes statistical analysis challenging. In particular, naive application of functional data analysis for the \tilde{f}^i_{active} is known to yield results that are often difficult to interpret, as these methods do not respect the inherent constraints possessed by probability density functions. Thus, we will work under the same framework outlined in [178], based on the Wasserstein distance of optimal transport [283]. This metric has theoretical appeal, has given intuitive results in a variety of applications, and possesses many computational advantages due to its connection to quantile functions, as will be seen below. Moreover, due to the mixed nature of the physical activity level distributions, the Wasserstein geometry is even more attractive as it accommodates such distributions without any special adaptation.

Next, we define the space of the physical activity distributional representations. Let $A := \{f : (0, \infty) \rightarrow \mathbb{R}^+ : \int_0^\infty f(x) dx < 1 \text{ and } \int_0^\infty x^2 f(x) dx < \infty\}$. Then the activity distributions constitute the set $\mathcal{D} \subset [0, 1] \times A$, where $(c_f, f) \in \mathcal{D}$ if $f \in A$ and $c_f = 1 - \int_0^\infty f(x) dx$. Given two arbitrary inactive-active representations $\mathfrak{f} = (c_f, f)$ and $\mathfrak{g} = (c_g, g) \in \mathcal{D}$, the 2-Wasserstein (or simply Wasserstein) distance between them is

$$d_{\mathcal{W}_2}(\mathfrak{f}, \mathfrak{g}) = \sqrt{\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt}, \quad (4.2)$$

where F^{-1} and G^{-1} are the quantile functions corresponding to the distributions represented by \mathfrak{f} and \mathfrak{g} , respectively. Given a metric or distance d on \mathcal{D} , of which $d_{\mathcal{W}_2}$ is one example, and a random variable \mathfrak{f} defined on \mathcal{D} , the *Fréchet mean* of \mathfrak{f} [82] is

$$\mu_{\mathfrak{f}} = \arg \min_{g \in \mathcal{D}} E(d^2(\mathfrak{f}, g)).$$

The corresponding *Fréchet variance* of \mathfrak{f} is then

$$\sigma_{\mathfrak{f}}^2 = E(d^2(\mathfrak{f}, \mu_{\mathfrak{f}})).$$

With the particular choice $d = d_{\mathcal{W}_2}$, we have

$$\mu_{\mathfrak{f}} = \arg \min_{g \in \mathcal{D}} E \left[\int_0^1 (F^{-1}(t) - G^{-1}(t))^2 dt \right],$$

and, with $Q_{\mathfrak{f}}$ denoting the quantile function corresponding to $\mu_{\mathfrak{f}}$,

$$\sigma_{\mathfrak{f}}^2 = E \left[\int_0^1 (F^{-1}(t) - Q_{\mathfrak{f}}(t))^2 dt \right].$$

Given samples of accelerometer measures belonging to n individuals, we can follow the above steps to form empirical quantile functions $\tilde{Q}_i = \tilde{F}_i^{-1}$, where $\tilde{F}_i(x) = \tilde{\mathbb{P}}_{inactive}^i + \int_0^x \tilde{f}_{active}^i(s)ds$. Then, due to the Euclidean nature of (4.2), the empirical Fréchet mean and variance, written in terms of quantile functions, take the form of

$$\bar{Q}(t) = \tilde{Q}_{\bar{f}}(t) = \frac{1}{n} \sum_{i=1}^n \tilde{Q}_i(t), \quad t \in [0, 1], \text{ and}$$

$$\tilde{\sigma}_{\bar{f}}^2 = \frac{1}{n-1} \sum_{i=1}^n \int_0^1 (\tilde{Q}_i(t) - \bar{Q}(t))^2 dt.$$

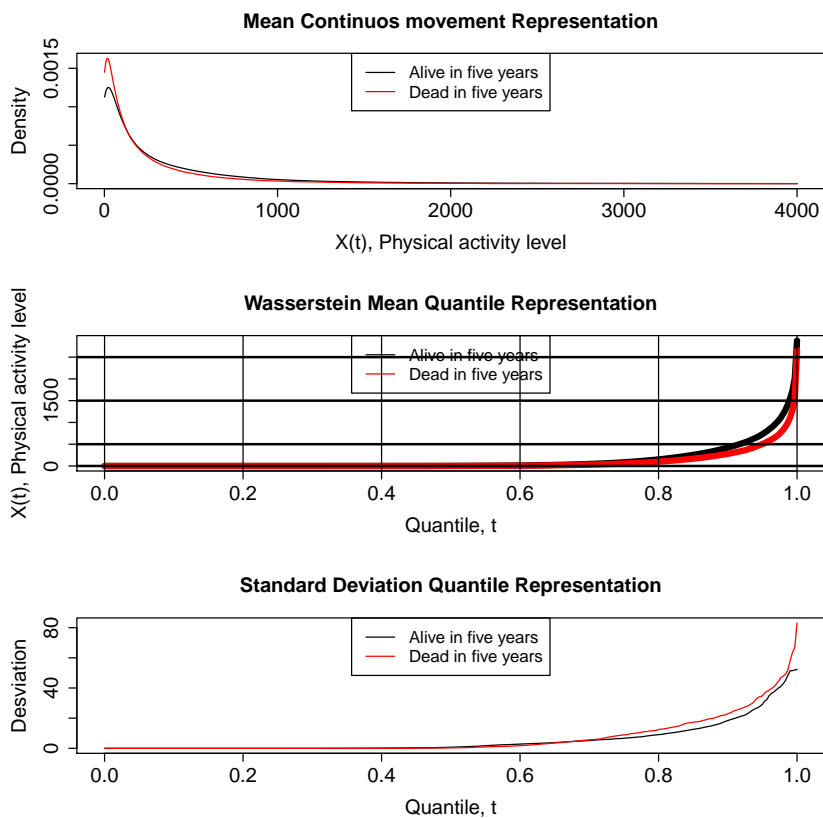


Figure 4.2: Summary curves of physical activity distributions in quantile space (mean and standard deviation) between alive and dead patients groups after five years together with the mean of continuous movement representation.

In this case, we also construct the estimated pointwise quantile variance curve $\tilde{\sigma}_{\bar{f}}^2(t)$, representing the sample variance of the values $\tilde{Q}_1(t), \dots, \tilde{Q}_n(t)$, for each $t \in [0, 1]$. Figure 4.1 illustrates the process of transforming raw data into our representation. In addition, Figure 4.2 shows the mean and variance curve of our representation, where patients are grouped by mortality status after five years.

4.2.3 Survey regression models

The individuals that we analyze from the NHANES database do not represent a simple random sample of the US population. Instead, they are the result of a structured sample of a complex survey design from a finite population of individuals. In order to perform inference correctly and obtain reliable results, we must account for the effects of the specific sample design when building a predictive model. Note that the information provided by this type of survey data is typically richer than those used in most medical studies that are of an observational nature [164]. In the latter case, the researcher does not explicitly control the sampling mechanism, so that obtaining a representative sample can be challenging and would often demand colossal data volumes.

Suppose that observations $\{(X_i, Y_i); i \in S\}$ are available, where X_i is a collection of covariates taking values in a metric space, and Y_i is a scalar response variable. The index set S represents a sample of n units from a finite population. To account for this sampling, each individual $i \in S$ will be associated with a positive weight w_i . In our analyses, these weights were taken to be the inverse of the probability $\pi_i > 0$ of being selected into the sample [134], i.e. $w_i = 1/\pi_i$ [165]. When performing estimation with survey data, a common approach is to use the w_i to define weighted versions of usual estimates designed for random samples. For example, the normalization Horvitz-Thompson estimator [117, 223] for the population average of the Y_i is the weighted sample average

$$\bar{Y}_w = \sum_{i \in S} \frac{\frac{1}{\pi_i} Y_i}{\sum_{i \in S} \frac{1}{\pi_i}} = \sum_{i \in S} \frac{w_i Y_i}{\sum_{i \in S} w_i}. \quad (4.3)$$

In this chapter, we propose to use a general kernel smoother [286] for survey data with weights that are composed of both the sampling weights w_i as well as the usual local weights that appear in such kernel methods. One main advantage of this estimator is its flexibility, as it is valid for either regression or classification problems. In addition, we also extend kernel ridge-regression [284], but this method is only appropriate for a continuous response variable.

4.2.4 Kernel smoother for survey data

Suppose the mean regression model

$$Y = m(X) + \epsilon \quad (4.4)$$

holds, where ϵ is a random error term satisfying $E(\epsilon|X) = 0$. Hence, the value $m(X)$ represents the conditional mean of Y given X , where m is assumed to be a smooth function. Given a sample $\{(X_i, Y_i, w_i); i \in S\}$ of size n from the finite population as described above, an estimate of $m(x)$ for a generic input x may be obtained using the standard kernel estimator [286]

$$\tilde{m}(x) = \sum_{i \in S} s(X_i, x) Y_i, \quad (4.5)$$

where $s(X_i, x)$ is an appropriate weight function that provides more weight for predictors X_i with smaller distance to x . Furthermore, the constraint $\sum_{i \in S} s(X_i, x) = 1$ must be satisfied for all x to obtain a coherent estimator. Typical choices for s include Nadaraya-Watson weights $s(X_i, x) = \frac{\mathbb{K}(h^{-1}d(X_i, x))}{\sum_{i \in S} \mathbb{K}(h^{-1}d(X_i, x))}$, where d is a metric on the set of predictors, for example the Wasserstein distance defined in (4.2) if the covariate X represents a physical activity level distribution, and $h > 0$ is the smoothing parameter. The generalization of the standard Nadaraya-Watson estimator, which was originally proposed for scalar or vector predictors, to more abstract data types has been used to handle functional predictors [76] as well as predictors and responses in more general spaces that possess a metric [257]. Here, we will utilize the quantile functional representation of the accelerometer data along with the Wasserstein metric to incorporate these complex objects as predictors of relevant outcomes.

Due to the survey design, we make the necessary adjustment to the usual Nadaraya-Watson weights by scaling them according to the survey weights w_i . Specifically, we set $s(X_i, x) = \frac{\mathbb{K}(h^{-1}d(X_i, x))w_i}{\sum_{i \in S} \mathbb{K}(h^{-1}d(X_i, x))w_i}$. This definition reflects that an observation should be given higher weight when the probability of selection is lower (larger values of w_i), consistent with the principles outlined in [117], and when the observed input X_i is closer to the input x at which one desires an estimate of the conditional mean. In general, the kernel smoother in (4.5) corresponds to a convex combination of the observed responses, a property that is not shared by similar smoothers, for example local linear regression estimators. An important consequence of this convexity property is in the case of a binary response variable $Y \in \{0, 1\}$. In this case, $m(x) \in [0, 1]$ represents a probability, so that (4.5) yields an estimate $\tilde{m}(x) \in [0, 1]$ that can be interpreted properly as a probability, or used in classification tasks, for example, without any *post hoc* modification.

When Y represents a categorical variable that can assume more than two values, a simple modification of (4.5) can still be used to produce valid estimates of the various probabilities.

4.2.5 Kernel ridge regression for survey data

The Reproducing Kernel Hilbert Space (RKHS) learning paradigm provides a unique and rich framework to create new and more flexible predictive models that can handle abstract variables X as predictors by assuming that the regression function m in (4.4) is an element of a space of functions \mathcal{H} on \mathcal{D} that is an RKHS. This section focuses attention on a method known as kernel ridge regression that leverages the properties of an RKHS to produce estimates that can be viewed as generalizations of the usual ridge regression estimator for linear models. In the following, we summarize the necessary components of the RKHS-based model and its estimator, and then adapt the estimator to the case of survey data.

For each input value $x \in \mathcal{D}$, one way of defining an RKHS is to begin with a kernel $k : \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R}$ that is symmetric and positive definite. Observe that the use of the term

kernel is distinct from that of the previous section. For clarity, a distinct notation has been introduced for the bivariate kernel of the current section. Beginning with functions of the type $\phi_x(\cdot) = k(x, \cdot)$ as basic elements, one can construct a Hilbert space of functions by taking linear combinations and, finally, by taking the usual metric completion. The constructed Hilbert space \mathcal{H} can be shown to have the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ with the property that $\langle \phi_x, \phi_y \rangle_{\mathcal{H}} = k(x, y)$. Furthermore, for any $f \in \mathcal{H}$, one has $f(x) = \langle \phi_x, f \rangle_{\mathcal{H}} = \langle k(x, \cdot), f \rangle_{\mathcal{H}}$, so that k is often referred to as a reproducing kernel, or the kernel that generates \mathcal{H} . Importantly, in our application, we must build a kernel on the space \mathcal{D} of physical activity distributions. To do so, consider a nonincreasing univariate function $\kappa : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ and, for $x, y \in \mathcal{D}$, define $k(x, y) = \kappa(d_{\mathcal{W}_2}(x, y))$, where $d_{\mathcal{W}_2}$ is the Wasserstein distance defined in (4.2). In making comparisons with models that use scalar TAC or vectors as the predictor, so that the input space is simply \mathbb{R}^q , we will use the same approach to constructing an RKHS by defining $k(x, y) = \kappa(\|x - y\|)$, with $\|\cdot\|$ being the Euclidean norm.


Considering the model defined Equation 4.4, an alternative to the smoothing method of the previous section is to assume that the regression function $m \in \mathcal{H}$. Given the infinite-dimensional nature of \mathcal{H} , estimation of m through the use of least squares, i.e.

$$\tilde{m} = \arg \min_{m \in \mathcal{H}} \sum_{i \in S} (Y_i - m(X_i))^2 \quad (4.6)$$

is ill-defined. Specifically, there are many different solutions to (4.6) that attain zero empirical error. Naturally, overfitting the model in this way results in poor predictive capacity for new observations. In the RKHS framework, it is common to introduce a norm-based penalty on m in the optimization procedure to induce regularization. The kernel ridge regression estimator then becomes

$$\tilde{m} = \arg \min_{m \in \mathcal{H}} \sum_{i \in S} (Y_i - m(X_i))^2 + \lambda \|m\|_{\mathcal{H}}^2, \quad (4.7)$$

where λ is the regularization parameter that controls the usual trade-off between bias and variance, which in turn determines the capacity of the model to generalize to new observations. By the classical Representer Theorem [238], the solution to (4.7) is known to take the form $\tilde{m}(\cdot) = \sum_{i \in S} \alpha_i k(\cdot, X_i)$, so that the estimator is a linear combination of the kernel features $k(\cdot, X_i)$ with coefficients α_i . Solving (4.7) under this restricted form of m results in the coefficient estimates $\tilde{\alpha} = (K + \lambda I)^{-1} Y$, where



$$K = \begin{pmatrix} k(X_1, X_1), \dots, k(X_1, X_n) \\ k(X_2, X_1), \dots, k(X_2, X_n) \\ \dots \\ k(X_n, X_1), \dots, k(X_n, X_n) \end{pmatrix}, \quad Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix},$$

and I is the $n \times n$ identity matrix.

The survey design can be incorporated into the optimization problem by utilizing the Horvitz–Thompson version of the estimator, namely

$$\tilde{m} = \arg \min_{m \in \mathcal{H}} \sum_{i \in S} w_i (Y_i - m(X_i))^2 + \lambda \|m\|_{\mathcal{H}}^2. \quad (4.8)$$

As (4.8) remains a convex objective function, the Representer Theorem holds and the solution will retain the same structure, $\tilde{m}(\cdot) = \sum_{i \in S} \alpha_i k(\cdot, X_i)$. However, the coefficients take the form of regularized weighted least squares estimates $\tilde{\alpha} = (WK + \lambda I)^{-1} WY$, with W being a diagonal matrix with the weights w_i constituting the diagonal elements. A notable advantage of the kernel ridge regression is that it preserves some computational advantages of linear models, as the optimal solution can be calculated via weighted least squares. This fact also simplifies selection of the tuning parameter λ , for example using Leave One Out Cross-Validation (LOOCV), given that explicit leave-one-out formulas are available for linear estimators [95].

Besides the regularization parameter, a crucial choice that determines the model's empirical performance is that of the RKHS learning space V or, equivalently, the function κ that determines the kernel K . In this work, the physical activity distributional representations are estimated from a mixed stochastic process in continuous time that is not smooth, as there are transitions between activity and inactivity time. Consequently, the probability distributions and, more importantly, the mechanism by which the data arise cannot be very smooth. In RKHS methods, there is a natural connection between smoothness of the data space and that of the relevant operator [152], which is a regression operator in this case. Therefore, we expect that statistical learning using very smooth functionals spaces, such as the RKHS spaces induced by a Gaussian kernel, would yield inferior results. Specifically, the empirical risk resulting from a kernel that induces a very smooth functional space may overestimate the true risk. Thus, it is critical to select a kernel function according to the expected smoothness of the proper regression function m , using both expert knowledge about the problem and an understanding of the data properties. In this regard, a recent reference about the performance of kernel ridge regressions in terms of smoothness of the function m is [276].

In our preliminary test, as we hypothesized, the best results are obtained with non-smooth kernels such as the Laplace kernel induced by $\kappa(u) = e^{-u/\sigma}$. Hence, we consider this kernel in the following. Here, $\sigma > 0$ is a scale parameter that can be chosen heuristically as in [88] by considering the distribution of pairwise distances $d(X_i, X_j)$, $1 \leq i < j \leq n$. When the inputs are physical activity distributions, the distance is taken to be $d_{\mathcal{W}_2}$, while it is taken to be the usual Euclidean metric when the inputs are vectors or scalar variables. Specifically, we use the scalar value $\tilde{\sigma}$ corresponding to the median of the discrete distribution with point masses at the values $d(X_i, X_j)$ of size $w_i w_j / \sum_{1 \leq i < j \leq n} w_i w_j$.

4.3 Results

To show the potential of the new representation of accelerometer data and the application of new nonparametric survey regression estimators, the analyses described below were performed with the sample described in Section 4.1.

1. Four different response variables were analyzed, namely age, Body Mass Index (BMI), blood pressure, and cholesterol levels. Kernel ridge regression was used to compare our functional representation with TAC, the most relevant and commonly used physical activity variable in different studies with this database [149, 254]. In addition, we further compare our methods against recognized accelerometer metrics [149], including sleep, sedentary behavior (SB), a measure of moderate-to-vigorous physical activity (MVPA), and wear time (WT), in combination with the TAC metric.
2. The ability of our functional representation to estimate the risk of mortality after five years was assessed with the Nadaraya-Watson survey estimator. Furthermore, an in-depth clinical analysis of the results provided by the algorithm is given.

4.3.1 Comparison of distributional representation vs. TAC

This section shows that our representation contains more valuable information about patient health than the standard TAC metric, which has been shown to be the most relevant physical activity variable in the NHANES database. Indeed, the TAC value can be seen as a particular scalar summary of our distributional representation, as this already contains information about total energy expenditure. According to the steps followed in [149], we define the TAC variable formally as follows. First, for each participant i with d_i days of valid accelerometer data, and each day j , we consider the set of raw accelerometer data indices by $C_{i,j}$. Then, we define the total daily energy for the day j as $TAC_{i,j} = \sum_{k \in C_{i,j}} A_{ik}$. Finally, we define the overall TAC metric for individual i as $TAC_i = \frac{1}{d_i} \sum_{j=1}^{d_i} TAC_{i,j}$. Here, to obtain a reliable estimation of the TAC metric, the data preprocessing introduced in Section 4.1 is critical.

One way to illustrate the benefits of our method is to compare its ability to capture essential biomarkers associated with the health and decline of physiological function to that of TAC and other commonly used scalar summaries of physical activity such as MVPA, SB, and WT that measure the proportion of the time that patients spend at different exercise intensities. For this purpose, we select age, Body mass index (BMI), blood pressure, and cholesterol as response variables. As a regression estimator, we select the kernel ridge-regression introduced in Section 4.2.5, with the Laplacian kernel. To make a conservative assessment of physical

Table 4.2: R-squared was computed for each representation used in kernel ridge-regression models with continuous variables being examined. MVPA is a compositional metric with a cut-off equal to 2020 counts. SB is the proportion of the time that an individual has an energy expenditure lesser than 100 counts. WT is an estimation of the proportion of the time that the individual wear the accelerometer device

	New representation	TAC	TAC+MVPA+SB+WT
Age	0.15	0.07	0.08
BMI	0.05	0.01	0.01
Blood pressure	0.02	-0.01	-0.01
Cholesterol total	0.034	0.016	0.018

activity levels, we calculate a survey-weighted leave-one-out version of R -square, defined as

$$R^2 = 1 - \frac{\sum_{i \in S} w_i (Y_i - \tilde{f}^{-i}(X_i))^2}{\sum_{i \in S} w_i (Y_i - \bar{Y}_w)^2}, \quad (4.9)$$

where w_i 's are the survey weights, \bar{Y}_w is defined in (4.3), and $\tilde{f}^{-i}(\cdot)$ is a generic regression estimate obtained after deleting the i -th observation. As the models are nonlinear and leave-one-out estimators are used, R^2 as defined in (4.9) can be negative, as seen for blood pressure as response with TAC as predictors in Table 4.2, where all results are compiled.

These results demonstrate that the statistical association is low to modest for all models, with age and BMI being the most predictable variables. In all cases, it is clear that our representation outperforms the summary metric TAC and the TAC variable in combinations with another compositional measures. As the new representation retains more information than summary metrics, we hypothesize that the advantages of the former may be even more significant in larger databases.

4.3.2 Estimating five-year mortality risk

A Nadaraya-Watson estimator was used to estimate the five-year mortality risk either from the distributional representation or from the TAC metric. In order to tune the bandwidth parameter h for each model, leave-one-out probabilities were estimated and used to provide an intermediate classification for each patient by comparing the probability of death to 0.5; the bandwidth was then chosen to minimize this intermediate classification error. The best choice of h for the distributional representation yielded 59 probabilities above 0.5, 23 of which corresponded to patients that died in fact within 5 years. For the TAC metric, regardless of the chosen smoothing parameter, leave-one-out estimates of probability of death was always below 0.5 for all subjects, suggesting that TAC possesses little to no information about risk for this cohort.

To appreciate the potential clinical advantages of our representation in this analysis, we compare for each patient the estimated probability of death within five years by means of the following models: 1) Nadaraya-Watson with our novel representation, 2) Nadaraya-Watson

with TAC variable, 3) functional logistic using quantile functions as predictors. This last model was fitted using functional principal component analysis, with the number of components chosen to explain at least 95% of the overall variability, a common practical approach. Figure 4.3 shows that the novel representation is the only one to appropriately assign moderate to high risk of death, and the discrimination capacity of probabilities assigned to each patient is more reasonable than the competitors' models. The choice of competing models thus reveals that: i) distributional information beyond TAC, such as tail behavior and variability, and ii) non-linear effect of the distribution, are essential to obtain a more reasonable probability estimates.

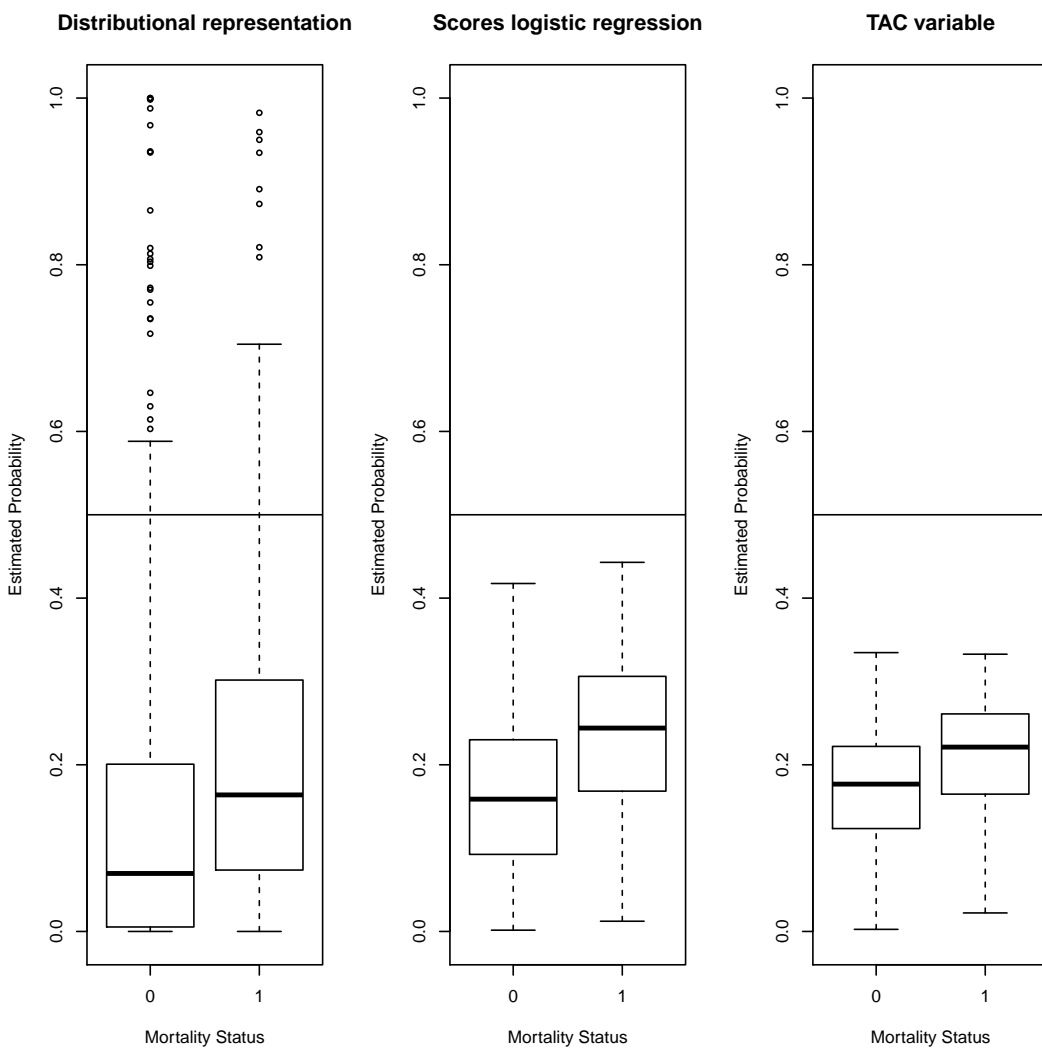


Figure 4.3: Boxplot of estimated probabilities drawn by the models in 5-year mortality prediction according to mortality status. (Left) Nadaraya-Watson distributional representation. (Center) Logistic regression introducing the five scores PCA quantile analysis. (Right) Nadaraya-Watson TAC variable.

In order to provide further clinical validation, we examined the long-term survival of patients in this cohort. First, the patients were stratified by age (68-75, 76-80, and 81-85); next, the patients in each strata were categorized into a risk and non-risk groups according to whether or

not the estimated probability of death within 5 years (by the distributional model) was greater than or less than 0.5. The age stratification is designed to illustrate that the model's ability to identify at-risk patients is relevant even in relatively younger subjects in the cohort. Overall, the results (Figure 4.4) confirm the high clinical sensitivity of the probabilities estimated with the Nadaraya-Watson model to stratify patients based on their long-term survival in more than 12 years of follow-up.

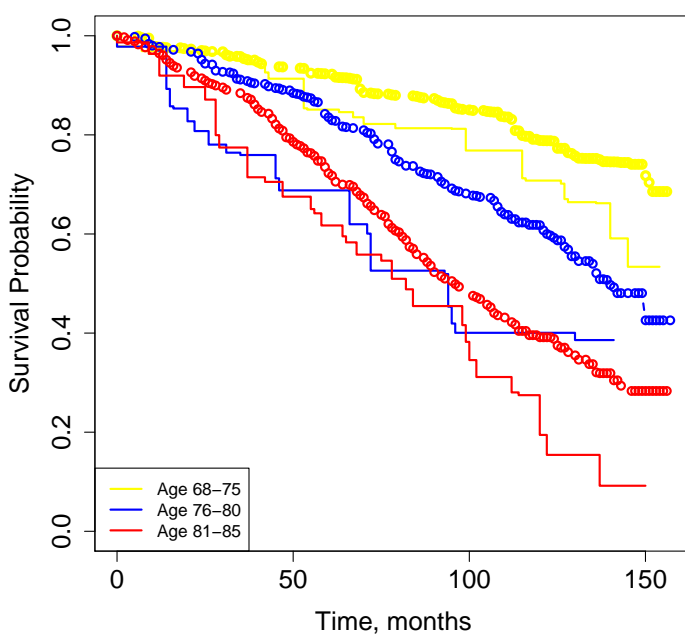


Figure 4.4: Survival curves for risk and non-risk groups, according to age stratification. The curves with only lines identify the risk group and, in another case, the non-risk group.

4.4 Discussion

In this work, we have introduced a new functional profile of an individual's physical activity to quantify more comprehensively the energy expenditure over a given period among a group of individuals monitored in free-living conditions. Our procedure can be seen as a functional extension of the so-called compositional metrics that constitute the most popular approach to date in the accelerometer field. A fundamental advantage of the new method is that it automatically captures information from compositional metrics, regardless of the cutoffs used to define them. As such, one loses no information compared to these metrics, and simultaneously avoids the need to define a-priori different cut-off points [30, 65]. Even with expert knowledge, such a selection introduces subjectivity into the analysis, with the cutoffs inevitably depending on sample characteristics or the analysis task at hand.

In the different regression tasks considered in this chapter, we have seen that the new representation possesses stronger associations than other common summary measures such as TAC, which has been shown to be the most successful variable, for example, in predicting 5-year mortality in the NHANES database in other studies [149, 254]. Overall, the strength of the association between our representation and clinically relevant covariates, as quantified by the leave-one-out R^2 metric is modest, indicating that there is a large amount of variability in many biomarkers associated with patient health, such as cholesterol, blood pressure, or BMI. Several epidemiological studies have used multivariate regression models with these biomarkers as response variables and summary measures of physical activity as covariates [18, 163]. Although some have found physical activity to be a statistically significant factor, many of these models only assess the statistical significance of the variable, rather than its practical significance or predictive capacity [209]. Assessing only statistical significance fails to accurately assess the magnitude of impact physical activity has on biomarker prediction, and obscures the conclusions and reproducibility of findings reached. As mentioned previously, the evaluation of a clinical diagnostic model aimed at predicting 5-year mortality by including physical activity levels as predictors is not straightforward. An accurate evaluation must be done from a clinical point of view using the long-term survival of the patients, a basic illustration of which is given by the survival curve shown in Figure 4.4, though ideally with a follow-up of more than 12-years.

Despite the large number of studies that have analyzed the impact of physical activity in the NHANES cohort against different biomarkers associated with health or with mortality and survival, few studies incorporates the complex survey mechanism in the analyses, which is very crucial for obtaining reliable and reproducible results that be a representative of American population structure. To increase the reliability of the analyses, we model relationships between the covariates and response variables nonparametrically. To the best of our knowledge, this work is the first to combine survey methods for the estimation of nonparametric regression models involving complex predictors such as the physical activity distributions we consider. In particular, we have demonstrated how to implement the Nadaraya-Watson kernel smoother as well as kernel ridge regression models in this context.

We believe that the proposed model constitutes an important step forward in the use of complex objects with this type of data, which appear naturally in some important physical activity cohorts, to obtain data representative of the physical activity patterns of a population, and not only in NHANES as a particular case. It is likely that in the current technological revolution the use of biosensors and smartphones in medical surveys will become increasingly common to characterize the population health [224] and, in particular, the physical activity levels of a population.

The analyses introduced here possess some potential limitations, and the methodology may be improved and extended in the following directions discussed below. First, sample selection is

essential in distributional representation validation, which may drive more predictive gains with other criteria. Second, we are using accelerometer data from NHANES 2003-2006; however, modern accelerometer devices can quantify the physical exercise profiles in light and high intensity more accurately [91, 258], and our functional representations may prove to be more accurate in different predictive tasks and risk analysis. This will be addressed in the next section. Third, missing data in raw time series is another critical problem, particularly in young and middle-aged populations. In the wearable data analysis literature, several works have proposed methods for missing data to address this problem with accelerometer data (see, for example, [3]). In this direction, if the MAR missing data mechanism holds, we can perform a more refined quantile and distribution function estimation introducing inverse probability weight into the estimator. In this sense, for the young and middle-aged population, maybe more extensive periods are necessary to create a reliable physical activity profile of the individual. Fourth, to minimize the curse of dimensionality in multivariate data, we could consider semi-parametric functional models where we introduce the functional information non-parametrically. At the same time, we can incorporate the rest of the variables in the linear model component terms, exploiting the linear models' inferential, robustness, and interpretation advantages. Finally, an essential point in this discussion is that the mixed nature of physical activity probability distribution implies that using a functional basis presents theoretical inconveniences due to the discontinuity of the quantile functions in the transition between inactivity and activity time. As a consequence, the already difficult task of interpreting output from even a basic functional model, such as functional linear regression model, is exacerbated by the constrained origin of the probability distribution. A straightforward procedure for doing so in terms of the specific biological nature of the problem is not obvious. We also note that the novel distributional representation has several applications that go beyond predictive tasks. For example, we can use the novel functional profiles graphical tool in order to build models for quantifying and assessing individual physical changes in physical exercise patterns over time in a context of longitudinal clinical trials or in order to monitor the individual in the physical activity prescription.

4.5 Application in NHANES 2011-2014: Discovering clinical physical activity phenotypes in the U.S. population

Precision medicine is based on the idea of defining clinical phenotypes [44] or clusters of people who share a similar prognosis or response either to treatment or to other clinical interventions. These patient phenotypes are also helpful to define the different transitions or changes in individual health characteristics and classify the expected patient evolution more accurately. Unfortunately, to date few contributions that propose physical activity phenotypes using accelerometer data exist [118]. A better understanding of the health consequences of

individual profiles of physical activity, using the full spectrum of accelerometry intensity across the day, would arguably help inform public health recommendations to promote the health of the population.

Benefiting from the abundant and unique information provided in the 2011-2014 National Health and Nutrition Examination Survey (NHANES) study, including the availability of high resolution accelerometry data, the current work aims to define new physical activity phenotypes using an unsupervised clustering analysis in people aged 65 to 80 years old. The secondary aim of this study is to ascertain the prospective associations of these phenotypes with 5-year survival probability and mortality. To achieve these aims, we capitalize on previously proposed distributional representations of accelerometry-based physical activity, which allows the quantification of time spent across the full spectrum of physical activity intensity without limiting to collapse the whole information into a few intensity intervals, as previously done using more traditional compositional metrics [65].

4.5.1 Materials and Methods

We use physical activity data from the NHANES waves 2011-2014, recorded with a more modern accelerometer device ActiGraph GT3X+ (ActiGraph of Pensacola, FL), that allows to measure light and high-intensity target zones with more precision.

A total of 2023 older adults aged 65 to 80 years old (with physical activity monitoring available at least hours per day for days) were included in the analysis. For the multivariate analysis, supported by additional biochemical, grip strength, and comorbidities variables, participants, were included due to missing data on covariates. In both cases, specific re-weight techniques on raw NHANES survey data were applied to handle the specific sampling mechanisms properly.

Sociodemographic and clinical data

Age (both as a categorical and continuous variable), race, gender, diagnosis of cancer or diabetes (as categorical variables), and blood pressure, combined grip strength measure, body mass index (BMI) and biochemical biomarkers including cholesterol and triglycerides (as continuous variables), were considered in the analysis. Age was divided into three ranges (65 – 70, 70 – 75 and 75 – 80, respectively) for age-stratified analysis. Race variable was coded as 1 = Mexican American; 2 = Other Hispanic; 3 = Non-Hispanic white; 4 = Non-Hispanic black; 5 = Non-Hispanic Asian; and 6 = Other Race, including multi-racial.

Physical activity monitoring

Physical activity signals were pre-processed by staff from the National Center for Health Statistics (NCHS) to determine signal patterns that were unlikely to be a result of human

movement. Then, acceleration measurements were summarized at the minute level by using Monitor-Independent Movement Summary (MIMS) units, an open-source, device-independent universal summary metric [129].

This new distributional representation allows us to measure the difference between physical activity profiles of different individuals by quantifying more comprehensively the amount of movement (i.e., acceleration, which resonates energy expenditure) over a given period and across the full spectrum of physical activity intensity.

Mortality and survival

NHANES data can be linked to the National Death Index (NDI), enabling the study of the association between acceleration data, mortality status and survival time. To this end, we accessed the 2015 Public-Use Linked Mortality Files [78], and included a binary variable indicating survival (or death) five years later, and the censored time to death.

Statistical analysis

The primary goal was to identify a reduced set of clinically relevant phenotypes of physical activity supported by the new distributional representation and evaluate their impact on health. To this aim, we performed a clustering analysis using the kernel-group algorithm [80]. We assessed the clinical relevance of these phenotypes to predict five-year mortality and survival, and compared their clinical sensitivity and accuracy with age. We performed logistic and Cox regression on survey data. We then implemented the Kaplan-Meier estimator and included the phenotype as a categorical predictor. Odds ratios and hazard ratios, and graphical survival plots were used to quantify the prospective associations of these phenotypes on mortality and survival in the study sample. Then, in order to remove the effect of potential confounding variables, we fitted again the logistic and Cox regression models and included also comorbidities, gender, race, cholesterol and triglycerides as predictors in the models. All statistical analyses were conducted using R software. Cluster analysis was performed using the Energy package, and survey analysis was performed using the Survey package.

4.5.2 Results

Physical activity phenotypes

Five clinical phenotypes were identified by means of a cluster analysis based on Euclidean energy distance. The optimal number of clusters was selected according to the rule-of-thumb [259]. Figure 4.5 displays the mean quantile curves and the standard deviation quantile curves for the distributional representation of physical activity of each phenotype. The proportion of individuals who died after five years is also shown. We observed three phenotypes (Phenotypes

2, 3 and 5) with low mortality rate (less than 8%) and two phenotypes (Phenotypes 1 and 4) with a mortality rate of 27.3% and 12.8%, respectively. The average distributional profiles of Phenotypes 1 and 4 showed a distinctive inactivity pattern: more than 80% of the time of participants in these two clusters is spent in sedentary behaviors (90% time vs. 80% time), with also important differences in the proportion of time spent in light and moderate to vigorous physical activity (MVPA) (5% vs. 10% and 2.9% vs. 6.5% respectively). Participants in Phenotypes 3 and 5 spent similar amount of time in sedentary (72% vs. 73%, respectively) and in light intensity (10% vs. 8%, respectively) activities but Phenotype 3 had 5% more time in MVPA. Finally, participants in the Phenotype 2, with the lowest mortality rate, only spent 62% percent of time sedentary, 10% in light intensity, 15% in MVPA and 13% in higher intensities.

Marginal survival analysis

Figure 4.6 displays a comparison of the survival curves for the different phenotypes and for the different age ranges. Participants in Phenotype 1 (the most inactive group) showed a lower survival compared with older individuals (75 - 80 years old). Figure 4.5 shows the 5-year mortality associated with each phenotype. Phenotypes 2 - 5 showed more than 90% less risk of mortality compared with Phenotype 1.

Multivariate Analysis

Population-based characteristics of the participants included in the multivariate analysis are shown in Table 4.3. Participants in Phenotype 1 were older on average than participants in the rest of phenotypes, had a higher BMI, higher triglyceride level and higher blood pressure. Phenotype 4, the second phenotype with more mortality rate, had a higher rate of diabetes and cancer, and the second higher BMI and age. Phenotype 1 (mortality rate of 27.3%) presented significant lower values of combined grip strength. However, Phenotype 4 (mortality rate of 12.8%) presented similar values of combined grip strength than the rest of physical activity phenotypes. Table 4.5 shows the multivariate estimated coefficients (hazard and odds ratios) for mortality associated with physical activity phenotypes. Results remained consistent with univariate models presented in Table 4.4. Importantly, the confidence intervals for odds and hazard ratios do not cross 1, suggesting statistical significance.

4.5.3 Discussion

This study reveals new physical activity phenotypes for the U.S. older population using novel distributional representations of accelerometer-derived physical activity. The new clinical phenotypes yield a higher clinical sensitivity for predicting 5-year mortality and survival outcomes

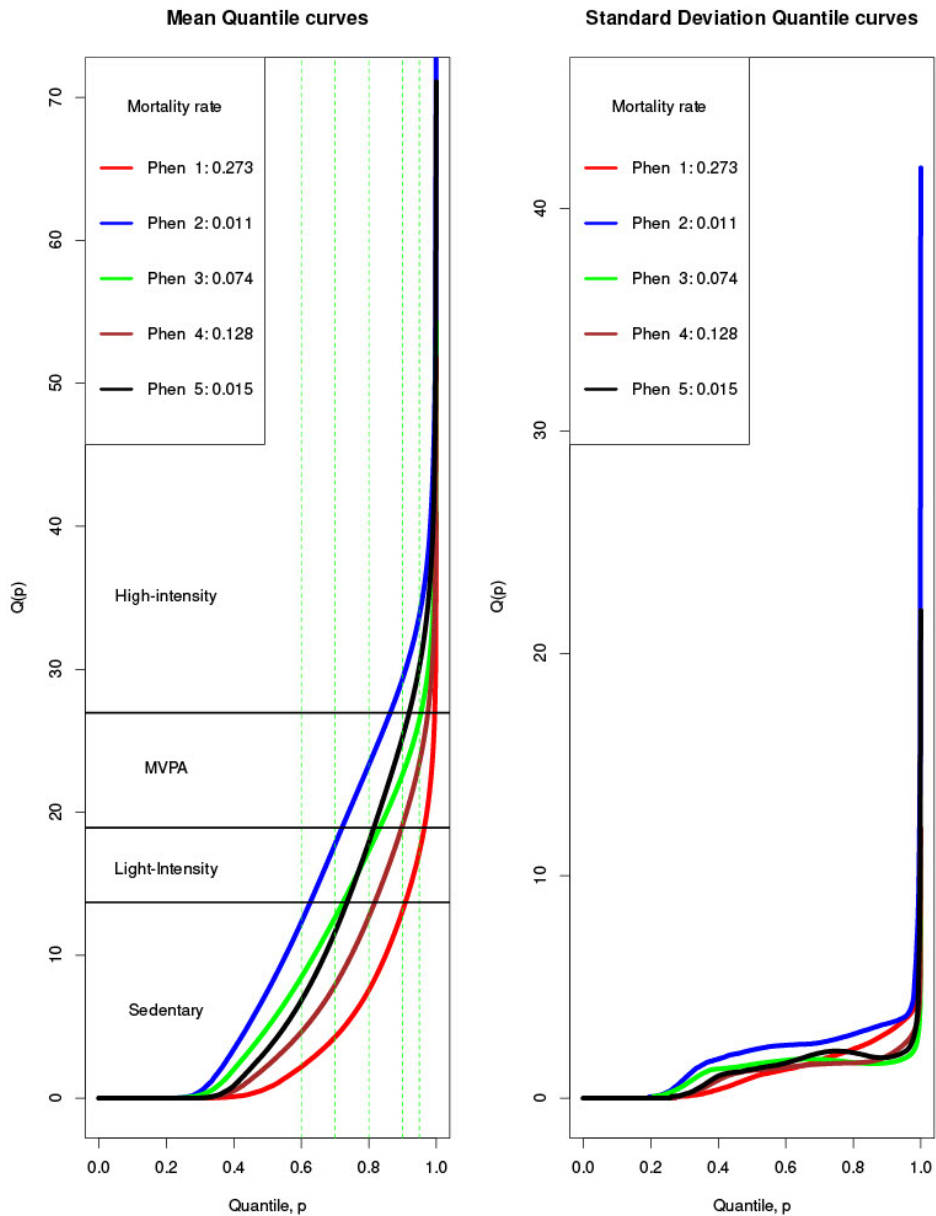


Figure 4.5: Mean and standard deviation of distributional representations for the five phenotypes together with their mortality rate.

than age alone. Our results show that the most inactive physical activity phenotype has a much lower survival probability than the oldest participants in our sample.

Our findings reinforce the idea that information related to physical activity is a key non-pharmacological biomarker of functional decline status and general health [103]. Previous studies [254] have shown the greater clinical sensitivity of physical activity to predict 5-year mortality with the NHANES data 2003-2006 (compared to age), although such level of performance was not observed in the UK-Biobank study [150]. This discrepancy is likely due to the limitation of UK-Biobank study design and the selection bias. Our results were confirmed in

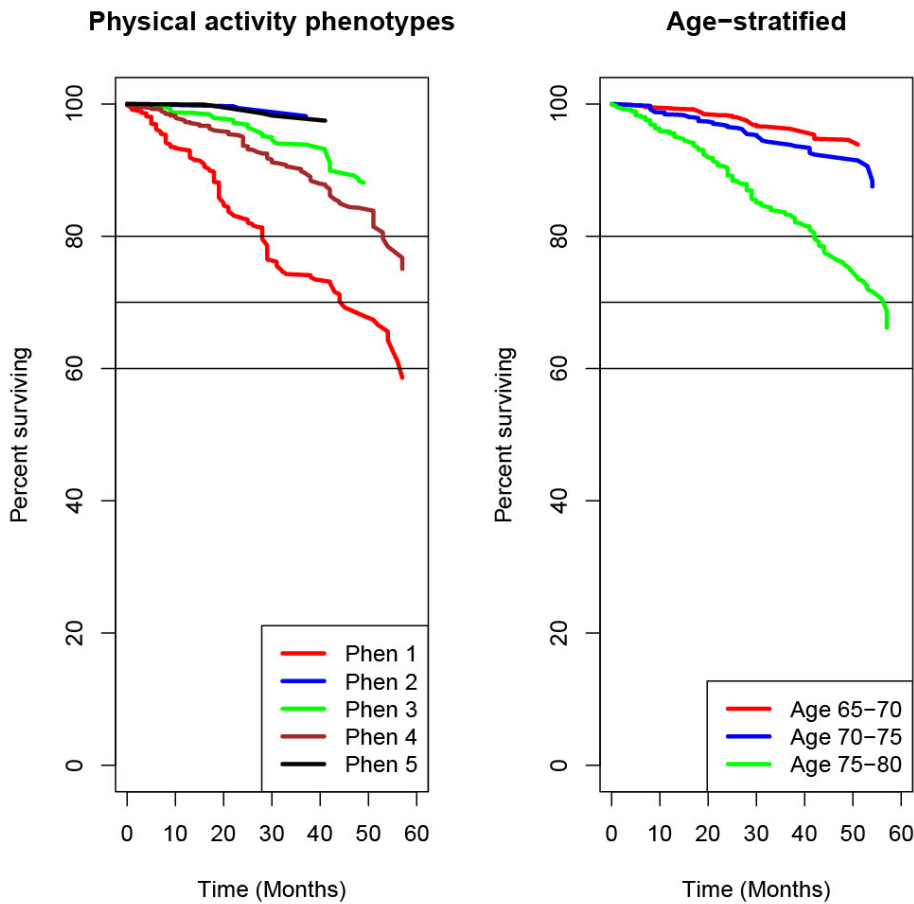


Figure 4.6: Kaplan-Meier curves for each phenotype and age group strata.

multivariate analyses adjusting for potential confounders, such as age, race, sex, comorbidities, or biochemical variables such as cholesterol or triglycerides. We also derived specific weights for the sample included in the analysis, thereby reinforcing the generalizability of our results.

The introduction of new clinical phenotypes with the novel distributional representations allowed us to assess the amount of movement along each intensity recorded by the accelerometer monitor, unlike other existing compositional metrics used in the literature [130]. The summary functional curves (mean and variance) derived from the cluster analysis done in our study show differentiated patterns of physical activity, with remarkable differences across the intensity spectrum from inactivity; and highlight the need to monitor and quantify physical activity more precisely, also to detect the impact on health of intensities often hidden in previous, threshold-based monitoring of physical activity. The phenotypes generated in this study may serve as a formal framework to assess activity changes, for example, with an intervention. In this sense, it is worth mentioning that a reduction in mortality risk between two of the phenotypes might only be due to an increase in the MVPA duration. In addition, the generated phenotypes could be used as a prognosis and monitoring tool. Our work adds to the

	Phenotype 1	Phenotype 2	Phenotype 3	Phenotype 4	Phenotype 5
Gender (Men)	0.65	0.376	0.507	0.633	0.4664
Hight Blood Presure	0.748	0.52	0.62	0.68	0.607
Cancer (yes)	0.273	0.224	0.258	0.286	0.162
Diabetes (yes)	0.531	0.5	0.484	0.591	0.492
No Alcohol consume	0.064	0.039	0.095	0.243	0.307
Middle Alcohol Consume	0.07	0.097	0.175	0.408	0.661
Hight Alcohol Consume	0	0	0.008	0.016	0.026
Mexican American	0.042	0.096	0.054	0.042	0.068
Other Hispanic	0.021	0.136	0.072	0.062	0.11
Non-Hispanic white	0.669	0.488	0.552	0.615	0.618
Non-Hispanic black	0.197	0	0.303	0.216	0.126
Non-Hispanic Asian	0.049	0.144	0.014	0.039	0.068
Other Race, including Multi-Racial	0.021	0.136	0.005	0.026	0.01
Age (years)	75.53 ± 4.87	69.38 ± 4.5	72.35 ± 5.21	73.1 ± 5.22	70.65 ± 4.76
BMI (Kg/m^2)	29 ± 5.75	26 ± 4.85	28 ± 6.15	29 ± 5.69	28 ± 4.76
Combined grip strength (kg)	50.6 ± 18.9	55.7 ± 15.6	58.2 ± 19.1	59.5 ± 19.5	59.8 ± 18.4
Triglycerides (mg/dL)	1.84 ± 0.93	1.47 ± 0.92	1.59 ± 1.04	1.68 ± 1.06	1.74 ± 0.98
Cholesterol (mg/dL)	181.49 ± 43.88	191.46 ± 38.47	193.75 ± 39.69	179.07 ± 41.28	198.79 ± 43.84

Table 4.3: Summary clinical characteristics of participants in each cluster. In binary variables, we show the rate, and in continuous variables, we show the mean and standard deviation.

	Hazard ratio	2.5%	97.5%	Odds ratio	2.5%	97.5%
Phenotype 2	0.07	0.02	0.27	0.06	0.02	0.23
Phenotype 3	0.34	0.19	0.59	0.34	0.19	0.62
Phenotype 4	0.54	0.39	0.75	0.52	0.35	0.75
Phenotype 5	0.09	0.04	0.23	0.09	0.04	0.23
Age (years)	1.12	1.08	1.16	1.14	1.09	1.18
Gender	0.86	0.64	1.16	0.87	0.63	1.21

Table 4.4: Hazard ratios and odds ratios (95% confidence interval) of mortality outcomes associated with different physical activity phenotypes (Reference: Group 1 - Inactivity phenotype.)

(yet scarce) number of works that have explored the idea of physical activity phenotypes as a health monitoring tool [118].

A recent review indicated that there may not exist solid evidence of the benefits of physical activity in patient prognosis in some diseases such as cardiovascular problems [53]. However, it is remarkable to note the sizeable individual response of patients to physical activity and that patients with standardized training programs improve fitness and not necessarily maximal oxygen uptake [116, 181, 186, 233]. Several investigations have shown the relationship between maximal oxygen uptake and the prognosis of these patients and their survival and risk of mortality [116]. Thus, monitoring patient profiles at a high level of resolution is essential to ensure the optimal prescription of physical activity. Indeed, some recent works showed the protective role of light intensity activity for longevity [52, 68]. In addition, the health impact of the optimal intensity-volume coupling is the result of a complex process influenced by many factors such as genetic and environmental, that must be considered in exercise prescription [91, 258]. In this regard, the new patient stratification methods may provide a framework for analyzing these factors and guiding training prescription.

	Hazard ratio	2.5%	97.5%	Odds ratio	2.5%	97.5%
Phenotype 2	0.12	0.02	0.66	0.10	0.02	0.62
Phenotype 3	0.29	0.11	0.75	0.30	0.11	0.80
Phenotype 4	0.55	0.31	0.98	0.49	0.25	0.98
Phenotype 5	0.07	0.01	0.61	0.07	0.01	0.62
Age	1.10	1.04	1.17	1.12	1.05	1.19
Gender (woman)	0.92	0.57	1.47	0.99	0.61	1.63
Other Hispanic	0.81	0.14	4.49	0.69	0.11	4.38
Non-Hispanic white	0.79	0.35	1.80	0.60	0.25	1.43
Non-Hispanic black	0.43	0.15	1.21	0.35	0.12	1.05
Non-Hispanic Asian	1.13	0.31	4.12	1.17	0.27	5.06
Other Race, inc. multi-racial	1.21	0.35	4.14	1.18	0.29	4.76
Blood pressure hight	0.84	0.40	1.80	0.77	0.35	1.71
BMI	0.99	0.93	1.06	0.99	0.93	1.06
Middle Alcohol	0.55	0.33	0.92	0.60	0.33	1.10
High Alcohol	0.97	0.12	8.17	1.68	0.15	18.64
Cancer (no)	0.84	0.48	1.47	0.95	0.51	1.75
Diabetes (no)	0.89	0.59	1.35	0.91	0.58	1.44
Tryglycerides	0.80	0.54	1.19	0.79	0.54	1.17
Cholesterol	1.00	0.99	1.00	1.00	0.99	1.00

Table 4.5: Results of logistics and Cox survey regression model in terms of hazard ratios and odds ratios (95% confidence interval). Reference: Group 1-Inactivity phenotype.

The main strength of this study is that the data used is a random sample from a complex survey design, unlike a significant fraction of physical activity studies that use observational data. Thanks to the NHANES survey design we can obtain more general conclusions about the impact of physical activity on health profiles of the U.S. population. The sample size is another strength; although other cohorts such as the UK Biobank have a more significant number of participants; yet its experimental design has inherent limitations.

Distributional representations provide further advantages in statistical modeling, since they intrinsically capture the information represented by compositional metrics [39, 90, 177] and lead to more refined physical activity profiles which expand along the continuous spectrum of intensity. In addition, the new and more sophisticated pre-processing of accelerometer data [129] leads to greater sensitivity, especially for detecting differences in light and high intensity physical activity.

An inherent limitation of this study is the non-incorporation of potential confounders such as genetic variables, but this is present also in other observational studies. In addition, with a more extensive physical activity monitoring period, we could have drawn more reliable conclusions about the impact of individual physical activity patterns on health. However, in this paper, we analyzed older individuals with lower functional capacity, and this could limit the impact of intraday variability in physical activity patterns (i.e., our population may show more consistent patterns of physical activity than younger and fitter populations). Similarly, the non-

inclusion of the temporal component of distribution representations is another added problem that may lead to new findings of the role of physical activity on health. For example, recent studies have shown the effects of the chronobiology differences in physical activity on health [184].

In summary, this study provides new phenotypes in the aging U.S. population and shows their clinical utility to predict the mortality and survival outcomes in the study sample. Following the principles of precision medicine [138], and according to the phenotypes obtained, differences in light and high-intensity physical activity are relevant for health. The use of distributional representations could be advantageous over more traditional threshold-based analytical approaches to explore the effects of physical activity on human health.

Part III
**Missing data modeling from
complex statistical objects**

5 Statistical independence, variable selection and conformal inference with missing responses in long-term glucose modeling using distributional representations

Missing data are common in epidemiological and medical studies. On the face of it, the extended practice of excluding participants with only partially available data on the variables of interest results in ignoring valuable information, thereby leading to biased estimates which often rely on unrealistic assumptions [123, 157, 210]. To draw reliable conclusions, principled methods are imperative by appropriate modeling of the missingness mechanism [272].

Kernel methods are a class of effective pattern recognition algorithms that are well suited to model nonlinear relations between the response and predictors. These are built on the notion of a kernel function as a similarity function between a new instance and those included in the training set [113, 239]. One of the most significant achievements of kernel methods is the proposal of appropriate kernel functions for managing complex statistical objects such as graphs, strings, or probability distributions [191]. Thus, kernel methods are expanding the range of possible applications for machine learning in the health domain, challenged with the rapid increase in new complex medical data.

The main purpose of this Chapter is to propose a set of kernel methods (Section 5.2) to handle missing responses for statistical independence testing (Section 5.2.2), variable selection (Section 5.2.3), and conformal inference (Section 5.2.4). One major advantage of these methods is their ability to operate as a sequence of predictive stages which increasingly filter out irrelevant information, while also providing an evaluation of the limits of the ensuing predictions. In particular, the present proposal is based on the reproducing kernel Hilbert space (RKHS) framework, providing a Hilbert space of functions that is fully characterized by a reproducing kernel. Importantly, every function in an RKHS that minimizes an empirical risk function can be written as a linear combination of the kernel function evaluated at the training data, and it is ensured that a solution for a machine learning problem that is close to the true solution and also generalizes well to the test data can be obtained. An essential

property of the RKHS framework is that it overcomes the limitations of previous proposals focused on Euclidean functional representations [73].

The proposed methods are motivated by the need to explore the limits of predicting long-term glucose changes in a five-year longitudinal population-based study, including both healthy and diabetic individuals, where a subsample of participants underwent continuous glucose monitoring procedures at the beginning of the study. As expected, a substantial number of participants withdrew from the study, and therefore, an analysis robust to missing values in the response variable is required in order to maintain the validity of the statistical inferences [158]. We include a novel distributional representation for CGM data as a predictor (see Figure 5.1) [178]. Among the different biomarkers, we select the glycated haemoglobin (A1c) as the response variable. A1c is a measure of the average blood glucose level over the past three months, and it is the preferred option because it provides more reproducible values in the laboratory and is subject to less measurement error [242]. Furthermore, we aimed to assess and discuss the residuals and predictive capacity of several variables associated with the evolution of A1c in the long term, providing interpretable clinical phenotypes for large uncertainty cases.

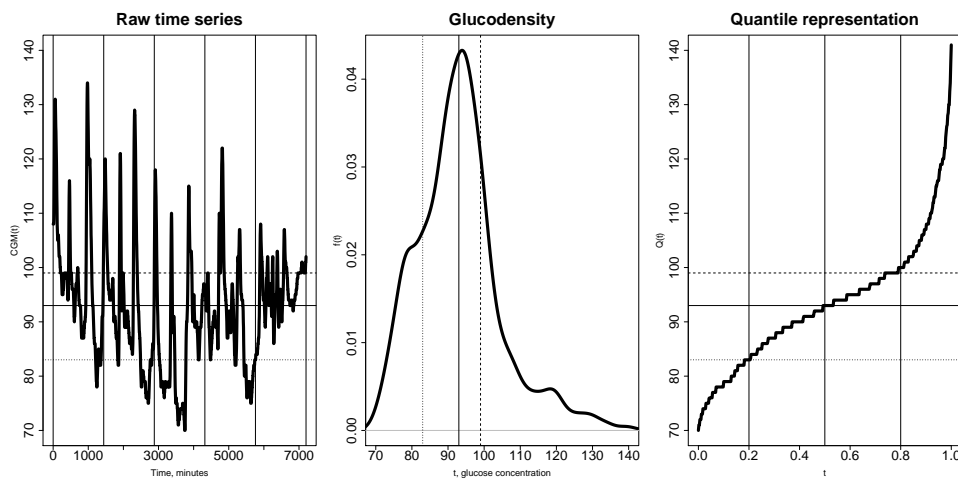


Figure 5.1: (Left) The 5-day CGM recording from a normoglycemic patient is shown. (Center) Glucodensity designates a distributional representation that estimates the proportion of time the patient spent at each glucose concentration. (Right) Quantile representation. Dotted, solid and dashed lines represent concentrations for 20 percent, 50 percent and 80 percent quantiles, respectively.

5.1 Data analysis outline

This chapter presents a data analysis framework designed as a pipeline of kernel methods for predictive problems with missing data. Their subsequent application to diabetes mellitus will allow us to examine the relationship between the baseline characteristics of participants in a five-year study and A1c as the response variable. The proposed framework comprises the following steps:

1. To measure the statistical association between each predictor and the response variable with an efficient statistical independence test. If the response is proven to be independent of a predictor, it can be screened out from further consideration. To this end, we adapted a previous kernel independence test and designed a new bootstrap method to perform test calibration. The test was applied to check the association between some diabetes biomarkers and five-year changes in the A1c variable, $A1c_{5years} - A1c_{initial}$.
2. To identify the best subset of predictors revealing higher-order interactions with the response variable in order to improve the prediction. To this end, we adapted a previous kernel variable selection method and applied it to find the best subset of diabetes biomarkers most strongly associated with $A1c_{5years}$.
3. To explore the prediction ability of a set of explanatory variables through a non-linear regression method. To this end, we adapted a previous kernel ridge regression method and applied it to predict $A1c_{5years}$.
4. To estimate the uncertainty of the predictions. To this end, we designed a new method to provide a prediction interval for the response variable, based on conformal inference. Using this method, we can measure the limits of the regression models previously obtained and, significantly, identify specific patient subpopulations that do not fit the expected behaviour, which is a key issue for clinical decision-making.

5.2 Methods

5.2.1 Preliminaries

We first pose the problem in general terms. Let $(X, Y, \delta) \in \mathcal{X} \times \mathbb{R} \times \{0, 1\}$ be a random vector such that $X = (X^1, \dots, X^p)$ denotes the predictor variables, Y is the response variable, and δ is a binary random variable that indicates whether the response is missing. \mathcal{X} denotes a general topological space, meaning that it can be arbitrary, discrete, continuous, or structured.

Let $\mathcal{D}_n = \{(X_i, Y_i, \delta_i)\}_{i=1}^n$ be a random sample of independent, identically distributed observations, where Y_i is missing if $\delta_i = 0$. We assume that δ conditioned to X is distributed according to $\delta|X \sim Ber(\pi(X))$, with $\pi(\cdot) = P(\delta = 1|X = \cdot)$; hence, some of the predictors can have an impact on the mechanism of missing data $\pi(\cdot) = P(\delta = 1|X = \cdot)$. For instance, in our example, older patients are more reluctant to perform a second CGM monitoring, so the probability of not observing a patient increases with age. We also assume a missing at random (MAR) mechanism in the response Y , namely, δ and Y are conditionally independent given X or, in short, δ is independent $Y|X$.

Consider the following relation between X and Y :

$$Y = m(X) + \epsilon, \quad (5.1)$$

where ϵ denotes a random noise with $E(\epsilon|X) = 0$, and m is the true regression function. Our goal is to predict Y by proposing a new data analysis framework that is robust to datasets in which some values for Y are not observed. To this end, we provide: 1) a method for univariate analysis based on testing the statistical independence between each predictor variable and the response variable; 2) a method for selecting the subset of predictor variables that best predicts the response variable; and 3) methods for predicting the response variable and inferring the uncertainty in the predictions.

These methods are based on the reproducing kernel Hilbert space (RKHS) learning paradigm. The core element of this paradigm is a positive definite kernel function $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ which allows us to measure the similarity between any x and y , with $x, y \in \mathcal{X}$. The positive definiteness of the kernel function guarantees the existence of a dot product space \mathcal{H} and feature mapping $\phi : \mathcal{X} \rightarrow \mathcal{H}$, such that $k_{\mathcal{X}}(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$. Thus, we can express a broad spectrum of statistical modeling problems as linear [152], allowing computational algorithms to easily determine optimal solutions.

5.2.2 Testing statistical independence

We wish to test whether two random variables $X \sim P_X$ and $Y \sim P_Y$ are independent, that is, if we can reject the null hypotheses $H_0 : X$ is independent Y , from n samples $\{(X_i, Y_i)\}_{i=1}^n$. To do this, we must calibrate the test under the null hypothesis to determine the results that are expected to occur with a certain probability if the null hypothesis holds. In our specific case, we must consider the effects of the mechanism of missing data in the response variable Y . We propose a methodology to address this problem based on kernel mean embeddings, which is valid when both covariate and response variables live in a separable Hilbert space. In addition, we introduce a new bootstrap procedure to perform the test calibration adapted to kernel mean embeddings.

Hilbert space embeddings of distributions or, in short, kernel mean embeddings [191], allow us to map distributions into a reproducing kernel Hilbert space (RKHS), in which kernel methods can be extended to probability measures. Kernel mean embeddings can be used to define a metric for distributions, the maximum mean discrepancy (MMD), which in turn can be applied to define an independence test, the Hilbert-Schmidt Independence criterion (HSIC), a non-parametric test of independence with the important property that it does not make any assumption as to the nature of the possible dependence among the two variables [98]. We extended this test to a missing data setting.

A reproducing kernel of \mathcal{H} is a kernel function that satisfies (1) $\forall x \in \mathcal{X}, k_{\mathcal{X}}(\cdot, x) \in \mathcal{H}$, and (2) $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \langle f, k_{\mathcal{X}}(\cdot, x) \rangle_{\mathcal{H}} = f(x)$. If \mathcal{H} has a reproducing kernel, it is said to be an RKHS, $\mathcal{H}_{k_{\mathcal{X}}}$. Kernel mean embedding results from extending the mapping ϕ to the space of probability distributions by representing each distribution as a mean function $\phi(P) = \int_{\mathcal{X}} k(\cdot, x) P(dx)$, resulting in the transformation of a distribution P into an element of the RKHS $\mathcal{H}_{k_{\mathcal{X}}}$. Given two probability measures P and Q , the RKHS distance between their embeddings can be defined as the MMD [97]:

$$MMD(P, Q)_k = \|\phi(P) - \phi(Q)\|_{\mathcal{H}_{k_{\mathcal{X}}}}. \quad (5.2)$$

For the class of *characteristic* kernels, the embeddings are injective, that is, $MMD_k(P, Q) = 0$, if and only if $P = Q$. MMD can then be applied to measure the degree of dependence between the random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ with marginal distributions P_X and P_Y and jointly distributed as $P_{X,Y}$. Note that testing the null hypothesis $H_0 : X$ is independent Y is equivalent to testing $H_0 : P_{X,Y} = P_X P_Y$. We denote by $\phi_X(\cdot)$, $\phi_Y(\cdot)$ and $\phi_{X,Y}(\cdot)$ the kernel mean embeddings of P_X , P_Y , and $P_{X,Y}$, respectively. Assuming \mathcal{H}_{k_z} is a RKHS over $\mathcal{X} \times \mathcal{Y}$ with kernel $k_z((x, y), (x', y')) = k_x(x, x') k_y(y, y')$, so that \mathcal{H}_{k_z} is a direct product $\mathcal{H}_{k_x} \otimes \mathcal{H}_{k_y}$ (with \otimes being the tensor product), then a natural way of testing independence is measuring the MMD distance between the functions $\phi_{X,Y}(\cdot)$ and $\phi_Y(\cdot) \otimes \phi_X(\cdot)$, which can be written as the Hilbert-Schmidt Independence Criterion (HSIC) between X and Y [97], defined as

$$HSIC(P_{X,Y}, P_X P_Y) = \|\phi_{X,Y} - \phi_X \otimes \phi_Y\|_{\mathcal{H}_{k_x} \otimes \mathcal{H}_{k_y}}^2. \quad (5.3)$$

It can be shown that when k_x and k_y are characteristic kernels, $HSIC(P_{X,Y}, P_X P_Y) = 0$ if and only if X is independent of Y . Expanding Equation 5.3, we have

$$\begin{aligned} HSIC(P_{X,Y}, P_X P_Y) &= \langle \phi_{X,Y} - \phi_X \otimes \phi_Y, \phi_{X,Y} - \phi_X \otimes \phi_Y \rangle_{\mathcal{H}_{k_x} \otimes \mathcal{H}_{k_y}} \\ &= \langle \phi_{X,Y}, \phi_{X,Y} \rangle + \langle \phi_X \otimes \phi_Y, \phi_X \otimes \phi_Y \rangle - 2\langle \phi_{X,Y}, \phi_X \otimes \phi_Y \rangle \end{aligned} \quad (5.4)$$

where $\mathcal{H}_{k_x} \otimes \mathcal{H}_{k_y}$ is dropped in the subscript for brevity. From the reproducing property, $E_P[f(x)] = \langle f, \phi(P) \rangle_{\mathcal{H}}$, $\forall f \in \mathcal{H}$, and Fubini's theorem, we obtain

$$\begin{aligned} HSIC(P_{X,Y}, P_X P_Y) &= E_{X,Y,X',Y'} [k_x(X, X') k_y(Y, Y')] \\ &\quad + E_{X,X'} [k_x(X, X')] E_{Y,Y'} [k_y(Y, Y')] \\ &\quad - 2E_{X,Y} [E_{X'} [k_x(X, X')] E_{Y'} [k_y(Y, Y')]], \end{aligned} \quad (5.5)$$

where X' and Y' are independent copies of random variables X and Y , respectively. Ultimately, testing independence involves calculating the squared distance between two mean functions in the appropriate RKHS space, resulting from transforming the original data to capture all distributional differences between both random variables.

In practice, a limited number of random elements, $\{(X_i, Y_i, \delta_i)\}_{i=1}^n$, are observed. Therefore, we must replace the population mean with the sample mean, defined through its empirical distribution. Then, the Hilbert-Schmidt independence criterion can be estimated as

$$\begin{aligned} \widehat{HSIC}(\tilde{P}_{X,Y}, \tilde{P}_X \tilde{P}_Y) &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n k_x(x_i, x_j) k_y(y_i, y_j) \\ &+ \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n k_x(x_i, x_j) \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n k_y(y_i, y_j) \\ &- \frac{1}{n^3} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n k_x(x_i, x_j) k_y(y_i, y_k). \end{aligned} \quad (5.6)$$

Under the MAR assumption, we observe $\{(X_i, Y_i, \delta_i)\}_{i=1}^n$, and we must estimate the missing data mechanism given by the function $\pi(\cdot) = P(\delta = 1 | X = \cdot)$. Several procedures have been proposed in the literature for this purpose, such as logistic regression, lasso, random forest, and ensemble methods. Subsequently, we re-weighted the dataset, taking into account the difficulty of observing the response of the i^{th} datum. In particular, we associate weight w_i with the i^{th} datum via an inverse probability weighting (IPW) estimator [272] given by

$$w_i = \frac{\delta_i}{n\pi(x_i)}, \quad i = 1, \dots, n, \quad (5.7)$$

which results in assigning large w_i values as the probability of observing a response decreases. Using this procedure, we obtain an asymptotic unbiased estimator that balances the sampling mechanism and allows us to make a proper inference according to the target population examined.

We define the normalized weight of w_i as

$$w_i^* = \frac{w_i}{\sum_{i=1}^n w_i}, \quad i = 1, \dots, n. \quad (5.8)$$

We denote the estimated and normalized i^{th} weight as \tilde{w}_i and \tilde{w}_i^* , respectively, after estimating $\tilde{\pi}(\cdot)$.

To obtain an estimator of HSIC with missing data, it is sufficient to replace the uniform weight $1/n$ of the empirical distribution with normalised weights $\tilde{W}^* = (\tilde{w}_1^*, \dots, \tilde{w}_n^*)$ in Equation 5.6. Thus, we obtain

$$\begin{aligned} \widehat{HSIC}(\tilde{P}_{X,Y}, \tilde{P}_X \tilde{P}_Y) &= \sum_{i=1}^n \sum_{j=1}^n \tilde{w}_i^* \tilde{w}_j^* k_x(X_i, X_j) k_y(Y_i, Y_j) \\ &+ \sum_{i=1}^n \sum_{j=1}^n \tilde{w}_i^* \tilde{w}_j^* k_x(X_i, X_j) \sum_{i=1}^n \sum_{j=1}^n \tilde{w}_i^* \tilde{w}_j^* k_y(Y_i, Y_j) \\ &- \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \tilde{w}_i^* \tilde{w}_j^* \tilde{w}_k^* k_x(X_i, X_j) k_y(Y_i, Y_k). \end{aligned} \quad (5.9)$$



Calibration under the null hypothesis with the precedent statistic is not trivial, and the permutation approach is generally not valid because the response Y is not exchangeable due to the non-homogeneous missing data mechanism. To overcome this difficulty, we propose a novel bootstrap approach that properly deals with non-vectorial predictors [66].

Under null hypothesis $H_0 : P_{X,Y} = P_X P_Y$, it can be assumed that $\phi_{X,Y}(\cdot) - \phi_X(\cdot) \otimes \phi_Y(\cdot) = 0(\cdot)$. Therefore,

$$\begin{aligned} \widetilde{HSIC}(\widetilde{P}_{X,Y}, \widetilde{P}_X \widetilde{P}_Y) &= \langle \widetilde{\phi}_{X,Y} - \widetilde{\phi}_X \otimes \widetilde{\phi}_Y, \widetilde{\phi}_{X,Y} - \widetilde{\phi}_X \otimes \widetilde{\phi}_Y \rangle_{\mathcal{H}_x \otimes \mathcal{H}_y} \\ &= \langle \widetilde{\phi}_{X,Y} - \phi_{X,Y} + \phi_X \otimes \phi_Y - \widetilde{\phi}_X \otimes \widetilde{\phi}_Y, \\ &\quad \widetilde{\phi}_{X,Y} - \phi_{X,Y} + \phi_X \otimes \phi_Y - \widetilde{\phi}_X \otimes \widetilde{\phi}_Y \rangle. \end{aligned} \quad (5.10)$$

Then, a natural bootstrap procedure that allows us to estimate the p -value for the test of independence is developed as follows:

1. To randomly sample with replacement n elements from the original dataset \mathcal{D}_n , repeating m times. We denote by $\mathcal{D}_n^{j*} = \{(X_i^{j*}, Y_i^{j*}, \delta_i^{j*})\}_{i=1}^n$, $j = 1, \dots, m$ the j^{th} random sample obtained.
2. To calculate $\widetilde{HSIC}^{j*}(\widetilde{P}_{X,Y}, \widetilde{P}_X \widetilde{P}_Y)$ as

$$\begin{aligned} \widetilde{HSIC}^{j*}(\widetilde{P}_{X,Y}, \widetilde{P}_X \widetilde{P}_Y) &= \langle \widetilde{\phi}_{X,Y} - \widetilde{\phi}_{X,Y}^{j*} + \widetilde{\phi}_X^{j*} \otimes \widetilde{\phi}_Y^{j*} - \widetilde{\phi}_X \otimes \widetilde{\phi}_Y, \\ &\quad \widetilde{\phi}_{X,Y} - \widetilde{\phi}_{X,Y}^{j*} + \widetilde{\phi}_X^{j*} \otimes \widetilde{\phi}_Y^{j*} - \widetilde{\phi}_X \otimes \widetilde{\phi}_Y \rangle_{\mathcal{H}_x \otimes \mathcal{H}_y}, \end{aligned} \quad (5.11)$$

where $j = 1, \dots, m$, $\widetilde{\phi}_{X,Y}^{j*}(\cdot)$, $\widetilde{\phi}_X^{j*}(\cdot)$, and $\widetilde{\phi}_Y^{j*}(\cdot)$ are the kernel mean embeddings estimated from the j^{th} bootstrap sample $\mathcal{D}_n^{j*} = \{(X_i^{j*}, Y_i^{j*}, \delta_i^{j*})\}_{i=1}^n$.

3. To estimate the p -value as

$$p\text{-value} = \frac{1}{m} \sum_{j=1}^m \mathbf{1} \left(\widetilde{HSIC}^{j*}(\widetilde{P}_{X,Y}, \widetilde{P}_X \widetilde{P}_Y) \geq \widetilde{HSIC}(\widetilde{P}_{X,Y}, \widetilde{P}_X \widetilde{P}_Y) \right). \quad (5.12)$$

Bootstrap consistency with missing data can be proved by using standard tools of empirical process theory [278], and it is provided in the Appendix.

5.2.3 Variable selection

Independence screening methods select predictor variables based on individual prediction ability; hence, they are ineffective in selecting a subset of variables that are individually weak but strong in combination. Subset selection aims to overcome this drawback by considering and evaluating the prediction ability of a subset of variables as a whole. One popular approach to subset selection is to directly optimize an objective function consisting of two terms: a data

fitting term to attain prediction accuracy and a regularization term to penalize a large number of variables [104].

Subset selection has recently been approached using the RKHS paradigm with satisfactory results. Two strategies stand out: first, minimizing the trace of the conditional covariance operator [40] and second, identifying those variables with a non-zero gradient function [292]. The first strategy scales poorly with the number of variables used. The second strategy can be formulated in a more compact manner. Here, it is extended to missing data.

Following [292], we identify the relevant predictors by learning the gradient of regression function m . Thus, it is assumed that if variable X^r is not relevant for predicting Y , then $g_r = \partial m(X)/\partial X^r = 0$ for any value of X . Let us denote by $g(X) = \nabla m(X) = (g_1(X), \dots, g_p(X))^T$ the gradient function. In a small neighborhood of X_i we can use the Taylor expansion to approximate $m(X)$, so when X_j is sufficiently close to X_i , $m(X_j) \approx Y_i + g(X_i)(X_j - X_i)$. We then define the estimation error as a function of $g(\cdot)$:

$$E(g) = E_{X,Y,X',Y'} \left[\omega(X, X') (Y - Y' - g(X)^T (X - X')) \right]^2,$$

where X', Y' denote independent and random variables distributed as X and Y , respectively. Function $\omega(X_i, X_j)$ is an appropriate weight function that decreases as $\|X_i - X_j\|$ increases and ensures that the local neighbourhood of X_i contributes more to estimating the gradient $g(X_i)$. Typically, $\omega(X_i, X_j) = e^{-\|X_i - X_j\|^2/\tau_n^2}$, where τ_n^2 is a positive parameter which must be adjusted to ensure asymptotic estimation consistency.

Because only a limited number of samples $\{(X_i, Y_i, \delta_i)\}_{i=1}^n$ are observed, we approximate $E(g)$ using its empirical counterpart

$$\tilde{E}(g) = \frac{1}{n^2} \sum_{i,j=1}^n \omega_{ij} (Y_j - Y_i - g(X_i)^T (X_j - X_i))^2, \quad (5.13)$$

where $\omega_{ij} = \omega(X_i, X_j)$.

We can add a regularization term for enforcing a sparsity constraint on the gradient vector, with the aim of shrinking the partial derivatives g_r towards zero with respect to irrelevant variables. We then add the term $J(g) = \lambda_n \sum_{r=1}^p \eta_r J(g_r)$, where η_r are adaptive tuning parameters. On the other hand, we can define the estimation error in (5.13) as a functional in the RKHS \mathcal{H}_k^p , and thus $g \in \mathcal{H}_k^p$ and $\mathcal{E} : \mathcal{H}_k \times \dots \times \mathcal{H}_k \rightarrow \mathbb{R}^+$, induced by a pre-specified positive kernel k . Thus, we propose the following optimization formula to learn the gradient vector:

$$\arg \min_{g \in \mathcal{H}_k^p} \frac{1}{n^2} \sum_{i,j=1}^n \omega_{ij} (Y_i - Y_j - g(X_i)^T (X_i - X_j))^2 + J(g). \quad (5.14)$$

Under the MAR assumption, we propose substituting ω_{ij} weights with $\tilde{\omega}_{ij}^* = \tilde{\omega}_i^* \tilde{\omega}_j^* \omega_{ij}$, where $\tilde{\omega}_i^*$ and $\tilde{\omega}_j^*$ denote the estimated normalized weights associated with data i^{th} and j^{th} ,

respectively, according to (5.8). The variable selection expression can be rewritten as follows:

$$\operatorname{argmin}_{g \in \mathcal{H}_k^p} \frac{1}{n^2} \sum_{i,j=1}^n \tilde{\omega}_{ij}^* \left(Y_i - Y_j - g(X_i)^T (X_i - X_j) \right)^2 + J(g). \quad (5.15)$$

The representer theorem states that the minimizer of (5.15) can be represented as a finite linear combination of kernel products evaluated on the dataset samples [238]:

$$g_r(\cdot) = \sum_{i=1}^n \alpha_i^r k_{\mathcal{X}} f(\cdot, X_i), \quad r = 1, \dots, p, \quad (5.16)$$

where $\alpha^r \in \mathbb{R}^n$. Given this representation, $g_r(\cdot) = 0$ iff $\alpha^r = (\alpha_1^r, \dots, \alpha_n^r)^T = (0, \dots, 0)^T$, or more concisely, $\|\alpha^r\|_2 = 0$.

Several regularization terms have been considered in previous studies. We adopted the group lasso penalty [85, 292]:

$$J(g_r) = \inf \left\{ \|\alpha^r\|_2 : g_r(\cdot) = \sum_{i=1}^n \alpha_i^r k_{\mathcal{X}}(\cdot, X_i) \right\}, \quad (5.17)$$

which encourages all α_i^r , $i = 1, \dots, n$ to be selected or shrunk to zero together to achieve the purpose of variable selection. Thus, our optimization problem can be rewritten as:

$$\operatorname{argmin}_{\alpha^1, \dots, \alpha^p} \sum_{i,j=1}^n \hat{\omega}_{ij}^* (y_i - f^*(x_i, x_j))^2 + \lambda_n \sum_{r=1}^p \eta_r \|\alpha^r\|_2, \quad (5.18)$$

where $f^*(X_i, X_j) = y_j - \sum_{r=1}^p k_i^T \alpha^r (x_i^r - X_j^r)$, $k_i = (k(X_i, X_1), \dots, k(X_i, X_n))^T$ is the i^{th} row of $K = (k(X_i, X_j))_{n \times n}$, and λ_n are tuning parameters. This last expression simplifies the original optimization framework (5.14) from a functional space to a vector space, and it can be solved in $O(|U|^2 p^2)$ using a block coordinate descent algorithm [292].

5.2.4 Prediction and uncertainty analysis

Let us recall that the ultimate goal is to predict Y using the information provided by predictor variables X . To achieve this aim, we adopt the kernel ridge regression approach proposed by Liu and Goldberg [159]. However, we draw on the linear regression theory to efficiently compute the leave-one-out cross-validation regularization parameter. This class of regularization parameters has been proven to largely shape the model performance [154]. Furthermore, estimating the uncertainty of predictions by providing robust confidence intervals is a valuable tool for subsequent decisions. Thus, we compute intervals with good finite sample coverage using advances in conformal inference recently exploited in causal theory [147].

Let us assume a linear regression model:

$$Y_i = m(X_i) + \epsilon = X_i^T \beta + \epsilon \quad i = 1, \dots, n, \quad (5.19)$$

where β is the vector of coefficients of the linear model. Given the original dataset $\mathcal{D}_n = \{(X_i, Y_i, \delta_i)\}_{i=1}^n$, kernel ridge regression is based on solving the following optimization problem:

$$\tilde{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (Y_i - X_i^T \beta)^2 + \lambda \|\beta\|_2^2, \quad (5.20)$$

which is solved by $\tilde{\beta} = (X^T X + \lambda I)^{-1} X^T Y$, where $X = (X_1, \dots, X_n)^T$, $Y = (Y_1, \dots, Y_n)^T$, and $\lambda > 0$ is the smoothing parameter of the regularization term.

Let \mathcal{H}_k be an RKHS with a kernel k_X . Then, by replacing every X_i with $\phi(X_i)$ and assuming that $\beta = \sum_{i=1}^n \alpha_i \phi(X_i)$, we obtain an analogue solution to that of Equation (5.20) by exploiting the linear structure of the problem but changing the usual dot product by the inner product of the selected RKHS. In particular, $\tilde{\alpha} = (K + \lambda I)^{-1} Y$, where $K = (k(X_i, X_j))_{n \times n}$.

In [159], the authors proposed two estimators for the missing data. In both cases, the solution has the same closed-form expression, given by the representer theorem. The first is $\tilde{\alpha} = (\lambda I + WK)^{-1} WY$, where the missing data mechanism is handled using the IPW estimator. The second is obtained through doubly robust estimation, combining a preliminary imputation of the missing response with the IPW estimator:

$$\tilde{\alpha} = (K + \lambda I)^{-1} (WY + (I - W)\mu(X)), \quad (5.21)$$

where $W = \text{diag}(w_1, \dots, w_n)$ denotes a diagonal matrix containing the weights (see Equation 5.7) and $\mu(X) = (\mu(X_1), \dots, \mu(X_n))^T$ denotes the imputation function.

Doubly robust estimators achieve optimal asymptotic variance when their weights w_1, \dots, w_n and imputation function $\mu(\cdot)$ are correctly specified, and only one of them needs to be correctly specified to achieve consistency. However, when any of them fails, the regression model performance can deteriorate dramatically with a finite sample [132, 282], thereby failing to provide real advantages with respect to the IPW estimator.

The impact of the smoothing parameter on model generalization is an essential issue for the ensuing performance and is strongly related to the minimum-norm interpolation problem in the context of RKHS. Therefore, we propose the selection of the smoothing parameter through *leave-one-out* cross-validation by adapting estimators to missing data [154].

To supply a prediction interval for the response with a confidence level of $1 - \alpha$, we provide a novel algorithm for performing conformal inference [146, 147], which is valid for handling missing responses and heteroscedastic noise.

We randomly split the dataset $\mathcal{D}_n = \{(X_i, Y_i, \delta_i)\}_{i=1}^n$ into training and test sets $\mathcal{D}^{train} = \{(X_i^{train}, Y_i^{train}, \delta_i^{train})\}_{i=1}^{n_1}$ and $\mathcal{D}^{test} = \{(X_i^{test}, Y_i^{test}, \delta_i^{test})\}_{i=1}^{n_2}$, where $n = n_1 + n_2$.

For a given new observation X_{n+1} we go through the following steps:

1. Fit the mean regression function $\tilde{m}(\cdot)$ from the set \mathcal{D}^{train} , according to Equation 5.21.

2. Compute the residuals $\tilde{\epsilon}_i = |Y_i^{test} - \tilde{m}(X_i^{test})|/\tilde{\sigma}(X_i^{test})$, for every $i = 1, \dots, n_2$ with $\delta_i^{test} = 1$. The value $\tilde{\sigma}(X_i^{test})$ is estimated by a regression function that predicts the absolute deviation of the residuals fitted with the training sample.
3. Estimate the empirical distribution as follows:

$$\tilde{F}_{n_2+1}^\epsilon(x) = \frac{1}{\sum_{i=1}^{n_2+1} \tilde{w}_i^{test}} \left(\sum_{i=1}^{n_2} 1\{\tilde{\epsilon}_i \leq x\} \tilde{w}_i^{test} + \tilde{w}_{n_2+1}^{test} \right), \quad (5.22)$$

where we also incorporate the weights of X_{n+1} and $\tilde{w}_{n_2+1}^{test}$ into the estimation.

4. Compute the $1 - \alpha$ quantile, $\tilde{q}_{1-\alpha}$, from $\tilde{F}_{n_2+1}^\epsilon$.
5. Finally, return $[\tilde{f}(X_{n+1}) - \tilde{q}_{1-\alpha} \tilde{\sigma}(X_{n+1}), \tilde{f}(X_{n+1}) + \tilde{q}_{1-\alpha} \tilde{\sigma}(X_{n+1})]$ as the required prediction interval.

5.3 An application in modelling long-term changes in glucose levels

Diabetes mellitus is one of the most critical public health problems and the ninth major cause of mortality worldwide [299]. At present, over 416 million and 47 million patients have type 2 (T2D) and type 1 (T1D) diabetes, respectively, [234]. Significantly, around 50% of patients with diabetes are undiagnosed [234]. Considering the impact of this pandemic among the general population, there is a need for new health policies and guidelines to enable early recognition of at-risk patients and improvement in the methodology of disease diagnosis in the standard clinical routine [120].

Some previous studies have focused on developing predictive models for patient stratification. Thus, the Finnish FINDRISC provides a diabetes score to predict the probability of developing diabetes within ten years using logistic regression [170]. In addition, the German GDRS provides a different score to predict the time to becoming a diabetic person using a survival model based on Cox regression [192]. In contrast, some authors argue against using thresholds and categorising patients into different ranges of glucose levels, and hence, against defining diabetes as a homogeneous disease [87].

The availability and rapid adoption of new digital medical devices have enabled an emerging clinical paradigm based on precision medicine, which will be called to raise early diagnosis and guide subsequent clinical decision-making through the intensive use of statistical models and machine learning techniques [46, 138, 240, 269]. In the case of diabetes, the latest advances in sensing technology allow for the assessment of glucose metabolism at a high-resolution level by capturing the individual differences in glucose fluctuations at different time scales via continuous glucose monitoring (CGM) [295]. In this sense, although T1D cannot be prevented at present, monitoring is of utmost importance. Recent studies have shown improved glycemic

control and decreased rates of hypoglycaemia in T1D patients using CGM, leading both the Endocrine Society and the American Diabetes Association to state that CGM use represents the standard of care for T1D [10, 212]. With respect to T2D, strong scientific evidence shows that it can be prevented by regular exercise, healthy eating, and the control of blood pressure and lipids [203], spurring innovation in wearable technology to enable its prediction and prevention in the general population [131].

Still, few studies have explored the use of CGM data from healthy populations to draw new conclusions regarding glucose homeostasis. It is worth mentioning [105], which provides some remarkable insights into the heterogeneity of glucose dysregulation, highlighting the inadequacy of a common designation as T2D for categorising and subsequently managing predictably different conditions. Importantly, this study refutes the assumption of similar glucose excursions for the same amount and composition of food. Ultimately, the specific glucose profiles observed for each patient depend on the complex interplay between the pathophysiological mechanisms of insulin resistance and insulin secretion [105, 295], thus enabling the treatment of the glycemic profile of an individual as a personal signature of glucose homeostasis. Accordingly, an appropriate interpretation of CGM data could help identify early stages of glucose dysregulation in apparently healthy individuals, with the possibility of providing early and tailored interventions. In this sense, there is a need for further research on the predictive value of CGM data [295].

This study aimed to examine the predictability of long-term changes in glucose levels by using a random sample of the general population. The exploration of the predictive value of the information provided by CGM data is of particular interest. For this purpose, we use the glucodensity representation. Intuitively, glucodensity is more sensitive than the previous CGM summary metrics. We then explored its use in modeling long-term glucose changes and compared it with TBR, TIR, and TAR measures. In addition, we also considered different summary metrics derived from CGM data [96, 230]: CONGA (continuous overall net glycemic action), MAGE (mean amplitude of glycemic excursions), and MODD (mean of daily differences).

5.3.1 The AEGIS diabetes study

Let us recall that the AEGIS diabetes population study, conducted in the Spanish town of A Estrada (Galicia), aimed to analyze the longitudinal changes in some clinical features related to circulating glucose in 1516 patients over 5 years. In addition, non-routine medical tests, such as CGM, are performed every five years on a randomized subset composed of 581 patients. At the beginning of this study [101], 581 participants were randomly selected to wear a CGM device for 3-7 days. Of the 581 participants, 68 were diagnosed with diabetes before the start of the study and 22 were diagnosed during the study. Table 3.1 lists the baseline characteristics of the 581 patients grouped by sex. After a five-year follow-up, a significant fraction of those

individuals did not agree to perform a second glucose monitoring, while some five-year relevant outcomes such as A1c could only be measured in 339 patients. Complete details of the study design and measurement methodology protocol can be found in [101].

5.3.2 Integrating multiple data sources

RKHS offers a powerful and natural data analysis paradigm that can cope with data of different natures [34]. A crucial issue is to select a suitable kernel that accurately captures the differences and specific characteristics of each information source examined. In our particular case, we take into account a continuous probability distribution and certain real-valued and categorical data $X = (X^{gluco}, X^{real}, X^{categ})$. A reasonable choice commonly used in the literature is the Laplacian kernel, $K = (k(X_i, X_j))_{n \times n}$, where $k(X_i, X_j) = e^{-\frac{\|X_i - X_j\|}{\sigma}}$. Here, we propose using the Laplacian kernel with the standard Euclidean distance as a universal and characteristic kernel in a real vector space. Moreover, it can be shown that the Laplacian kernel retains these properties considering the set of continuous density functions endowed with 2–Wasserstein distance, providing theoretical guarantees that we can approximate a large variety of regression functions. Based on the connection between positive kernels and negative-type metrics [24, 241], we propose using a simple and global Laplacian kernel that integrates these three sources:

$$k_{\mathcal{X}}(X_i, X_j) = e^{-\left(a \frac{\|X_i^{gluco} - X_j^{gluco}\|}{\sigma_{gluco}} + b \frac{\|X_i^{real} - X_j^{real}\|}{\sigma_{real}} + c \frac{\|X_i^{categ} - X_j^{categ}\|}{\sigma_{categ}}\right)}, \quad (5.23)$$

where $a, b, c, \sigma_{gluco}, \sigma_{real}, \sigma_{categ} > 0$ and we assume for the sake of simplicity that $(a, b, c) \in \mathcal{S}^2$, where \mathcal{S}^2 is a 2-simplex, $\mathcal{S}^2 = \{(a, b, c) \in \mathbb{R}^3 : a + b + c = 1; 0 \leq a \leq 1, 0 \leq b \leq 1, 0 \leq c \leq 1\}$.

5.4 Results

The present framework of predictive tools allows us to answer some open clinical questions concerning long-term glucose changes from the analysis of data in the AEGIS study.

1. Glycated haemoglobin A1c is a haemoglobin-glucose combination formed within the cell; it is a useful indicator of long-term blood glucose control and is considered the standard biomarker for diabetes diagnosis and management. *Is there a prognostic variable that can be used to predict future A1c changes in healthy individuals?*
2. Current medical literature assigns a considerable relevance to all of the predictor variables listed in Table 3.1 for characterizing the evolution and impact of glucose homeostasis on health. However, from a biological perspective, these variables are well known to be

Variable	$p - value$
Age	0.32
Sex	0.16
FPG	0.50
HOMA-IR	0.52
BMI	0.42
A1c	0.03
CONGA	0.24
MAGE	0.68
MODD	0.16
Glucodensity	< 0.001

Table 5.1: Estimated raw p-values of A1c total variation vs each biomarker using the method proposed in Section 5.2.2 with normoglycemic patients.

highly correlated. *Can we identify a reduced subset of relevant explanatory variables to predict five-year A1c changes?*

- CGM technology may provide a more suitable tool for assessing glucose homeostasis than traditional diabetes biomarkers. *How does CGM data improve our ability to predict future A1c changes?*
- An increased uncertainty in predictions for a specific region of the feature space may suggest a subpopulation that has not been properly modeled. *Can we provide a characterization of individuals for whom the model yields a less accurate prediction?*

5.4.1 Is there a prognostic variable that can be used to predict future A1c changes in healthy individuals?

To answer this question, we studied whether there was any evidence of a univariate statistical association for normoglycemic patients ($A1c < 5.7\%$ and $FPG < 100 \text{ mg/dL}$) between glucose variation measured by $A1c_{5years} - A1c_{initial}$ and the predictor variables shown in Table 3.1.

For this purpose, we use the Hilbert-Schmidt independence criterion proposed in the context of missing data (Section 5.2.2) together with a specific bootstrap approach designed for this task. The underlying mechanism of missing data was estimated using univariate logistic regression.

The results in Table 5.1 show that the only statistically significant variables with a p-value of less than 5% are glucodensity and basal A1c. Figure 5.2 illustrates that the marginal relationships with other variables, if any, are weak.

5.4.2 Can we identify a reduced subset of relevant variables to predict five-year

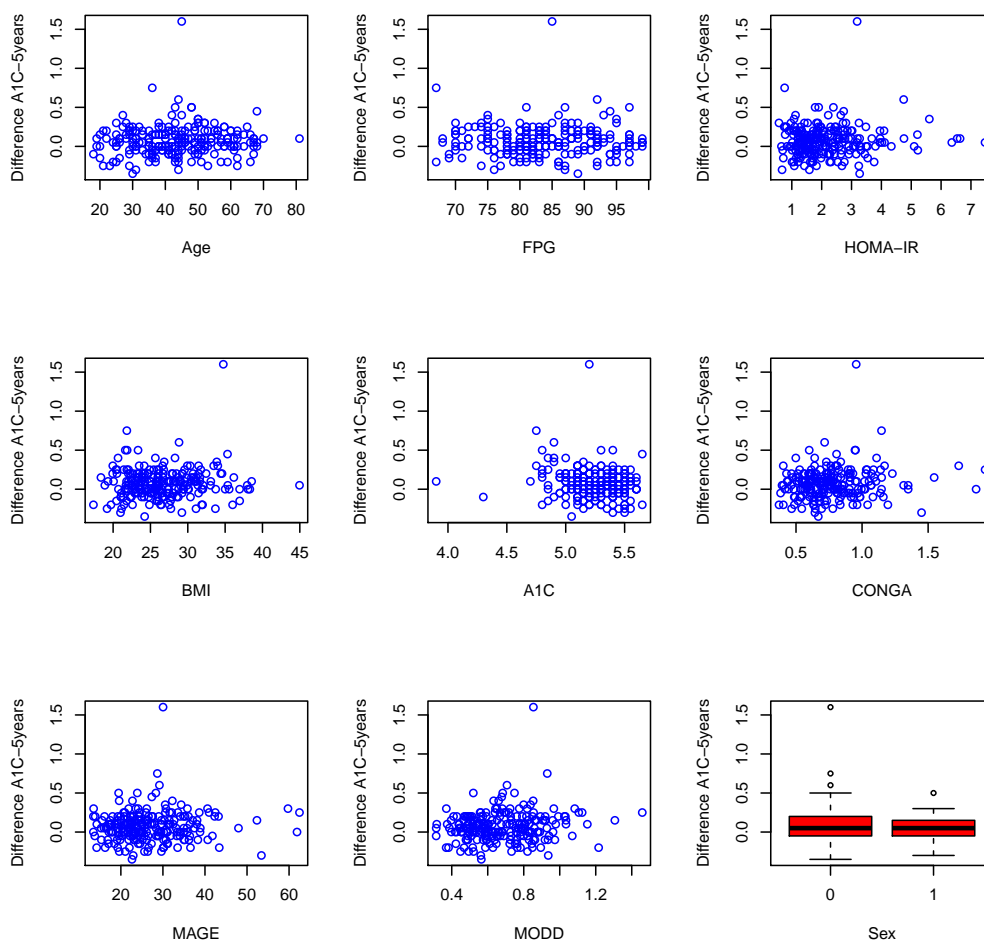


Figure 5.2: Marginal dependence relation between examined variables in the AEGIS database.

A1c changes?

Multivariate models can exploit higher-order interactions between the predictors and the response to improve predictions. However, a key point in increasing the interpretability and generalization ability of the model is to identify a subset of the variables that capture the essential information in the dataset, thus removing redundancy. We adjusted the method proposed in Section 5.2.3 to find the subset of variables most strongly associated with $A1c_{5years}$. For this purpose, both diabetic and non-diabetic patients were analysed, and we considered all the variables in Table 3.1 except for sex. We also included the TBR, TIR, and TAR measures specified in Section 6.1. To avoid overfitting and improve the reproducibility of the results, we selected model parameters by cross-validation. We estimated the underlying missing data mechanism using lasso logistic regression.

Finally, the explanatory variables selected by the algorithm were age, $A1c_{initial}$, FPG, BMI, and MAGE. Notably, the CGM contribution is made through the specific MAGE index, leaving

aside time in ranges.

5.4.3 How does CGM data improve our ability to predict future A1c changes?

To answer this question, we fit several kernel ridge regression models (Section 5.2.4) for predicting $A1c_{5years}$: 1) excluding CGM data as a predictor; 2) including CGM data through the MAGE index; 3) including CGM data through the above-mentioned time in ranges; and 4) including CGM data through glucodensity representation. Both share age, $A1c_{initial}$, FPG, and BMI as covariates. The kernel selection and parameter tuning were calibrated as described in Section 5.3.2.

To compare the performance of these regression models, we used R^2 after including the specific missing data mechanism:

$$R^2 = 1 - \frac{\sum_{i=1}^n \left(Y_i - \frac{\sum_{j=1, j \neq i}^n w_j Y_j}{\sum_{j=1, j \neq i}^n w_j} \right)^2}{\sum_{i=1}^n \left(Y_i - \tilde{f}_{-i}(X_i) \right)^2}, \quad (5.24)$$

where $\tilde{f}_{-i}(\cdot)$, is the regression function fitted to $\{(X_j, Y_j, \delta_j)\}_{j \neq i}^n$, i.e. excluding the i^{th} -datum.

The performance results, obtained using leave-one cross-validation, are: 1) $R_{noCGM}^2 = 0.61$; 2) $R_{MAGE}^2 = 0.65$, 3) $R_{TIR}^2 = 0.64$; and 4) $R_{gluco}^2 = 0.71$. Figure 5.3 depicts the residuals versus the $A1c_{initial}$ values. As can be seen, the highest residuals are found in diabetic patients; otherwise, the distribution of residuals is heterogeneous. Ultimately, CGM data represented by glucodensities provide valuable information for predicting long-term A1c changes.

5.4.4 Can we provide a characterization of individuals for whom the model yields a less accurate prediction?

Figure 5.4 depicts prediction intervals at a confidence level of 90%, after applying conformal inference (Section 5.2.4) to measure the uncertainty of the predictions performed by the above regression model (CGM data included as a covariate).

We regard an $A1c_{5year}$ prediction as significantly affected by uncertainty if the length of the interval is greater than 0.7 because a deviation greater than this threshold can entail a change in the glyceic state of the patient, for example, from normoglycemic to diabetes. Hence, we can identify certain clinical features that allow us to assign each patient to high- or low-variability groups based on the uncertainty of future glucose values. This can be useful to phenotypically characterize some subpopulations for which the model provides an unreliable prediction, and therefore, a more personalised follow-up is advisable. In particular, Figure 5.5

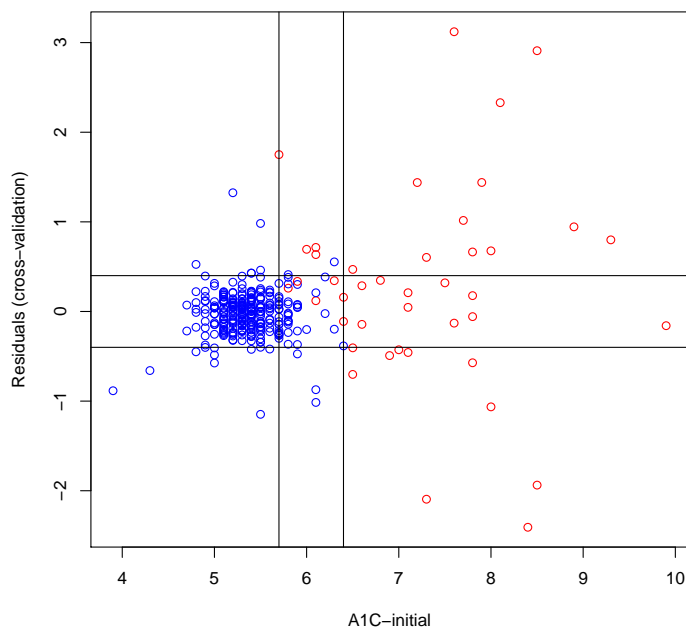


Figure 5.3: Residuals vs. $A1c_{initial}$ for the model that includes glucodensity as a covariate in the AEGIS database. Red circles correspond to diabetic patients

shows that long-term changes cannot be adequately predicted for individuals with elevated FPG levels. The same holds true for individuals with FPG levels in the normoglycemic range and overweight. More refined decision rules can be established at higher measurement costs.

5.5 Discussion

The above analysis aimed to explore the predictability of glucose regulation in the general population by studying the relationship between patient basal characteristics at the start of a longitudinal study and A1c values obtained five years later. Specifically, we intend to exploit the ability of CGM data to effectively capture a personal signature of glucose homeostasis through the inclusion of glucodensity, a novel distributional representation of glucose excursions, as a predictor.

The AEGIS study makes it possible to assess the predictive capacity of glucodensity in the context of well-known biomarkers for diabetes diagnosis and control, providing some interesting findings. First, glucodensity shows a significant association with A1c changes, using statistical dependence measures in normoglycemic subjects. Nevertheless, the weak marginal association of biomarkers with $A1c_{5years}$ suggests the need for a multivariate approach to capture the complexity of long-term glucose changes. The application of a variable selection procedure supplies a subset of relevant biomarkers (age, $A1c_{initial}$, FPG, BMI, and MAGE) resulting from the detection of higher-order interactions with $A1c_{5years}$. Then, the ability to predict

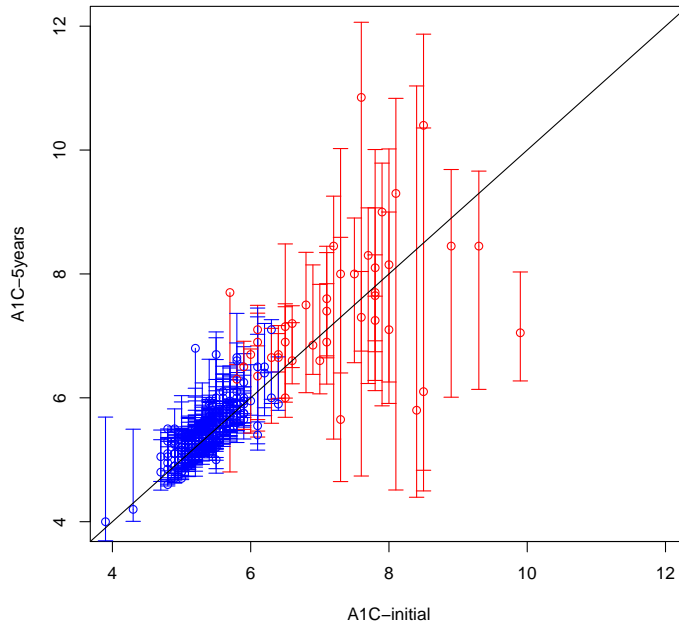


Figure 5.4: Prediction intervals for each response observed in the AEGIS database (90% confidence level). The red circles correspond to patients with diabetes.

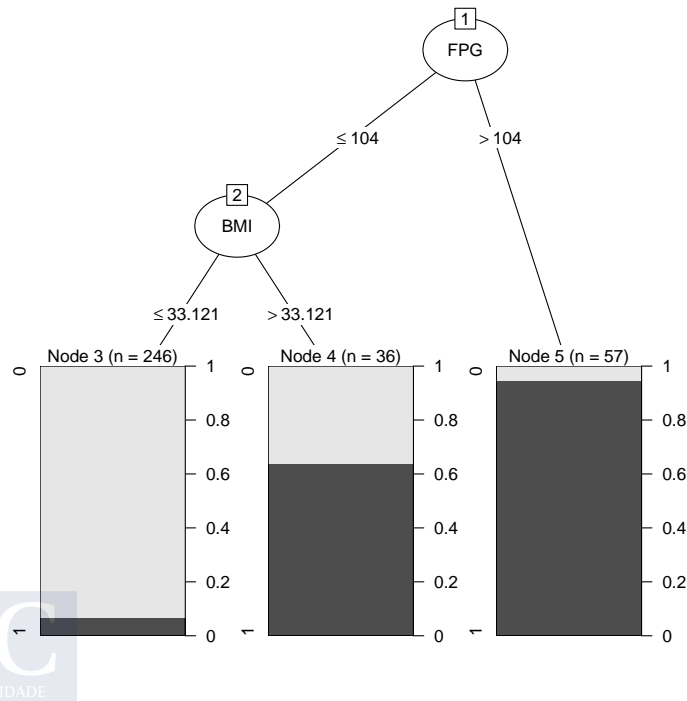


Figure 5.5: Clinical decision rules that allow us to identify those patients with a significant uncertainty in their $A1c_{5year}$ predictions.

$A1c_{5years}$ from this subset of biomarkers is analysed using several regression models that differ in terms of including CGM data as a predictor. As a result, the R_{gluco}^2 value, corresponding to the model which adopts a glucodensity-based representation for CGM data, shows a good proportion of variance explained by the model and is similar to that reported by other authors for short-term predictions [89, 296]. Moreover, glucodensity has a positive impact on improving accuracy in predicting $A1c_{5years}$ by expanding the model expressiveness along the continuous spectrum of glucose concentrations.

Some recent studies have proposed different machine learning methods for predicting the progression to diabetes from a healthy or prediabetic state with relatively good performance [37, 290]. The strength of both studies lies in their inclusion of a large number of subjects. Both of them proposed a classification strategy relying on a threshold-based categorization upon different ranges of glucose or A1c. This allows them to obtain robust results but at the expense of making an inappropriate interpretation in physiological terms [87, 295]. Furthermore, both studies are of observational nature and subjects with only partially available data were excluded from the analysis. This suggests that caution should be exercised when utilising these results for decision-making.

The present line of research assigns a key role to the analysis of glucose excursions from CGM data in search of better phenotyping and corresponding progress towards the implementation of a personalized intervention [105, 273]. An interesting asset of the present proposal is the proper evaluation of the limits of predictive models by estimating the uncertainty of the predictions for each new subject. Thus, a careful analysis of the results that exhibit significant discrepancies with the model predictions provides the opportunity to identify certain patient phenotypes that need to be followed up more closely. These discrepancies can be explained by many different factors (lifestyle, diet, disease, pharmacological treatments, etc.) over time. The present study shows that these discrepancies can be promptly recognized using routine clinical practice biomarkers.

An inherent limitation of the AEGIS study was its modest sample size. In this respect, kernel methods have proven effective in coping with a distributional representation, but at the cost of a substantial amount of data to show a significant advantage in high uncertainty settings. A larger sample size would refine the predictive model and enable the inclusion of stratification effects in future studies [4, 5]. Another limitation can be found in the 3-7-days CGM recording period of this study. An extension of this period to 14 days would probably limit possible intraday variations in glycemic profile representation; however, the discomfort from wearing a CGM device for such a long period is not a minor issue.

Ultimately, our findings enforce the prominent role of CGM data in providing a comprehensive picture of glucose metabolism [100] and allow us to envision new research on further characterizing glucose dynamics by devising new methods for (1) measuring the variability of

glucose excursion, (2) clustering different glucose profiles, or (3) unveiling patterns of glucose excursions related to specific pathophysiological mechanisms. In particular, the inclusion of both CGM-based information and longitudinal multi-omics information in the analysis may provide deep insight into the underlying mechanisms involved in the onset and progression of the disease [300]. Lastly, further research is needed on new glycemic outcomes, beyond average measures like A1c, in order to capture a more accurate picture of glycemic dynamics, and glucodensity might be exploited as a new source of information for more robust predictions.

6 Hypothesis testing in the presence of complex paired missing data by maximum mean discrepancy: An application to continuous glucose monitoring

A common experimental design in clinical studies, especially longitudinal ones, is the matched pairs design where observations are made from the same subjects under two different conditions, often at two points in time. Testing the null hypothesis that observations come from the same distribution represents an essential step before performing any modeling task. However, a typical issue when dealing with paired data is the occurrence of missing data.

The literature on matched pairs with missing data has primarily focused on one-dimensional, continuous, discrete or ordinal variables, aimed at detecting changes in location/mean [67, 102, 173, 291], scale/variance [55], and distribution [86]. Some of the proposals apply multiple imputation techniques [6, 7, 281], but they often require large sample sizes for being correct. Other proposals rely on specific model assumptions such as symmetry or bivariate normality [67, 235, 291], but they exhibit a non-robust behavior against deviations. The common approach recently adopted in literature results from combining in a non parametric approach separate test statistics for the paired and unpaired observations, by using either weighted test statistics [12, 77, 86, 136, 173, 236], a multiplication combination test [11], or combined p-values [13, 143, 221, 294]

The recent scientific and technological progress in measuring biological processes has enabled monitoring of patient's condition with a growing level of detail and complexity. Thus, beyond the ongoing identification of univariate biomarkers, new complex data structures are being incorporated into the analysis, as is the case of population ages and mortality distributions [32], distributions of functional connectivity patterns in the brain [62, 215], post-intracerebral hemorrhage hematoma densities [213], graph-based representations of connectivity and functional brain activity [264], and glucose distributions from continuous monitoring [178].

The aim of the present chapter is to provide a statistical test for matched pairs with missing data which does not require any parametric assumptions and uses all observations available.

We propose new maximum mean discrepancy (MMD) estimators to achieve this aim [97]. The energy distance and the MMD are two equivalent statistical metrics with the ability to detect distributional differences between random samples [247, 260]. Moreover, MMD-based statistics can also be seen as a natural generalization of the ANOVA test to cases where the distributions are not necessary Gaussian [228]. MMD overcomes Gaussian assumptions by representing distances between distributions as distances between mean embeddings in a reproducing kernel Hilbert space (RKHS). MMD has been successfully applied to independence testing [261], two-sample testing [97], survival analysis [75], or clustering analysis [80].

Besides conducting an extensive simulation study, the new testing procedures are applied to the AEGIS diabetes dataset, resulting from a longitudinal population-based study [101]. This dataset includes data from continuous glucose monitoring (CGM), performed at the beginning of the study and five years later. Let us recall that there is a substantial loss to follow up. A distributional representation of glucose concentration summarizes several days of monitoring, providing a personal signature of glucose homeostasis [178]. The present approach allows us to address some interesting questions related to the possible changes in CGM profile with ageing, or the relation between obesity and diabetes. Furthermore, an adaption of a previous clustering method to matched pairs with missing data allows us to find out specific patient phenotypes, with potential applications in patient stratification [80].

The rest of this chapter is outlined as follows. In Section 6.1 we provide a motivation for the new methods from the distributional representation of CGM data. In Section 6.2 we define the problem in general terms and introduce the statistical model based on the MMD metric, providing weighted test statistics for dealing with missing data under MCAR mechanism (Section 6.2.1) and under MAR mechanism (Section 6.2.2). A proof presenting theoretical guarantees of the proposed methods is delivered in the Appendix. In Section 6.2.3 the choice of kernel functions and corresponding hyperparameters is discussed. Then we present the results of an extensive simulation study in Section 6.2.4. In Section 6.2.5 a previous clustering method is adapted to missing data under the MAR mechanism. We present in Section 6.3 some applications of both hypothesis testing and clustering analysis to the AEGIS study, by exploiting the distributional representation of CGM data. We close with a discussion in Section 6.4.

6.1 Motivation from glucodensity representation

Figure 6.1 contains an example of the glucodensity representation for the continuous glucose monitoring performed on two different individuals, both in a prediabetes and later diabetes status. This figure immediately poses the challenge of defining new statistical methods to compare two sets of glucodensity measurements to assess whether some population statistics

Chapter 6. Hypothesis testing in the presence of complex paired missing data by maximum mean discrepancy: An application to continuous glucose monitoring differ. This can be useful to compare the glucose homeostasis before and after a treatment, or after a certain period of time.

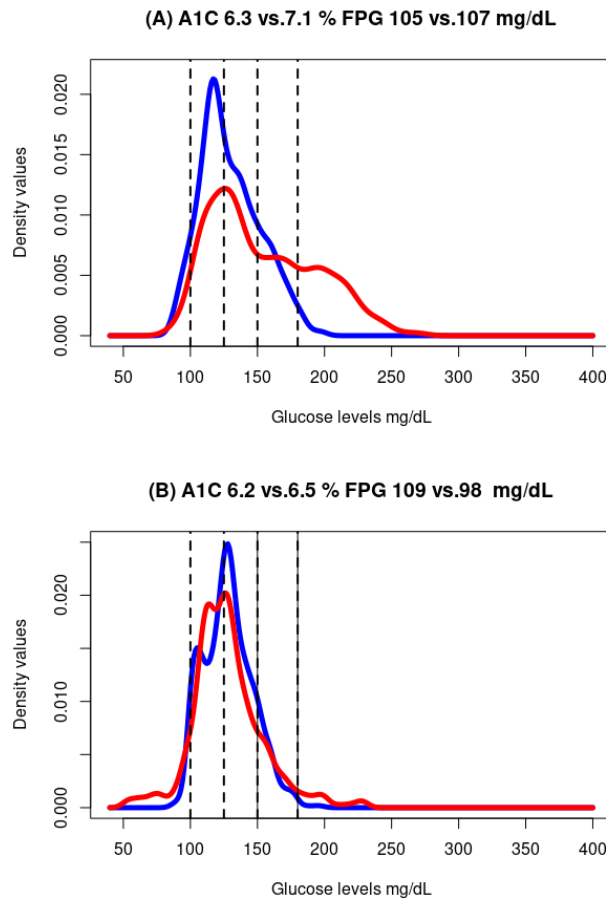


Figure 6.1: Glucodensity changes in prediabetic patients (blue) who develop diabetes after 5 years (red).

6.2 Hypotheses and statistics

Let \mathcal{D} be a separable Hilbert space and $(X_1, X_2)^T \in \mathcal{D}^2$ a random pair representing two different measurements on a subject at two different time points. Let us consider a general matched pairs design given by i.i.d. random variables

$$\mathbf{X}_j = \begin{pmatrix} X_{1j} \\ X_{2j} \end{pmatrix}, j = 1, \dots, n. \quad (6.1)$$

X_{1j} can represent the glucodensity at the beginning of a certain study for the j -th patient, and X_{2j} the glucodensity at the end of the study for the same patient. Let assume that both $\{X_{1j}\}_{j=1}^n$ and $\{X_{2j}\}_{j=1}^n$ are drawn from probability measures P_1 and P_2 , respectively. We are interested in testing the equality of distributions as null hypothesis $H_0 : \{P_1 = P_2\}$ against the alternative $H_1 : \{P_1 \neq P_2\}$, i.e., to check whether there are systematical differences between the outcomes at different time points.

6.2.1 Missing Completely at Random (MCAR) mechanism

When some of the elements of the matched pairs are missing completely at random the available data can be sorted as:

$$\mathbf{X} = \underbrace{\begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix} \dots \begin{pmatrix} X_{1n_1} \\ X_{2n_1} \end{pmatrix}}_{\substack{\text{Complete data } \mathbf{X}^{\text{com}} \\ n_1 \text{ observations}}} \underbrace{\begin{pmatrix} X_{1n_1+1} \\ - \end{pmatrix} \dots \begin{pmatrix} X_{1n_1+n_2} \\ - \end{pmatrix}}_{\substack{\text{Incomplete data } \mathbf{X}_1^{\text{inc}} \\ n_2 \text{ observations}}} \underbrace{\begin{pmatrix} - \\ X_{2n_1+n_2+1} \end{pmatrix} \dots \begin{pmatrix} - \\ X_{2n_1+n_2+n_3} \end{pmatrix}}_{\substack{\text{Incomplete data } \mathbf{X}_2^{\text{inc}} \\ n_3 \text{ observations}}}, \quad (6.2)$$

where $n = n_1 + n_2 + n_3$. For ease of notation, we denote $\mathbf{X}_1^{\text{com}} = \{X_{1j}\}_{j=1}^{n_1}$, $\mathbf{X}_2^{\text{com}} = \{X_{2j}\}_{j=1}^{n_1}$, $\mathbf{X}_1^{\text{inc}} = \{X_{1j}\}_{j=n_1+1}^{n_1+n_2}$ and $\mathbf{X}_2^{\text{inc}} = \{X_{2j}\}_{j=n_1+n_2+1}^n$. Additionally, a missingness status variable can be defined $\delta_{ij} \in \{0, 1\}$, $i = 1, 2$, $j = 1, \dots, n$, so $\delta_{ij} = 1$ if the element is missing and $\delta_{ij} = 0$ otherwise.

A natural way of testing the equality of distributions is measuring the distance between them. We propose two test statistics: \mathcal{T}_1 for the complete data sets $\mathbf{X}_1^{\text{com}}$ and $\mathbf{X}_2^{\text{com}}$, and \mathcal{T}_2 for the incomplete data sets $\mathbf{X}_1^{\text{inc}}$ and $\mathbf{X}_2^{\text{inc}}$, which are then combined in one weighted test statistic:

$$\mathcal{T}(\mathbf{X}) = \alpha \mathcal{T}_1(\mathbf{X}_1^{\text{com}}, \mathbf{X}_2^{\text{com}}) + (1 - \alpha) \mathcal{T}_2(\mathbf{X}_1^{\text{inc}}, \mathbf{X}_2^{\text{inc}}), \quad (6.3)$$

for some weighting parameter $\alpha \in [0, 1]$. Both \mathcal{T}_1 and \mathcal{T}_2 are based on the maximum mean discrepancy (MMD) to measure the empirical distance between the marginal distributions [97]. Let $k : \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R}^+$ be a symmetric definite positive kernel. The existence of a dot product space \mathcal{H} and feature mapping $\phi : \mathcal{D} \rightarrow \mathcal{H}$ is guaranteed, such that $k(X, X') = \langle \phi(X), \phi(X') \rangle_{\mathcal{H}}$. Kernel mean embedding results from extending the mapping ϕ to the space of probability distributions by representing each distribution as a mean function $\phi(F) = \mathbf{E}[k(\cdot, X)] = \int_{\mathcal{D}} k(\cdot, X) dP$. The kernel mean embedding can be empirically estimated by $\tilde{\phi} = \frac{1}{n} \sum_{i=1}^n k(\cdot, X_i)$. Then, we can measure the distance between random samples as follows:

$$\begin{aligned} \frac{n_1}{n_1 + n_2 + n_3} \mathcal{T}_1(\mathbf{X}_1^{\text{com}}, \mathbf{X}_2^{\text{com}}) &= \|\tilde{\phi}_1^{\text{com}} - \tilde{\phi}_2^{\text{com}}\|_{\mathcal{H}}^2 \\ &= \left\langle \frac{1}{n_1} \sum_{i=1}^{n_1} k(\cdot, X_{1i}) - \frac{1}{n_1} \sum_{i=1}^{n_1} k(\cdot, X_{2i}), \frac{1}{n_1} \sum_{i=1}^{n_1} k(\cdot, X_{1i}) - \frac{1}{n_1} \sum_{i=1}^{n_1} k(\cdot, X_{2i}) \right\rangle \\ &= \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} k(X_{1i}, X_{1j}) + \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} k(X_{2i}, X_{2j}) - \frac{2}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} k(X_{1i}, X_{2j}). \end{aligned}$$

Analogously,

$$\begin{aligned} \frac{n_2 + n_3}{n_1 + n_2 + n_3} \mathcal{T}_2(\mathbf{X}_1^{\text{inc}}, \mathbf{X}_2^{\text{inc}}) &= \|\tilde{\phi}_1^{\text{inc}} - \tilde{\phi}_2^{\text{inc}}\|_{\mathcal{H}}^2 \\ &= \frac{1}{n_2^2} \sum_{i=n_1+1}^{n_1+n_2} \sum_{j=n_1+1}^{n_1+n_2} k(X_{1i}, X_{1j}) + \frac{1}{n_3^2} \sum_{i=n_1+n_2+1}^{n_1+n_2+n_3} \sum_{j=n_1+n_2+1}^{n_1+n_2+n_3} k(X_{2i}, X_{2j}) \end{aligned}$$

$$-\frac{2}{n_2 n_3} \sum_{i=n_1+1}^{n_1+n_2} \sum_{j=n_1+n_2+1}^{n_1+n_2+n_3} k(X_{1i}, X_{2j}).$$

Importantly, for the class of *characteristic* kernels, the embeddings are injective, and hence $\|P_1 - P_2\|_{\mathcal{H}}^2 = 0$, if and only if $P_1 = P_2$ [255].

In order to calibrate the tests under the null hypothesis it should be pointed out that both \mathcal{T}_1 and \mathcal{T}_2 do not follow a free asymptotic distribution. The empirical estimate of MMD is a one-sample V-statistic and hence asymptotic distribution is difficult to obtain due to the degeneracy of the V-statistic, which incorporates a correlation structure for the complete paired observations \mathbf{X}^{com} [97]. To address this issue we propose a wild bootstrap procedure for the first n_1 observations, while the remaining $n_2 + n_3$ observations can be properly handled by permutations methods, that can achieve an exact type I error control. For each $b = 1, \dots, B$, it proceeds as follows:

1. For the first n_1 complete paired observations, take random weights w_i^b , $i = 1, \dots, n_1$, with

$$w_i^b = e^{-1/l_{n_1}} w_{i-1}^b + \sqrt{1 - e^{-2/l_{n_1}}} \epsilon_i,$$

where $w_0^b, \epsilon_1, \dots, \epsilon_{n_1}$ are independent standard normal variables, and l_{n_1} is a bootstrap parameter used to mimic the dependence structure, such that $l_{n_1} = o(n_1)$ but $\lim_{n_1 \rightarrow \infty} l_{n_1} = \infty$. Then,

$$\mathcal{T}_1^b(\mathbf{X}_1^{\text{com}}, \mathbf{X}_2^{\text{com}}) = \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} w_i^b w_j^b [k(X_{1i}, X_{1j}) + k(X_{2i}, X_{2j}) - 2k(X_{1i}, X_{2j})].$$

2. The remaining $n_2 + n_3$ observations belonging to $\mathbf{X}_1^{\text{inc}}$ and $\mathbf{X}_2^{\text{inc}}$ are randomly permuted, i.e. each observation is randomly assigned to new $\mathbf{X}_1^{\text{inc}(\pi)}$ or $\mathbf{X}_2^{\text{inc}(\pi)}$ sets, resulting in new $\mathcal{T}_2^b(\mathbf{X}_1^{\text{inc}(\pi)}, \mathbf{X}_2^{\text{inc}(\pi)})$.
3. Then, calculate

$$\mathcal{T}^b = \alpha \mathcal{T}_1^b(\mathbf{X}_1^{\text{com}}, \mathbf{X}_2^{\text{com}}) + (1 - \alpha) \mathcal{T}_2^b(\mathbf{X}_1^{\text{inc}(\pi)}, \mathbf{X}_2^{\text{inc}(\pi)})$$

Finally, return $p\text{-value} = \frac{1}{B} \sum_{b=1}^B 1\{\mathcal{T}^b \geq \mathcal{T}(\mathbf{X})\}$.

Theorem 5. Let $\mathbf{X}^{\text{com}} = \{(X_{1i}, X_{2i})^T\}_{i=1}^{n_1}$ be a set of i.i.d. complete paired samples, and $\mathbf{X}_1^{\text{inc}} = \{X_{1i}\}_{i=n_1+1}^{n_1+n_2}$, and $\mathbf{X}_2^{\text{inc}} = \{X_{2i}\}_{i=n_1+n_2+1}^{n_1+n_2+n_3}$ two sets of i.i.d. incomplete paired samples. Let suppose that $n_1/(n_1 + n_2 + n_3) \rightarrow \kappa_1 \in (0, 1)$ and $n_2/(n_1 + n_2 + n_3) \rightarrow \kappa_2 \in (0, 1)$ as $n_1, n_2, n_3 \rightarrow \infty$; then, the test statistic given by (6.3) is consistent against the alternative $H_1 : \{P_1 \neq P_2\}$; we can detect a difference in distribution with the sample size growing to infinity. Furthermore, the calibration strategy described above is also consistent in the same sense.

A proof is provided in the Appendix.

6.2.2 Missing at Random (MAR) mechanism

We assume a MAR mechanism where the probability of being missing on the second time point is based on the corresponding value on the first time point, which can be described as follows

$$\mathbf{X} = \underbrace{\begin{pmatrix} X_{11} \\ X_{21} \end{pmatrix} \dots \begin{pmatrix} X_{1n_1} \\ X_{2n_1} \end{pmatrix}}_{\substack{\text{Complete data } \mathbf{X}^{\text{com}} \\ n_1 \text{ observations}}} \underbrace{\begin{pmatrix} X_{1n_1+1} \\ - \end{pmatrix} \dots \begin{pmatrix} X_{1n_1+n_2} \\ - \end{pmatrix}}_{\substack{\text{Incomplete data } \mathbf{X}_1^{\text{inc}} \\ n_2 \text{ observations}}}, \quad (6.4)$$

where $n = n_1 + n_2$. We denote by $\pi(\cdot) = P(\delta_{2j} = 1 | X_{1j} = \cdot)$, the conditional probability that the observation X_{2j} will be missing given X_{1j} . A natural way to incorporate the missing data mechanism in the test statistic is to associate weight ω_j with the j -th observation via an inverse probability weighting (IPW) estimator [272], given by

$$\omega_j = \frac{\delta_{2j}}{n\pi(X_{1j})}, \quad j = 1, \dots, n. \quad (6.5)$$

We define the normalized weight of ω_j as $\hat{\omega}_j = \omega_j / \sum_{j=1}^n \omega_j$, $j = 1, \dots, n$. In practice, we estimate the probability $\pi(\cdot)$ by means of a binary classification algorithm, and denote by $\tilde{\omega}_j$ the ensuing estimated normalized weight. We propose the following test statistic

$$\begin{aligned} \mathcal{T}(\mathbf{X}) &= \mathcal{T}(\mathbf{X}_1^{\text{com}}, \mathbf{X}_2^{\text{com}}) = \left\| \tilde{\phi}_1^{\text{com}} - \tilde{\phi}_2^{\text{com}} \right\|_{\mathcal{H}}^2 \\ &= \left\langle \sum_{j=1}^{n_1} \tilde{\omega}_j k(\cdot, X_{1j}) - \sum_{j=1}^{n_1} \tilde{\omega}_j k(\cdot, X_{2j}), \sum_{j=1}^{n_1} \tilde{\omega}_j k(\cdot, X_{1j}) - \sum_{j=1}^{n_1} \tilde{\omega}_j k(\cdot, X_{2j}) \right\rangle \\ &= \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \tilde{\omega}_i \tilde{\omega}_j k(X_{1i}, X_{2j}) + \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \tilde{\omega}_i \tilde{\omega}_j k(X_{1i}, X_{2j}) - 2 \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \tilde{\omega}_i \tilde{\omega}_j k(X_{1i}, X_{2j}). \end{aligned}$$

In this scenario, we propose to calibrate the test under the null hypothesis in an analogous manner to the MCAR mechanism. Specifically, for each bootstrap iteration we propose to use the following estimator

$$\mathcal{T}^b(\mathbf{X}) = \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} w_i^b w_j^b \tilde{\omega}_i \tilde{\omega}_j [k(X_{1i}, X_{1j}) + k(X_{2i}, X_{2j}) - 2k(X_{1i}, X_{2j})].$$

6.2.3 Kernel choice and kernel hyperparameters

We propose using the Gaussian kernel $k(X, Y) = e^{-\|X-Y\|^2/\sigma}$ for $X, Y \in \mathbb{R}$, and $k(X, Y) = e^{-d_{w_2}^2(X, Y)/\sigma}$ for $X, Y \in \mathcal{D}$, where $\sigma > 0$. Importantly, the Gaussian kernel is a characteristic kernel, and thus we can detect asymptotically any difference in distribution. The kernel

Chapter 6. Hypothesis testing in the presence of complex paired missing data by maximum mean discrepancy: An application to continuous glucose monitoring
bandwidth σ was estimated through the median heuristic $\sigma^2 = \text{median}\{\|X_i - X_j\|^2 : 1 \leq i < j \leq n\}$.

6.2.4 Simulation study

We investigate the finite sample behavior of the above methods in extensive simulations. A total of 2,000 simulations were performed for both MCAR and MAR scenarios. Methods were examined with respect to their Type-I error rate control at level 5%. A total of 2,000 bootstrap runs and permutation replicas were held. The wild bootstrap parameter l_{n_1} was selected according to $l_{n_1} = \sqrt{n_1}$.

The observations were generated by mimicking the sort of distributional representations commonly obtained from CGM data. Since the 2-Wasserstein distance depends only on quantile functions, observations were sampled from the following location-scale model on quantile functions [213]: let $Z \in \mathbb{R}^p$ be a random vector of predictor variables and let Q_0 be a fixed quantile function; here we considered the age as the only predictor variable and fixed $Q_0(t) = 70 + 240t$ in the range of glucose values expected from type-2 diabetes; let $\eta(z) = a_0 + a_1 z_1$ and $\tau(z) = b_0 + b_1 z_1$ be the location and scale components of the model, respectively, where $a = (a_0, a_1)$ and $b = (b_0, b_1)$ are the corresponding coefficients and we assume that $\tau(Z) > 0$ almost surely; let V_1 and V_2 two random variables that satisfy $E(V_1|Z) = 0, E(V_2|Z) = 1$, and $V_2 > 0$ almost surely; the model is given by

$$Q(t) = V_1 + V_2\eta(Z) + V_2\tau(Z)Q_0(t), \quad (6.6)$$

MCAR scenario

We fixed $n_1 = n_2 = n_3 = 150$. In order to introduce correlation structure into the quantile functions for $\mathbf{X}_1^{\text{com}}$ and $\mathbf{X}_2^{\text{com}}$, we sampled variables V_1^* and V_2^* from bivariate uniform distributions with correlation given by $\rho \in \{0.00, 0.20, 0.40, 0.60, 0.80\}$. The location-scale model is given by $V_1 = -20 + 40V_1^*$ and $V_2 = 0.8 + 0.4V_2^*$, and fixed parameters $a_0 = b_0 = 0$, $a_1 = 0.3$ and $b_1 = 0.005$. The observations for $\mathbf{X}_1^{\text{inc}}$ and $\mathbf{X}_2^{\text{inc}}$ were i.i.d. generated and then we applied the same location-scale model than before. A total of 2,000 simulations were performed assuming that the age was distributed as $Z_1, Z_2 \sim \mathcal{U}_{[30,50]}$ both at the beginning and at the end of the study, that is, for all the variables in \mathbf{X} . Another 2,000 simulations were performed assuming that the age was distributed as $Z_1 \sim \mathcal{U}_{[30,50]}$ at the beginning of the study, that is, for all the variables in $\mathbf{X}_1^{\text{com}}$ and $\mathbf{X}_1^{\text{inc}}$, and was distributed as $Z_2 \sim \mathcal{U}_{[50,70]}$ at the end of the study, that is, for all the variables in $\mathbf{X}_2^{\text{com}}$ and $\mathbf{X}_2^{\text{inc}}$.

MAR scenario

We fixed $n = 300$. The missing mechanism is given by $P(\delta_{2j} = 1|Y_1, Y_2) = (1 + e^{-1+Y_1+Y_2})^{-1}$, $j = 1, \dots, n$, where $Y_1, Y_2 \sim \mathcal{N}(0, 1)$ are two independent random variables. We introduced correlation structure into the quantile functions for $\mathbf{X}_1^{\text{com}}$ and $\mathbf{X}_2^{\text{com}}$ as we did in the MCAR scenario. We used the same location-scale model. The same methodology as in the MCAR scenario was applied for sampling the age.

Results

Table 6.1 shows the results of the simulation study. We can see that calibration under the null hypothesis provides acceptable results. However, some biases under the two missingness mechanisms can be noted. As long as the difference between the sample data and the null hypothesis increases, the test rejects more frequently the null hypothesis.

ρ	Z_1	Z_2	MCAR	MAR
0.00	$\mathcal{U}_{[30,50]}$	$\mathcal{U}_{[30,50]}$	0.03	0.03
0.20	$\mathcal{U}_{[30,50]}$	$\mathcal{U}_{[30,50]}$	0.04	0.03
0.40	$\mathcal{U}_{[30,50]}$	$\mathcal{U}_{[30,50]}$	0.05	0.03
0.60	$\mathcal{U}_{[30,50]}$	$\mathcal{U}_{[30,50]}$	0.03	0.04
0.80	$\mathcal{U}_{[30,50]}$	$\mathcal{U}_{[30,50]}$	0.04	0.04
0.00	$\mathcal{U}_{[30,50]}$	$\mathcal{U}_{[50,70]}$	0.98	0.88
0.20	$\mathcal{U}_{[30,50]}$	$\mathcal{U}_{[50,70]}$	0.99	0.90
0.40	$\mathcal{U}_{[30,50]}$	$\mathcal{U}_{[50,70]}$	0.99	0.91
0.60	$\mathcal{U}_{[30,50]}$	$\mathcal{U}_{[50,70]}$	0.99	0.93
0.80	$\mathcal{U}_{[30,50]}$	$\mathcal{U}_{[50,70]}$	0.99	0.96

Table 6.1: The proportion of simulations rejecting the null hypothesis under MCAR and MAR mechanisms is shown.

6.2.5 Paired missing data clustering

Let $\mathbf{X} = \{(X_{1j}, X_{2j}, \delta_{2j})\}_{j=1}^n$, be a dataset of i.i.d. random variables obtained under a MAR mechanism, where we denote again by $\pi(\cdot) = P(\delta_{2j} = 1|X_{1j} = \cdot)$, the conditional probability that the observation X_{2j} will be missing given X_{1j} . We associate a weight $\tilde{\omega}_j$ with the j -th observation via an IPW estimator, by applying equation (6.5). Let $\mathbf{X}_j = (X_{1j}, X_{2j})^T$, $\mathbf{X}_h = (X_{1h}, X_{2h})^T \in \mathbf{X}^{\text{com}}$ be two different complete paired samples. We define the following bivariate kernel $k(\mathbf{X}_j, \mathbf{X}_h) = e^{-(d_{W_2}^2(X_{1j}, X_{1h}) + d_{W_2}^2(X_{2j}, X_{2h}))/\sigma}$, where $\sigma^2 = \text{median}\{d_{W_2}^2(X_{1j}, X_{1h}) + d_{W_2}^2(X_{2j}, X_{2h}) : 1 \leq j < h \leq n\}$.

Consider a disjoint partition $\mathbf{X}^{\text{com}} = \bigcup_{i=1}^k C_i$, with $C_i \cap C_l = \emptyset$, for all $i \neq l$. Following [80], we aim to build a new partition $\tilde{C}_1, \dots, \tilde{C}_k$ by maximizing an objective function given by

$$(\tilde{C}_1, \dots, \tilde{C}_k) = \arg \max_{(C_1, \dots, C_k)} \sum_{i=1}^k \frac{1}{v_i} \sum_{\mathbf{X}_j, \mathbf{X}_h \in C_i} \tilde{\omega}_j \tilde{\omega}_h k(\mathbf{X}_j, \mathbf{X}_h), \quad (6.7)$$

where $v_i = \sum_{\mathbf{X}_j \in C_i} \tilde{\omega}_j$. We can iteratively solve this optimization problem by measuring the impact of moving each observation to another cluster. Let denote by $S_i = \sum_{\mathbf{X}_j, \mathbf{X}_h \in C_i} \tilde{\omega}_j \tilde{\omega}_h k(\mathbf{X}_j, \mathbf{X}_h)$ the internal similarity of cluster C_i , and $S_i(\mathbf{X}_j) = \sum_{\mathbf{X}_h \in C_i} \tilde{\omega}_j \tilde{\omega}_h k(\mathbf{X}_j, \mathbf{X}_h)$ the internal similarity with respect to the observation \mathbf{X}_j . By moving the observation \mathbf{X}_j from cluster C_i to C_l we change the result of the objective function by

$$\Delta S^{i \rightarrow l}(\mathbf{X}_j) = \frac{S_l^+}{v_l + \tilde{\omega}_j} + \frac{S_i^-}{v_i - \tilde{\omega}_j} - \frac{S_l}{v_l} - \frac{S_i}{v_i}, \quad (6.8)$$

where $S_l^+ = S_l + 2S_l(\mathbf{X}_j) + \tilde{\omega}_j \tilde{\omega}_j k(\mathbf{X}_j, \mathbf{X}_j)$ is the internal similarity of the new cluster C_l after the addition of the observation \mathbf{X}_j , and $S_i^- = S_i - 2S_i(\mathbf{X}_j) + \tilde{\omega}_j \tilde{\omega}_j k(\mathbf{X}_j, \mathbf{X}_j)$ is the internal similarity of the new cluster C_i after removing the observation \mathbf{X}_j . Ultimately, we compute $i^* = \arg \max_{l=1, \dots, k | l \neq i} \Delta Q^{i \rightarrow l}(\mathbf{X}_j)$, and if $\Delta S^{i \rightarrow i^*}(\mathbf{X}_j) > 0$ we move \mathbf{X}_j to cluster C_{i^*} , otherwise we keep it in C_i .

6.3 Illustrative data analysis

As a practical application, we consider again the AEGIS study, aimed at analyzing the evolution of different clinical biomarkers related to circulating glucose in a initial random sample of 1516 patients. In addition, a CGM are performed every five years on a randomized subset of patients. Specifically, at the beginning of the study, 581 participants were randomly selected for wearing a CGM device for 3-7 days. Out of the total of 581 participants, 68 were diagnosed with diabetes before the study and 22 during the first five years. Table 3.1 shows the baseline characteristics of these 581 patients grouped by sex. After a five-year follow-up, only 161 participants agreed to perform a second glucose monitoring.

The AEGIS study raises some interesting questions that can be addressed with the present approach.

Changes in CGM profile with ageing. Some recent works explore the important role of ageing in glucose dysregulation, and the difficulties inherent in maintaining glucose homeostasis as close to normal as possible [42]. The proposed \mathcal{T} -test gives us the opportunity to examine if there exist statistical differences after five years at a distributional level. We estimate the missing data mechanism by means of logistic regression, using as predictors the age and glycaemic status (normoglycemic, prediabetes or type-2 diabetes) at the beginning of the study and sex of each participant. We applied the \mathcal{T} -test considering glucodensities at both time

points to check the null hypothesis of equality of distributions. We obtained a p -value = 0.048, identifying significant differences at both time points.

Obesity in diabetes. Obesity is a critical risk factor for the development of type-2 diabetes [148]. In order to further characterize this risk subpopulation, we analyzed those normoglycemic subjects with overweight in the AEGIS dataset, by examining again if there exist statistical differences after five years at a distributional level. We applied the \mathcal{T} -test to check the null hypothesis in the following two subgroups of the normoglycemic population: i) individuals with a body mass index less than $22Kg/m^2$ (low body mass index); ii) individuals with a body mass index higher than $22Kg/m^2$ (overweight and obesity). In the first case we obtained a p -value = 0.36, providing no evidence against the null hypothesis, while in the second case we obtained a p -value = 0.056, which can be interpreted as borderline. Figure 6.2 shows the difference between the quantile curves in these two subgroups.

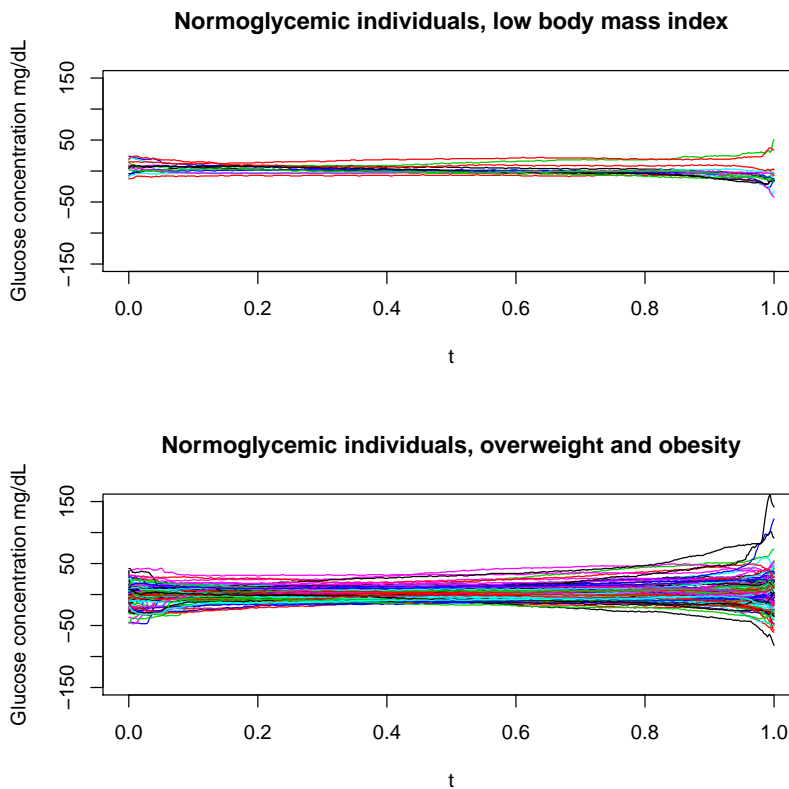


Figure 6.2: Difference between the quantile curves (before and after) in normoglycemic individuals according to body mass status. The dispersion is more significant for the overweight and obesity subgroup, consistent with an increasing glycemic risk.

Patient stratification. Clustering analysis can be a useful tool for providing distinctive and meaningful patient phenotypes and, consequently, in guiding patient stratification for delivering more personalized care [138]. We applied a clustering analysis to those individuals for whom CGM has been performed at both time points. Figure 6.3 shows the resulting two clusters. The individuals in cluster 1 do not present significant changes between both

time points, while some significant differences are noted in cluster 2. Table 6.2 shows the baseline clinical characteristics of each cluster. Both groups of individuals have important differences in insulin resistance and glycaemic variability metrics. Importantly, in cluster 2 the average glycaemic characteristics in terms of glycated hemoglobin and fasting plasma glucose are consistent with prediabetes ($5.7\% \leq A1c \leq 6.4\%$ or $100 \text{ mg/dl} \leq \text{FPG} \leq 125 \text{ mg/dl}$ according to American Diabetes Association guidelines). In contrast, cluster 1 is composed of normoglycemic individuals. Ultimately, clustering results effectively correlates with a significant change in the glycaemic status.

Finally, we performed stepwise logistic regression with forward selection to identify which baseline characteristics independently predicted the corresponding group, resulting age, FPG and CONGA. We checked the null hypothesis that each coefficient is equal to zero. Table 6.3 shows the results of this analysis, identifying FPG and CONGA as the subset of characteristics that best predicted the outcome.

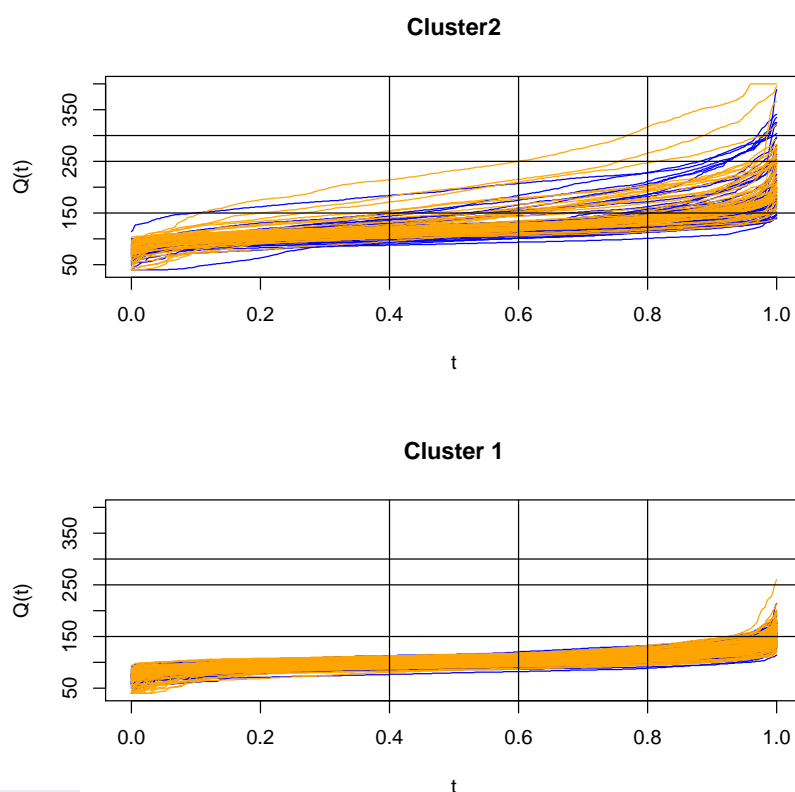
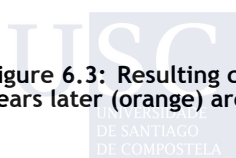


Figure 6.3: Resulting clusters are shown. Both quantile curves at the beginning of the study (blue) and five years later (orange) are shown for each cluster.



	cluster 1	cluster 2
Age (years decimal)	43.66 ± 12.80	53.11 ± 12.34
A1c, %	5.27 ± 0.25	6.20 ± 1.07
FPG, mg/dl	84.83 ± 9.93	108.39 ± 32.61
HOMA-IR, mg/dl.μ IU/ml	2.28 ± 1.16	4.97 ± 8.50
BMI, kg/m ²	27.09 ± 4.87	29.29 ± 4.83
Waist, cm	87.29 ± 13.60	95.18 ± 14.43
CONGA, mg/dl	0.75 ± 0.20	1.21 ± 0.52
MAGE, mg/dl	26.16 ± 7.16	45.80 ± 24.58
MODD	0.66 ± 0.18	1.05 ± 0.48

Table 6.2: Clinical baseline characteristics for the individuals belonging to each cluster. Mean and standard deviation are shown.

Coefficients	
(Intercept)	-10.01 (1.82) ^{***}
age	0.04 (0.02)
FPG	0.05 (0.02) ^{**}
CONGA	3.60 (0.91) ^{***}
Quality measures	
AIC	140.37
BIC	152.69
Log Likelihood	-66.18
Deviance	132.37

^{***} $p < 0.001$; ^{**} $p < 0.01$

Table 6.3: Coefficients obtained from logistic regression. Results from some different model selection criteria for the fitted model are shown.

6.4 Discussion

The analysis of paired data with missing values is becoming critical in longitudinal studies, particularly when comparing the participants' condition across different time points. The available methods in the literature are not applicable when data adopt non-vectorial representations, better suited to capture functional, structural or other complex forms of information increasingly common in current medicine. To overcome this limitation we have provided novel methods for hypothesis testing in the presence of complex paired missing data under both MCAR and MAR mechanisms. They are not based on any parametric assumption and use all observations within the matched pairs design. The methods are based on the notion of maximum mean discrepancy, a metric between mean embeddings in a RKHS that can be applied to both Euclidean and non-Euclidean data, with different structured, functional and distributional representations, by an appropriate design of the reproducing kernel. Specifically, the space of probability density functions has been used throughout the text to test the feasibility of this approach.

The asymptotic validity of the methods was proven and can be found in the appendix. In an extensive simulation study, the type-I error rate control of the tests has been examined under both MCAR and MAR mechanisms, performing well with different correlation coefficients. The sample size affects the behavior of the tests, since inference in a functional space customarily demands more data than in a vectorial space. Hence a worsening of performance is expected for very small sample sizes.

The application of these methods to a real longitudinal, population-based, diabetes study has highlighted some of their capabilities and advantages to explore new clinical findings, by exploiting monitoring information along the continuous range of glucose values. It should be emphasized the robustness of the results, even in an scenario with an important proportion of missing data. Furthermore, a complementary clustering analysis has revealed the effectiveness of this approach to provide an early risk identification with the potential to enable a personalized strategy. This chapter thus adds to recent debate on the identification of novel subgroups of diabetes for a proper stratification of treatment and progression to complications [5, 54, 109], here emphasizing the crucial role of continuous glucose monitoring for this purpose.

Part IV

Conclusions

7 Conclusions

The present work proposes a new representation for time series data in metric spaces and the corresponding descriptive and predictive methods stemming from the confluence of statistics and machine learning research. The distributional nature of this new representation allows us to go beyond traditional compositional representations by using the full spectrum of time series values across a given period of time. The proposed methodology finds a direct application in biosensor data analysis and possesses the potential to contribute to support the sort of reasoning that characterizes the personalized medicine paradigm. The effectiveness of these methods has been tested in the domains of diabetes and physical activity, for which we show their advantages in different predictive tasks as compared to existing data analysis methodology.

The new distributional representation is introduced in Chapter 3 for a continuous stochastic process, with an application to CGM data in the form of a new statistical object which we call glucodensity [178]. The intuitive idea behind glucodensity representation is to deliver the density function of the CGM time series, which takes values in a non-linear space. Consequently, a list of non-linear methods based on distances and kernels is provided to perform hypothesis testing, cluster analysis, regression modeling, and to overcome the technical difficulties derived from the unavailability of the rich properties given by Euclidean geometries.

The methods provided avoid technical difficulties that appear in compositional data -as the problems with zeros- and that naturally emerge in biosensor data analysis with compositional metrics. Importantly, these methods work in real-world scenarios under minimal theoretical conditions. In particular, for the methods based on kernels and energy distance, only the existence of moments of second order is required. On the other hand, Hölder and Sobolev regularity conditions are required in the case of smoothing methods such as Nadaraya-Watson kernel regression estimators, which are often held by biosensor data.

The validation of such methods is conducted in the diabetes domain even though they are also deemed applicable to other domains (ECG, fMRI, ...). In the case of diabetes, the results on the AEGIS study clearly show that glucodensity captures more information about the glucose homeostasis signature than traditional diabetes biomarkers. An essential point of the validation is that we use a random sample drawn from the general population, including

normoglycaemic, prediabetes, and diabetes individuals. In future work, an extension of the present framework can be provided as a multilevel distributional representation for capturing information about different resolution scales. In addition, we are also interested in the notion of multidimensional glucodensity, in which we simultaneously analyze the glucose concentration, speed, and acceleration in a tridimensional density function. These new representations would minimize the loss of information due to the current removal of the order of events in the representation. In the case of multidimensional glucodensity, we can use the methods developed here. However, we must resort to computational intensive numerical methods to calculate the distances between densities based on optimal transportation theory [218].

Chapter 4 introduces a distributional representation for certain mixed stochastic processes, motivated by the need for analysing the physical activity recorded by an accelerometer device in the NHANES 2003-2006 cohort. To this end, we propose an extension of the Nadaraya-Watson kernel estimator and of the kernel ridge regression method for complex survey designs. The analysis of the NHANES dataset shows the advantages of the distributional representations over existing accelerometer summary measures in predicting different clinical outcomes, including long-term survival. One of the advantages of our validation process is the very design of the NHANES dataset. Still, the use of more modern accelerometer devices, enabling the recording at a high level of resolution of the amount and intensity of physical activity, made us feel confident of achieving better predictive results. Therefore, we used the prior methods in the NHANES 2011-2014 cohort to define clinical physical activity phenotypes in the elderly American population, and evaluated their clinical meaningfulness in terms of survival and mortality of the patients [174]. The results show that the new phenotypes of patients possess a high statistical association with survival and mortality, which can be stronger than that shown by the variable age. From a clinical point of view, these phenotypes are promising tools to improve the clinical interventions according to the principles of personalized medicine.

The following chapters focus on developing new statistical methods for statistical complex objects in longitudinal studies suffering from missing data.

Chapter 5 introduces a new framework for coping with missing data in different predictive tasks under the RKHS paradigm [175]. New methods are provided for independence testing, variable selection and uncertainty quantification. Furthermore, we adapt the notions of distance correlation and Hilbert Schmidt independence criterion, two measures of dependence between two paired random vectors, for missing responses. Moreover, we propose a new calibration test strategy based on Efron's bootstrap that works in non-Euclidean spaces and overcomes the problems and limitations cautioned by Arcones and Gine for U- or V-statistics [15]. The new methods are motivated by the opportunity of including the glucodensity as a predictor variable for modeling long-term glucose changes. For independence testing, we consider a proper mapping based on the 2-Wasserstein distance after embedding the data in a

proper quantile space endowed with a Hilbertian structure. We also adapt to the missing data setting a previous variable selection method based on the gradient of the conditional mean regression function. Finally, we extend the ideas of conformal inference to missing responses and heteroscedastic noise, aimed at addressing the important level of uncertainty typically present in biomedical applications. As a result, we supply a prediction interval for the response. The results obtained after applying these techniques to model long-term glucose changes prove the advantages of distributional representations with regard to the state-of-the-art diabetes diagnosis and control biomarkers.

On the other hand, a closer inspection of the level of uncertainty in the prediction results can provide a crucial information for clinical management. Thus, a phenotypically characterization of those subpopulations for which the model provides an unreliable prediction may be used to guide a specific approach, built on new assumptions, measurement procedures and interventions. As we have highlighted in Chapter 5 we can then promote a more personalized follow-up engaged with precision medicine principles.

Chapter 6 is motivated by the need to assess the possible changes in distributional representations from the same subjects under two different conditions, often at two temporal points, for example, before and after administering a treatment. Besides, it is assumed the general setting of longitudinal studies where there is a loss to follow-up. We propose new estimators of the maximum mean discrepancy to handle complex matched pairs with missing data. These estimators can detect differences in data distributions under different missingness mechanisms, MCAR and MAR. New calibration test strategies combining permutation and wild bootstrap are provided in order to incorporate the correlation between two temporal points of the repeated measure.

CGM data from the AEGIS study are used to illustrate the application of this approach. By employing the new distributional representations new clinical criteria on how glucose changes vary on the long term can be explored. It should be emphasized the robustness of the results, even in an scenario with an important proportion of missing data. Furthermore, a complementary clustering analysis has revealed the effectiveness of this approach to provide an early risk identification with the potential to enable a personalized strategy.

From a clinical point of view, estimating the distributional representations using more extensive time periods may be necessary to increase the reliability in the construction of functional profiles. An exciting task for further research would be to evaluate the representation's effectiveness with a more considerable number of patients in fundamental medical questions. For example, with diabetes complications such as retinopathy or to predict the development of diabetes mellitus in prediabetes populations incorporating CGM information. Answering these research questions and developing new statistical and machine learning models are the following steps to guide and address many future challenges of digital medicine.

Beyond the merits of the methodological contributions, the findings of the present thesis strengthen the role of the new biosensors in providing further insight into the onset and progression of the disease, since they supply a characterization of the underlying pathophysiological processes in increasing levels of resolution, and importantly, in free-living conditions. Ultimately, the proposed distributional representations and the corresponding methods are intended to reveal new information hidden in monitoring data and, therefore, be useful for practitioners to move from a diseased-centered approach towards a more focused patient-centered one.

With the aim of facilitating the reproducibility of the results, the proposed core methods were implemented under Open Source Licenses in the R package *Biosensors.usc*.

7.1 New opportunities

The validation processes in this work have involved several longitudinal studies, the AEGIS in the diabetes domain, and the NHANES, in its waves from years 2003 through 2006 and from years 2011 through 2014, in the physical activity domain. All of them include different sort of data types and, interestingly, also the continuous recording for several physiological variables provided by wearable monitoring devices, aimed at making increasingly objective the characterization of the status and evolution of the participants.

Indeed, according to the availability of new biosensors, a considerable effort to compile extensive data collections from patients monitored with these technologies is recently arising. For instance, the UK Biobank, created in 2006, is a large-scale biomedical database providing global access to medical and genetic data from half a million volunteer participants to improve our understanding of the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses. On the other hand, in 2016, the National Institutes of Health (NIH), as part of the Precision Medicine Initiative, has launched the study known as the "All of Us" research project [202], intended to involve a cohort of at least one million volunteers from around the United States over ten years, with a prominent role of digital health technology. In addition, a multi-institutional research initiative called "Project Baseline", aimed at creating a map of human health, has been enrolling volunteers since June 2017 seeking a ten thousand sample size [16]. Also, similar initiatives have been promoted in Israel and other first-world countries. Undoubtedly, these studies will be a key part of the expected progress in extracting new clinical knowledge from large population cohorts. Regarding the present thesis, the high sample size of these databases outlines a new opportunity to use distributional representations with more reliable and conclusive results.

A critical point in the translation of the new findings to the clinical practice is the approval of digital biomarkers as new measures for the control and diagnosis of diseases and the prescription of clinical treatments by health institutions such as the American Food and Drug

Administration (FDA) and European Medicine Agency (EMA). As an example, in the last few years, the use of CGM as a confirmatory measure has been validated to assess the efficacy of diabetes treatments [178], and new additions are expected in the next years.

7.2 Future work

To conclude, we would like to sketch some open problems raised in the course of this work.

- The global Fréchet regression method [214] presents important drawbacks in many real problems, by solely capturing those linear relations between the predictors and the response. New semi-parametric methods can be designed in the event that the response variable lies in metric spaces as well as new non linear regression methods based on the RKHS paradigm. As an example, an extension of the popular Wahba's classical representer theorem [238] with Euclidean predictors to the setting of separable metric spaces can be provided as a continuation of the global Fréchet method.
- Wearable technology simultaneously captures several physiological, environmental, and biomechanical variables. Therefore, a clear research direction is an extension of distributional representations for multidimensional/multimodal data.
- The estimation of distributional representations in practice is subject to a good deal of uncertainty. Adapting the methods discussed here to the Bayesian approach can lead to the quantification of the uncertainty in the response and most parameters in a more refined way than the frequentist approach. In this sense, the uncertainty quantification in density estimation and posterior aggregation in the regression model as a plug-in requires new double complex bootstrap strategies. Bayesian approach easily introduces aggregations in classical models and similarly, uncertainty quantification in the posterior distribution can be done.
- Causal inference in digital medicine [1, 232] is one of the most promising research directions to produce reliable scientific knowledge and improve the dynamic assignment of optimal treatment. In the setting of complex data, few proposals are available in the literature [156]. The extension of the methods discussed here to the field of causal inference is an exciting challenge that should be further addressed. We should advance that the estimators proposed in Chapters 5 and 6 can be applied to counterfactual inference because the IPW estimator also works in this setting.

Part V

Appendices

A Theory of U and V-statistics

This section is based on [137, 244].

A.1 The notions of U-and V-statistics to derive limit distributions

Definition 10. Let $m \in \mathbb{N}$, $q \in \{1, \dots, m\}$, \mathcal{X} a metric space, $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space, $X : \Omega \rightarrow \mathcal{X}$ a random variable with law P_X and $(X_i)_{i \in \mathbb{N}} \subset \mathcal{X}$ of iid copies of X , i.e. $(X_i)_{i \in \mathbb{N}} \stackrel{\text{iid}}{\sim} P_X$.

Furthermore, consider the following sets

- $C_q(m) := \{(i_1, \dots, i_q) \in \{1, \dots, m\} : i_1 < \dots < i_q\}$,
- $M_q(m) := \{1, \dots, m\}^q$ (all mappings).

Observe that $|C_q(m)| = \binom{m}{q}$, $|P_q(m)| = \frac{m!}{(m-q)!}$ and $|M_q(m)| = m^q$.

Consider a measurable symmetric function $g : \mathcal{X}^q \rightarrow \mathbb{R}$, and suppose that we are interested in the statistical functional

$$\theta_g := \theta_g(P_X) := E(g(X_1, \dots, X_q)). \quad (\text{A.1})$$

We define the U – statistics estimator as follows

$$U_m(g) := \binom{m}{q}^{-1} \sum_{C_q(m)} g(X_{i_1}, \dots, X_{i_q}), \quad (\text{A.2})$$

while the V – statistics estimator as

$$V_m(g) := \frac{1}{m^q} \sum_{M_q(m)} g(X_{i_1}, \dots, X_{i_q}). \quad (\text{A.3})$$

Due to the symmetry of g , $U_m(g)$ is unbiased estimators of the statistical functional θ_g . The V – statistics on the other hand has a bias due to the occurrence of equal indices in $M_q(m)$. The function g in the literature is also know as kernel function.

Now, we consider, for each $c \in \{1, \dots, q-1\}$, the function $g_c : \mathcal{X}^c \rightarrow \mathbb{R}$ by

$$g_c(X_1, \dots, X_c) := E(g(X_1, \dots, X_c, X_{c+1}, \dots, X_q)), \quad (\text{A.4})$$

and we define for each $c \in \{1, \dots, q\}$,

$$\epsilon_c := \text{Var}(g_c(X_1, \dots, X_c)).$$

Theorem 6. *The variance of $U_m(g)$ is given by*

$$\text{Var}(U_m(g)) = \binom{m}{q}^{-1} \sum_{c=1}^q \binom{q}{c} \binom{m-q}{q-c} \epsilon_c. \quad (\text{A.5})$$

Definition 11. *The Hayék projection of the first order of $U_m(g)$ is given by*

$$\tilde{U}_m(g) = \sum_{j=1}^m E(U_m(g) | X_j) - (m-1)\theta_g. \quad (\text{A.6})$$

Proposition 7. *The center Hayék projection of the first order can be written as*

$$\tilde{U}_m(g) - \theta_g = \frac{q}{m} \sum_{j=1}^m \tilde{g}_1(X_j). \quad (\text{A.7})$$

Theorem 8. *Assume that $E(g^2(X_1, \dots, X_q)) < \infty$. Then*

$$E\left[(U_m(g) - \tilde{U}_m(g))^2\right] = \mathcal{O}(m^{-2}). \quad (\text{A.8})$$

as $m \rightarrow \infty$.

Theorem 9. *Assume that $E(|g(X_1, \dots, X_q)|) < \infty$. Then*

$$U_m(g) \rightarrow \theta_g \quad (\text{A.9})$$

as $m \rightarrow \infty$.

Definition 12. *(Degenerate and non degenerate U-statistics) A U-statistics is called degenerate if $\epsilon_1 = \text{Var}(g_1(X_1)) = 0$ and non-degenerate if $\epsilon_1 > 0$.*

Theorem 10. *(Central Limit theorem of U-statistic) Assume that $E(g^2(X_1, \dots, X_q)) < \infty$. Then*

- if $\epsilon_1 > 0$ it is hold that

$$\sqrt{m}(U_m(g) - \theta_g) \rightarrow^d \mathcal{N}(0, q^2 \epsilon_1). \quad (\text{A.10})$$

as $m \rightarrow \infty$

- $\epsilon_1 = 0$ it is hold that

$$m(U_m(g) - \theta_g) \rightarrow^d \binom{q}{2} \sum_{j=1}^{\infty} \lambda_j (Z_j^2 - 1) \quad (\text{A.11})$$

as $m \rightarrow \infty$, and where for any $j \in \mathbb{N}$, $Z_j \sim \mathcal{N}(0, 1)$, and λ_j 's are solution of the following integral equation

$$(T_{\tilde{g}_2}(f))(x) = \int_{\mathcal{X}} \tilde{g}_2(x, y) f(y) P_X(dy) = E(\tilde{g}_2(x, Y)). \quad (\text{A.12})$$

Theorem 11. (Central Limit theorem of V-statistic) Assume that $E(g^2(X_1, \dots, X_q)) < \infty$. Then

- if $\epsilon_1 > 0$ it is hold that

$$\sqrt{m}(V_m(g) - \theta_g) \rightarrow^d \mathcal{N}(0, q^2 \epsilon_1). \quad (\text{A.13})$$

as $m \rightarrow \infty$

- $\epsilon_1 = 0$ it is hold that

$$m(V_m(g) - \theta_g) \rightarrow^d \binom{q}{2} \sum_{j=1}^{\infty} \lambda_j Z_j^2 \quad (\text{A.14})$$

as $m \rightarrow \infty$, and where for any $j \in \mathbb{N}$, $Z_j \sim \mathcal{N}(0, 1)$, and λ_j 's are solution of the following integral equation

$$(T_{\tilde{g}_2}(f))(x) = \int_{\mathcal{X}} \tilde{g}_2(x, y) f(y) P_X(dy) = E(\tilde{g}_2(x, Y)). \quad (\text{A.15})$$

B Proofs: Chapters 5 and 6

B.1 Chapter 5

Definition 13. (*Gaussian process*) A Gaussian process is a stochastic time continuous process $\{X(t) : t \in [0, 1]\}$ with state space \mathbb{R} such that any finite-dimensional projection have a joint Gaussian distribution.

A Gaussian process is centered if $E(X(t)) = 0$ for all $t \in [0, 1]$ and its covariance function is the symmetric covariance function $cov(X(s), X(t)) = E[(X(s) - E(X(s)))(X(t) - E(X(t)))]$

Proposition 12. A centered Gaussian process $X = \{X(t) : t \in [0, 1]\}$, with $X(\cdot) \in L^2[0, 1]$, satisfies the following property:

$$\|X\|^2 = \int_0^1 X(t)^2 dt \stackrel{d}{=} \sum_{n=0}^{\infty} \lambda_n Z_n^2,$$

where $Z_n \sim N(0, \lambda_n)$ and the values of sequence $\{\lambda_n\}_{n=0}^{\infty}$ are strictly positive.

Proof. Let $\{X(t) : t \in [0, 1]\}$ be a centered Gaussian process with covariance function $k(s, t) = E(X(s)X(t))$. Let $\{\phi_n\}_{n=0}^{\infty}$ be an orthonormal basis of $L^2[0, 1]$, then by Parseval's theorem any $X(\cdot) \in L^2[0, 1]$ has the representation

$$X = \sum_{n=0}^{\infty} \langle X, \phi_n \rangle \phi_n, \quad \text{where} \quad \langle f, g \rangle = \int_0^1 f(t)g(t)dt. \quad (\text{B.1})$$

Consequently,

$$\|X\|^2 = \sum_{n=0}^{\infty} \langle X, \phi_n \rangle^2. \quad (\text{B.2})$$

Let $\mathcal{K} : L^2[0, 1] \rightarrow L^2[0, 1]$ be an operator such that $\mathcal{K}f(t) = \int_0^1 k(s, t)f(s)ds$ and let $\{\gamma_n\}_{n=0}^{\infty}$ be the orthonormal basis induced by the spectral problem related with \mathcal{K} , satisfying $\mathcal{K}\gamma_n = \lambda_n\gamma_n, \forall n \in \mathbb{N}$. We compute X_n as the orthogonal projection of X on the subspace spanned by functions $\{\gamma_n\}_{n=0}^{\infty}$

$$X_n = \langle X, \gamma_n \rangle = \int_0^1 X(t)\gamma_n(t)dt, \quad (\text{B.3})$$

which satisfies

$$E(X_n) = \int_0^1 E(X(t)\gamma_n(t)) dt = 0 \tag{B.4}$$

and

$$\begin{aligned} Var(X_n) &= E(X_n^2) = E\left(\int_0^1 \int_0^1 X(s)X(t)\gamma_n(s)\gamma_n(t) ds dt\right) \\ &= \int_0^1 \int_0^1 [k(s,t)\gamma_n(s)ds]\gamma_n(t) dt = \lambda_n \|\gamma_n\|^2 = \lambda_n. \end{aligned} \tag{B.5}$$

Thus, $X_n \sim N(0, \lambda_n)$. In addition, $Cov(X_n, X_m) = 0$ if $n \neq m$, and, as a consequence the random variables of the sequence $\{X_n\}_{n=1}^\infty$ are independent Gaussian.

On the other hand, since $\frac{X_n}{\sqrt{\lambda_n}} \sim N(0, 1)$, then, $\frac{X_n^2}{\lambda_n} \sim \chi_1^2$, and finally

$$\|X\|^2 = \int_0^1 X(t)^2 dt = \sum_{n=1}^\infty \langle X, \gamma_n \rangle^2 = \sum_{n=1}^\infty X_n^2 = \sum_{n=1}^\infty \lambda_n Z_n^2,$$

where $\{Z_n\}_{n=1}^\infty$ are independent Gaussian, i.e. $Z_n \sim N(0, 1)$. □

In order to adapt the Hilbert-Schmidt Independence Criterion to missing responses, we must introduce specific assumptions for the missing data mechanism through propensity scores.

Assumption 4. (*Conditioning missing data mechanism*) Let $\pi(x, \theta)$ be a missing data mechanism, with $x \in \mathcal{X}$ and $\theta \in \Theta$, being Θ the space of parameters, which is a closed subset of \mathbb{R}^p . The following assumptions are made

1. $\pi(x, \theta) > c$ where $c > 0$.
2. $\pi(x, \theta)$ is twice continuously differentiable almost everywhere, with bounded derivatives with respect to $\theta \in \Theta$.
3. The family of functions $\mathcal{F} = \{\frac{1}{\pi(x, \theta)} : \theta \in \Theta\}$ satisfies the uniform entropy condition, that is,

$$\int_0^\infty \sup_Q \sqrt{N(\epsilon, \mathcal{F}, L^2(Q))} d\epsilon < \infty \tag{B.6}$$

where $N(\epsilon, \mathcal{F}, L^2(Q))$ stands for the covering number of the family \mathcal{F} with respect to the $L^2(Q)$ -norm and the supremum is taken over all finitely discrete probability measures Q on \mathcal{X} .

4. Let X_1, X_2, \dots, X_n be independent observations, the estimator $\tilde{\theta}$ admits a Bahadur representation, $\sqrt{n}(\tilde{\theta} - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h(X_i) + o_P(1)$, where $h(\cdot)$ is an influence function that satisfies $E(h(X)) = 0$, and $E(h(X)^2) < \infty$.

Assumption 4.1 restricts the minimum value of the missing data mechanism in order to guarantee that it is far from zero. Such a condition is necessary to ensure that the limit of the IPW estimator will be Gaussian (see [169] for further details). Assumption 4.2 is a standard regularity condition for consistency and allows us to derive the asymptotic distribution of the parameter θ as well as to construct new statistics involving the missing data mechanism. Assumption 4.3 is a more technical condition from empirical process theory that guarantees the empirical consistency of the IPW estimator. Finally, assumption 4.4 is a simplified condition that guarantees that the asymptotic expansion of $\tilde{\theta}$ follows a linear structure, and, as a consequence, the central limit theorem is satisfied.

Here we pay attention to the case that the missing data mechanism is specified by a finite-dimensional parameter $\theta \in \Theta$, e.g., through a logistic regression. However, we could introduce more general assumptions on the space in which we perform the statistical learning for the missing data mechanism. From now on, we denote $\pi(\cdot, \theta)$ simply by $\pi(\cdot)$. Moreover, before proving the omnibus character of the test statistics introduced in Section 5.2.2 together with the bootstrap consistency of the test calibration strategy, we suppose that $\pi(\cdot)$ is known beforehand.

We establish the following key lemma to prove the final result.

Lemma 13. *Given $E(k_X(X, X')^2) < \infty$, $E(k_Y(Y, Y')^2) < \infty$, and a missing data mechanism $\pi(\cdot) = P(\delta = 1 | X = \cdot)$ satisfying Assumptions 4.1 and 4.2. Then, the empirical and bootstrap Hilbert-Schmidt calibration strategy statistics defined in Section 5.2.2 are consistent for detecting all second-order finite-moment alternatives.*

Proof. Let $\mathcal{D}_n = \{(X_i, Y_i, \delta_i)\}_{i=1}^n$ be a random sample of independent, identically distributed observations. Let us recall the equations 5.7 and 5.8 for calculating the weights w_i and the normalized weights w_i^* :

$$w_i = \frac{\delta_i}{n\pi(X_i)}, \quad i = 1, \dots, n,$$

$$w_i^* = \frac{w_i}{\sum_{i=1}^n w_i}, \quad i = 1, \dots, n.$$

Let $\tilde{\mathcal{F}} = \left\{ \frac{\delta}{\pi(X)} f(Y), f \in \mathcal{F} \right\}$ be a class of functions where \mathcal{F} is a Donsker class of those functions including the events that define the empirical distribution $\{\mathbf{1}(Y \leq t) : t \in \mathbb{R}\}$. Using Assumptions 4.1 and 4.2 together with the fact that $\pi(\cdot)$ is a fixed function we infer that $\tilde{\mathcal{F}}$ is also a Donsker class [280].

We introduce the empirical measure associated to the variable Y as

$$\mathbb{P}_{Y,n}^\pi = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{\pi(X_i)} \mathbf{1}\{Y_i\} \quad (\text{B.7})$$

where $1\{Y_i\}$ denotes the Dirac measure centered at the observation Y_i . The corresponding empirical process is $\sqrt{n}(\mathbb{P}_{Y,n}^\pi - P_Y)$. Since $\tilde{\mathcal{F}}$ is a Donsker class, by applying the central limit theorem we obtain the following asymptotic convergence [280]

$$\sqrt{n}(\mathbb{P}_{Y,n}^\pi - P_Y) \xrightarrow{d} \mathbb{B}(P_Y),$$

where $\mathbb{B}(P_Y)$ is a Brownian process with a specific covariance structure determined by function $\pi(\cdot)$.

Analogously, we obtain for the empirical process $\sqrt{n}(\mathbb{P}_{X,n} - P_X)$ the next asymptotic convergence,

$$\sqrt{n}(\mathbb{P}_{X,n} - P_X) \xrightarrow{d} \mathbb{B}(P_X),$$

and for the empirical process $\sqrt{n}(\mathbb{P}_{X,Y,n}^\pi - P_{X,Y})$,

$$\sqrt{n}(\mathbb{P}_{X,Y,n}^\pi - P_{X,Y}) \xrightarrow{d} \mathbb{B}(P_{X,Y}).$$

According to [280], we can obtain the same asymptotic convergence for the corresponding bootstrap empirical processes,

$$\sqrt{n}(\mathbb{P}_{Y,n}^{*,\pi} - \mathbb{P}_{Y,n}^\pi) \xrightarrow{d} \mathbb{B}(P_Y),$$

and we get analogous results for $\mathbb{P}_{X,Y,n}^{*,\pi}$ and $\mathbb{P}_{X,n}^*$.

In Section 5.2.2 we propose the following test statistic (Equation 5.10),

$$\widetilde{HSIC}(\tilde{P}_{X,Y}, \tilde{P}_X \tilde{P}_Y) = \langle \tilde{\phi}_{X,Y} - \tilde{\phi}_Y \otimes \tilde{\phi}_X, \tilde{\phi}_{X,Y} - \tilde{\phi}_Y \otimes \tilde{\phi}_X \rangle,$$

and its bootstrap counterpart (Equation 5.11),

$$\begin{aligned} \widetilde{HSIC}^{j^*}(\tilde{P}_{X,Y}, \tilde{P}_X \tilde{P}_Y) &= \langle \tilde{\phi}_{X,Y} - \tilde{\phi}_{X,Y}^{j^*} + \tilde{\phi}_X^{j^*} \otimes \tilde{\phi}_Y^{j^*} - \tilde{\phi}_X \otimes \tilde{\phi}_Y, \\ &\quad \tilde{\phi}_{X,Y} - \tilde{\phi}_{X,Y}^{j^*} + \tilde{\phi}_X^{j^*} \otimes \tilde{\phi}_Y^{j^*} - \tilde{\phi}_X \otimes \tilde{\phi}_Y \rangle, \end{aligned}$$

Since $\tilde{\phi}_Y = \sum_{i=1}^n w_i k(\cdot, Y_i) \in \mathcal{H}$, we can write $\tilde{\phi}_Y = \psi_Y(\mathbb{P}_{Y,n}^\pi)$ and $\phi_Y = \psi_Y(P_Y)$, where $\psi_Y(\cdot)$ is an appropriate mapping satisfying certain regularity conditions such as Hadamard Differentiability or Quasi Hadamard Differentiability, and where subscript Y denotes dependence on variable Y . Here we assume that $\psi_Y(\cdot)$ is Hadamard Differentiability and, as a consequence, we can directly apply functional delta method:

$$\sqrt{n}(\tilde{\phi}_Y - \phi_Y) = \sqrt{n}(\psi_Y(\mathbb{P}_{Y,n}^\pi) - \psi_Y(P_Y)) \xrightarrow{d} \mathbb{B}^k(\phi_Y), \quad (\text{B.8})$$

or more specifically,

$$\sqrt{n} \left(\sum_{i=1}^n w_i k(\cdot, Y_i) - \int k(\cdot, y) P(dy) \right) \xrightarrow{d} \mathbb{B}^k(\phi_Y), \quad (\text{B.9})$$

where $\mathbb{B}^k(\phi_Y)$ is a Brownian process but, in this case, it also depends on the function $k(\cdot, y)$. We can obtain the same asymptotic convergence for the corresponding bootstrap process.

By combining the previous results with the functional delta methods, we can establish that

$$\sqrt{n}(\tilde{\phi}_{X,Y} - \tilde{\phi}_Y \otimes \tilde{\phi}_X) \xrightarrow{d} \mathbb{B}^k(\phi_{HSIC}), \quad (\text{B.10})$$

where again $\mathbb{B}^k(\phi_{HSIC})$ is a Brownian process.

According to the convention introduced in the notation, we can write the proposed test statistics as

$$\begin{aligned} \widetilde{HSIC}(\tilde{P}_{X,Y}, \tilde{P}_X \tilde{P}_Y) &= \\ &= \langle \psi_{X,Y}(\mathbb{P}_{X,Y,n}^\pi) - \psi_Y(\mathbb{P}_{Y,n}^\pi) \otimes \psi_X(\mathbb{P}_{X,n}^\pi), \\ &\quad \psi_{X,Y}(\mathbb{P}_{X,Y,n}^\pi) - \psi_Y(\mathbb{P}_{Y,n}^\pi) \otimes \psi_X(\mathbb{P}_{X,n}^\pi) \rangle \\ &= \|\psi_{X,Y}(\mathbb{P}_{X,Y,n}^\pi) - \psi_Y(\mathbb{P}_{Y,n}^\pi) \otimes \psi_X(\mathbb{P}_{X,n}^\pi)\|^2 \end{aligned} \quad (\text{B.11})$$

and

$$\begin{aligned} \widetilde{HSIC}^*(\tilde{P}_{X,Y}, \tilde{P}_X \tilde{P}_Y) &= \\ &= \langle \psi_{X,Y}(\mathbb{P}_{X,Y,n}^{*,\pi}) - \psi_{X,Y}(\mathbb{P}_{X,Y,n}^\pi) + \psi_Y(\mathbb{P}_{Y,n}^{*,\pi}) \otimes \\ &\quad \psi_X(\mathbb{P}_{X,n}^{*,\pi}) - \psi_Y(\mathbb{P}_{Y,n}^\pi) \otimes \psi_X(\mathbb{P}_{X,n}^\pi), \\ &\quad \psi_{(X,Y)}(\mathbb{P}_{X,Y,n}^{*,\pi}) - \psi_{(X,Y)}(\mathbb{P}_{X,Y,n}^\pi) + \psi_Y(\mathbb{P}_{Y,n}^{*,\pi}) \otimes \\ &\quad \psi_X(\mathbb{P}_{X,n}^{*,\pi}) - \psi_Y(\mathbb{P}_{Y,n}^\pi) \otimes \psi_X(\mathbb{P}_{X,n}^\pi) \rangle \\ &= \|\psi_{X,Y}(\mathbb{P}_{X,Y,n}^{*,\pi}) - \psi_{X,Y}(\mathbb{P}_{X,Y,n}^\pi) + \psi_Y(\mathbb{P}_{Y,n}^{*,\pi}) \otimes \\ &\quad \psi_X(\mathbb{P}_{X,n}^{*,\pi}) - \psi_Y(\mathbb{P}_{Y,n}^\pi) \otimes \psi_X(\mathbb{P}_{X,n}^\pi)\|^2, \end{aligned} \quad (\text{B.12})$$

where for keeping the notation uncluttered we have removed the j superscript corresponding to each bootstrap instance. We can write

$$\widetilde{HSIC}(\tilde{P}_{X,Y}, \tilde{P}_X \tilde{P}_Y) = \|\varphi(\psi_{X,Y}(\mathbb{P}_{X,Y,n}^\pi), \psi_Y(\mathbb{P}_{Y,n}^\pi) \otimes \psi_X(\mathbb{P}_{X,n}^\pi))\|^2 \quad (\text{B.13})$$

and

$$\widetilde{HSIC}^*(\tilde{P}_{X,Y}, \tilde{P}_X \tilde{P}_Y) = \|\varphi'(\psi_{X,Y}(\mathbb{P}_{X,Y,n}^{*,\pi}), \psi_Y(\mathbb{P}_{Y,n}^{*,\pi}) \otimes \psi_X(\mathbb{P}_{X,n}^{*,\pi}))\|^2 \quad (\text{B.14})$$

where φ and φ' are two mappings that allow us to write the test statistics in a more compact form. Under the null hypothesis $H_0 : P_{X,Y} = P_X P_Y$, and according to proposition 12, we have

$$n(\widetilde{HSIC}(\tilde{P}_{X,Y}, \tilde{P}_X \tilde{P}_Y)) \xrightarrow{d} \chi^2(\phi_{HSIC}), \quad (\text{B.15})$$

where $\chi^2(\phi_{HSIC})$ is an infinite combination of weighted χ^2 distributions. Moreover, also under the null hypothesis, the centered bootstrap process satisfies the same property

$$n\left(\widetilde{HSIC}^*(\tilde{P}_{X,Y}, \tilde{P}_X \tilde{P}_Y)\right) \xrightarrow{d} \chi^2(\phi_{HSIC}). \quad (\text{B.16})$$

It can be easily proven the convergence $\widetilde{HSIC}(\tilde{P}_{X,Y}, \tilde{P}_X \tilde{P}_Y) \xrightarrow{p} HSIC(P_{X,Y}, P_X P_Y)$, by direct application of continuous mapping theorem to expression B.13, and noting that the joint and marginal empirical measures weakly converge to population measures, that is,

$$\|\mathbb{P}_{Y,n}^\pi - P_Y\|_\infty \xrightarrow{p} 0, \quad \|\mathbb{P}_{X,n} - P_X\|_\infty \xrightarrow{p} 0 \quad \text{and} \quad \|\mathbb{P}_{X,Y,n}^\pi - P_{X,Y}\|_\infty \xrightarrow{p} 0,$$

by the Glivenko–Cantelli theorem.

Let us note that by definition, $HSIC(P_{X,Y}, P_X P_Y) \geq 0$, where the equality holds if and only if X and Y are independent. Thus, when the null hypothesis is not true the test statistic diverges asymptotically, due to statistics consistency. In addition, since the bootstrap test calibration strategy under the null hypothesis mimics the limit distribution, the bootstrap consistency of the proposed calibration strategy is guaranteed [126]. \square

Importantly, this test statistic can be written as a V-statistic, but the Efrón bootstrap is no longer consistent in general [15], and it is necessary to resort to other strategies such as subsampling [219] or to provide a different centered statistic.

Finally, even though we apply Efrón Bootstrap to an HSIC-based statistic for missing data, also providing theoretical guarantees, the arguments remain valid in the complete data case.

Theorem 14. *Given $E(k_X(X, X')^2) < \infty$, $E(k_Y(Y, Y')^2) < \infty$, and an estimates of a missing data mechanism $\tilde{\pi}(\cdot) = \mathbb{P}(R = 1 | X = \cdot)$ satisfying Assumptions 4.1-4.4. Then, the empirical and bootstrap Hilbert-Schmidt calibration strategy statistics defined in Section 5.2.2, are consistent for detecting all second-order finite-moment alternatives.*

Proof. Consider the following simple decomposition:

$$\begin{aligned} n\left(\widetilde{HSIC}(\mathbb{P}_{n,X,Y}^{\tilde{\pi}}, \mathbb{P}_{n,X} \mathbb{P}_{n,Y}^{\tilde{\pi}}) - HSIC(P_{X,Y}, P_X P_Y)\right) &= \\ &= n\left(\widetilde{HSIC}(\mathbb{P}_{n,X,Y}^{\tilde{\pi}}, \mathbb{P}_{n,X} \mathbb{P}_{n,Y}^{\tilde{\pi}}) - \widetilde{HSIC}(\mathbb{P}_{n,X,Y}^\pi, \mathbb{P}_{n,X} \mathbb{P}_{n,Y}^\pi)\right) \\ &\quad + \widetilde{HSIC}(\mathbb{P}_{n,X,Y}^\pi, \mathbb{P}_{n,X} \mathbb{P}_{n,Y}^\pi) - HSIC(P_{X,Y}, P_X P_Y) \end{aligned} \quad (\text{B.17})$$

By using the notation introduced in equation B.13

$$\begin{aligned} \widetilde{HSIC}(\mathbb{P}_{n,X,Y}^{\tilde{\pi}}, \mathbb{P}_{n,X} \mathbb{P}_{n,Y}^{\tilde{\pi}}) - \widetilde{HSIC}(\mathbb{P}_{n,X,Y}^\pi, \mathbb{P}_{n,X} \mathbb{P}_{n,Y}^\pi) &= \\ &= \|\varphi\left(\psi_{X,Y}(\mathbb{P}_{X,Y,n}^{\tilde{\pi}}), \psi_Y(\mathbb{P}_{Y,n}^{\tilde{\pi}}) \otimes \psi_X(\mathbb{P}_{X,n})\right)\|^2 \\ &\quad - \|\varphi\left(\psi_{X,Y}(\mathbb{P}_{X,Y,n}^\pi), \psi_Y(\mathbb{P}_{Y,n}^\pi) \otimes \psi_X(\mathbb{P}_{X,n})\right)\|^2, \end{aligned} \quad (\text{B.18})$$

where in the first term on the right-hand side the missing data mechanism $\tilde{\pi}(\cdot)$ is estimated, while in the second one $\pi(\cdot)$ it is known.

Assumptions 4.1-4.4 allow us to establish the weak convergence given by $\mathbb{P}_{X,Y,n}^{\tilde{\pi}} \rightarrow \mathbb{P}_{X,Y,n}^{\pi}$ and $\mathbb{P}_{Y,n}^{\tilde{\pi}} \rightarrow \mathbb{P}_{Y,n}^{\pi}$. By virtue of the continuous mapping theorem we have

$$\widehat{HSIC}(\mathbb{P}_{n,X,Y}^{\tilde{\pi}}, \mathbb{P}_{n,X} \mathbb{P}_{n,Y}^{\tilde{\pi}}) - \widehat{HSIC}(\mathbb{P}_{n,X,Y}^{\pi}, \mathbb{P}_{n,X} \mathbb{P}_{n,Y}^{\pi}) \rightarrow 0. \quad (\text{B.19})$$

Finally, by invoking the Slutsky theorem and using the results established in the prior Lemma, we deduce the asymptotic χ^2 distribution and test consistency under the null hypothesis. By replicating the same arguments as before, we obtain the counterpart for the bootstrap process. \square

B.2 Chapter 6

B.2.1 Proof of Theorem 5

Theorem. *Let $\mathbf{X}^{com} = \{(X_{1i}, X_{2i})^T\}_{i=1}^{n_1}$ be a set of i.i.d. complete paired samples, and $\mathbf{X}_1^{inc} = \{X_{1i}\}_{i=n_1+1}^{n_1+n_2}$, and $\mathbf{X}_2^{inc} = \{X_{2i}\}_{i=n_1+n_2+1}^{n_1+n_2+n_3}$ two sets of i.i.d. incomplete paired samples. Let suppose that $n_1/(n_1 + n_2 + n_3) \rightarrow \kappa_1 \in (0, 1)$ and $n_2/(n_1 + n_2 + n_3) \rightarrow \kappa_2 \in (0, 1)$ as $n_1, n_2, n_3 \rightarrow \infty$; then, the test statistic given by (6.3) is consistent against the alternative $H_1 : \{P_1 \neq P_2\}$; we can detect a difference in distribution with the sample size growing to infinity. Furthermore, the calibration strategy described above is also consistent in the same sense.*

Proof. The test statistics given by (6.3) is a convex combination of two independent statistics \mathcal{T}_1 and \mathcal{T}_2 .

We must note that \mathcal{T}_1 is a degenerate one-sample V-statistic with kernel: $h(z_i, z_j) = k(x_{1i}, x_{1j}) - k(x_{1i}, x_{2j}) - k(x_{1j}, x_{2i}) + k(x_{2i}, x_{2j})$, where $z_i = (x_{1i}, x_{2i})$. It follows from the weak law of large numbers of V-statistics that we have convergence in probability to the expected value

$$\frac{n_1}{n_1 + n_2 + n_3} \mathcal{T}_1(\mathbf{X}_1^{com}, \mathbf{X}_2^{com}) \xrightarrow{p} \mathbf{E}[h(Z, Z')], \quad (\text{B.20})$$

where Z' is a i.i.d copy from Z , and under the null hypothesis, and $\mathbf{E}[h(Z, Z')] = 0$. The wild bootstrap consistency for the calibration strategy easily derives from results by [151].

Similarly, \mathcal{T}_2 is a degenerate two sample V-statistic; following [97] we note the point-wise consistency

$$\frac{n_2 + n_3}{n_1 + n_2 + n_3} \mathcal{T}_2(\mathbf{X}_1^{inc}, \mathbf{X}_2^{inc}) \xrightarrow{p} \mathbf{E}[h((X, X'), (Y, Y'))], \quad (\text{B.21})$$

where X, X' from $\mathbf{X}_1^{\text{inc}}$ and Y, Y' from $\mathbf{X}_2^{\text{inc}}$ are i.i.d copies, and $h((x_i, x_j), (y_l, y_m)) = k(x_i, x_j) - k(x_i, y_l) - k(x_j, y_m) + k(y_l, y_m)$. Moreover, under the null hypothesis $E[h((X, X'), (Y, Y'))] = 0$. Furthermore, permutation calibration strategy under random permutation of the pooled sample has been proved consistent by [262].

We must note that if the null hypothesis is not true then both statistics diverge. Moreover, the combination test is omnibus. □

C R package: Biosensors.usc

The R package biosensor.usc aims to provide a unified and user-friendly framework for using new distributional representations of biosensors data in different statistical modeling tasks: regression models, hypothesis testing, cluster analysis, visualization, and descriptive analysis. Distributional representations are a functional extension of compositional time-in-range metrics and we have used them successfully so far in modeling glucose profiles and accelerometer data. However, these functional representations can be used to represent any biosensor data such as ECG or medical imaging such as fMRI.

C.1 Installation

You can install this package from source code using the devtools library:

```
devtools::install_github("glucodensities/biosensors.usc@main",  
type = "source")
```

C.1.1 Quick start

The purpose of this section is to give users a general sense of the package, including the components, what they do and some basic usage. We will briefly go over the main functions, see the basic operations and have a look at the outputs. Users may have a better idea after this section what functions are available. More details are available in the package documentation.

First, we load the biosensors.usc package:

```
library(biosensors.usc)
```

C.1.2 Package example

This example is extracted from the paper [105].

We include part of this data set in the inst/exdata folder. This data set has two different types of files. The first one contains the functional data, which csv files must have long format

with, at least, the following three columns: id, time, and value. The id identifies the individual, the time indicates the moment in which the data was captured, and the value is a monitor measure:

```
file1 = system.file("extdata", "data_1.csv", package = "biosensors.usc")
```

The second type contains the clinical variables. This csv file must contain a row per individual and must have a column id identifying this individual:

```
file2 = system.file("extdata", "variables_1.csv", package = "biosensors.usc")
```

From these files, biosensor data can be loaded as follow:

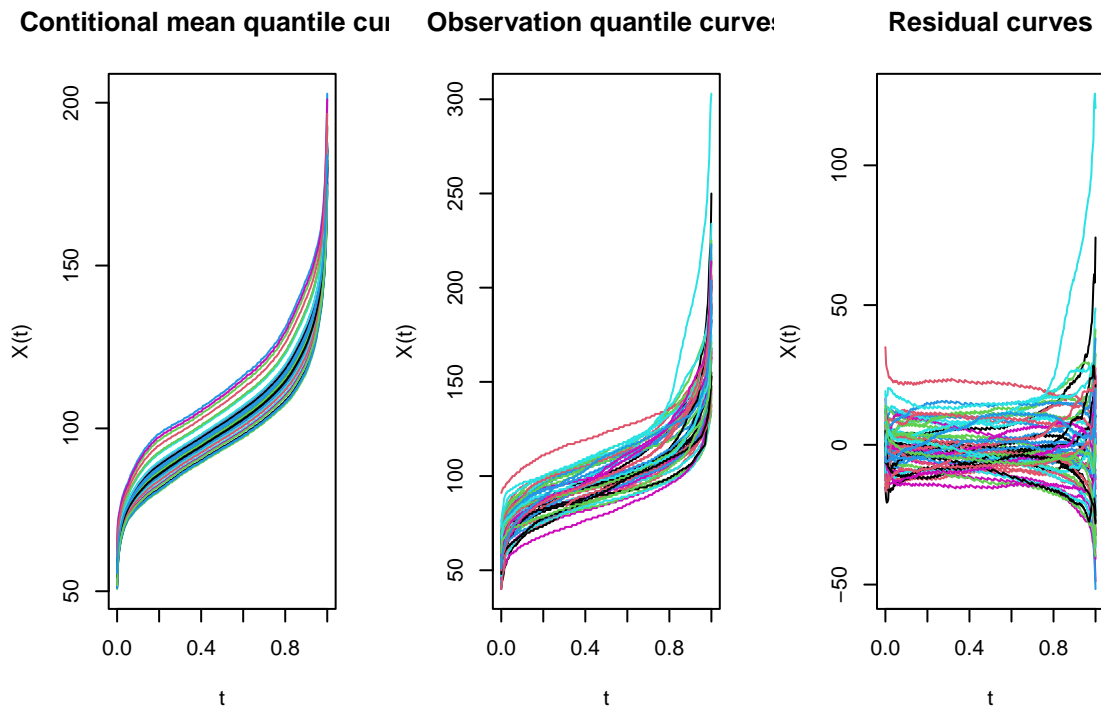
```
data1 = load_data(file1, file2)
class(data1)
#> [1] "biosensor"
names(data1)
#> [1] "data"      "densities" "quantiles" "variables"
```

The `load_data` function returns a biosensor object. This object contains a data frame with biosensor raw data, a functional data object (fdata) with a non-parametric density estimation, a functional data object (fdata) with the empirical quantile estimation, and a data frame with the covariates.

C.1.2.1 Wasserstein regression and prediction

You can call the Wasserstein regression, using as predictor the distributional representation and as response a scalar outcome. In this example, we use the previously loaded biosensor data and the BMI covariate:

```
regm = regmod_regression(data1, "BMI")
```

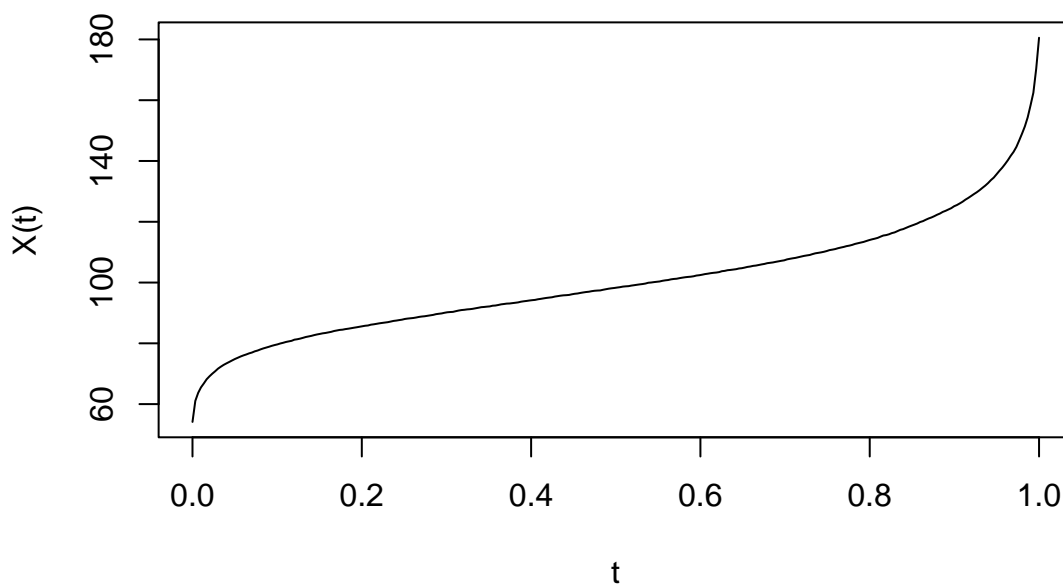


As result, this function returns the fitted regression and plots the residuals of the curves against the fitted values. In addition, the function plots the confidence band of the mean values.

We can obtain the regression prediction from a $k \times p$ matrix of input values for regressors for prediction, where k is the number of points we do the prediction and p is the dimension of the input variables:

```
xpred = as.matrix(25)
pred = regmod_prediction(regm, xpred)
```

Wasserstein prediction



C.1.2.2 Ridge regression

Call the ridge regression as follows, using as predictor the distributional representation and as response a scalar outcome:

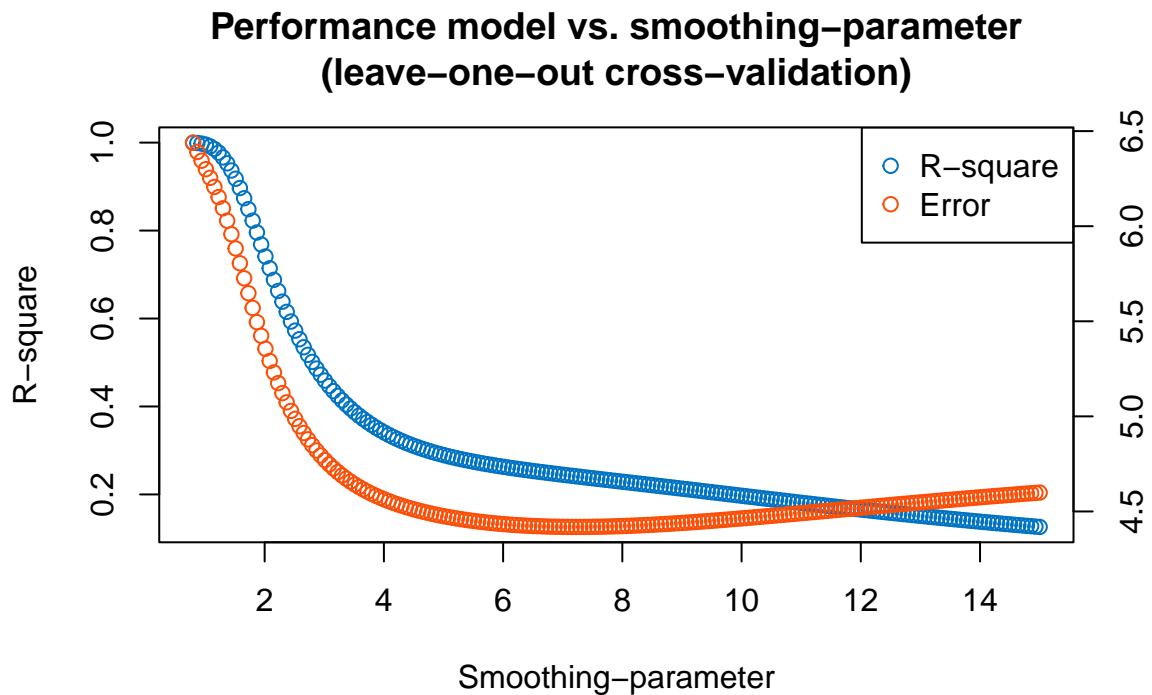
```
ridg = ridge_regression(data1, "BMI")
```

C.1.2.3 Nadaraya-Watson regression and prediction

Use the following function to obtain the functional non-parametric Nadaraya-Watson regression with 2-Wasserstein distance, using as predictor the distributional representation and as response a scalar outcome:



```
nada = nadayara_regression(data1, "BMI")
```

Use the previously computed Nadaraya-Watson regression to obtain the regression prediction given the quantile curves:

```
npre = nadayara_prediction(nada, t(colMeans(data1quantilesdata)))
```

C.1.2.4 Hypothesis testing

We can perform hypothesis testing between two random samples of distributional representations to detect differences in scale and localization (ANOVA test) or distributional differences (energy distance).

Let's load first another sample:

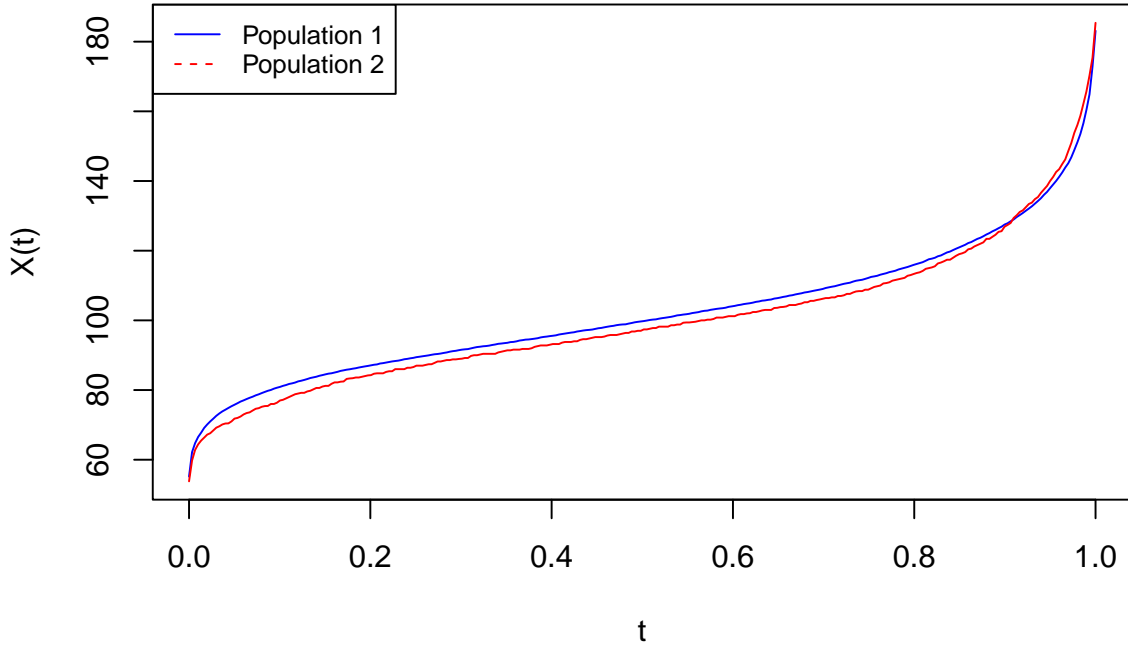
```
file3 = system.file("extdata", "data_2.csv", package = "biosensors.usc")
file4 = system.file("extdata", "variables_2.csv", package = "biosensors.usc")
data2 = load_data(file3, file4)
```



Then call the following function:

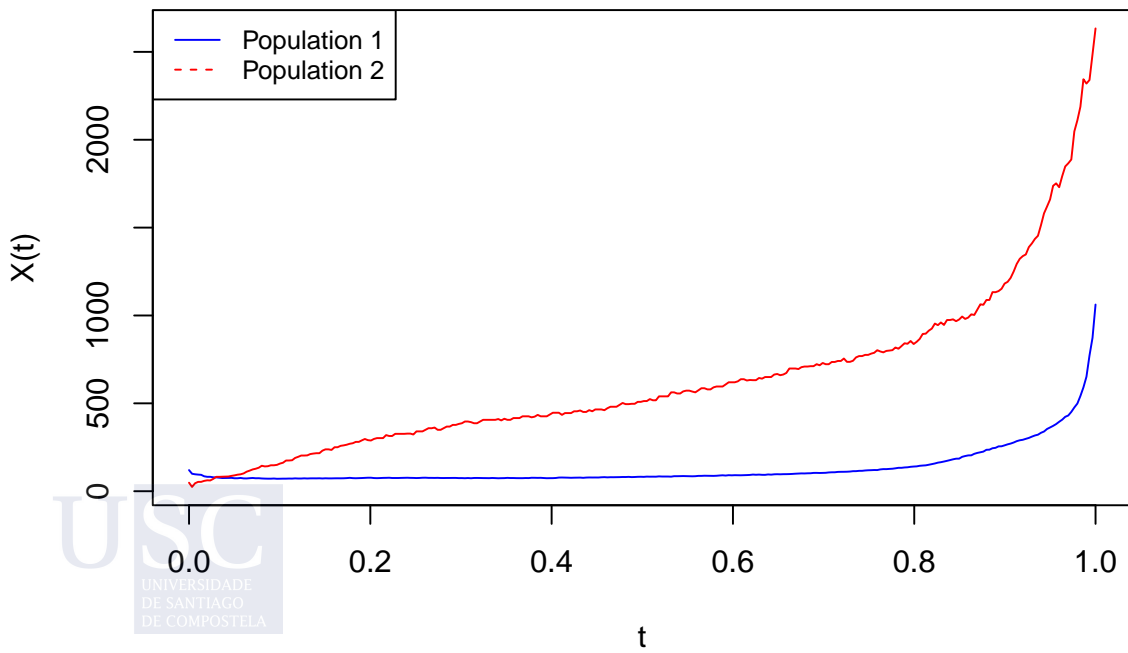
```
htest = hypothesis_testing(data1, data2)
```

Quantile mean



#> Warning: executing %dopar% sequentially: no parallel backend registered

Quantile variance



The function will plot the quantile mean and the quantile variance of the two populations. The corresponding p-values of the ANOVA test and distributional differences are stored in the following names:

```
print(htestenergy_pvalue)
#> [1] 0.00990099
print(htestanova_pvalue)
#> [1] 0.0003094763
```

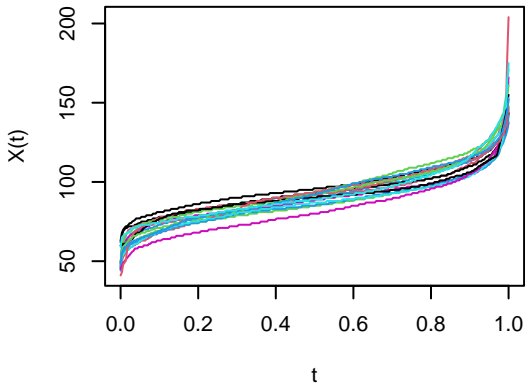
C.1.2.5 Clustering

Call the energy clustering with Wasserstein distance using quantile distributional representations as covariates:

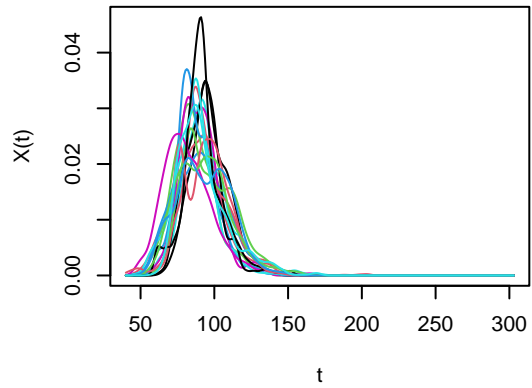


```
clus = clustering(data1, clusters=3)
```

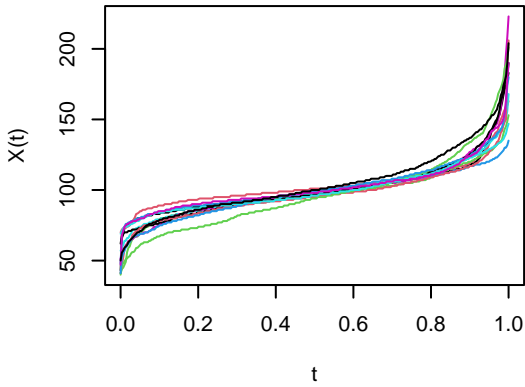
Cluster 1 (quantiles)



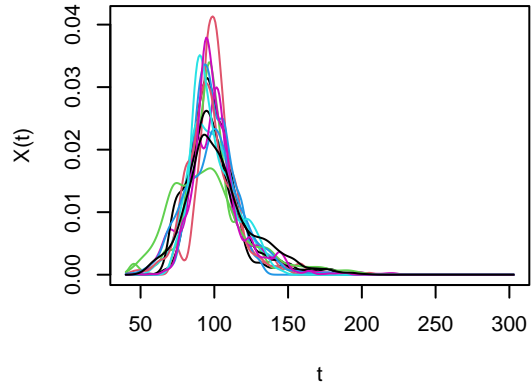
Cluster 1 (densities)



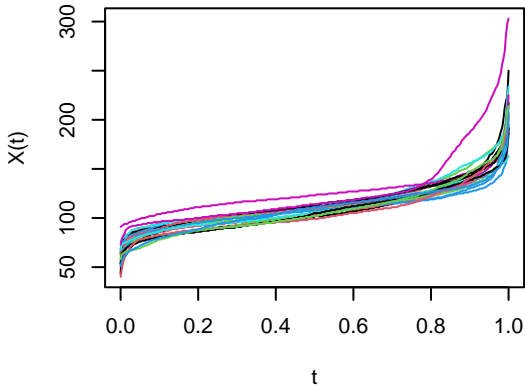
Cluster 2 (quantiles)



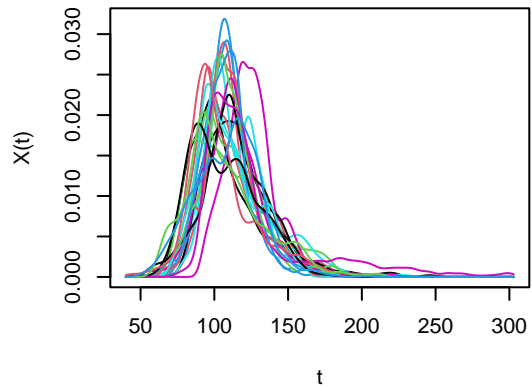
Cluster 2 (densities)



Cluster 3 (quantiles)



Cluster 3 (densities)



The function also plots the clusters of quantiles and densities.

We can also use the previously computed clustering to obtain the clusters of another set of objects calling the following function:

```
assignments = clustering_prediction(clus, data1quantilesdata)
print(assignments)
#> [1] 1 1 1 1 1 2 2 1 1 1 1 2 2 3 1 2 1 3 2
#> [20] 3 1 2 1 3 2 1 3 1 3 2 2 2 1 1 1 1 2 1
#> [39] 3 2 3 3 3 3 2 3 2 1 3 2 3 1
```


Bibliography

- [1] Causality in digital medicine. *Nature Communications*, 12(1):5471, Sep 2021.
- [2] Benjamin Ackerman, Catherine R Lesko, Juned Siddique, Ryoko Susukida, and Elizabeth A Stuart. Generalizing randomized trial findings to a target population using complex survey population data. *Statistics in Medicine*, 40(5):1101–1120, 2021.
- [3] Jung Ae Lee and Jeff Gill. Missing value imputation for physical activity data measured by accelerometer. *Statistical Methods in Medical Research*, 27(2):490–506, 2018.
- [4] Emma Ahlqvist, Petter Storm, Annemari Käräjämäki, Mats Martinell, Mozghan Dorkhan, Annelie Carlsson, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: A data-driven cluster analysis of six variables. *The Lancet Diabetes & Endocrinology*, 6:361–369, 03 2018.
- [5] Emma Ahlqvist, Tiinamaija Tuomi, and Leif Groop. Clusters provide a better holistic view of type 2 diabetes than simple clinical features. *The Lancet Diabetes & Endocrinology*, 7(9):668–669, 2019.
- [6] Michael G Akritas, Efi S Antoniou, and Jouni Kuha. Nonparametric analysis of factorial designs with random missingness: bivariate data. *Journal of the American Statistical Association*, 101(476):1513–1526, 2006.
- [7] Michael G Akritas, Jouni Kuha, and D Wayne Osgood. A nonparametric approach to matched pairs with missing data. *Sociological Methods & Research*, 30(3):425–454, 2002.
- [8] Osvaldo P Almeida, Karim M Khan, Graeme J Hankey, Bu B Yeap, Jonathan Golledge, and Leon Flicker. 150 minutes of vigorous physical activity per week predicts survival and successful ageing: a population-based 11-year longitudinal study of 12,201 older Australian men. *British Journal of Sports Medicine*, 48(3):220–225, 2014.
- [9] Luigi Ambrosio and Nicola Gigli. A user’s guide to optimal transport. In *Modelling and Optimisation of Flows on Networks*, pages 1–155. Springer, 2013.

- [10] American Diabetes Association. 7. Diabetes technology: Standards of medical care in diabetes-2019. *Diabetes Care*, 42:S71–S80, 2019.
- [11] Lubna Amro, Frank Konietzschke, and Markus Pauly. Multiplication-combination tests for incomplete paired data. *Statistics in Medicine*, 38:3243–3255, 2019.
- [12] Lubna Amro and Markus Pauly. Permuting incomplete paired data: a novel exact and asymptotic correct randomization test. *Journal of Statistical Computation and Simulation*, 87(6):1148–1159, 2017.
- [13] Lubna Amro, Markus Pauly, and Burim Ramosaj. Asymptotic-based bootstrap approach for matched pairs with missingness in a single arm. *Biometrical Journal*, 63(7):1389–1405, 2021.
- [14] A. Antoniadis. Wavelets in statistics: A review. *Journal of the Italian Statistical Society*, 6(2):97, Aug 1997.
- [15] Miguel A Arcones and Evarist Gine. On the bootstrap of U and V statistics. *The Annals of Statistics*, pages 655–674, 1992.
- [16] Kristine Arges, Themistocles Assimes, Vikram Bajaj, Suresh Balu, Mustafa R Bashir, Laura Beskow, et al. The project baseline health study: a step towards a broader mission to map human health. *NPJ digital medicine*, 3(1):1–10, 2020.
- [17] American Diabetes Association. 6. Glycemic targets: standards of medical care in diabetes—2018. *Diabetes Care*, 41(Supplement 1):S55–S64, 2018.
- [18] Audie A Atienza, Richard P Moser, Frank Perna, Kevin Dodd, Rachel Ballard-Barbash, Richard P Troiano, et al. Self-reported and objectively measured activity related to biomarkers using NHANES. *Medicine and Science in Sports and Exercise*, 43(5):815–821, 2011.
- [19] Karlee J. Ausk, Edward J. Boyko, and George N. Ioannou. Insulin resistance predicts mortality in nondiabetic individuals in the u.s. *Diabetes Care*, 33(6):1179–1185, 2010.
- [20] Tadej Battelino, Thomas Danne, Richard M. Bergenstal, Stephanie A. Amiel, Roy Beck, Torben Biester, et al. Clinical targets for continuous glucose monitoring data interpretation: Recommendations from the international consensus on time in range. *Diabetes Care*, 2019.
- [21] Roy W. Beck, Richard M. Bergenstal, Tonya D. Riddlesworth, Craig Kollman, Zhaomian Li, Adam S. Brown, et al. Validation of time in range as an outcome measure for diabetes clinical trials. *Diabetes Care*, 42(3):400–405, 2019.

- [22] Roy W. Beck, Crystal G. Connor, Deborah M. Mullen, David M. Wesley, and Richard M. Bergenstal. The fallacy of average: How using hba1c alone to assess glycemic control can be misleading. *Diabetes Care*, 40(8):994–999, 2017.
- [23] Orly Ben-Yacov, Anastasia Godneva, Michal Rein, Smadar Shilo, Dmitry Kolobkov, Netta Koren, et al. Personalized postprandial glucose response–targeting diet versus mediterranean diet for glycemic control in prediabetes. *Diabetes Care*, 44(9):1980–1991, 2021.
- [24] Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic analysis on semigroups: theory of positive definite and related functions*, volume 100. Springer, 1984.
- [25] Richard M. Bergenstal, Andrew J. Ahmann, Timothy Bailey, Roy W. Beck, Joan Bissen, Bruce Buckingham, et al. Recommendations for standardizing glucose reporting and analysis to optimize clinical decision making in diabetes: The ambulatory glucose profile (agp). *Diabetes Technology & Therapeutics*, 15(3):198–211, 2013. PMID: 23448694.
- [26] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.
- [27] Dimitris Bertsimas, Nathan Kallus, Alexander M Weinstein, and Ying Daisy Zhuo. Personalized diabetes management using electronic medical records. *Diabetes Care*, 40(2):210–217, 2017.
- [28] Dimitris Bertsimas, Agni Orfanoudaki, and Rory B Weiner. Personalized treatment for coronary artery disease: A machine learning approach. *Circulation*, 138(Suppl_1):A11213–A11213, 2018.
- [29] Satarupa Bhattacharjee and Hans-Georg Mueller. Concurrent object regression. *ArXiv Preprint*, 2021.
- [30] Lyvia Biagi, Arthur Bertachi, Marga Giménez, Ignacio Conget, Jorge Bondia, Josep Antoni Martín-Fernández, et al. Individual categorisation of glucose profiles using compositional data analysis. *Statistical Methods in Medical Research*, 28(12):3550–3567, 2019.
- [31] Lyvia Biagi, Arthur Bertachi, Marga Giménez, Ignacio Conget, Jorge Bondia, Josep Antoni Martín-Fernández, et al. Individual categorisation of glucose profiles using compositional data analysis. *Statistical Methods in Medical Research*, 28(12):3550–3567, 2019. PMID: 30380996.

- [32] Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. Geodesic pca in the wasserstein space by convex pca. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 53, pages 1–26. Institut Henri Poincaré, 2017.
- [33] Kristian Bolin. Physical inactivity: productivity losses and healthcare costs 2002 and 2016 in sweden. *BMJ Open Sport & Exercise Medicine*, 4(1):e000451, 2018.
- [34] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.
- [35] Thomas W Buford, Michael D Roberts, and Timothy S Church. Toward exercise as personalized medicine. *Sports Medicine*, 43(3):157–165, 2013.
- [36] Johannes Burtscher and Martin Burtscher. Run for your life: tweaking the weekly physical activity volume for longevity, 2020.
- [37] Avivit Cahn, Avi Shoshan, Tal Sagiv, Rachel Yescharim, Ran Goshen, Shalev Varda, and Itamar Raz. Prediction of progression from pre-diabetes to diabetes: Development and validation of a machine learning model. *Diabetes Metabolism. Research and Reviews*, 36(2):e3252, 2020.
- [38] Louis Capitaine, Robin Genuer, and Rodolphe Thiébaud. Fréchet random forests, 2019.
- [39] Sebastien FM Chastin, Javier Palarea-Albaladejo, Manon L Dontje, and Dawn A Skelton. Combined effects of time spent in physical activity, sedentary behaviors and sleep on obesity and cardio-metabolic health markers: a novel compositional data analysis approach. *PLoS ONE*, 10(10):e0139984, 2015.
- [40] Jianbo Chen, Mitchell Stern, Martin J Wainwright, and Michael I Jordan. Kernel feature selection via conditional covariance minimization. *Advances in Neural Information Processing Systems (NIPS 2017)*, 30:6946–6955, 2017.
- [41] Yaqing Chen, Paromita Dubey, Hans-Georg Müller, Muriel Bruchhage, Jane-Ling Wang, and Sean Deoni. Modeling sparse longitudinal data in early neurodevelopment. *NeuroImage*, 237:118079, 2021.
- [42] Chee W Chia, Josephine M Egan, and Luigi Ferrucci. Age-related changes in glucose metabolism, hyperglycemia, and cardiovascular risk. *Circulation*, 123(7):886–904, September 2018.
- [43] Andreas Christmann and Ingo Steinwart. *Support vector machines*. Springer, 2008.

- [44] Kian Fan Chung. Defining phenotypes in asthma: a step towards personalized medicine. *Drugs*, 74(7):719–728, 2014.
- [45] Wendy K Chung, Karel Erion, Jose C Florez, Andrew T Hattersley, Marie-France Hivert, Christine G Lee, et al. Precision medicine in diabetes: a consensus report from the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). *Diabetes Care*, 43(7):1617–1635, 2020.
- [46] Davide Cirillo and Alfonso Valencia. Big data analytics for personalized medicine. *Current Opinion in Biotechnology*, 58:161–167, 2019.
- [47] Dan Tsir Cohen and Aryeh Kontorovich. Metric-valued regression. *ArXiv Preprint*, 2022.
- [48] Philip E. Cryer. Glycemic goals in diabetes: Trade-off between glycemic control and iatrogenic hypoglycemia. *Diabetes*, 63(7):2188–2195, 2014.
- [49] Paul Dagum. Digital biomarkers of cognitive function. *NPJ Digital Medicine*, 1(1):1–3, 2018.
- [50] Xiongtao Dai, Zhenhua Lin, and Hans-Georg Müller. Modeling sparse longitudinal data on Riemannian manifolds. *Biometrics*, 77(4):1328–1341, 2021.
- [51] H el ene De Canni ere, Christophe JP Smeets, Melanie Schoutteten, Carolina Varon, Chris Van Hoof, Van Huffel, et al. Using biosensors and digital biomarkers to assess response to cardiac rehabilitation: Observational study. *Journal of Medical Internet Research*, 22(5):e17326, 2020.
- [52] Borja del Pozo Cruz, Stuart JH Biddle, Paul A Gardiner, and Ding Ding. Light-intensity physical activity and life expectancy: National Health and Nutrition Survey. *American Journal of Preventive Medicine*, 61(3):428–433, 2021.
- [53] Borja del Pozo Cruz, Duncan E McGregor, Jes us del Pozo Cruz, Matthew P Buman, Javier Palarea-Albaladejo, Rosa M Alfonso-Rosa, et al. Integrating sleep, physical activity, and diet quality to estimate all-cause mortality risk: A combined compositional clustering and survival analysis of the national health and nutrition examination survey 2005–2006 cycle. *American Journal of Epidemiology*, 189(10):1057–1064, 2020.
- [54] John Dennis. Precision medicine in type 2 diabetes: Using individualized prediction models to optimize selection of treatment. *Diabetes*, 69:2075–2085, 2020.
- [55] Ben Derrick, Annalise Ruck, Deirdre Toher, and Paul White. Tests for equality of variances between two samples which contain both paired observations and independent observations. *Journal of Applied Quantitative Methods*, 13(2):36–47, 2018.

- [56] Chong-Zhi Di, Ciprian M Crainiceanu, Brian S Caffo, and Naresh M Punjabi. Multilevel functional principal component analysis. *The Annals of Applied Statistics*, 3(1):458, 2009.
- [57] Junrui Di, Charmaine Demanuele, Anna Kettermann, F Isik Karahanoglu, Joseph C Cappelleri, Andrew Potter, et al. Considerations to address missing data when deriving clinical trial endpoints from digital health technologies. *Contemporary Clinical Trials*, 113:106661, 2022.
- [58] Persi Diaconis and Julia Salzman. Projection pursuit for discrete data. *Probability and Statistics: Essays in Honor of David A. Freedman*, 2:265–288, 2008.
- [59] Linda A DiMeglio, Carmella Evans-Molina, and Richard A Oram. Type 1 diabetes. *The Lancet*, 391(10138):2449 – 2462, 2018.
- [60] Ding Ding, Andrea Ramirez Varela, Adrian E Bauman, Ulf Ekelund, I-Min Lee, and othersl. Towards better evidence-informed global action: lessons learnt from the Lancet series and recent developments in physical activity and public health. *British journal of Sports Medicine*, 54(8):462–468, 2020.
- [61] Ian L Dryden and Kanti V Mardia. *Statistical shape analysis: with applications in R*, volume 995. John Wiley & Sons, 2016.
- [62] Paromita Dubey and Hans-Georg Müller. Functional models for time-varying random objects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):275–327, 2020.
- [63] Paromita Dubey and Hans-Georg Müller. Fréchet analysis of variance for random objects. *Biometrika*, 106(4):803–821, 10 2019.
- [64] Dorothea Dumuid, Željko Pedišić, Tyman Everleigh Stanford, Josep-Antoni Martín-Fernández, Karel Hron, Carol A Maher, et al. The compositional isotemporal substitution model: a method for estimating changes in a health outcome for reallocation of time between sleep, physical activity and sedentary behaviour. *Statistical Methods in Medical Research*, 28(3):846–857, 2019.
- [65] Dorothea Dumuid, Tyman E Stanford, Josep-Antoni Martín-Fernández, Željko Pedišić, Carol A Maher, Lucy K Lewis, et al. Compositional data analysis for physical activity, sedentary time and sleep research. *Statistical Methods in Medical Research*, 27(12):3726–3738, 2018.
- [66] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

- [67] Gunnar Ekbohm. Comparing means in the paired case with missing data on one response. *Biometrika*, 63(1):169–172, 1976.
- [68] Ulf Ekelund, Jakob Tarp, Morten W Fagerland, Jostein Steene Johannessen, Bjørge H Hansen, Barbara J Jefferis, et al. Joint associations of accelerometer-measured physical activity and sedentary time with all-cause mortality: a harmonised meta-analysis in more than 44 000 middle-aged and older individuals. *British Journal of Sports Medicine*, 54(24):1499–1506, 2020.
- [69] Ulf Ekelund, Jakob Tarp, Jostein Steene-Johannessen, Bjørge H Hansen, Barbara Jefferis, Morten W Fagerland, et al. Dose-response associations between accelerometry measured physical activity and sedentary time and all cause mortality: systematic review and harmonised meta-analysis. *BMJ*, 366, 2019.
- [70] Roger Eston. Use of ratings of perceived exertion in sports. *International Journal of Sports Physiology and Performance*, 7(2):175–182, 2012.
- [71] Sean M Ewings, Sujit K Sahu, John J Valletta, Christopher D Byrne, and Andrew J Chipperfield. A Bayesian network for modelling blood glucose concentration and exercise in type 1 diabetes. *Statistical Methods in Medical Research*, 24(3):342–372, 2015. PMID: 24492795.
- [72] Manuel Febrero-Bande and Manuel de la Fuente. Statistical computing in functional data analysis: The R package *fda.usc*. *Journal of Statistical Software*, 51(4):1–28, 2012.
- [73] Manuel Febrero-Bande, Pedro Galeano, and Wenceslao González-Manteiga. Estimation, imputation and prediction for the functional linear model with scalar response with responses missing at random. *Computational Statistics & Data Analysis*, 131:91–103, 2019.
- [74] Denice S Feig, Lois E Donovan, Rosa Corcoy, Kellie E Murphy, Stephanie A Amiel, Katharine F Hunt, Elizabeth Asztalos, et al. Continuous glucose monitoring in pregnant women with type 1 diabetes (CONCEPTT): a multicentre international randomised controlled trial. *The Lancet*, 390(10110):2347 – 2359, 2017.
- [75] Tamara Fernandez and Arthur Gretton. A maximum-mean-discrepancy goodness-of-fit test for censored data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2966–2975. PMLR, 2019.
- [76] Frédéric Ferraty and Philippe Vieu. *Nonparametric Functional Data Analysis: Theory and Practice (Springer Series in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

- [77] Youyi Fong, Ying Huang, Maria P Lemos, and M Juliana McElrath. Rank-based two-sample tests for paired data with missing values. *Biostatistics*, 19(3):281–294, 2018.
- [78] National Center for Health Statistics et al. Office of analysis and epidemiology. *The Linkage of National Center for Health Statistics Survey Data to the National Death Index—2015 Linked Mortality File (LMF): Methodology Overview and Analytic Considerations*, 2006.
- [79] Committee for Medicinal Products for Human Use. Guideline on clinical investigation of medicinal products in the treatment or prevention of diabetes mellitus. *London, European Medicines Society*, 2012.
- [80] Guilherme França, Maria L Rizzo, and Joshua T Vogelstein. Kernel k-groups via Hartigan’s method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4411–4425, 2020.
- [81] Paul W Franks and N Atabaki-Pasdar. Causal inference in obesity research. *Journal of internal medicine*, 281(3):222–232, 2017.
- [82] Maurice René Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l’institut Henri Poincaré*, 10(4):215–310, 1948.
- [83] Janine Freeman and Lynne Lyons. The use of continuous glucose monitoring to evaluate the glycemic response to food. *Diabetes Spectrum*, 21(2):134–137, 2008.
- [84] Christine M Friedenreich, Charlotte Ryder-Burbidge, and Jessica McNeil. Physical activity, obesity and sedentary behavior in cancer etiology: epidemiologic evidence and biologic mechanisms. *Molecular Oncology*, 15(3):790–800, 2021.
- [85] K. Fukumizu and C. Leng. Gradient-based kernel method for feature extraction and variable selection. In *Advances in Neural Information Processing Systems*, NIPS’12, pages 2114–2122, 2012.
- [86] Daniel Gaigall. Testing marginal homogeneity of a continuous bivariate distribution with possibly incomplete paired data. *Metrika*, 83(4):437–465, 2020.
- [87] E.A. Gale. Is type 2 diabetes a category error? *The Lancet*, 381:1956–1957, 2013.
- [88] Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic. *ArXiv Preprint*, 2017.
- [89] Irina Gaynanova, Naresh Punjabi, and Ciprian Crainiceanu. Modeling continuous glucose monitoring (CGM) data during sleep. *Biostatistics*, 2020.

- [90] Rahul Ghosal, Vijay R Varma, Dmitri Volfson, Inbar Hillel, Jacek Urbanek, Jeffrey M Hausdorff, et al. Distributional data analysis via quantile functions and its application to modelling digital biomarkers of gait in Alzheimer's disease. *ArXiv Preprint*, 2021.
- [91] Jason MR Gill. Linking volume and intensity of physical activity to mortality. *Nature Medicine*, 26(9):1332–1334, 2020.
- [92] Cedric E Ginestet, Jun Li, Prakash Balachandran, Steven Rosenberg, and Eric D Kocaczyk. Hypothesis testing for network data in functional neuroimaging. *The Annals of Applied Statistics*, pages 725–750, 2017.
- [93] Yair Goldberg and Michael R Kosorok. Q-learning with censored data. *Annals of Statistics*, 40(1):529, 2012.
- [94] Jeff Goldsmith, Xinyue Liu, Judith Jacobson, and Andrew Rundle. New insights into activity patterns in children, found using functional data analyses. *Medicine and Science in Sports and Exercise*, 48(9):1723, 2016.
- [95] Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.
- [96] Ana María Gómez, Diana Cristina Henao, Angelica Imitola Madero, Lucia B Taboada, Viviana Cruz, Maria Alejandra Robledo Gomez, et al. Defining high glycemic variability in type 1 diabetes: comparison of multiple indexes to identify patients at risk of hypoglycemia. *Diabetes Technology & Therapeutics*, 21(8):430–439, 2019.
- [97] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.
- [98] Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. *Advances in Neural Information Processing Systems*, 20:585–592, 2007.
- [99] Arthur Gretton, Bharath K. Sriperumbudur, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, and Kenji Fukumizu. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 25, pages 1205–1213, 2012.
- [100] Beyond A1C Writing Group. Need for Regulatory Change to Incorporate Beyond A1C Glycemic Metrics. *Diabetes Care*, 41(6):e92–e94, 05 2018.

- [101] Francisco Gude, Pablo Díaz-Vidal, Cintia Rúa-Pérez, Manuela Alonso-Sampedro, Carmen Fernández-Merino, Jesús Rey-García, et al. Glycemic variability and its association with demographics and lifestyles in a general adult population. *Journal of diabetes science and technology*, 11(4):780–790, 2017.
- [102] Beibei Guo and Ying Yuan. A comparative review of methods for comparing means using partially paired data. *Statistical Methods in Medical Research*, 26(3):1323–1340, 2017.
- [103] Regina Guthold, Gretchen A Stevens, Leanne M Riley, and Fiona C Bull. Worldwide trends in insufficient physical activity from 2001 to 2016: a pooled analysis of 358 population-based surveys with 1·9 million participants. *The Lancet Global Health*, 6(10):e1077–e1086, 2018.
- [104] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
- [105] Heather Hall, Dalia Perelman, Alessandra Breschi, Patricia Limcaoco, Ryan Kellogg, Tracey McLaughlin, and Michael Snyder. Glucotypes reveal new patterns of glucose dysregulation. *PLoS Biology*, 16(7):1–23, 07 2018.
- [106] Peter Hall, Soumendra Nath Lahiri, and Jörg Polzehl. On bandwidth choice in non-parametric regression with both short-and long-range dependent errors. *The Annals of Statistics*, 23(6):1921–1936, 1995.
- [107] Kyunghye Han, Hans-Georg Müller, and Byeong U. Park. Additive functional regression for densities as responses. *Journal of the American Statistical Association*, 0(0):1–24, 2019.
- [108] Torsten Harms and Pierre Duchesne. On kernel nonparametric regression designed for complex survey data. *Metrika*, 72(1):111–138, 2010.
- [109] Christian Herder and Michael Roden. A novel diabetes typology: towards precision diabetology from pathogenesis to treatment. *Diabetology*, 2022.
- [110] Irl B Hirsch. Glycemic variability and diabetes complications: does it matter? of course it does! *Diabetes Care*, 38(8):1610–1614, 2015.
- [111] Irl B. Hirsch, Jennifer L. Sherr, and Korey K. Hood. Connecting the dots: Validation of time in range metrics with microvascular outcomes. *Diabetes Care*, 42(3):345–348, 2019.

- [112] Wieland Hoelzel, Cas Weykamp, Jan-Olof Jeppsson, Kor Miedema, John R. Barr, Ian Goodall, Tadao Hoshino, et al. Ifcc reference system for measurement of hemoglobin A1c in human blood and the national standardization schemes in the United States, Japan, and Sweden: A method-comparison study. *Clinical Chemistry*, 50(1):166–174, 2004.
- [113] Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 2008.
- [114] Susan Holmes. Bootstrapping phylogenetic trees: theory and methods. *Statistical Science*, 18(2):241–255, 2003.
- [115] Susan Holmes. Statistics for phylogenetic trees. *Theoretical population biology*, 63(1):17–32, 2003.
- [116] Hans Hoppeler. Deciphering ν o2, max: limits of the genetic approach. *Journal of Experimental Biology*, 221(21):jeb164327, 2018.
- [117] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.
- [118] Erin K Howie, Anne L Smith, Joanne A McVeigh, and Leon M Straker. Accelerometer-derived activity phenotypes in young adults: a latent class analysis. *International Journal of Behavioral Medicine*, 25(5):558–568, 2018.
- [119] Karel Hron, Alessandra Menafoglio, Matthias Templ, K Hruuzova, and Peter Filzmoser. Simplicial principal component analysis for density functions in Bayes spaces. *Computational Statistics & Data Analysis*, 94:330–350, 2016.
- [120] Frank B Hu, Ambika Satija, and JoAnn E Manson. Curbing the diabetes pandemic: the need for global policy solutions. *JAMA*, 313(23):2319–2320, 2015.
- [121] Stephan Huckemann, Thomas Hotz, and Axel Munk. Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric lie group actions. *Statistica Sinica*, pages 1–58, 2010.
- [122] Stephan F Huckemann and Benjamin Eltzner. Data analysis on nonstandard spaces. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(3):e1526, 2021.
- [123] Rachael A Hughes, Jon Heron, Jonathan A C Sterne, and Kate Tilling. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *International Journal of Epidemiology*, 48(4):1294–1304, 03 2019.

- [124] Alan Julian Izenman. Review papers: Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, 86(413):205–224, 1991.
- [125] Nicholas C Jacobson, Berta Summers, and Sabine Wilhelm. Digital biomarkers of social anxiety severity: digital phenotyping using passive smartphone sensors. *Journal of Medical Internet Research*, 22(5):e16875, 2020.
- [126] Arnold Janssen, Thorsten Pauls, et al. How do bootstrap and permutation tests work? *The Annals of Statistics*, 31(3):768–806, 2003.
- [127] María-Dolores Jiménez-Gamero, MV Alba-Fernández, and Francisco Javier Ariza-López. Approximating the null distribution of a class of statistics for testing independence. *Journal of Computational and Applied Mathematics*, 354:131–143, 2019.
- [128] Wittawat Jitkrittum, Zoltán Szabó, Kacper Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In *NeurIPS*, 2016.
- [129] Dinesh John, Qu Tang, Fahd Albinali, and Stephen Intille. An open-source monitor-independent movement summary for accelerometer data processing. *Journal for the measurement of physical behaviour*, 2(4):268–281, 2019.
- [130] Clifford Leroy Johnson, Sylvia M Dohrmann, Vicki L Burt, and Leyla Kheradmand Mohadjer. *National health and nutrition examination survey: sample design, 2011-2014*. US Department of Health and Human Services, 2014.
- [131] Lucy Johnston, Gonglei Wang, Kunhui Hu, Chungeng Qian, and Guozhen Liu. Advances in biosensors for continuous glucose monitoring towards wearables. *Frontiers in Bioengineering and Biotechnology*, 9, 2021.
- [132] Joseph DY Kang, Joseph L Schafer, et al. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539, 2007.
- [133] Eric S Kilpatrick. Glycated haemoglobin in the year 2000. *Journal of Clinical Pathology*, 53(5):335–339, 2000.
- [134] Leslie Kish. *Survey sampling*. Wiley Classics Library, 1965.
- [135] David C Klonoff. Continuous glucose monitoring: roadmap for 21st century diabetes therapy. *Diabetes Care*, 28(5):1231–1239, 2005.
- [136] Frank Konietzschke, Solomon W Harrar, Katharina Lange, and Edgar Brunner. Ranking procedures for matched pairs with missing data—asymptotic theory and a small sample approximation. *Computational Statistics & Data Analysis*, 56(5):1090–1102, 2012.

- [137] Vladimir S Korolyuk and Yu V Borovskich. *Theory of U-statistics*, volume 273. Springer Science & Business Media, 2013.
- [138] Michael R Kosorok and Eric B Laber. Precision medicine. *Annual Review of Statistics and Its Application*, 6:263–286, 2019.
- [139] Michael R Kosorok and Erica EM Moodie. *Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine*. SIAM, 2015.
- [140] Lampros C Kourtis, Oliver B Regele, Justin M Wright, and Graham B Jones. Digital biomarkers for Alzheimer’s disease: the mobile/wearable devices opportunity. *NPJ Digital Medicine*, 2(1):1–9, 2019.
- [141] Boris P Kovatchev. Metrics for glycaemic control—from HbA1c to continuous glucose monitoring. *Nature Reviews Endocrinology*, 13(7):425–436, 2017.
- [142] Boris P. Kovatchev, Marc Breton, Chiara Dalla Man, and Claudio Cobelli. In silico preclinical trials: A proof of concept in closed-loop control of type 1 diabetes. *Journal of Diabetes Science and Technology*, 3(1):44–55, 2009. PMID: 19444330.
- [143] Pen Fei Kuan and Bo Huang. A simple and robust method for partially matched samples using the p-values pooling approach. *Statistics in Medicine*, 32:3247–3259, 2013.
- [144] Joseph C Kvedar, Alexander L Fogel, Eric Elenko, and Daphne Zohar. Digital medicine’s march on chronic disease. *Nature Biotechnology*, 34(3):239, 2016.
- [145] Jun Yang Lee and Shaun Wen Huey Lee. Telemedicine cost–effectiveness for diabetes management: a systematic review. *Diabetes Technology & Therapeutics*, 20(7):492–500, 2018.
- [146] Jing Lei, Max G’Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.
- [147] Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B*, 83(5):911–938, 2021.
- [148] King Sun Leong and John P Wilding. Obesity and diabetes. *Best Practice & Research Clinical Endocrinology & Metabolism*, 13(2):221–237, 1999.
- [149] Andrew Leroux, Junrui Di, Ekaterina Smirnova, Elizabeth J MCGuffey, Quy Cao, Elham Bayatmokhtari, Lucia Tabacu, Vadim Zipunnikov, Jacek K Urbanek, and Ciprian Crainiceanu. Organizing and analyzing the activity data in NHANES. *Statistics in Biosciences*, 11(2):262–287, 2019.

- [150] Andrew Leroux, Shiyao Xu, Prosenjit Kundu, John Muschelli, Ekaterina Smirnova, Nilanjan Chatterjee, and Ciprian Crainiceanu. Quantifying the predictive performance of objectively measured physical activity on mortality in the UK Biobank. *The Journals of Gerontology: Series A*, 2020.
- [151] Anne Leucht and Michael H Neumann. Dependent wild bootstrap for degenerate U- and V-statistics. *Journal of Multivariate Analysis*, 117:257–280, 2013.
- [152] Bing Li. Linear operator-based statistical analysis: A useful paradigm for big data. *Canadian Journal of Statistics*, 46(1):79–103, 2018.
- [153] Xiao Li, Jessilyn Dunn, Denis Salins, Gao Zhou, Wenyu Zhou, Sophia Miryam Schüssler-Fiorenza Rose, Dalia Perelman, Elizabeth Colbert, Ryan Runge, Shannon Rego, et al. Digital health: tracking physiomes and activity using wearable biosensors reveals useful health-related information. *PLoS Biology*, 15(1):e2001402, 2017.
- [154] Tengyuan Liang, Alexander Rakhlin, et al. Just interpolate: Kernel “ridgeless” regression can generalize. *Annals of Statistics*, 48(3):1329–1347, 2020.
- [155] Tong Lin and Hongbin Zha. Riemannian manifold learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):796–809, 2008.
- [156] Zhenhua Lin, Dehan Kong, and Linbo Wang. Causal inference on distribution functions. *arXiv preprint arXiv:2101.01599*, 2021.
- [157] Roderick J. Little, Ralph D’Agostino, Michael L. Cohen, Kay Dickersin, Scott S. Emerson, John T. Farrar, et al. The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367(14):1355–1360, 2012.
- [158] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.
- [159] Tiantian Liu, Yair Goldberg, et al. Kernel machines with missing responses. *Electronic Journal of Statistics*, 14(2):3766–3820, 2020.
- [160] Jingyi Lu, Xiaojing Ma, Jian Zhou, Lei Zhang, Yifei Mo, Lingwen Ying, Wei Lu, et al. Association of time in range, as assessed by continuous glucose monitoring, with diabetic retinopathy in type 2 diabetes. *Diabetes Care*, 41(11):2370–2376, 2018.
- [161] Jingyi Lu, Chunfang Wang, Yun Shen, Lei Chen, Lei Zhang, Jinghao Cai, Wei Lu, et al. Time in range in relation to all-cause and cardiovascular mortality in patients with type 2 diabetes: A prospective cohort study. *Diabetes Care*, 2020.

- [162] Daniel J. Lockett, Eric B. Laber, Anna R. Kahkoska, David M. Maahs, Elizabeth Mayer-Davis, and Michael R. Kosorok. Estimating dynamic treatment regimes in mobile health using V-learning. *Journal of the American Statistical Association*, 115(530):692–706, 2020. PMID: 32952236.
- [163] Amy Luke, Lara R Dugas, Ramon A Durazo-Arvizu, Guichan Cao, and Richard S Cooper. Assessing physical activity and its relationship to cardiovascular risk factors: Nhanes 2003-2006. *BMC Public Health*, 11(1):1–11, 2011.
- [164] Thomas Lumley. *Complex surveys: a guide to analysis using R*, volume 565. John Wiley & Sons, 2011.
- [165] Thomas Lumley and Alastair Scott. Fitting regression models to survey data. *Statistical Science*, pages 265–278, 2017.
- [166] Brigid M Lynch, David W Dunstan, Genevieve N Healy, Elisabeth Winkler, Elizabeth Eakin, and Neville Owen. Objectively measured physical activity and sedentary time of breast cancer survivors, and associations with adiposity: findings from NHANES (2003–2006). *Cancer Causes & Control*, 21(2):283–288, 2010.
- [167] Russell Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3284–3305, 2013.
- [168] Russell Lyons. Strong negative type in spheres. *Pacific Journal of Mathematics*, 307(2):383–390, 2020.
- [169] Xinwei Ma and Jingshen Wang. Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, 115(532):1851–1860, 2020.
- [170] K Makrilakis, S Liatis, S Grammatikou, D Perrea, C Stathi, P Tsiligros, and N Katsilambros. Validation of the finnish diabetes risk score (FINDRISC) questionnaire for screening for undiagnosed type 2 diabetes, dysglycaemia and the metabolic syndrome in greece. *Diabetes & Metabolism*, 37(2):144–151, 2011.
- [171] Kanti V Mardia, Charles C Taylor, and Ganesh K Subramaniam. Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics*, 63(2):505–512, 2007.
- [172] J Steve Marron and Andrés M Alonso. Overview of object oriented data analysis. *Biometrical Journal*, 56(5):732–753, 2014.
- [173] Pablo Martínez-Camblor, Norberto Corral, and Jesus María de la Hera. Hypothesis test for paired samples in the presence of missing data. *Journal of Applied Statistics*, 40(1):76–87, 2013.

- [174] Marcos Matabuena, Paulo Félix, Ziad Akram Ali Hammouri, Jorge Mota, and Borja del Pozo Cruz. Physical activity phenotypes and mortality in older adults: a novel distributional data analysis of accelerometry in the NHANES. *Aging Clinical and Experimental Research*, Oct 2022.
- [175] Marcos Matabuena, Paulo Félix, Carlos García-Meixide, and Francisco Gude. Kernel machine learning methods to handle missing responses with complex predictors. application in modelling five-year glucose changes using distributional representations. *Computer Methods and Programs in Biomedicine*, 221:106905, 2022.
- [176] Marcos Matabuena, Philip R Hayes, and Luis Puente-Maestu. Prediction of maximal oxygen uptake from submaximal exercise testing in chronic respiratory patients. new perspectives. *Archivos de Bronconeumología*, 55(10):507, 2019.
- [177] Marcos Matabuena and Alex Petersen. Distributional data analysis with accelerometer data in a NHANES database with nonparametric survey regression models. *ArXiv Preprint*, 2021.
- [178] Marcos Matabuena, Alexander Petersen, Juan C Vidal, and Francisco Gude. Gluco-densities: a new representation of glucose profiles using distributional data analysis. *Statistical Methods in Medical Research*, 30(6):1445–1464, 2021.
- [179] Marcos Matabuena and Rosana Rodríguez-López. An improved version of the classical Banister model to predict changes in physical condition. *Bulletin of Mathematical Biology*, 81(6):1867–1884, 2019.
- [180] Marcos Matabuena, JC Vidal, Oscar Hernan Madrid Padilla, and Dino Sejdinovic. Kernel biclustering algorithm in Hilbert spaces. *ArXiv Preprint*, 2022.
- [181] Marcos Matabuena, Juan C Vidal, Philip R Hayes, and Fernando Huelin Trillo. A 6-minute sub-maximal run test to predict VO2 max. *Journal of Sports Sciences*, 36(22):2531–2536, 2018.
- [182] DR Matthews, JP Hosker, AS Rudenski, BA Naylor, DF Treacher, and RC Turner. Homeostasis model assessment: insulin resistance and β -cell function from fasting plasma glucose and insulin concentrations in man. *Diabetologia*, 28(7):412–419, 1985.
- [183] C.M. McDonnell, S.M. Donath, S.I. Vidmar, G.A. Werther, and F.J. Cameron. A novel approach to continuous glucose analysis utilizing glycemic variation. *Diabetes Technology & Therapeutics*, 7(2):253–263, 2005. PMID: 15857227.

- [184] Erin I McDonnell, Vadim Zipunnikov, Jennifer A Schrack, Jeff Goldsmith, and Julia Wrobel. Registration of 24-hour accelerometric rest-activity profiles and its application to human chronotypes. *Biological Rhythm Research*, pages 1–21, 2021.
- [185] Romeu Mendes, Nelson Sousa, António Almeida, Paulo Subtil, Fernando Guedes-Marques, Victor Machado Reis, et al. Exercise prescription for patients with type 2 diabetes—a synthesis of international recommendations: narrative review. *British Journal of Sports Medicine*, 50(22):1379–1381, 2016.
- [186] Samuel Meyler, Lindsay Bottoms, and Daniel Muniz-Pumares. Biological and methodological factors affecting response variability to endurance training and the influence of exercise intensity prescription. *Experimental Physiology*, 106(7):1410–1424, 2021.
- [187] Jairo H Migueles, Eivind Aadland, Lars Bo Andersen, Jan Christian Brønd, Sebastien F Chastin, Bjarne H Hansen, et al. Granada consensus on analytical approaches to assess associations with accelerometer-determined physical behaviours (physical activity, sedentary behaviour and sleep) in epidemiological studies. *British Journal of Sports Medicine*, 56(7):376–384, 2022.
- [188] GD Molnar, WF Taylor, and MM Ho. Day-to-day variation of continuously monitored glycaemia: a further measure of diabetic instability. *Diabetologia*, 8(5):342–348, 1972.
- [189] Louis Monnier, Claude Colette, and David R Owens. Glycemic variability: the third component of the dysglycemia in diabetes. is it important? how to measure it? *Journal of diabetes science and technology*, 2(6):1094–1100, 2008.
- [190] Louis Monnier, Claude Colette, and David R Owens. Glycemic variability: the third component of the dysglycemia in diabetes. is it important? how to measure it? *Journal of diabetes science and technology*, 2(6):1094–1100, 2008.
- [191] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.
- [192] Kristin Mühlenbruch, Rebecca Paprott, Hans-Georg Joost, Heiner Boeing, Christin Heidemann, and Matthias B Schulze. Derivation and external validation of a clinical version of the german diabetes risk score (GDRS) including measures of HbA1c. *BMJ Open Diabetes Research and Care*, 6(1):e000524, 2018.
- [193] Hans-Georg Müller and Alexander Petersen. Density estimation including examples. *Wiley StatsRef: Statistics Reference Online*, pages 1–12, 2014.

- [194] Melissa A Napolitano, Kelley E Borradaile, Beth A Lewis, Jessica A Whiteley, Jaime L Longval, Alfred F Parisi, Albrecht, et al. Accelerometer use in a physical activity intervention trial. *Contemporary Clinical Trials*, 31(6):514–523, 2010.
- [195] Anthony J Nastasi, Alka Ahuja, Vadim Zipunnikov, Eleanor M Simonsick, Luigi Ferrucci, and Jennifer A Schrack. Objectively measured physical activity and falls in well-functioning older adults: findings from the Baltimore longitudinal study of aging. *American Journal of Physical Medicine & Rehabilitation*, 97(4):255, 2018.
- [196] Aravind Natarajan, Alexandros Pantelopoulos, Hulya Emir-Farinas, and Pradeep Natarajan. Heart rate variability with photoplethysmography in 8 million individuals: a cross-sectional study. *The Lancet Digital Health*, 2(12):e650–e657, 2020.
- [197] DM Nathan, H Turgeon, and S Regan. Relationship between glycated haemoglobin levels and mean glucose levels over time. *Diabetologia*, 50(11):2239–2244, 2007.
- [198] David Nerini and Badih Ghattas. Classifying densities using functional regression trees: Applications in Oceanology. *Computational Statistics & Data Analysis*, 51(10):4984 – 4993, 2007.
- [199] Michelle Nguyen, Julia Han, Elias K. Spanakis, Boris P. Kovatchev, and David C. Klonoff. A review of continuous glucose monitoring-based composite metrics for glycemic control. *Diabetes Technology & Therapeutics*, 22(8):613–622, 2020.
- [200] Scott P. Nichols, Ahyeon Koh, Wesley L. Storm, Jae Ho Shin, and Mark H. Schoenfisch. Biocompatible materials for continuous glucose monitoring devices. *Chemical Reviews*, 113(4):2528–2549, Apr 2013.
- [201] Tom MW Nye, Xiaoxian Tang, Grady Weyenberg, and Ruriko Yoshida. Principal component analysis and the locus of the Fréchet mean in the space of phylogenetic trees. *Biometrika*, 104(4):901–922, 2017.
- [202] All of Us Research Program Investigators. The “All of Us” research program. *New England Journal of Medicine*, 381(7):668–676, 2019.
- [203] World Health Organization. *Global report on diabetes*. World Health Organization, 2016.
- [204] Subhadip Pal and Jeremy Gaskins. Modified Pólya-Gamma data augmentation for Bayesian analysis of directional data. *Journal of Statistical Computation and Simulation*, pages 1–22, 2022.

- [205] Victor M Panaretos and Yoav Zemel. Statistical aspects of Wasserstein distances. *ArXiv Preprint*, 2018.
- [206] Theodosios Pavlidis. A review of algorithms for shape analysis. *Computer Graphics and Image Processing*, 7(2):243–258, 1978.
- [207] Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. *Modeling and analysis of compositional data*. John Wiley & Sons, 2015.
- [208] Bente Klarlund Pedersen and B Saltin. Evidence for prescribing exercise as therapy in chronic disease. *Scandinavian Journal of Medicine & Science in Sports*, 16(S1):3–63, 2006.
- [209] Margaret Sullivan Pepe, Holly Janes, Gary Longton, Wendy Leisenring, and Polly Newcomb. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology*, 159(9):882–890, 2004.
- [210] Neil J Perkins, Stephen R Cole, Ofer Harel, Eric J Tchetgen Tchetgen, BaoLuo Sun, Emily M Mitchell, et al. Principled Approaches to Missing Data in Epidemiologic Studies. *American Journal of Epidemiology*, 187(3):568–575, 11 2017.
- [211] Usama Pervaiz, Diego Vidaurre, Mark W Woolrich, and Stephen M Smith. Optimising network modelling methods for fMRI. *Neuroimage*, 211:116604, 2020.
- [212] AL Peters, AJ Ahmann, T Battelino, A Evert, IB Hirsch, MH Murad, WE Winter, and H Wolpert. Diabetes technology-continuous subcutaneous insulin infusion therapy and continuous glucose monitoring in adults: An endocrine society clinical practice guideline. *The Journal of Clinical Endocrinology and Metabolism*, 101(11):3922–3937, 2016.
- [213] Alexander Petersen, Xi Liu, and Afshin A. Divani. Wasserstein F -tests and confidence bands for the Fréchet regression of density response curves. *The Annals of Statistics*, 49(1):590 – 611, 2021.
- [214] Alexander Petersen and Hans-Georg Müller. Fréchet regression for random objects with Euclidean predictors. *The Annals of Statistics*, 47(2):691–719, 2019.
- [215] Alexander Petersen and Hans-Georg Müller. Functional data analysis for density functions by transformation to a Hilbert space. *Annals of Statistics*, 44(1):183–218, 02 2016.
- [216] Alexander Petersen and Hans-Georg Müller. Fréchet regression for random objects with Euclidean predictors. *Annals of Statistics*, 47(2):691–719, 04 2019.

- [217] Arthur Pewsey and Eduardo García-Portugués. Recent advances in directional statistics. *Test*, 30(1):1–58, 2021.
- [218] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.
- [219] Dimitris N Politis and Joseph P Romano. Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, pages 2031–2050, 1994.
- [220] David Pollard. *Empirical processes: theory and applications*. Institute of Mathematical Statistics, 1990.
- [221] Qianya Qi, Li Yan, and Lili Tian. Testing equality of means in partially paired data with incompleteness in single response. *Statistical Methods in Medical Research*, 28:1508–1522, 2019.
- [222] Rui Qiu, Zhou Yu, and Ruoqing Zhu. Random forests weighted local Fréchet regression with theoretical guarantees. *ArXiv Preprint*, 2022.
- [223] Sophia Rabe-Hesketh and Anders Skrondal. Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4):805–827, 2006.
- [224] Ali Rafei, Carol A C Flannagan, and Michael R Elliott. Big Data for Finite Population Inference: Applying Quasi-Random Approaches to Naturalistic Driving Data Using Bayesian Additive Regression Trees. *Journal of Survey Statistics and Methodology*, 8(1):148–180, 02 2020.
- [225] David A Raichlen, Herman Pontzer, Theodore W Zderic, Jacob A Harris, Audax ZP Mabulla, Marc T Hamilton, et al. Sitting, squatting, and the evolutionary biology of human inactivity. *Proceedings of the National Academy of Sciences*, 117(13):7115–7121, 2020.
- [226] J. Ramsay, J. Ramsay, and B.W Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, 2005.
- [227] Jesús D. Arroyo Relión, Daniel Kessler, Elizaveta Levina, and Stephan F. Taylor. Network classification with applications to brain connectomics. *The Annals of Applied Statistics*, 13(3):1648 – 1677, 2019.
- [228] Maria L Rizzo and Gábor J Székely. Disco analysis: A nonparametric extension of analysis of variance. *The Annals of Applied Statistics*, 4(2):1034–1055, 2010.

- [229] David Rodbard. Continuous glucose monitoring: a review of successes, challenges, and opportunities. *Diabetes Technology & Therapeutics*, 18(S2):S2–3, 2016.
- [230] David Rodbard. Glucose variability: a review of clinical applications and research developments. *Diabetes Technology & Therapeutics*, 20(S2):S2–5, 2018.
- [231] Yves Rolland, Gabor Abellan van Kan, and Bruno Vellas. Physical activity and Alzheimer's disease: from prevention to therapeutic perspectives. *Journal of the American Medical Directors Association*, 9(6):390–405, 2008.
- [232] Sherri Rose and Dimitris Rizopoulos. Machine learning for causal inference in Biostatistics. *Biostatistics*, 21(2):336–338, 2019.
- [233] Robert Ross, Bret H Goodpaster, Lauren G Koch, Mark A Sarzynski, Wendy M Kohrt, Neil M Johannsen, James S Skinner, Alex Castro, Brian A Irving, Robert C Noland, et al. Precision exercise medicine: understanding exercise response variability. *British Journal of Sports Medicine*, 53(18):1141–1153, 2019.
- [234] Pouya Saeedi, Inga Petersohn, Paraskevi Salpea, Belma Malanda, Suvi Karuranga, Nigel Unwin, Stephen Colagiuri, Leonor Guariguata, Ayesha A Motala, Katherine Ogurtsova, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes Research and Clinical Practice*, 157:107843, 2019.
- [235] Hani M Samawi, Amal Helu, and Robert Vogel. A nonparametric test of symmetry based on the overlapping coefficient. *Journal of Applied Statistics*, 38(5):885–898, 2011.
- [236] Hani M Samawi and Robert Vogel. Notes on two sample tests for partially correlated (paired) data. *Journal of Applied Statistics*, 41(1):109–117, 2014.
- [237] Matteo C Sattler, Johannes Jaunig, Christoph Tösch, Estelle D Watson, Lidwine B Mokkink, Pavel Dietz, et al. Current evidence of measurement properties of physical activity questionnaires for older adults: An updated systematic review. *Sports Medicine*, 50(7):1271–1315, 2020.
- [238] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International Conference on Computational Learning Theory*, pages 416–426. Springer, 2001.
- [239] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press, 2001.

- [240] Nicholas J Schork. Personalized medicine: time for one-person trials. *Nature*, 520(7549):609–611, 2015.
- [241] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.
- [242] Elizabeth Selvin, Ciprian M Crainiceanu, Frederick L Brancati, and Josef Coresh. Short-term variability in measures of glycemia and implications for the classification of diabetes. *Archives of Internal Medicine*, 167(14):1545–1551, 2007.
- [243] Stephen Senn. Mastering variation: variance components and personalised medicine. *Statistics in Medicine*, 35(7):966–977, 2016.
- [244] Robert J Serfling. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.
- [245] F. John Service. Glucose variability. *Diabetes*, 62(5):1398–1404, 2013.
- [246] F John Service, George D Molnar, John W Rosevear, Eugene Ackerman, Lael C Gatewood, and William F Taylor. Mean amplitude of glycemic excursions, a measure of diabetic instability. *Diabetes*, 19(9):644–655, 1970.
- [247] Cencheng Shen and Joshua T Vogelstein. The exact equivalence of distance and kernel methods in hypothesis testing. *AStA Advances in Statistical Analysis*, 105(3):385–403, 2021.
- [248] Lauren B Sherar, Pippa Griew, Dale W Esliger, Ashley R Cooper, Ulf Ekelund, Ken Judge, and Chris Riddoch. International children’s accelerometry database (ICAD): design and methods. *BMC Public Health*, 11(1):1–13, 2011.
- [249] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986.
- [250] Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.
- [251] RBAM Singh, Anne Barden, Trevor Mori, and Lawrence Beilin. Advanced glycation end-products: a review. *Diabetologia*, 44(2):129–146, 2001.
- [252] John R Sirard and Russell R Pate. Physical activity assessment in children and adolescents. *Sports Medicine*, 31(6):439–454, 2001.

- [253] Jan Škrha, Jan Šoupal, and Martin Prázný. Glucose variability, HbA1c and microvascular complications. *Reviews in Endocrine and Metabolic Disorders*, 17(1):103–110, 2016.
- [254] Ekaterina Smirnova, Andrew Leroux, Quy Cao, Lucia Tabacu, Vadim Zipunnikov, Ciprian Crainiceanu, et al. The predictive performance of objective measures of physical activity derived from accelerometry data for 5-year all-cause mortality in older adults: National health and nutritional examination survey 2003–2006. *The Journals of Gerontology: Series A*, 2019.
- [255] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.
- [256] Eberhard Standl, Kamlesh Khunti, Tina Birgitte Hansen, and Oliver Schnell. The global epidemics of diabetes in the 21st century: Current situation and perspectives. *European Journal of Preventive Cardiology*, 26(2_suppl):7–14, 2019.
- [257] Florian Steinke, Matthias Hein, and Bernhard Schölkopf. Nonparametric regression between general Riemannian manifolds. *SIAM Journal on Imaging Sciences*, 3:527–563, 2010.
- [258] Tessa Strain, Katrien Wijndaele, Paddy C Dempsey, Stephen J Sharp, Matthew Pearce, Justin Jeon, et al. Wearable-device-measured physical activity and future health risk. *Nature Medicine*, 26(9):1385–1391, 2020.
- [259] David J Strauss. A model for clustering. *Biometrika*, 62(2):467–475, 1975.
- [260] Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.
- [261] Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.
- [262] Gábor J Székely, Maria L Rizzo, et al. Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272, 2004.
- [263] Gábor J. Székely and Maria L. Rizzo. The energy of data. *Annual Review of Statistics and Its Application*, 4(1):447–479, 2017.
- [264] Sylvain Takerkart, Guillaume Auzias, Bertrand Thirion, and Liva Ralaivola. Graph-based inter-subject pattern analysis of fMRI data. *PLoS ONE*, 9(8):e104586, 2014.

- [265] R Talská, Alessandra Menafoglio, Jitka Machalová, Karel Hron, and E Fiserov. Compositional regression with functional response. *Computational Statistics & Data Analysis*, 123:66–85, 2018.
- [266] Ian F Tannock, John A Hickman, et al. Limits to personalized cancer medicine. *New England Journal of Medicine*, 375(13):1289–1294, 2016.
- [267] Jakob Tarp, Børge Herman Hansen, Morten Wang Fagerland, Jostein Steene-Johannessen, Sigmund Alfred Anderssen, and Ulf Ekelund. Accelerometer-measured physical activity and sedentary time in a cohort of us adults followed for up to 13 years: the influence of removing early follow-up on associations with mortality. *International Journal of Behavioral Nutrition and Physical Activity*, 17(1):1–8, 2020.
- [268] Roy Taylor. Type 2 diabetes. *Diabetes Care*, 36(4):1047–1055, 2013.
- [269] Eric J Topol. Transforming medicine via digital innovation. *Science Translational Medicine*, 2(16):16cm4, 2010.
- [270] Arturo Tozzi, James F Peters, and Norbert Jaušovec. EEG dynamics on hyperbolic manifolds. *Neuroscience Letters*, 683:138–143, 2018.
- [271] Richard P Troiano, David Berrigan, Kevin W Dodd, Louise C Masse, Timothy Tilert, Margaret McDowell, et al. Physical activity in the united states measured by accelerometer. *Medicine and science in sports and exercise*, 40(1):181, 2008.
- [272] Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.
- [273] Anastasios A Tsiatis. *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. CRC press, 2019.
- [274] Jiaobing Tu, Rebeca M Torrente-Rodríguez, Minqiang Wang, and Wei Gao. The era of digital health: A review of portable and wearable affinity biosensors. *Advanced Functional Materials*, 30(29):1906713, 2020.
- [275] Danielle C. Tucker, Yichao Wu, and Hans-Georg Müller. Variable selection for global Fréchet regression. *Journal of the American Statistical Association*, 0(0):1–15, 2021.
- [276] Rui Tuo, Yan Wang, and CF Jeff Wu. On the improved rates of convergence for Matérn-type kernel ridge regression with application to calibration of computer models. *SIAM/ASA Journal on Uncertainty Quantification*, 8(4):1522–1547, 2020.
- [277] Sara van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

- [278] Sara A Van de Geer. *Applications of empirical process theory*, volume 91. Cambridge University Press Cambridge, 2000.
- [279] Karl Gerald Van den Boogaart, Juan José Egozcue, and Vera Pawlowsky-Glahn. Bayes Hilbert spaces. *Australian & New Zealand Journal of Statistics*, 56(2):171–194, 2014.
- [280] Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak Convergence and Empirical Processes*, pages 16–28. Springer, 1996.
- [281] Geert Verbeke and Geert Molenberghs. *Linear mixed models for longitudinal data*. Springer New York, NY, 2009.
- [282] Karel Vermeulen and Stijn Vansteelandt. Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110(511):1024–1036, 2015.
- [283] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Berlin, Heidelberg, 2009.
- [284] Vladimir Vovk. Kernel ridge regression. In *Empirical Inference*, pages 105–116. Springer, 2013.
- [285] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.
- [286] Matt P Wand and M Chris Jones. *Kernel smoothing*. CRC press, 1994.
- [287] Haonan Wang and JS Marron. Object oriented data analysis: Sets of trees. *The Annals of Statistics*, 35(5):1849–1873, 2007.
- [288] Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016.
- [289] Zeyi Wang, Haris I Sair, Ciprian Crainiceanu, Martin Lindquist, Bennett A Landman, Susan Resnick, et al. On statistical tests of functional connectome fingerprinting. *Canadian Journal of Statistics*, 49(1):63–88, 2021.
- [290] Yang Wu, Haofei Hu, Jinlin Cai, Runtian Chen, Xin Zuo, Heng Cheng, et al. Machine learning for predicting the 3-year risk of incident diabetes in Chinese adults. *Frontiers in Public Health*, 9, 2021.
- [291] Jin Xu and Solomon W Harrar. Accurate mean comparisons for paired samples with missing data: An application to a smoking-cessation trial. *Biometrical journal*, 54(2):281–295, 2012.

- [292] Lei Yang, Shaogao Lv, and Junhui Wang. Model-free variable selection in reproducing kernel Hilbert space. *The Journal of Machine Learning Research*, 17(1):2885–2908, 2016.
- [293] Ruriko Yoshida. Tropical data science over the space of phylogenetic trees. In *Proceedings of SAI Intelligent Systems Conference*, pages 340–361. Springer, 2021.
- [294] Donghyeon Yu, Johan Lim, Feng Liang, Kyunga Kim, Byung Soo Kim, and Woncheol Jang. Permutation test for incomplete paired data with application to cDNA microarray data. *Computational Statistics & Data Analysis*, 56(3):510–521, 2012.
- [295] Francesco Zaccardi and Kamlesh Khunti. Glucose dysregulation phenotypes — time to improve outcomes. *Nature Reviews Endocrinology*, 14(11):632–633, Nov 2018.
- [296] Aleksandr Zaitcev, Mohammad R Eissa, Zheng Hui, Tim Good, Jackie Elliott, and Mohammed Benaissa. A deep neural network application for improved prediction of HbA1c in Type 1 diabetes. *IEEE Journal of Biomedical and Health Informatics*, 2020.
- [297] David Zeevi, Tal Korem, Niv Zmora, David Israeli, Daphna Rothschild, Adina Weinberger, et al. Personalized nutrition by prediction of glycemic responses. *Cell*, 163(5):1079–1094, 2015.
- [298] Yufan Zhao, Donglin Zeng, Mark A Socinski, and Michael R Kosorok. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4):1422–1433, 2011.
- [299] Yan Zheng, Sylvia H Ley, and Frank B Hu. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nature Reviews Endocrinology*, 14(2):88–98, 2018.
- [300] Wenyu Zhou, M Reza Sailani, Kévin Contrepois, Yanjiao Zhou, Sara Ahadi, Shana R Leopold, et al. Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature*, 569(7758):663–671, 2019.

Publications

The contents of this thesis are now in the process of being published. By the time being, there are three published article.

- **Title:** Glucodensities: A new representation of glucose profiles using distributional data analysis.

Year: 2021.

Author 1: Marcos Matabuena, Universidade de Santiago de Compostela.

Author 2: Alexander Petersen, Department of Statistics, Brigham Young University, Provo, UT, USA.

Author 3: Juan Vidal, Universidade de Santiago de Compostela.

Author 4: Francisco Gude, Unidad de Epidemiología Clínica, Hospital Clínico Universitario de Santiago de Compostela, Santiago de Compostela, Spain.

Reference: Statistics Methods in Medical Research. DOI: <https://doi.org/10.1177/0962280221998064>.

Editorial: SAGE.

Quality indexes: Impact factor: 2.494 (2021) Q1. Five year impact factor: 3.234 (2021). Q1

Journal authorisation: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (see for further details <https://doi.org/10.1177/0962280221998064>).

Corresponding chapter: Chapter 3.

Contributions of the Ph.D. candidate: Main author and the main contributor. MM conceptualizes the manuscript, develops the methods, obtains the results, writes the early draft, and reviews the final version.

- **Title:** Kernel machine learning methods to handle missing responses with complex predictors. Application in modelling five-year glucose changes using distributional representations

Year: 2022.

Author 1: Marcos Matabuena, Universidade de Santiago de Compostela.

Author 2: Paulo Félix, Universidade de Santiago de Compostela.

Author 3: Carlos Meixide, ETH Zürich.

Author 4: Francisco Gude, Unidad de Epidemiología Clínica, Hospital Clínico Universitario de Santiago de Compostela, Santiago de Compostela, Spain.

Reference: Computer Methods and Programs in Biomedicine
DOI: <https://doi.org/10.1016/j.cmpb.2022.106905>.

Editorial: Elsevier.

Quality indexes: Impact factor: 7.521 (2021) Q1. Five year impact factor: 4.284 (2021). Q1

Journal authorisation: This is an open access article distributed under the terms of the Creative Commons CC-BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (see for further details <https://doi.org/10.1016/j.cmpb.2022.106905>).

Corresponding chapter: Chapter 5.

Contributions of the Ph.D. candidate: Main author and the main contributor. MM conceptualizes the manuscript, develops the methods, obtains the results, writes the early draft, and reviews the final version.

- **Title:** Physical activity phenotypes and mortality in older adults: a novel distributional data analysis of accelerometry in the NHANES

Year: 2022.

Author 1: Marcos Matabuena, Universidade de Santiago de Compostela.

Author 2: Paulo Félix, Universidade de Santiago de Compostela.

Author 3: Ziad Akram Ali Hammouri, Universidade de Santiago de Compostela.

Author 4: Jorge Mota. Universidade de Oporto.

Author 4: Borja del Pozo Cruz University of Southern Denmark, Odense, Denmark

Reference: Aging Clinical and Experimental Research DOI: <https://doi.org/10.1007/s40520-022-02260-3>

Editorial: Springer.

Quality indexes: Impact factor: 4.481 (2021) Q1. Five year impact factor: 4.075 (2021) Q2.

Journal authorisation: Open Access, This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made (see for further details <https://doi.org/10.1007/s40520-022-02260-3>).

Corresponding chapter: Chapter 4.

Contributions of the Ph.D. candidate: Main author and the main contributor. MM conceptualizes the manuscript, develops the methods, obtains the results, writes the early draft, and reviews the final version.

List of Figures

Fig. 2.1	Laplacian conditional matrix mean estimation from a person with schizophrenia.	29
Fig. 2.2	The three component blood vessel trees from a given patient.	30
Fig. 3.1	Glucodensities are estimated from a random sample of the AEGIS study with diabetic and normoglycemic patients. For each patient, this glucose representation estimates the proportion of time spent at each glucose concentration over a continuum, representing a more sophisticated approach to assess glucose metabolism.	47
Fig. 3.2	Real values vs estimated values when glucodensity is predictor.	56
Fig. 3.3	Residuals in quantile space.	57
Fig. 3.4	Real values vs. estimated values when time-in-range metric is the predictor. (Blue) time-in-range metric with cut-offs calculated with normoglycemic patients of AEGIS database. (Red) time-in-range metric using the cut-offs suggested by ADA.	58
Fig. 3.5	(Left two panels) Glucodensities for women and men of the AEGIS study, plotted as quantile functions; (Third panel) 2-Wasserstein mean quantile functions for each group; (Fourth Panel) Cross-sectional standard deviation curves for quantile functions in each group.	62
Fig. 3.6	Clustering analysis of diabetes patients in AEGIS study	63

Fig. 4.1 Example of transforming the raw accelerometer signal distributional profile for a randomly selected individual: (top) Physical activity recordings in real time; (middle) density function for active movement; and (bottom) quantile representation. 72

Fig. 4.2 Summary curves of physical activity distributions in quantile space (mean and standard deviation) between alive and dead patients groups after five years together with the mean of continuous movement representation. 77

Fig. 4.3 Boxplot of estimated probabilities drawn by the models in 5-year mortality prediction according to mortality status. (Left) Nadaraya-Watson distributional representation. (Center) Logistic regression introducing the five scores PCA quantile analysis. (Right) Nadaraya-Watson TAC variable. 84

Fig. 4.4 Survival curves for risk and non-risk groups, according to age stratification. The curves with only lines identify the risk group and, in another case, the non-risk group. 85

Fig. 4.5 Mean and standard deviation of distributional representations for the five phenotypes together with their mortality rate. 91

Fig. 4.6 Kaplan-Meier curves for each phenotype and age group strata. 92

Fig. 5.1 (Left) The 5-day CGM recording from a normoglycemic patient is shown. (Center) Glucodensity designates a distributional representation that estimates the proportion of time the patient spent at each glucose concentration. (Right) Quantile representation. Dotted, solid and dashed lines represent concentrations for 20 percent, 50 percent and 80 percent quantiles, respectively. 100

Fig. 5.2 Marginal dependence relation between examined variables in the AEGIS database. 113

Fig. 5.3 Residuals vs. $A1c_{initial}$ for the model that includes glucodensity as a covariate in the AEGIS database. Red circles correspond to diabetic patients . . . 115

Fig. 5.4 Prediction intervals for each response observed in the AEGIS database (90% confidence level). The red circles correspond to patients with diabetes. . . 116

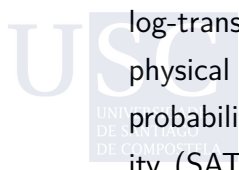
Fig. 5.5 Clinical decision rules that allow us to identify those patients with a significant uncertainty in their $A1c_{5year}$ predictions. 116

Fig. 6.1 Glucodensity changes in prediabetic patients (blue) who develop diabetes after 5 years (red). 121

- Fig. 6.2 Difference between the quantile curves (before and after) in normoglycemic individuals according to body mass status. The dispersion is more significant for the overweight and obesity subgroup, consistent with an increasing glycemic risk. 128
- Fig. 6.3 Resulting clusters are shown. Both quantile curves at the beginning of the study (blue) and five years later (orange) are shown for each cluster. 129

List of Tables

Tab. 3.1	Characteristics of AEGIS study participants with CGM monitoring by sex. Means and standard deviations are shown. A1c: glycated haemoglobin; FPG, fasting plasma glucose; HOMA-IR, homeostasis model assessment-insulin resistance; BMI, body mass index; CONGA, glycemic variability in terms of continuous overall net glycemic action; MAGE, mean amplitude of glycemic excursions; MODD, mean of daily difference.	48
Tab. 3.2	Cut-offs for time-in-range metrics using own estimations through normoglycemic individuals of AEGIS study	50
Tab. 3.3	Cut-offs for time-in-range metrics following ADA guidelines [20]	50
Tab. 3.4	Clinical importance of biomarkers used in the statistical analysis	55
Tab. 3.5	R^2 estimated with time-in-range metrics under consideration	57
Tab. 4.1	Variable summaries for the chosen cohort grouped as survivors/decedents. The reported values are mean (standard deviation) for continuous variables and counts (%) for categorical variables. Total activity count (TAC); total log-transformed activity count (TLAC); total minutes of moderate/vigorous physical activity (MVPA); active to sedentary/sleep/non-wear transition probability (ASTP); sedentary/sleep/non-wear to active transition probability (SATP); Coronary heart disease CHD (CHD); Congestive heart failure (CHF)	73



Tab. 4.2 R-squared was computed for each representation used in kernel ridge-regression models with continuous variables being examined. MVPA is a compositional metric with a cut-off equal to 2020 counts. SB is the proportion of the time that an individual has an energy expenditure lesser than 100 counts. WT is an estimation of the proportion of the time that the individual wear the accelerometer device 83

Tab. 4.3 Summary clinical characteristics of participants in each cluster. In binary variables, we show the rate, and in continuous variables, we show the mean and standard deviation. 93

Tab. 4.4 Hazard ratios and odds ratios (95% confidence interval) of mortality outcomes associated with different physical activity phenotypes (Reference: Group 1 - Inactivity phenotype.) 93

Tab. 4.5 Results of logistics and Cox survey regression model in terms of hazard ratios and odds ratios (95% confidence interval). Reference: Group 1-Inactivity phenotype. 94

Tab. 5.1 Estimated raw p-values of A1c total variation vs each biomarker using the method proposed in Section 5.2.2 with normoglycemic patients. 112

Tab. 6.1 The proportion of simulations rejecting the null hypothesis under MCAR and MAR mechanisms is shown. 126

Tab. 6.2 Clinical baseline characteristics for the individuals belonging to each cluster. Mean and standard deviation are shown. 130

Tab. 6.3 Coefficients obtained from logistic regression. Results from some different model selection criteria for the fitted model are shown. 130





Esta tesis tiene como objetivo proponer nuevas representaciones distribucionales y métodos estadísticos en espacios métricos para modelar de forma eficaz los datos procedentes de la monitorización continua de los pacientes durante las actividades propias de su vida diaria. Proponemos nuevas pruebas de hipótesis para datos emparejados, modelos de regresión, algoritmos de cuantificación de la incertidumbre, pruebas de independencia estadística y algoritmos de análisis de conglomerados para las nuevas representaciones distribucionales y otros objetos estadísticos complejos. Los diferentes resultados recogidos a lo largo de la tesis muestran las ventajas en términos de predicción, interpretabilidad y capacidad de modelización de las nuevas propuestas frente a los métodos existentes.