TESIS DOCTORAL

ANÁLISIS ESTADÍSTICO DE FORMAS 3D CON APLICACIONES ANTROPOMÉTRICAS

Presentada por **Sònia Barahona Albiol**

y dirigida por **Dra. Amelia Simó Vidal**

Dra. Mª Victoria Ibáñez Gual

Junio 2018

Programa de Doctorado en Ciencias

Escuela de Doctorado de la Universitat Jaume I

ANÁLISIS ESTADÍSTICO DE FORMAS 3D CON APLICACIONES ANTROPOMÉTRICAS

Memoria presentada por Sònia Barahona Albiol para

optar al grado de doctora por la Universitat Jaume I

Doctoranda:

Sònia Barahona Albiol

Directoras de la tesis:

Dra. Amelia Simó Vidal

Dra. Mª Victoria Ibáñez Gual

Castellón de la Plana, junio 2018

A mis padres

Esta tesis doctoral ha sido realizada gracias a la financiación de los siguientes proyectos:

- Ayuda pre-docotal para la formación de personal investigador concedida por la UJI. Ref. PREDOC/2014/18.
- Ayuda para la realización de actividades formativas de la Escuela de Doctorado de la UJI (participación en congresos) convocatoria 2017.

Además, el proyecto de investigación

Herramientas para la predicción de la talla y el ajuste de ropa infantil a partir de la reconstrucción 3D del cuerpo y de técnicas 'big data'

en el que se enmarca la tesis, ha sido financiado por el Ministerio de Economía y Competitividad con las ayudas a Proyectos de I+D+i «RETOS INVESTIGACIÓN». Ref. DPI2013-47279-C2-1-R.

Agradecimientos

El presente trabajo se ha realizado en el Departamento de Matemáticas de la Universitat Jaume I de Castellón.

Me siento en la necesidad de agradecer de corazón la dedicación que Ximo Gual, María Victoria Ibáñez y Amelia Simó han puesto en guiarme en este trabajo. Mencionar que aunque Ximo Gual no figura como director de la tesis ha ejercido como tal. Vuestra complicidad, cariño y sencillez ha hecho muy agradable el desarrollo de este proyecto. He aprendido mucho con vosotros y de vosotros y habéis hecho constantemente que despierte mi curiosidad por las Matemáticas. *Moltes gràcies pel vostre temps i implicació.*

Agradecer también a Pablo Centella por su indispensable ayuda en la programación de los modelos teóricos desarrollados. Gracias a Juanjo Nuño por su colaboración en el trabajo relacionado con la Estereología.

Mencionar a los compañeros de proyecto en el que se enmarca este trabajo (investigadores del IBV y de la UV), con quienes hemos afrontado problemas interesantes desde diferentes campos de investigación.

Me siento muy agradecida a todos los miembros del Departamento de Matemáticas de la UJI porque me acogieron con los brazos abiertos y me hicieron sentir como en casa durante mi estancia en el Departamento.

Quiero recordar a aquellos compañeros y profesores con los que disfruté de las matemáticas, desde el instituto hasta la facultad. En particular, quería agradecer a Pablo Galindo su acompañamiento, comprensión y humanidad en tantas explicaciones matemáticas y consejos. Gracias por tu tiempo.

A mi compañero en la vida, Hilario, le agradezco su ilusión y compromiso en todos los proyectos que comenzamos.

Finalmente, muchas gracias a mi familia. Por haberme enseñado, a través del ejemplo, que la felicidad está en poner todo el esfuerzo y dedicación en cada una de las cosas que hacemos en la vida. Gracias también por la "herencia matemática" recibida.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

El presente trabajo de tesis doctoral se ha desarrollado dentro del grupo "Análisis de Imagen Médica y Estereología" del Departamento de Matemáticas de la Universitat Jaume I de Castellón. Como consecuencia, esta memoria consta de dos partes claramente diferenciadas: la primera de ellas, dentro del ámbito del Análisis de Imagen (Capítulos 5, 6, 7), la segunda parte se encuentra ubicada en el marco de la Estereología (Capítulo 8).

Como se ha mencionado, el trabajo recogido en esta memoria se ha llevado a cabo en el seno de un grupo de investigación del Departamento de Matemáticas de la UJI. Este mantiene una estrecha colaboración con otros dos grupos de investigación. El primero está formado por miembros del Departamento de Estadística e Investigación Operativa de la Universidad de Valencia, mientras que el segundo lo componen investigadores del Instituto de Biomecánica de Valencia especializados en tratamiento de imagen y antropometría aplicada a productos de consumo.

Esta asociación se realiza en torno a un proyecto del Ministerio de Economía y Competitividad dentro del marco de Ayudas a Proyectos de I+D+i *«RETOS INVESTIGACIÓN»*. El objetivo principal del proyecto es desarrollar y validar metodologías y herramientas que apoyen la toma de decisiones en los procesos de compra de artículos de moda, en particular, que permitan seleccionar la talla adecuada de prendas de ropa infantil para un niño o una niña sin necesidad de que se pruebe la ropa.

La selección de la talla adecuada constituye un problema en la compra de este tipo de artículos tanto en tienda como, en especial, por Internet. El temor a un mal ajuste es la principal barrera de la compra de ropa por Internet, y el elevado porcentaje de devoluciones por este motivo, uno de los principales costes de este canal de ventas para distribuidores y fabricantes. Un ajuste adecuado de la prenda es fundamental en la selección de talla para

cualquier usuario; pero lo es aún más en el caso de la indumentaria infantil, ya que la talla seleccionada debe permitir al niño o niña la realización de gran cantidad de actividad física a diario y de manera confortable.

Para superar estas barreras y alcanzar el objetivo principal del proyecto se han desarrollado una serie de herramientas que permiten capturar la forma del cuerpo del niño en $3D$ y seleccionar la talla óptima de acuerdo a una predicción del ajuste basada en la forma del cuerpo, el patrón de la prenda y las preferencias del comprador. El desarrollo de dichas herramientas requiere la combinación de métodos y técnicas procedentes de distintas disciplinas (estadística, matemáticas, computación gráfica, minería de datos y biomecánica entre otras).

Por tanto, la primera parte de la presente memoria, tiene como objetivo diseñar un modelo teórico y una metodología que permita afrontar los problemas de asignación y predicción propuestos. Además, durante el estudio, observamos que los sistemas de tallaje actuales para prendas infantiles no proporcionan un buen ajuste a una gran parte de la población, por lo que también nos planteamos proponer nuevos sistemas de tallaje, es decir, queremos resolver un problema de clasificación.

Cada una de estas tres cuestiones se resuelven en los Capítulos 5, 6, 7. Para ello, establecemos un modelo teórico que consiste en caracterizar el contorno de cuerpos geométricos mediante una herramienta matemática llamada *current*. Los *currents* son objetos matemáticos complejos, pero recientemente (Durrleman, 2010) propuso identificar, por dualidad, un subespacio de *currents* con un espacio de Hilbert con prácticas propiedades: un *vector-valued Reproducing Kernel Hilbert Space (RKHS)*. Por ello, tras ciertas asociaciones, podemos caracterizar el contorno de un cuerpo geométrico mediante una función en un espacio RKH. Así, se plantea un nuevo escenario, en el que queremos aplicar técnicas estadísticas (como algoritmos de clasificación o métodos de regresión) a una base de datos formada por funciones en un espacio de Hilbert. La teoría de Análisis de Datos Funcionales (FDA) nos ayuda a manejar estos elementos, que tienen dificultades inherentes por estar en un espacio infinito-dimensional. Tras adaptar los métodos estadísticos al espacio de trabajo, podemos aplicar estos modelos teóricos a la base de datos del proyecto formada por cuerpos de niños y niñas (ver sección 4.2).

A continuación, se indica cuál es la organización de la presente memoria. El Capítulo 2 tiene como objetivo recordar al lector conceptos que serán de utilidad para entender y manejar los desarrollos teóricos posteriores. Está estructurado siguiendo el hilo argumental que existe desde que caracterizamos los cuerpos por *currents*, hasta que utilizamos Análisis de Datos Funcionales para generalizar las técnicas estadísticas utilizadas. El último punto de este

capítulo explica qué es la Estereología y qué son las fórmulas rotacionales y describe con un sencillo ejemplo los pasos que se siguen en la cuarta aportación (Capítulo 8), que es la relacionada con este ámbito. En el Capítulo 3, se detallan todas las asociaciones que se realizan para caracterizar el contorno de un cuerpo geométrico mediante un campo vectorial en un espacio RKH. La base de datos de escáneres de niños y niñas asociada al proyecto se describe en el Capítulo 4, y también se especifican las características de bases de datos sintéticas que se usan para valorar los métodos diseñados. La Parte II es la principal de la memoria, ya que, en ella se presentan las aportaciones desarrolladas en el trabajo de tesis doctoral (Capítulos 5, 6, 7 y 8). El Capítulo 8 contiene el trabajo desarrollado en el ámbito de la Esterología. La Estereología es un conjunto de métodos que permiten la estimación insesgada y eficiente de cantidades de objetos geométricos. En particular, es una herramienta útil para el análisis de imágenes obtenidas de pruebas médicas, se utiliza también en ciencias como la microscopía, biología, etc. Finalmente, aparecen los Capítulos 9 y 10 en los que se extraen conclusiones del trabajo y se indican posibles líneas futuras de investigación.

Parte I

Fundamentos

Capítulo 2

Preliminares

En este capítulo pretendemos hacer un repaso a los conceptos y resultados fundamentales que luego utilizaremos a lo largo de la memoria.

Hemos intentado dar las ideas intuitivas, con esto intentamos no ser formalistas. Para el lector que esté interesado en las definiciones formales daremos las referencias de interés a consultar.

Este capítulo se estructura, como se indica en la introducción, siguiendo el hilo argumental que existe desde la caracterización de los cuerpos por *currents*, hasta la adaptación de las técnicas estadísticas utilizadas. Cada punto corresponde aproximadamente a uno de los pasos a seguir en este proceso y en cada uno de ellos se sientan las bases teóricas necesarias para el trabajo desarrollado.

El último punto del capítulo muestra los conceptos y procedimientos básicos para trabajar con fórmulas rotacionales en Estereología. Esta presentación resulta conveniente como lectura previa al Capítulo 8.

2.1. Análisis de Formas

El Análisis de Formas se basa en el estudio estadístico de la forma, el tamaño de un objeto y su variabilidad, con el fin de poder identificarlo y/o caracterizarlo. A menudo, al estudiar muestras de objetos geométricos, estamos interesados en clasificar los objetos según su forma, o dada una clasificación ya hecha, asignar un nuevo objeto a la clase a la que pertenezca. Siguiendo la definición de Kendall (Kendall and Stuart, 1977), forma es toda aquella información geométrica de un objeto que queda invariante bajo traslaciones, rotaciones y/o cambios de escala del mismo. Así pues, dos objetos tienen la misma forma si podemos encontrar una transformación, que mediante tras-

laciones, rotaciones y/o cambios de escala, haga que los dos objetos encajen perfectamente. Diremos que dos objetos tienen la misma forma y el mismo tamaño si podemos encontrar una transformación basada en la composición de rotaciones y/o traslaciones que convierta uno de los objetos exactamente en el otro.

Existe una gran variedad de objetos matemáticos para caracterizar la forma (o la forma y el tamaño) de un objeto geométrico, dependiendo del tipo de objeto matemático elegido, la metodología estadística a utilizar será diferente. Como en la mayoría de casos los espacios en los que se encuentran son diferentes, podemos indicar los siguientes procedimientos.

En primer lugar, las funciones pueden representar los contornos cerrados de los objetos (curvas en *2D* y superficies en *3D*); en segundo lugar, los objetos geométricos también pueden ser tratados como subconjuntos de $\mathbb{R}^m$ y, finalmente, esos objetos geométricos pueden ser descritos como sucesiones de puntos que vienen dados por ciertas propiedades anatómicas o geométricas (*landmarks*). Las formas, en todos los enfoques anteriores, están embebidas en un espacio que no es espacio vectorial (en un gran número de casos variedades suaves) y donde no hay definida una métrica natural. Esto hace particularmente difícil la definición de la mayoría de los estadísticos más habituales; por ejemplo, no hay una manera explícita simple de calcular una media (Pennec, 2006).

Una aproximación más reciente considera el espacio de formas planas como el espacio de las curvas simples cerradas con una métrica de tipo Sobolev (Gual-Arnau *et al.*, 2015). Este espacio tiene la propiedad de ser isométrico a una variedad Grasmaniana de dimensión infinita de subespacios 2-dimensionales, y esta isometría fue utilizada para calcular geodésicas, distancias entre formas, y forma media. La teoría correspondiente al espacio de las formas tridimensionales fue generalizada por (Bauer *et al.*, 2011), donde las formas se modelaron como superficies en una variedad de superficies, y se calcularon las distancias a partir de las geodésicas de esta variedad. Sin embargo, estas aproximaciones no son invariantes frente a reparametrizaciones y tienen una gran complejidad computacional (especialmente el caso *3D*).

A lo largo de este trabajo trataremos de describir un modelo teórico con el que caracterizar la forma y el tamaño de objetos geométricos con fronteras acotadas, a los que llamaremos **cuerpos** a partir de ahora.

La propuesta consiste en representar el contorno de cada cuerpo (curva en $\mathbb{R}^2$, superficie en $\mathbb{R}^3$, o hipersuperficie en $\mathbb{R}^n$), mediante una estructura matemática llamada *current*, la cual permite identificar cada forma con un campo vectorial perteneciente a un espacio particular de Hilbert.

Esta estructura no se limita a un tipo particular de datos. De hecho, pro-

porciona un método unificado para procesar cualquier conjunto de puntos, curvas y superficies o una mezcla de ellos. No asume hipótesis en la topología y es robusta al cambio de conectividad de las estructuras. Además, es poco sensible al muestreo de las formas y no depende de la elección de la parametrización. Sin embargo, la principal ventaja de este enfoque es que las formas quedan embebidas en un espacio vectorial provisto de un producto interior; por lo que, podemos usar herramientas estadísticas fácilmente; en particular es fácil calcular medias y distancias. Otro punto fuerte del método es que no se requiere una correspondencia punto a punto de las formas ni una aplicación entre ellas; sin embargo una debilidad es que si se aplica un movimiento rígido a una forma y no a la otra, las distancias entre ellas se ven afectadas; esta es la razón por la que es muy importante registrar las formas previamente.

2.2. Currents

Antes de definir los *currents* (estructura matemática que, como se ha mencionado, va a permitir la caracterización de cuerpos) vamos a recordar al lector algunos conceptos básicos de cálculo vectorial. Este repaso será de utilidad debido a la relación de dualidad existente entre las formas diferenciales y los *currents*. Una vez definida esta herramienta de modelización de cuerpos, particularizaremos en la definición de *current geométrico*, que será el que utilizaremos a la largo de nuestro trabajo.

Como los *currents* se definen a través de integrales, este apartado incluye un repaso de integración sobre hipersuperficies. Finalmente, existe una equivalencia entre las formas diferenciales y los campos vectoriales, por lo que veremos cómo queda la expresión del *current* tras aplicar esta correspondencia.

2.2.1. Campos vectoriales, espacio tangente y formas diferenciales

Comenzamos el apartado definiendo algunos conceptos básicos de cálculo vectorial con la finalidad principal de recordar la definición de forma diferencial.

Definición 2.2.1. *El espacio tangente en un punto $p \in \mathbb{R}^n$ es el conjunto*

$$\mathbb{R}^n_p := \{(p, v) : v \in \mathbb{R}^n\}.$$

Es decir, es una copia del espacio euclídeo $\mathbb{R}^n$ con base en el punto p. Así, podemos entender el espacio tangente como el espacio de n-vectores cuyo punto incial, en lugar de estar ubicado en el orgien, se sitúa en el punto p.

Definición 2.2.2. *Se llama haz tangente a la unión puntual de los espacios tangentes en cada punto de* $\mathbb{R}^n$, $\cup_{p\in\mathbb{R}^n}\mathbb{R}^n_p$ *y se denota por* $T\mathbb{R}^n$.

Utilizando las definiciones previas, podemos recordar el concepto de campo vectorial.

Definición 2.2.3. *Un campo vectorial es una función* $w : \mathbb{R}^n \longrightarrow T\mathbb{R}^n$ *tal que para cada* $p \in \mathbb{R}^n$, $w(p) \in \mathbb{R}^n_p$.

En otras palabras, el campo vectorial w asigna a cada punto p un vector con inicio en p.

Para cada $p \in \mathbb{R}^n$, consideramos el espacio dual de $\mathbb{R}^n_p$

$$(\mathbb{R}^n_p)^* := \{\varphi : \mathbb{R}^n_p \longrightarrow \mathbb{R} : \varphi \text{ es lineal}\}.$$

Definición 2.2.4. *Se llama haz cotangente a* $\cup_{p\in\mathbb{R}^n}(\mathbb{R}^n_p)^*$ *y se denota por* $T^*\mathbb{R}^n$.

Finalmente, para poder definir las k-formas diferenciales en $\mathbb{R}^n$, es necesario conocer el espacio de las transformaciones multilineales alternantes.

Definición 2.2.5. *Sea* V *un espacio vectorial real de dimensión* n. *Decimos que la función* $T : V^k \longrightarrow \mathbb{R}$ *es multilineal si es lineal en cada coordenada y es alternante si*

$$T(v_1, \cdots, \overbrace{v_i}^{i}, \cdots, \overbrace{v_j}^{j}, \cdots, v_k) = -T(v_1, \cdots, \overbrace{v_j}^{i}, \cdots, \overbrace{v_i}^{j}, \cdots, v_k),$$

para cada $i, j = 1, 2, \cdots, k$, $i \neq j$.
Denotaremos el conjunto de funciones multilineales alternantes en V^k *por* $\Lambda^k(V)$.

Definición 2.2.6. *Una* k*-forma diferencial en* $\mathbb{R}^n$ *es una función*

$$\omega : \mathbb{R}^n \longrightarrow \cup_{p\in\mathbb{R}^n}\Lambda^k(\mathbb{R}^n_p).$$

Es decir, si ω es una k-forma diferencial, para cada $p \in \mathbb{R}^n$, $\omega(p)$ es una transformación multilineal alternante en $(\mathbb{R}^n_p)^k$, donde $\mathbb{R}^n_p$ es el espacio tangente para p.

Veamos, en particular, la expresión de una 1-forma diferencial en $\mathbb{R}^2$ y la de una 2-forma diferencial en $\mathbb{R}^3$, ya que se utilizarán en las siguientes secciones.

Denotamos $dx_1, dx_2, \cdots, dx_n$ la base de $(\mathbb{R}^n_p)^*$, dual de la base canónica.

Una **1-forma diferencial** en $\mathbb{R}^2$ es una función $\omega : \mathbb{R}^2 \longrightarrow T^*\mathbb{R}^2$, y se puede expresar:

$$\omega(p) = \omega_1(p)dx_1 + \omega_2(p)dx_2,$$

donde $\omega_i : \mathbb{R}^2 \longrightarrow \mathbb{R}$, $i = 1, 2$.

Una **2-forma diferencial** en $\mathbb{R}^3$ es una función $\omega : \mathbb{R}^3 \longrightarrow \Lambda^2(\mathbb{R}^3_p)$, y se puede expresar:

$$\omega(p) = \omega_{12}(p)(dx_1 \wedge dx_2) + \omega_{13}(p)(dx_1 \wedge dx_3) + \omega_{23}(p)(dx_2 \wedge dx_3),$$

donde $\omega_{12}, \omega_{13}, \omega_{23} : \mathbb{R}^3 \longrightarrow \mathbb{R}$ y $\wedge$ denota el producto vectorial definido para $\varphi_1, \varphi_2 \in (\mathbb{R}^3_p)^*$ como

$$\varphi_1 \wedge \varphi_2(u, v) = det \begin{pmatrix} \varphi_1(u) & \varphi_1(v) \\ \varphi_2(u) & \varphi_2(v) \end{pmatrix}.$$

2.2.2. Definición de current

Tras el repaso de conceptos básicos de Geometría, podemos definir los *currents*.

Definición 2.2.7. *Si denotamos Ω_k el espacio de las k-formas en $\mathbb{R}^n$. El espacio de k-currents en $\mathbb{R}^n$ es su dual topológico Ω'_k, es decir, el espacio de formas lineales y continuas en Ω_k.*

Existe una clasificación de subespacios de *currents* según su regularidad. Sin embargo en esta tesis sólo trabajamos en un subespacio de *currents* concreto que son los *currents* geométricos, que pasamos a definir.

Dada una subvariedad a trozos m-dimensional en $\mathbb{R}^n$, X, podemos asociarle un *k-current* mediante la integración sobre variedades. El *current* C_X, es una aplicación que integra las k-formas sobre la variedad X,

$$\begin{aligned} C_X : \quad & \Omega_k \longrightarrow \mathbb{R} \\ & w \longrightarrow \int_X w. \end{aligned} \tag{2.1}$$

Esta asociación nos permite modelizar elementos geométricos, ya que la aplicación $X \to C_X$ es inyectiva. En cambio, hay *currents* que no proceden

de la integración de ninguna variedad, es decir, no hay sobreyectividad en la aplicación $X \to C_X$.

Por otra parte, en nuestras aplicaciones, estaremos interesados en *currents* que proceden de la integración sobre una única subvariedad y no sobre subvariedades a trozos. El inconveniente que nos encontramos, es que el subconjunto de *currents* que procede de una única subvariedad no es un espacio vectorial. Esto se debe a que la suma de *currents* "equivale" a la unión de variedades. Así, por ejemplo, al calcular un estadístico básico como la media de *currents*, el resultado sería el *current* asociado a la unión de todas las subvariedades escaladas. Este hecho contradice a lo que uno espera de la idea de media de cuerpos geométricos. Cuando calculamos el cuerpo medio de un niño (relacionado con nuestra aplicación), esperamos obtener una figura similar al cuerpo de un niño y no la unión de todos ellos. Por todo lo expuesto, la metodología de los *currents* nos resultará útil a la hora de caracterizar cuerpos geométricos y trabajar con ellos en un espacio vectorial, pero no trataremos de recuperar cuerpos geométricos dado un *current*, pues, en caso de poder hacerlo, probablemente no representará la figura que buscamos.

En cambio, como ya se mencionaba en la parte final de la sección 2.1, es una herramienta que tiene excelentes propiedades para representar y determinar la forma y tamaño de un cuerpo geométrico (misma estructura para cualquier conjunto de puntos, curvas y superficies; poco sensible al muestreo de las formas; no requiere correspondencia punto a punto; es robusta al cambio de conectividad de las estructuras; etc.). Además, observamos que la definición de la integral anterior Ec. (2.1) no depende de la parametrización de la variedad y por tanto, el *current* es invariante bajo parametrizaciones, lo que supone una ventaja de esta metodología.

Particularizando, el interés de este trabajo reside en caracterizar un tipo muy concreto de subvariedades que son las hipersuperficies en $\mathbb{R}^n$. Por tanto, queremos integrar $(n-1)$-formas en $\mathbb{R}^n$ sobre la hipersuperficie S, dando como resultado el $(n-1)$-*current* geométrico siguiente, al que llamaremos $(n-1)$-*current* por comodidad:

$$C_S(\omega) = \int_S \omega, \ \forall \omega \in \Omega_{n-1}. \tag{2.2}$$

2.2.3. Integración sobre hipersuperficies

Con motivo de la definición de *current* Ec. (2.2) que usaremos a lo largo de la memoria, es interesante, recodar y detallar cómo integrar sobre hipersuperficies. Principalmente, estudiamos cómo queda la expresión del *current*

asociado a una curva en $\mathbb{R}^2$ y a una superficie en $\mathbb{R}^3$, porque serán los casos que desarrollaremos a nivel práctico.

Veamos en primer lugar cómo queda la expresión general de un *current* asociado a una hipersuperficie parametrizada.

Sea $\omega \in \Omega_{n-1}$ y S la hipersuperficie parametrizada por

$$r : D \subset \mathbb{R}^{n-1} \longrightarrow \mathbb{R},$$

con $r(D) = S$, entonces $r(x) = r(x_1, x_2, \cdots, x_{n-1}) \in S$ y

$$C_S(\omega) = \int_S \omega = \int_D \omega(r(x)) \left(\frac{\partial r}{\partial x_1} \wedge \cdots \wedge \frac{\partial r}{\partial x_{n-1}} \right) dx_1 \cdots dx_{n-1}.$$

Casos particulares:

- Particularizando en la expresión anterior, podemos observar que si hacemos referencia al *current* asociado a una **curva** L simple y orientada en $\mathbb{R}^2$, la aplicación *current* actuará sobre 1-**formas** diferenciales expresadas como

  $$\omega = \omega_1 dx_1 + \omega_2 dx_2,$$

  donde $\omega_i : \mathbb{R}^2 \longrightarrow \mathbb{R}$, $i = 1, 2$.

  Así, si $\alpha(t) = (\alpha^1(t), \alpha^2(t))$ con $t \in [a, b]$ es una parametrización regular de la curva L, $\dfrac{\partial \alpha}{\partial t} = \left(\dfrac{\partial \alpha^1}{\partial t}, \dfrac{\partial \alpha^2}{\partial t} \right)$ es el vector tangente a la curva L en el punto $\alpha(t)$ de la curva,

  $$C_L(\omega) = \int_L \omega = \int_a^b \omega(\alpha(t)) \left(\frac{\partial \alpha}{\partial t} \right) dt =$$
  $$= \int_a^b \left[\omega_1(\alpha^1(t), \alpha^2(t)) \frac{\partial \alpha^1}{\partial t} + \omega_2(\alpha^1(t), \alpha^2(t)) \frac{\partial \alpha^2}{\partial t} \right] dt.$$

- En el caso de un *current* asociado a una **superficie** S en $\mathbb{R}^3$, actuará sobre 2-**formas** expresadas como sigue:

  $$\omega = \omega_{12}(dx_1 \wedge dx_2) + \omega_{13}(dx_1 \wedge dx_3) + \omega_{23}(dx_2 \wedge dx_3),$$

  donde $\omega_{12}, \omega_{13}, \omega_{23} : \mathbb{R}^3 \longrightarrow \mathbb{R}$.

Así, si $\beta(u,v) = (\beta^1(u,v), \beta^2(u,v), \beta^3(u,v)), (u,v) \in U \subset \mathbb{R}^2$ es una parametrización de la superficie orientada $S = \beta(U)$, entonces $\frac{\partial \beta}{\partial u}(u,v) \wedge \frac{\partial \beta}{\partial v}(u,v) = \left(\frac{\partial(\beta^1,\beta^2)}{\partial(u,v)}, \frac{\partial(\beta^2,\beta^3)}{\partial(u,v)}, \frac{\partial(\beta^3,\beta^1)}{\partial(u,v)} \right)$ es una base ortonormal del plano tangente a la superficie en $\beta(u,v)$,

$$C_S(\omega) = \int_S \omega = \int_U \omega(\beta(u,v)) \left(\frac{\partial \beta}{\partial u}(u,v) \wedge \frac{\partial \beta}{\partial v}(u,v) \right) dudv$$
$$= \int_U \omega_{12}(\beta(u,v)) \frac{\partial(\beta^1,\beta^2)}{\partial(u,v)} + \omega_{13}(\beta(u,v)) \frac{\partial(\beta^2,\beta^3)}{\partial(u,v)} + \omega_{23}(\beta(u,v)) \frac{\partial(\beta^3,\beta^1)}{\partial(u,v)} dudv$$

donde $\frac{\partial(\beta^1,\beta^2)}{\partial(u,v)} = det \begin{pmatrix} \frac{\partial \beta^1}{\partial u} & \frac{\partial \beta^1}{\partial v} \\ \frac{\partial \beta^2}{\partial u} & \frac{\partial \beta^2}{\partial v} \end{pmatrix}$ representa al Jacobiano.

2.2.4. Representación vectorial de currents

Gracias a la correspondencia que existe entre formas diferenciales y campos vectoriales, transformamos las expresiones anteriores, de modo que a partir de ahora trabajaremos con campos vectoriales.

Una forma $\omega \in \Omega_{n-1}$, puede ser representada por un campo vectorial $\overline{\omega}$ en $\mathbb{R}^n$ gracias a la isometría entre las $(n-1)$-formas en Ω_{n-1} y los vectores en $\mathbb{R}^n$. Entonces,

$$\begin{aligned} C_S(\omega) &= \int_D \omega(r(x)) \left(\frac{\partial r}{\partial x_1} \wedge \cdots \wedge \frac{\partial r}{\partial x_{n-1}} \right) dx_1 \ldots dx_{n-1} \\ &= \int_D Det \left(\frac{\partial r}{\partial x_1}, \ldots, \frac{\partial r}{\partial x_{n-1}}, \bar{\omega}(r(x)) \right) dx_1 \ldots dx_{n-1}. \end{aligned} \tag{2.3}$$

Formalmente, la asociación entre formas y vectores viene dada por dualidad y el operador *Hodge star* (Do Carmo, 2012).

Casos particulares:

- Si la **curva** L está parametrizada por α como en la sección anterior, el *current* asociado a ésta se puede expresar como

$$C_L(\omega) = \int_a^b \overline{\omega}(\alpha(t)) \cdot \frac{\partial \alpha}{\partial t} dt, \tag{2.4}$$

donde $\overline{\omega}$ es un campo vectorial en $\mathbb{R}^2$ y $\cdot$ es el producto interior en $\mathbb{R}^2$.

- Si la **superficie** orientada S en $\mathbb{R}^3$ está parametrizada mediante β como en la sección previa, el *current* asociado a la superficie se expresa:

$$C_S(\omega) = \int_U \overline{\omega}(\beta(u,v)) \cdot \left(\frac{\partial \beta}{\partial u}(u,v) \wedge \frac{\partial \beta}{\partial v}(u,v) \right) dudv, \tag{2.5}$$

  donde $\overline{\omega}$ es un campo vectorial en $\mathbb{R}^3$ y $\cdot$ es el producto interior en $\mathbb{R}^3$.

2.3. Conceptos básicos de Análisis Funcional

Siguiendo la estructura marcada por la introducción, una vez tenemos los cuerpos geométricos caracterizados mediante *currents*, transformaremos a través de una isometría estos funcionales en campos vectoriales en un *Reproducing Kernel Hilbert Space* (*RKHS*) (detallado en el Capítulo 3). Por ello, a partir de esa transformación, nuestros elementos estarán en un espacio de Hilbert con unas propiedades particulares. Consideramos interesante hacer un recordatorio de algunos resultados básicos y generales de Análisis Funcional, principalmente en espacios de Hilbert, que utilizaremos a lo largo de la presente memoria. En la siguiente sección 2.4, concretaremos las características particulares de un espacio RKH.

Comenzamos recordando el concepto de función definida positiva, que aparecerá en múltiples ocasiones.

Definición 2.3.1. *Sea E un conjunto. Una función k en $E \times E$ es definida positiva si para cualquier conjunto de números reales $\{a_j\}_{j=1}^n$ y cualquier conjunto $\{t_j\}_{j=1}^n$ de E, $n \in \mathbb{N}$,*

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

Desde este momento, denotaremos por $\mathbb{H}$ a un espacio de Hilbert con producto interior $\langle \cdot, \cdot \rangle$ y norma $\| \cdot \|$, y recordamos e indicamos notación y resultados relacionados con la ortogonalidad y las bases de funciones ortonormales.

Definición 2.3.2. *Diremos que $x_1, x_2 \in \mathbb{H}$ son ortogonales si $\langle x_1, x_2 \rangle = 0$. Una sucesión ortonormal es un conjunto contable de elementos $\{e_1, e_2, \ldots\}$ con $\|e_j\| = 1$ para todo j y los e_j son ortogonales dos a dos.*

Definición 2.3.3. *Una sucesión ortonormal $\{e_n\}$ en un espacio de Hilbert $\mathbb{H}$ es una base ortonormal o un sistema completo ortonormal (CONS) si*

$\overline{span\{e_n\}} = \mathbb{H}$, *donde el* $span\{e_n\}$ *está formado por todas las combinaciones lineales finitas de los elementos* $\{e_n\}$.

Teorema 2.3.1. *Todo elemento* x *de un espacio de Hilbert* $\mathbb{H}$ *con base ortonormal* $\{e_j\}$ *se puede expresar a través de la expansión de Fourier*

$$x = \sum_{j=1}^{\infty} \langle x, e_j \rangle e_j,$$

y

$$\|x\|^2 = \sum_{j=1}^{\infty} \langle x, e_j \rangle^2.$$

A continuación, aparecen algunas definiciones y resultados básicos de teoría de operadores en espacios de Hilbert.

Sean $\mathbb{X}_1$ y $\mathbb{X}_2$ espacios vectoriales con normas $\|\cdot\|_1$ y $\|\cdot\|_2$, respectivamente. Recordamos que si $\mathcal{B}(\mathbb{X}_1, \mathbb{X}_2)$ es el espacio de operadores lineales y acotados de $\mathbb{X}_1$ a $\mathbb{X}_2$, la **norma de un operador** $T \in \mathcal{B}(\mathbb{X}_1, \mathbb{X}_2)$ se define como sigue:

$$\|T\| = sup\{\|Tx\|_2 : x \in \mathbb{X}_1,\ \|x\|_1 = 1\}.$$

Entonces, para cualquier $x \in \mathbb{X}_1$,

$$\|Tx\|_2 \leq \|T\|\|x\|_1.$$

En caso de que $\mathbb{X}_1 = \mathbb{X}_2 = \mathbb{X}$, denotamos por $\mathcal{B}(\mathbb{X})$ el conjunto de operadores acotados en $\mathbb{X}$.

Denotamos por $\mathbb{H}^* = \mathcal{B}(\mathbb{H}, \mathbb{R})$ el espacio dual de $\mathbb{H}$, es decir, el espacio de funcionales lineales y acotados de $\mathbb{H}$ a $\mathbb{R}$.

Teorema 2.3.2 (*Representación de Riesz*). *Supongamos que* $\mathbb{H}$ *es un espacio de Hilbert con producto interior* $\langle \cdot, \cdot \rangle$ *y norma* $\|\cdot\|$ *y* $T \in \mathbb{H}^*$. *Existe un único elemento* $e_T \in \mathbb{H}$, *llamado representante de* T, *con la propiedad:*

$$T(x) = \langle x, e_T \rangle,$$

para todo $x \in \mathbb{H}$ *y* $\|T\| = \|e_T\|$.

Teorema 2.3.3. *Sean* $\mathbb{H}_1, \mathbb{H}_2$ *espacios de Hilbert con productos interiores* $\langle \cdot, \cdot \rangle_i$, $i = 1, 2$. *Sea* $T \in \mathcal{B}(\mathbb{H}_1, \mathbb{H}_2)$, *hay un único elemento* T^* *de* $\mathcal{B}(\mathbb{H}_2, \mathbb{H}_1)$ *determinado por la relación*

$$\langle Tx_1, x_2 \rangle_2 = \langle x_1, T^* x_2 \rangle_1,$$

para todo $x_1 \in \mathbb{H}_1$, $x_2 \in \mathbb{H}_2$.

Definición 2.3.4. *Para cualquier $T \in \mathcal{B}(\mathbb{H}_1, \mathbb{H}_2)$, el único operador T^* de $\mathcal{B}(\mathbb{H}_2, \mathbb{H}_1)$ que cumple la propiedad del teorema anterior se llama adjunto de T. Si $\mathbb{H}_1 = \mathbb{H}_2$, T se llama **operador autoadjunto** cuando $T^* = T$.*

Definición 2.3.5. *Sea T un operador en un espacio de Hilbert $\mathbb{H}$.*

- *Es **definido no negativo** si es autoadjunto y $\langle Tx, x\rangle \geq 0$ para todo $x \in \mathbb{H}$.*
- *Es **definido positivo** si es autoadjunto y $\langle Tx, x\rangle > 0$ para todo $x \in \mathbb{H}$.*
- *Para dos operadores T_1, T_2 escribimos $T_1 \leq T_2$ (respectivamente $T_1 < T_2$) si $T_2 - T_1$ es definido no negativo (respectivamente definido positivo).*

En cuanto a la inversibilidad de operadores, recordamos las siguientes nociones básicas.

Definición 2.3.6. *Sea un operador lineal T definido de un espacio vectorial $\mathbb{X}_1$ a un espacio vectorial $\mathbb{X}_2$.*

- *T es inyectivo si $Ker(T) = \{0\}$.*
- *T es sobreyectivo si $Im(T) = \mathbb{X}_2$.*
- *T es biyectivo si es inyectivo y sobreyectivo a la vez. En tal caso, existe la aplicación inversa T^{-1} de $\mathbb{X}_2$ a $\mathbb{X}_1$ cumpliendo $T^{-1}T = TT^{-1} = I$ (donde I es el operador identidad, de manera que $I(x) = x$, $\forall x \in \mathbb{X}_1$ o $\mathbb{X}_2$).*

Teorema 2.3.4. *Sean $\mathbb{X}_1$ y $\mathbb{X}_2$ espacios de Banach con $T \in \mathcal{B}(\mathbb{X}_1, \mathbb{X}_2)$. Si T es biyectivo, $T^{-1} \in \mathcal{B}(\mathbb{X}_2, \mathbb{X}_1)$.*

2.3.1. Operadores compactos y Teorema espectral

Ahora focalizamos el estudio en los operadores compactos en espacios de Hilbert. Tienen un interés especial porque es posible aproximarlos por operadores finito-dimensionales y tienen propiedades parecidas a las matrices.

Definición 2.3.7. *Un operador lineal T de un espacio normado $\mathbb{X}_1$ a otro espacio normado $\mathbb{X}_2$ es **compacto** si para cualquier sucesión acotada $\{x_n\} \in \mathbb{X}_1$, $\{Tx_n\}$ contiene una subsucesión convergente en $\mathbb{X}_2$.*

El siguiente teorema muestra algunas propiedades básicas de operadores compactos.

Teorema 2.3.5. *Consideramos operadores lineales compactos entre dos espacios normados.*

1. *Un operador compacto es acotado.*
2. *La clausura del rango de cualquier operador compacto es separable.*
3. *El operador resultante de la composición de dos operadores compactos es compacto.*

Hacemos a continuación un respaso de la teoría espectral de operadores, que resultará relevante en los Capítulos 6 y 7.

Definición 2.3.8. *Sea $T \in \mathcal{B}(\mathbb{H})$ y suponemos que hay un $\lambda \in \mathbb{R}$ y un elemento $e \in \mathbb{H}$ tal que*

$$Te = \lambda e.$$

Entonces, λ es un valor propio y e es el correspondiente vector propio (o función propia cuando $\mathbb{H}$ es un espacio de funciones) de T.

Teorema 2.3.6. *Sea $T \in \mathcal{B}(\mathbb{H})$ y suponemos que $e_j \in Ker(T - \lambda_j I)$ para $j = 1, 2, \ldots$ donde λ_j son valores propios de T no cero. Entonces,*

1. *Los elementos e_j son linealmente independientes, y*
2. *Si T es autoadjunto, los elementos e_j son mutuamente ortogonales.*

Teorema 2.3.7. *Sea $T \in \mathcal{B}(\mathbb{H})$ un operador compacto. Entonces,*

1. *$Ker(T - \lambda I)$ es de dimensión finita para $\lambda \neq 0$.*
2. *El número distinto de valores propios de T con valor absoluto más grande que cualquier número positivo es finito.*
3. *El conjunto de valores propios no cero es contable.*

El siguiente teorema, de gran relevancia en la teoría de Análisis Funcional, muestra la descomposicón en valores y vectores propios de un operador autoadjunto y compacto en un espacio de Hilbert.

Teorema 2.3.8 (***Teorema espectral***)**.** *Sea T un operador compacto y autoadjunto en $\mathbb{H}$. El conjunto de valores propios no cero para T es finito o está formado por una sucesión que tiende a 0. Cada valor propio no cero tiene multiplicidad finita y los valores correspondientes a diferentes valores propios son ortogonales. Sean $\lambda_1, \lambda_2, \ldots$ los valores propios ordenados tal*

que $|\lambda_1| \geq |\lambda_2| \geq \ldots$ *y sean* $e_1, e_2, \ldots$ *los correspondientes vectores propios ortonormales obtenidos utilizando el proceso de ortogonalización de Gram-Schmidt. Entonces* $\{e_j\}$ *es una base ortonormal para* $\overline{Im(T)}$ *y para todo* $x \in \mathbb{H}$*,*

$$Tx = \sum_{j \geq 1} \lambda_j \langle x, e_j \rangle e_j. \tag{2.6}$$

Si el operador T es autoadjunto, compacto y también definido no negativo, entonces todos los valores propios son no negativos y la representación de la Ec. (2.6) puede ser extendida a exponentes no enteros. Por ejemplo, esto nos permite definir el operador autoadjunto,

$$T^{1/2}x = \sum_{j=1}^{\infty} \sqrt{\lambda_j} \langle x, e_j \rangle e_j.$$

El cual satisface $T^{1/2}T^{1/2} = T$. Además, $T^{1/2}$ es un operador compacto de $\mathcal{B}(\mathbb{H})$.

También, como consecuencia del teorema, si además de las características exigidas al operador se requiere que este sea definido positivo, todos los valores propios serán positivos, y por tanto $Ker(T) = \{0\}$.

Propiedad 1. *Si un operador* T *es autoadjunto, compacto y definido positivo en* $\mathbb{H}$ *es inyectivo.*

Estas últimas observaciones serán útiles en la tercera aportación de la presente memoria (Capítulo 7).

Es importante señalar que, en general, la obtención de los valores propios de un operador no es sencilla. Sin embargo, en el caso de los operadores integrales, que se introducen en la siguiente sección, el cálculo de la expresión analítica de la descomposición espectral puede ser simple (por ejemplo, la descomposición del operador integral definido a partir del kernel $k(s,t) = min(s,t), t, s \in [0,1]$ se obtiene en Hsing and Eubank (2015)) o en su defecto, los valores propios pueden ser estimados por una aproximación matricial.

2.3.2. Operadores integrales y Teorema de Mercer

La descomposición espectral de operadores integrales nos ayudará a encontrar bases de funciones ortonormales en nuestro espacio de trabajo. En Estadística, cuando se trata con datos funcionales se suele hacer un tratamiento previo para paliar las dificultades que añade el hecho de que nuestros elementos sean infinito-dimensionales. Una de las técnicas más utilizadas es

la expresión de datos funcionales en bases de funciones ortonormales, como se estudia en la sección 2.5. Recordar estos resultados de teoría de operadores será útil para la mejor comprensión del desarrollo teórico realizado en los Capítulos 6 y 7.

Definición 2.3.9 (***Operador integral***). *Sea $(E, \mathfrak{B}, \mu)$ un espacio de medida para cualquier medida finita μ. Supongamos que k es una función medible en $E \times E$ tal que $\int\int_{E\times E} k^2(s,t)d\mu(s)d\mu(t)$ es finita y define el operador integral L_k por*

$$(L_k f)(\cdot) := \int_E k(s,\cdot)f(s)d\mu(s), \tag{2.7}$$

*para $f \in \mathbb{L}^2(E, \mathfrak{B}, \mu)$. A la función k la llamaremos **kernel** de L_k.*

Propiedades 1. *El operador L_k definido en Ec. (2.7) tiene las siguientes propiedades:*

- $L_k \in \mathcal{B}(\mathbb{L}^2(E, \mathfrak{B}, \mu))$.
- *L_k es compacto.*
- *L_k es definido no negativo si y solo si su kernel es definido no negativo.*

El siguiente resultado es el Teorema de Mercer, el cual afirma esencialmente que si L_k es un operador integral con un kernel simétrico y definido no negativo, podemos obtener series equivalentes al operador y su kernel.

Teorema 2.3.9. ***[Teorema de Mercer]*** *Sea el kernel continuo, simétrico y definido no negativo k y L_k el correspondiente operador integral. Si (λ_j, e_j) son los pares de valores y funciones propias de L_k, entonces k tiene la representación,*

$$k(s,t) = \sum_{j=1}^{\infty} \lambda_j e_j(s)e_j(t),$$

para todo s, t donde la suma converge absoluta y uniformemente.

Propiedad 2. *Bajo las condiciones del teorema anterior,*

$$\lim_{n\longrightarrow\infty} \sup_{s,t} \sum_{j=n+1}^{\infty} |\lambda_j e_j(s)e_j(t)| = 0.$$

El siguiente resultado describe una característica de optimalidad del truncamiento de la descomposición de k del Teorema de Mercer.
Decimos que un kernel $k(s,t)$ tiene rango r si el operador integral correspondiente tiene el rango r, es decir, si $dim(Im(T)) = r$.

Teorema 2.3.10. *Sea k un kernel simétrico y definido no negativo con la descomposición propia*

$$k(s,t) = \sum_{j=1}^{\infty} \lambda_j e_j(s) e_j(t).$$

Entonces, para cualquier entero r para el cual $\lambda_r > 0$,

$$\min_{rank(W)=r} \int\int_{E\times E} \{k(s,t) - W(s,t)\}^2 d\mu(s) d\mu(t) = \sum_{j=r+1}^{\infty} \lambda_j^2,$$

donde el mínimo se alcanza en $W(s,t) = \sum_{j=1}^{r} \lambda_j e_j(s) e_j(t)$.

2.3.3. Diagonalización simultánea de dos operadores definidos no negativos

Extendiendo el resultado de diagonalización simultánea de matrices a la teoría de operadores, se obtiene el siguiente teorema (Hsing and Eubank, 2015) en el que uno de los operadores ha de ser compacto.

Teorema 2.3.11. *Sean $\mathcal{C}$ y $\mathcal{W}$ operadores autoadjuntos en un espacio de Hilbert separable $\mathbb{H}$. Suponemos que $\mathcal{C}$ es compacto y definido no negativo, $\mathcal{W}$ es definido positivo, y $\mathcal{G} = \mathcal{C} + \mathcal{W}$ es invertible. Sean $\{(\eta_j, \upsilon_j)\}_{j=1}^{\infty}$ los valores y vectores propios de $\mathcal{G}^{-1/2}\mathcal{C}\mathcal{G}^{1/2}$, donde los valores η_j están necesariamente en $[0,1)$, y definimos*

$$\gamma_j := \eta_j/(1-\eta_j),$$
$$u_j := (1-\eta_j)^{-1/2}\mathcal{G}^{-1/2}\upsilon_j.$$

Entonces,

$$\mathcal{C}u_j = \eta_j \mathcal{G} u_j = \gamma_j \mathcal{W} u_j,$$

$$\langle u_i, \mathcal{W} u_j \rangle = \delta_{ij},$$

y

$$x = \sum_{j=1}^{\infty} \langle x, \mathcal{W} u_j \rangle u_j,$$

para todo x.

Este resultado se adapta a la situación presente en el Capítulo 7, con el fin de encontrar una base $\{u_j\}_{j=1}^{\infty}$ que relaciona a dos operadores, aportando propiedades interesantes.

2.3.4. Elementos aleatorios en un espacio de Hilbert

Podemos analizar los elementos aleatorios en un espacio de Hilbert desde una doble perspectiva. Por una parte, los datos funcionales son realizaciones de variables aleatorias que toman valores en un espacio de Hilbert. Por otra parte, los datos funcionales son caminos muestrales de un proceso estocástico (con media y función covarianza suaves). Ver Hsing and Eubank (2015) para más detalle.

Como nuestros datos (cuerpos geométricos aleatorios) quedan expresados como funciones en un espacio de Hilbert, estudiamos algunas propiedades según esta doble visión. Utilizaremos estos resultados en la tercera aportación de la memoria (Capítulo 7), para hallar una base de funciones en la que expresar nuestros datos muestrales.

Realizaciones de variables aleatorias en $\mathbb{H}$

El siguiente teorema es una caracterización para elementos medibles de un espacio de Hilbert.

Teorema 2.3.12. *Sea χ una aplicación de un espacio de probabilidad $(\Omega, \mathcal{F}, \mathcal{P})$ en $(\mathbb{H}, \mathcal{B}(\mathbb{H}))$. Entonces,*

- *χ es medible si $\langle \chi, f \rangle$ es medible para todo $f \in \mathbb{H}$.*
- *Si χ es medible, su distribución queda únicamente determinada por las distribuciones marginales de $\langle \chi, f \rangle$, $f \in \mathbb{H}$.*

Sea χ con $\mathbb{E}(\|\chi\|) < \infty$ un elemento aleatorio en un espacio de Hilbert separable $\mathbb{H}$ definido en un espacio de probabilidad $(\Omega, \mathcal{F}, \mathcal{P})$. La **media** m y **covarianza** $\mathcal{K}$ de χ se definen como sigue:

$$m = \mathbb{E}(\chi) = \int_\Omega \chi d\mathcal{P}, \text{ por tanto, } \langle m, f \rangle = \mathbb{E}[\langle \chi, f \rangle] \text{ para } f \in \mathbb{H}.$$

$$\mathcal{K} = \mathbb{E}[(\chi - m) \otimes (\chi - m)] = \int_\Omega (\chi - m) \otimes (\chi - m) d\mathcal{P}, \tag{2.8}$$

si $\mathbb{E}(\|\chi\|^2) < \infty$.

Las integrales anteriores son integrales de Bochner, la generalización natural de la integral de Lebesgue a espacios de Banach. Además, el operador, $\otimes$ actúa como sigue: $(x_1 \otimes x_2)(y) := \langle x_1, y \rangle x_2$. Por tanto,

$$\mathcal{K}(f) = \int_\Omega \langle \chi - m, f \rangle (\chi - m) d\mathcal{P}.$$

A continuación se muestran algunas propiedades del operador covarianza $\mathcal{K}$.

Proposición 2.3.1. *El operador $\mathcal{K}$ es compacto.*

Propiedades 2. *Supongamos $m = 0$ y $\mathbb{E}(\|\chi\|^2) < \infty$. Sean $f, g \in \mathbb{H}$*

- $\langle \mathcal{K}f, g\rangle_{\mathbb{H}} = \mathbb{E}[\langle \chi, f\rangle\langle \chi, g\rangle]$.
- *$\mathcal{K}$ es un operador definido no negativo.*
- $P(\chi \in \overline{Im(\mathcal{K})}) = 1$.

Aplicando el Teorema espectral (2.3.8) a $\mathcal{K}$, que es compacto y autoadjunto en $\mathbb{H}$, se obtiene el siguiente resultado.

Teorema 2.3.13. *El operador $\mathcal{K}$ admite la descomposición propia*

$$\mathcal{K} = \sum_{j=1}^{\infty} \lambda_j e_j \otimes e_j.$$

Las funciones propias $\{e_j\}_{j=1}^{\infty}$ forman una base ortonormal para $\overline{Im(\mathcal{K})}$ mientras los valores propios son no negativos con el conjunto $\{\lambda_j\}_{j=1}^{\infty}$ finito o formado por una sucesión que tiende a 0. Cada valor propio distinto de cero tiene multiplicidad finita.

$$\mathcal{K}(f) = \sum_{j=1}^{\infty} \lambda_j \langle e_j, f\rangle e_j.$$

Teorema 2.3.14. *Suponemos que $\mathcal{K}$ tiene la descomposición del teorema anterior. Entonces, con probabilidad uno,*

$$\chi = \sum_{j=1}^{\infty} \langle \chi, e_j\rangle e_j,$$

donde $\langle \chi, e_j\rangle$, $j \geq 1$ son variables aleatorias incorreladas con media 0 y varianzas λ_j.

Teorema 2.3.15. *Si $\{f_j\}_{j=1}^{\infty}$ es una base ortonormal de $\mathbb{H}$,*

$$\mathbb{E}\|\chi - \sum_{j=1}^{n} \langle \chi, f_j\rangle f_j\|^2 = \mathbb{E}\|\chi\|^2 - \sum_{j=1}^{n} \langle \mathcal{K}f_j, f_j\rangle,$$

el cual se minimiza tomando $f_j = e_j$, $1 \leq j \leq n$.

Caminos muestrales de un proceso estocástico

Sea E un espacio métrico general compacto y $X = \{X(t) : t \in E\}$ un proceso estocástico en un espacio de probabilidad $(\Omega, \mathcal{F}, \mathcal{P})$. X representa una función aleatoria que puede ser observada parcial o completamente. En un proceso estocástico se asume que $X(t)$ es una variable aleatoria, es decir, que es medible para cada t fija. Es importante notar que, sin exigir ninguna condición más, que $X(t)$ sea una variable aleatoria no implica que $X(\cdot)$ sea una variable aleatoria en $L^2(E, \mathcal{B}(E), \mu)$.

Se puede definir la **función media** m y la **función covarianza** γ del proceso como sigue:

- $m(t) = \mathbb{E}[X(t)]$
- $\gamma(s,t) = Cov(X(s), X(t))$, para $s, t \in E$

Definición 2.3.10. *Un proceso que tiene bien definidas la función media y la función covarianza se denomina proceso de segundo orden.*

Teorema 2.3.16. *La función γ es definida no negativa.*

Definición 2.3.11. *Sea X un proceso de segundo orden. Decimos que X es continuo en media cuadrática si*

$$\lim_{t_n \longrightarrow t} \mathbb{E}\left(X(t_n) - X(t)\right)^2 = 0.$$

Teorema 2.3.17. *Sea X un proceso de segundo orden. X es continuo en media cuadrática sii las funciones covarianzas son continuas.*

Definimos el siguiente operador integral en $L^2(E, \mathcal{B}(E), \mu)$

$$(L_\gamma f)(t) := \int_E \gamma(t,s) f(s) d\mu(s),$$

donde μ es una medida finita. Llamamos L_γ al operador covarianza de un proceso.

Aplicando el Teorema de Mercer,

$$\gamma(s,t) = \sum_{j=1}^{\infty} \lambda_j e_j(s) e_j(t),$$

donde la suma converge absoluta y uniformemente al soporte de μ, con (λ_j, e_j) los valores y funciones propias de L_γ.

Notar que el operador L_γ es distinto al dado en Ec. (2.8) y por ello también lo es su descomposición.

2.4. Scalar-valued RKHS

Aronszajn (1950) desarrolló la teoría de los *scalar-valued Reproducing Kernel Hilbert Spaces*. Un *scalar-valued RKHS* es un espacio de Hilbert de funciones $f : \mathbb{R}^n \longrightarrow \mathbb{R}$. En los últimos años, el estudio de RKHS se ha extendido a funciones vectoriales, donde el espacio contiene campos vectoriales de $\mathbb{R}^n$ a $\mathbb{R}^n$.

Como se ha mencionado anteriormente, nuestros datos geométricos se representan a través de *currents*, y tras ciertas asociaciones que veremos en el Capítulo 3, tendremos los cuerpos expresados como campos vectoriales en un *vector-valued RKHS*. Los *vector-valued RKHS*, son una extensión de los *scalar-valued RKHS*, por lo que es interesante conocer y describir brevemente las propiedades del caso escalar, ya que, muchas de ellas se heredarán en el caso vectorial.

Definición 2.4.1. *Sea $E \subset \mathbb{R}^n$ un conjunto y $\mathbb{H}$ un espacio de Hilbert de funciones reales definidas en E. Una función bivariante $k : E \times E \longrightarrow \mathbb{R}$ se llama kernel reproductivo (rk) para $\mathbb{H}$ si:*

1 $\forall t \in E$, $k(\cdot, t) \in \mathbb{H}$ y

2 k cumple la reproducing property, es decir, $\forall f \in \mathbb{H}$ y $t \in E$

$$f(t) = \langle f, k(\cdot, t) \rangle.$$

Definición 2.4.2. *Si $\mathbb{H}$ es un espacio de Hilbert de funciones reales que posee un rk, diremos que $\mathbb{H}$ es un scalar-valued Reproducing Kernel Hilbert Space (scalar-valued RKHS).*

Definición 2.4.3. *Llamamos kernels escalares a aquellas funciones $k : E \times E \longrightarrow \mathbb{R}$ que son definidas positivas y simétricas ($k(s,t) = k(t,s)\ \forall s, t \in E$).*

A continuación se muestran algunos ejemplos de kernels escalares. Consideremos $E \subset \mathbb{R}^n$,

- La función Gaussiana (o kernel Gaussiano)

$$k_G(s,t) := e^{\frac{-\|s-t\|^2_{\mathbb{R}^n}}{\lambda^2}},$$

donde $\lambda > 0$ es un parámetro de escala.

- El kernel de Epanechnikov

$$k_E(s,t) = (1 - \frac{\|s-t\|^2_{\mathbb{R}^n}}{\lambda^2})1_{\{\|s-t\| \leq \lambda\}},$$

  donde $1_{\{\dots\}}$ es la función indicatriz.

- El kernel de Cauchy

$$k_C(s,t) = (1 + \frac{\|s-t\|^2_{\mathbb{R}^n}}{\lambda^2})^{-1},$$

Los tres kernels definidos dependen de un parámetro λ que deberá ser escogido. Además, estos kernels son isotrópicos e invariantes ante traslaciones, ya que $k(s,t) = k(\|s-t\|_{\mathbb{R}^n})$.

Los siguientes teoremas son dos resultados fundamentales en la teoría de RKHS, relacionan los conceptos de rk y kernel.

Teorema 2.4.1. *El rk de un scalar-valued RKHS es único, simétrico y definido positivo (es decir, es un kernel).*

Inversamente,

Teorema 2.4.2. *Si k es un kernel (función bivariante en $E \times E$ simétrica y definida positiva), existe un único scalar-valued RKHS de funciones en E, denotado por H_k, tal que k es el rk asociado a H_k.*

Como consecuencia de este teorema, los kernels k_G, k_E y k_C generan RKHS de funciones reales H_{k_G}, H_{k_E} y H_{k_C}. Cada kernel genera un *scalar-valued RKHS* distinto, es más, dependiendo del valor del parámetro λ escogido en cada kernel, el RKHS que queda determinado es diferente. Por ello, la elección del parámetro λ del kernel dependerá de la aplicación práctica de la teoría desarrollada.

2.4.1. Relaciones entre los espacios RKHS y los kernels que los generan

Sean k_1 y k_2 dos funciones kernels.

Denotamos $k_1 \ll k_2$ si $k_2 - k_1$ es una función definida positiva.

Teorema 2.4.3. *Si $k_1 \ll k_2$, $H_{k_1} \subset H_{k_2}$ y las normas $\|\cdot\|_1$ y $\|\cdot\|_2$ para H_{k_1} y H_{k_2} satisfacen $\|f_1\|_2 \leq \|f_1\|_1 \; \forall f_1 \in H_{k_1}$.*

Teorema 2.4.4. *Sean k_1, k_2 kernels definidos positivos. Entonces, $H_{k_1} \subset H_{k_2}$ si existe una constante positiva B tan que $k_1 \ll Bk_2$.*

Teorema 2.4.5. *El producto de dos RKHS con rks k_1 y k_2 es también un RKHS con $k_1 k_2$ como su rk.*

2.4.2. Operadores integrales en RKHS

Cuando un operador integral está definido en un RKHS, no solo hablamos del kernel R que define el operador integral, sino que también tenemos el kernel reproductivo, k, propio del espacio. De este modo, encontramos propiedades que no se dan cuando el operador se define en un espacio de Hilbert cualquiera.

Teorema 2.4.6. *Un operador integral L_R en un RKHS es autoadjunto si su kernel R es simétrico.*

Teorema 2.4.7. *Un operador integral en un RKHS con rk k es definido no negativo si su kernel R es definido no negativo, en tal caso, existe una constante B tal que $Bk - R$ es definido no negativo.*

Debido a que la covarianza γ es una función definida no negativa (Teorema (2.3.16)), puede generar un espacio H_γ con γ su kernel reproductor. Además, si γ es la función covarianza en un espacio H_k, se cumple (Lukić and Beder, 2001):

$$H_\gamma \subset H_k. \tag{2.9}$$

Por otro lado, si k es un kernel continuo, simétrico y definido no negativo y consideramos el operador integral $L_k : L^2 \longrightarrow L^2$ con valores y funciones propias (λ_j, e_j), al aplicar el Teorema de Mercer (2.3.9) a este operador se verifica:

$$H_k = Im(L_k^{1/2}) = \{f \in L^2 : \sum_{j=1,\lambda_j>0}^{\infty} \frac{\langle f, e_j\rangle_{L^2}}{\lambda_j} < \infty\}.$$

En particular, esto implica que $L_k(f) \in H_k$, $\forall f \in L^2$ y $\{\sqrt{\lambda_j}e_j\}_{j=1}^{\infty}$ es una base ortonormal de H_k (Quang *et al.*, 2010).

Este resultado relaciona de algún modo los espacios H_k y L^2 y se utilizará en el Capítulo 7. Además, sus productos interiores cumplen:

$$\langle e_i, e_j\rangle_{L^2} = \delta_{ij}, \qquad \text{and} \qquad \langle e_i, e_j\rangle_{H_k} = \delta_{ij}/\lambda_i,$$

donde δ_{ij} es la delta de Kronecker.

2.4.3. Elementos aleatorios en un RKHS

En la sección anterior, estudiamos los elementos aleatorios en un espacio de Hilbert desde dos perspectivas: procesos estocásticos y variables aleatorias.

Ahora, veremos qué ocurre cuando el espacio de Hilbert es un RKHS H_k con k el kernel reproductor.

Recordemos que X es un proceso estocástico si $X(t)$ es una variable aleatoria para cada t fija y que X es una variable aleatoria en H_k si X es una aplicación medible del espacio de probabilidad a H_k.

Teorema 2.4.8. *Un elemento aleatorio X de H_k es un proceso estocástico. Inversamente, un proceso estocástico X que toma valores en H_k es una variable aleatoria de H_k.*

Teorema 2.4.9. *Sea X un elemento aleatorio de H_k con $\mathbb{E}\|X\|^2 < \infty$. Entonces, X es un proceso continuo en media cuadrática en E y el elemento media m y el operador covarianza $\mathcal{K}$ están relacionados con las funciones media y covarianza $m(t)$ y $\gamma(s,t)$ por*

$$m(t) = \mathbb{E}[X(t)] = \langle m, k(\cdot, t)\rangle$$

y

$$\gamma(s,t) = Cov(X(t), X(s)) = \langle \mathcal{K}k(\cdot,t), k(\cdot,s)\rangle. \tag{2.10}$$

Teorema 2.4.10. *Sea X un elemento aleatorio de H_k con media cero y $\mathbb{E}\|X\|^2 < \infty$. Si el operador covarianza $\mathcal{K}$ tiene la descomposición en valores propios*

$$\mathcal{K} = \sum_{j=1}^{\infty} \lambda_j e_j \otimes e_j,$$

entonces,

$$\gamma(s,t) = \sum_{j=1}^{\infty} \lambda_j e_j(s) e_j(t),$$

donde la suma converge absoluta y uniformemente, y

$$\lim_{n\longrightarrow\infty} \sup_{t\in E} \mathbb{E}[X(t) - X_n(t)]^2 = 0,$$

donde $X_n(t) := \sum_{j=1}^{n} \langle X, e_j\rangle e_j(t)$.

Este último es un resultado interesante pero no muy útil desde el punto de vista práctico, pues no se puede obtener de manera sencilla la descomposición espectral del operador $\mathcal{K}$. Por este motivo, en una de las aportaciones de la tesis (Capítulo 7), buscamos una descomposición de este tipo pero usando el operador integral definido a partir de la función covarianza.

2.5. Análisis de Datos Funcionales

Desde finales de la década de los 90, la popularidad de la teoría de datos funcionales ha ido aumentando hasta convertirse en uno de los campos más importantes de investigación en Estadística. Se emplea cuando el espacio muestral es un espacio funcional de dimensión infinita. Aunque esta teoría ha incorporado muchas herramientas de la Estadística Paramétrica Clásica o de la Estadística Multivariante, la naturaleza infinito dimensional del espacio muestral plantea problemas particulares. Los libros de Silverman and Ramsay (2005) y Ferraty and Vieu (2006) son referencias clave en la literatura del Análisis de Datos Funcionales (*Functional Data Analysis, FDA*).

Como ya se ha mencionado, muchas técnicas estadísticas tradicionales tienen dificultades para manejar los datos funcionales por su gran o infinita dimensión. En este contexto, se proponen técnicas para obtener representaciones finitas de datos funcionales. Los métodos de regularización y la expresión de los datos funcionales en una base ortonormal de funciones eliminan el ruido, reducen la dimensión y permiten usar métodos clásicos de Estadística Multivariante.

2.5.1. Método de regularización

En muchas ocasiones, a nivel práctico, cada dato funcional viene dado por un conjunto de observaciones $f_n = \{(x_i, y_i) \in X \times Y\}_{i=1}^n$, donde frecuentemente $X, Y \subset \mathbb{R}$. El número de datos que se puede registrar, n, es claramente finito, aunque si quisiéramos una descripción exacta de la función necesitaríamos un número infinito de observaciones. Por tanto, la primera tarea en la metodología de FDA suele ser transformar este conjunto de datos en una función $\overline{f} : X \longrightarrow Y$. El siguiente resultado garantiza la existencia de una función suave que aproxima a las observaciones muestrales.

Sea un conjunto compacto D en $\mathbb{R}$ en el que está contenida la malla $\{a_i\}_{i=1}^N$. Consideramos el espacio RKHS con rk k cuyas funciones parten de D y van a $\mathbb{R}$, que denotamos por $H_k(D, \mathbb{R})$. El *Teorema de Representación* (Cucker and Smale, 2001) asegura que para cada conjunto de valores $\{b_i\}_{i=1}^N$ existe una única aplicación $\overline{f} : \mathbb{R} \longrightarrow \mathbb{R}$ tal que:

$$\overline{f} = \underset{g \in H_k(D,\mathbb{R})}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} \|g(a_i) - b_i\|_{\mathbb{R}}^2 + \gamma \|g\|_{H_k}^2, \tag{2.11}$$

donde $\gamma > 0$ es un parámetro de regularización. De esta manera, partiendo de un conjunto de valores $\{(a_i, b_i)\}_{i=1}^N$, encontramos una función $\overline{f} \in$

$H_k(D, \mathbb{R})$ tan suave como queramos, que aproxima a estos puntos. La elección del kernel k en los métodos de regularización es un problema ampliamente estudiado en la literatura (Keerthi and Lin, 2003; Munoz and Moguerza, 2006).

En cambio, en nuestro trabajo, usamos este método con un objetivo distinto al habitual. La regularización suele utilizarse para transformar un conjunto de observaciones puntuales en una función suave. En cambio, nuestras observaciones muestrales ya son funciones en $H_k(D, \mathbb{R})$. Este resultado lo utilizamos pues, para conseguir que todas las funciones con las que trabajamos tengan una expresión similar, como se describe a continuación.

Si en el teorema anterior partimos de una función $f \in H_k(D, \mathbb{R})$ y consideramos $b_i := f(a_i)$, podemos obtener la función $\overline{f} \in H_k(D, \mathbb{R})$, tan suave como queramos (dependiendo de la elección de γ) que ajusta los datos, es decir, $\overline{f}(a_i)$ es cercano a $f(a_i)$.

Además, la única solución $\overline{f}$ a la Ec. (2.11) tiene la expresión

$$\overline{f}(\cdot) = \sum_{i=1}^{N} k(a_i, \cdot)(\beta_i), \tag{2.12}$$

donde los valores $\beta_i \in \mathbb{R}$ se han calculado resolviendo el siguiente sistema lineal:

$$(\gamma\, N\, \mathbb{I}_{N\times N} + K|_a)\beta = b,$$

con $K|_a$ la matriz definida por $K|_a(i,j) = k(a_i, a_j) \in \mathbb{R}$, $i, j = 1, \ldots, N$ y los vectores β y b son $\beta = (\beta_1, \beta_2, \ldots, \beta_N)'$; $b = (b_1, b_2, \ldots, b_N)'$.

Tras aplicar este teorema a todas las funciones muestrales $f_j \in H_k(D, \mathbb{R})$, conseguimos que todas tengan una misma expresión:

$$\overline{f_j}(\cdot) = \sum_{i=1}^{N} k(a_i, \cdot)\beta_{i,j} \in H_k(D, \mathbb{R}),\ j = 1, \ldots, m.$$

Esto será útil para aproximar los coeficientes de estas funciones expresadas en una base ortonormal, como veremos en el Capítulo 6.

2.5.2. Expresión en bases ortornormales

Como ya se ha hecho notar, en muchas ocasiones los datos funcionales se expresan a nivel práctico a partir de un conjunto de observaciones puntuales. Así como se utiliza el método de regularización para encontrar una función

suave que aproxima las observaciones, podemos transformar los datos a través de la expresión en una base de funciones (Cuevas, 2014).

Para ello, se suele asumir que la función de la que obtenemos las observaciones $\{(x_i, f(x_i) = y_i)\}_{i=1}^n$ está en un espacio general de funciones, por ejemplo, $f \in L^2$. Entonces, si $\{e_k\}$ es una base ortornormal del espacio, podemos aproximar los datos iniciales a través de una función $\overline{f}$. Fijamos un número J de elementos de la base, normalmente menor que n, y definimos

$$\overline{f} = \sum_{j=1}^{J} c_j e_j,$$

donde los coeficientes c_j se obtienen minimizando (Silverman and Ramsay, 2005)

$$\sum_{k=1}^{n} \left(f(x_k) - \sum_{j=1}^{J} c_j e_j(x_k) \right)^2 .$$

Así, conseguimos una representación de los datos con una dimensión menor (ya que J es menor que n) y además eliminamos ruido, pues $\overline{f}$ es una función suave que aproxima a f.

Las bases $\{e_k\}$ que se escogen más habitualmente son las funciones Fourier, Wavelets o B-splines (Silverman and Ramsay, 2005).

En el presente trabajo, contamos con un espacio muestral de funciones no discretizadas. Son elementos de un espacio RKHS definidos en los infinitos puntos. Además, a estas funciones ya se les ha aplicado el método de regularización para que todas tengan una expresión similar Ec. (2.12) y sean suaves. Así pues, necesitamos únicamente reducir la dimensión. Esta transformación la realizaremos expresando cada función en una base ortonormal formada por funciones propias de operadores integrales (ver Capítulos 6, 7), ordenada según una aproximación del criterio de optimalidad. De este modo, podremos trabajar con los primeros J coeficientes de las bases, como si de un problema multivariante se tratara.

2.6. Métodos estadísticos básicos

El objetivo de esta sección es recordar al lector métodos de Estadística Multivariante que se generalizarán en la Parte II para el estudio de la forma.

Sea $x_1, \ldots, x_m$ una muestra aleatoria de $X = (X_1, \ldots, X_d)^t$, vector aleatorio en $\mathbb{R}^d$.

2.6.1. Clasificación no supervisada: algoritmo k-medias

En este apartado se enuncian las características de un algoritmo de clasificación: el k-medias. El objetivo de este algoritmo es clasificar la muestra de m elementos en grupos, de forma que, por un lado, los elementos pertenecientes a un mismo grupo sean muy similares entre sí, y por otro, los objetos pertenecientes a grupos diferentes sean lo más distintos posibles, todo ello con respecto a las d variables analizadas. La similaridad en este contexto se mide a través de la distancia Euclídea. Es decir, partimos de un número prefijado de *clusters* o grupos k, se asigna cada objeto a uno y solo a un grupo y se busca una asignación que minimice la suma de las distancias entre todos los pares de elementos de cada grupo.

Dados $x_1, ..., x_m$ y una k-partición $\mathcal{C} = (C_1, \ldots, C_k)$ del conjunto $\mathcal{O} = \{1, \ldots, m\}$ de objetos subyacentes con clases no vacías, sea:

$$W(\mathcal{C}) = \sum_{i=1}^{k} \sum_{l \in C_i} \|x_l - \bar{x}_i\|^2,$$

donde $\bar{x}_i$ denota el centroide de los C_i (i-ésimo elemento de la partición).

Como es conocido, el clásico algooritmo de k-medias clasifica buscando una k-partición $\mathcal{C}$ de $\mathcal{O}$ en la que se alcance el mínimo valor de $W(\mathcal{C})$. Este problema de optimización uniparamétrico es equivalente al problema de optimización con dos parámetros:

$$W(\mathcal{C}, z_1, \ldots, z_k) = \sum_{i=1}^{k} \sum_{l \in C_i} \|x_l - z_i\|^2, \tag{2.13}$$

donde la minimización es también respecto a todos los vectores $Z = (z_1, \ldots, z_k)$ de k puntos $z_1, \ldots, z_k$ de $\mathbb{R}^d$ (clases representativas).

El algoritmo del k-medias trata de acercarse a una k-partición óptima iterando los siguientes pasos (Bock, 2007):

- **PASO 1.** Dada una particióon inicial $\mathcal{C}(0)$, obtenemos el vector centroide $Z(0) = (\bar{x}_1^0, \ldots, \bar{x}_k^0)$. Sea $i = 1$.
- **PASO 2.** Dado el vector centroide $Z(i-1)$, obtenemos $\mathcal{C}(i)$ minimizando Ec. (2.13) con respecto a $\mathcal{C}$, asignando cada punto a una clase cuyo centroide tiene la distancia Euclídea mínima a ello.
- **PASO 3.** Dado $\mathcal{C}(i)$, minimizamos Ec.(2.13) con respecto a Z, obteniendo el nuevo vector centroide $Z(i) = (\bar{x}_1^i, \ldots, \bar{x}_k^i)$, la media muestral.

- **PASO 4.** Sea $i = i + 1$ y volvemos al PASO 2 hasta que se alcance la convergencia.

Por construcción, este algoritmo mantiene una sucesión

$$Z(0), \mathcal{C}(1), Z(1), \mathcal{C}(2), Z(2), ...$$

de centroides y particiones con valores decrecientes de la función objetivo Ec. (2.13) que converge hacia el valor mínimo (habitualmente mínimo local).

El nuevo vector centroide que se obtiene en el PASO 3 de este algoritmo hace decrecer el valor de la función objetivo porque la media muestral minimiza la distancia Euclídea de cualquier elemento en el *cluster*.

El k-medias es un algoritmo de clasificación no supervisada, conviene aquí pues aclarar la diferencia entre clasificación supervisada y no supervisada. Los sistemas de clasificación supervisados son aquellos en los que, a partir de un conjunto de objetos clasificados (conjunto de entrenamiento), intentamos asignar una clasificación a un segundo conjunto de objetos. En contra, los sistemas de clasificación no supervisados son aquellos en los que no disponemos de una batería de objetos previamente clasificados, sino que únicamente a partir de las propiedades de los objetos intentamos dar una agrupación de los objetos según su similaridad.

En la siguiente sección se indica cómo funciona un algoritmmo de clasificación supervisada: análisis discriminante.

Ambos algoritmos se utilizan en la Parte II de aportaciones (Capítulos 5 y 6) para agrupar o hacer asignaciones de cuerpos geométricos según su forma y tamaño, utilizando *currents* y la teoría de FDA.

2.6.2. Clasificación supervisada: análisis discriminante lineal

En contraste con el algoritmo k-medias, el análisis discriminante es un procedimiento de clasificación supervisada. Supongamos que un conjunto de objetos está ya clasificado en una serie de grupos. El análisis discriminante tiene los siguientes objetivos:

- Elaborar una regla de decisión que permita asignar una nueva observación de origen desconocido, a uno de los grupos ya establecidos.

- Analizar qué variables de las consideradas influyen en la diferenciación de los grupos.

La idea básica del análisis discriminante, utilizando el método lineal de Fisher, consiste en proyectar las variables originales $X_1, \ldots, X_d$ en un subespacio de dimensión menor r, para clasificar los elementos de la muestra $x_1, \ldots, x_m$ en k grupos,

$$D_i = u_{i1}X_1 + \ldots + u_{id}X_d = u_i'X,\ i = 1, \ldots, r.$$

de forma que en este subespacio se consiga una buena separación entre poblaciones y donde $r = min(k-1, d)$, $corr(D_i, D_j) = 0$ para todo $i \neq j$. La decisión se toma a partir del valor que tomen estas funciones D_i, que vienen dadas por el vector de proyección u_i y se llaman funciones discriminantes lineales de Fisher o variedades canónicas.

Para obtener la primera variedad canónica, es decir, $D_1 = u_1'X$, buscamos el vector proyección u_1 que maximiza el cociente:

$$\frac{\text{variabilidad entre grupos}}{\text{variabilidad dentro de los grupos}}.$$

Usando las propiedades de las transformaciones lineales en la varianza y la media esto equivale a maximizar:

$$\varphi = \frac{u_1'Bu_1}{u_1'Wu_1}, \tag{2.14}$$

donde si denotamos por $\overline{x}$ la media muestral, n_g el tamaño del grupo g, $\overline{x}_g$ el vector medio de los elementos del grupo g y $x_j^{(g)}$ el elemento j-ésimo del grupo g, con $g = 1, \ldots, k$:

$B = \sum_{g=1}^{k}\sum_{j=1}^{n_g} n_g(\overline{x}_g - \overline{x})(\overline{x}_g - \overline{x})^t$ es la matriz de suma de cuadrados inter-grupos y

$W = \frac{1}{m-k}\sum_{g=1}^{k}\sum_{j=1}^{n_g} n_g(x_j^{(g)} - \overline{x}_g)(x_j^{(g)} - \overline{x}_g)^t$ es la matriz de suma de cuadrados intra-grupos

Lo que equivale (derivando e igualando a 0 la Ec. (2.14)) a buscar u_1 como el vector propio de $W^{-1}B$ asociado al mayor valor propio.

Para obtener la segunda variedad canónica, nos planteamos la misma maximización, pero exigiendo además que esta segunda variedad esté incorrelada con la primera. Esto se consigue al tomar el vector propio asociado al segundo valor propio más grande de $W^{-1}B$.

En general, si se quieren calcular r funciones discriminantes se obtienen como soluciones los r vectores propios de $W^{-1}B$ asociados a los r mayores valores propios de esta matriz $\lambda_1 \geq \lambda_2 \geq \ldots > 0$.

De este modo, el valor propio λ_i, $i = 1, \ldots, r$ mide el poder de discriminación de la i-ésima función discriminante. Decidimos a qué grupo pertenecerá la nueva observación en función de la distancia que hay entre las proyecciones del nuevo elemento y las proyecciones de los elementos de cada grupo. La nueva observación pertenecerá al grupo cuyas proyecciones sean más cercanas a las suyas.

2.6.3. Modelos lineales generalizados para regresión ordinal

Dada Y una variable aleatoria dependiente o respuesta y dadas $X_1, ..., X_d$ variables independientes (aleatorias o no), los modelos de regresión buscan modelizar la media de Y (μ) en función de $X_1, ..., X_d$.

Los modelos de regresión lineal asumen una distibución gaussiana para la variable dependiente y una relación lineal para la media. Los modelos lineales generalizados (GLMs) son la generalización natural de los modelos lineales. Los GLMs asumen que la variable dependiente tiene una distribución de probabilidad de la familia exponencial (no necesariamente la distribución Gaussiana), y se modeliza la relación entre los predictores $X_1, \ldots, X_d$ y la respuesta Y mediante una función link g como sigue:

$$g(\mu) = \alpha + \sum_{i=1}^{d} \beta_i X_i, \tag{2.15}$$

donde g puede ser cualquier función monótona diferenciable, y α y $\beta = (\beta_1, \ldots, \beta_d)$ son los parámetros a estimar.

Cuando la variable respuesta es categórica ordinal, es decir, cuando Y puede coger uno de los valores discretos ordenados indexados de $1, \ldots, J$ con probabilidades $\pi_1, \ldots, \pi_J$: $\sum_{j=1}^{J} \pi_j = 1$, esas probabilidades pueden ser modelizadas usando el modelo de *logits* acumulativos (Agresti, 2010):

$$logit[P(Y \leq j)] = \alpha_j + \sum_{i=1}^{d} \beta_i X_i, \qquad \forall j \in 1, \ldots, J-1. \tag{2.16}$$

Entonces,

$$P(Y \leq j) = \frac{\exp\left(\alpha_j + \sum_{i=1}^{d} \beta_i X_i\right)}{1 + \exp\left(\alpha_j + \sum_{i=1}^{d} \beta_i X_i\right)}, \qquad \forall j \in 1, \ldots, J-1.$$

Los GLMs se pueden generalizar al caso en el que las variables predictoras son funciones. En nuestro trabajo, a través de la herramienta de los *currents*, hemos caracterizado la forma y tamaño de cuerpos mediante funciones. Así, en el Capítulo 7 desarrollamos un modelo para predecir el ajuste de una prenda (variable categórica) según el contorno de cuerpos de niños y niñas (variable funcional).

2.7. Fórmulas rotacionales

La Geometría Integral es un campo de las Matemáticas cuyo objetivo general es la obtención de propiedades cuantitativas de objetos geométricos, y a la vez de probabilidades geométricas, a partir de la medida invariante de elementos geométricos como rectas, geodésicas, planos, conjuntos congruentes, movimientos, afinidades,... Por su parte, la Estereología trata de un conjunto de métodos para la exploración del espacio $n-$dimensional, cuando sólo es posible conocer secciones a través de elementos geométricos (rectas, geodésicas, planos,...) o sus proyecciones. Los principales métodos de la Estereología están estrechamente relacionados con la Geometría Integral (Santaló, 2004). En la última década se han desarrollado, dentro de la Geometría Integral y la Estereología, las denominadas fórmulas rotacionales, cuyo objetivo es la obtención de propiedades cuantitativas de objetos geométricos, a partir de secciones de estos objetos mediante subespacios (rectas, planos, subespacios totalmente geodésicos,...) que pasan por un punto fijo determinado O, (Gual-Arnau and Cruz-Orive, 2009; Gual-Arnau *et al.*, 2010; Auneau and Jensen, 2010; Thórisdóttir and Kiderlen, 2014; Thórisdóttir *et al.*, 2014; Cruz-Orive and Gual-Arnau, 2015; Gual-Arnau and Cruz-Orive, 2016)

Las motivaciones por desarrollar fórmulas rotacionales se remontan a algunos años atrás, por un lado en Tomografía Geométrica, existe una relación entre fórmulas obtenidas mediante proyecciones de un objeto y fórmulas obtenidas a partir de secciones a través de un punto de referencia, (Gardner,

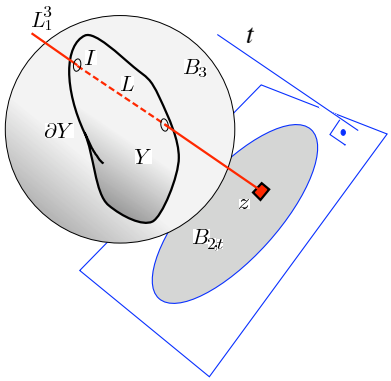
1995). Por otro lado, gracias a los nuevos avances en Microscopía, cuyas nuevas técnicas permiten obtener secciones ópticas virtuales a través de un punto de referencia, se ha desarrollado una rama de la Estereología, Estereología Local. El objetivo de la Estereología Local es la obtención de parámetros cuantitativos de estructuras espaciales a partir de secciones que pasan por un punto de referencia, como por ejemplo el nucleo de una célula, (Jensen, EBV, 1998).

Para explicar el modo de proceder en la obtención de una fórmula rotacional consideraremos que deseamos obtener el área de una superficie en $\mathbb{R}^3$, a partir de secciones mediante planos que pasan por un punto de referencia O. Sea por tanto $Y \subset \mathbb{R}^3$ un dominio compacto cuya frontera viene dada por la superficie diferenciable ∂Y. $S(\partial Y)$ denota el área de la superficie ∂Y.

El primer paso para la obtención de una fórmula rotacional para $S(\partial Y)$, a partir de planos que pasan por un punto fijo O, es disponer de una fórmula integral para $S(\partial Y)$. En este caso consideramos la clásica fórmula tipo Crofton Santaló (2004):

$$S(\partial Y) = \frac{1}{\pi} \int_{Y \cap L_1^3 \neq \emptyset} I(\partial Y \cap L_1^3) dL_1^3, \tag{2.17}$$

donde I denota el número de puntos de intersección y dL_1^3 es la medida de rectas en $\mathbb{R}^3$ invariante bajo la acción del grupo de movimientos en $\mathbb{R}^3$. Esta medida se obtiene fijando un plano $L_{2[0]}^3$ que pasa por un punto fijo O y por cada punto de este plano se considera la recta L_1^3 que pasa por este punto y es perpendicular al plano.



$$S(\partial Y) = \frac{1}{\pi} \int_{B_3 \cap L_1^3 \neq \emptyset} I\left(\partial Y \cap L_1^3\right) \mathrm{d}L_1^3$$

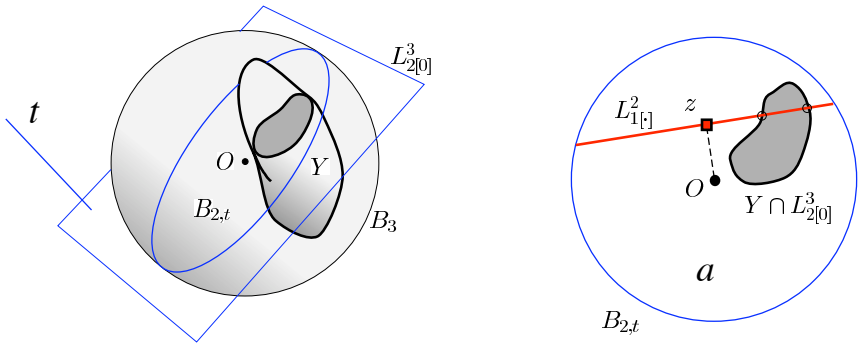
$$V(Y) = \frac{1}{2\pi} \int_{B_3 \cap L_1^3 \neq \emptyset} L\left(Y \cap L_1^3\right) \mathrm{d}L_1^3$$

Como a nosotros nos interesa "ver" lo que ocurre dentro del plano $L_{2[0]}^3$ y la recta L_1^3 es perpendicular al mismo, el segundo paso será considerar un

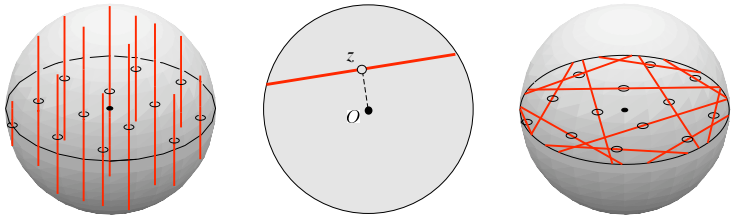
cambio de variables para expresar la medida dL_1^3 de una forma distinta; así obtenemos una nueva densidad:

$$dL_1^3 = \rho dL_1^2 dL_{2[0]}^3, \tag{2.18}$$

donde ahora L_1^2 se corresponde con una recta en el plano $L_{2[0]}^3$ y ρ es la distancia desde el punto O a la recta L_1^2.



Por tanto, como vemos en la siguiente figura, ahora en lugar de tener rectas perpendiculares al plano $L_{2[0]}^3$, estas rectas vienen determinadas dentro de este plano:



El tercer paso consiste en obtener la fórmula rotacional para el área $S(\partial Y)$, a partir de las ecuaciones Ec.(2.17) y Ec.(2.18):

$$\begin{aligned} \pi S(\partial Y) &= \int_{Y \cap L_{2[0]}^3 \neq \emptyset} \beta_1 dL_{2[0]}^3 \\ &= \int_{Y \cap L_{2[0]}^3 \neq \emptyset} \left(\int_{L_1^2} I((\partial Y \cap L_{2[0]}^3) \cap L_1^2)\, \rho dL_1^2 \right) dL_{2[0]}^3. \end{aligned} \tag{2.19}$$

El último paso consiste en dar interpretaciones geométricas de la función integrando β_1.

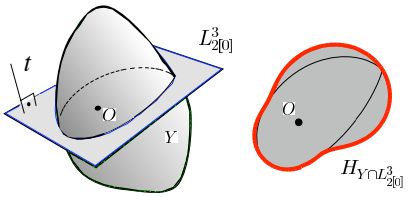
En la primera interpretación consideramos que Y es un dominio convexo, entonces, teniendo en cuenta la siguiente relación entre el elemento de área en $L^3_{2[0]} \sim \mathbb{R}^2$ y la medida invariante de rectas en $\mathbb{R}^2$, dL^2_1,

$$dz = \rho dL^2_1,$$

obtenemos que $I((\partial Y \cap L^3_{2[0]}) \cap L^2_1) = 2$ si z pertenece al conjunto soporte o conjunto pedal $H_{Y\cap L^3_{2[0]}}$ de $Y \cap L^3_{2[0]}$ y 0 en el resto; por tanto

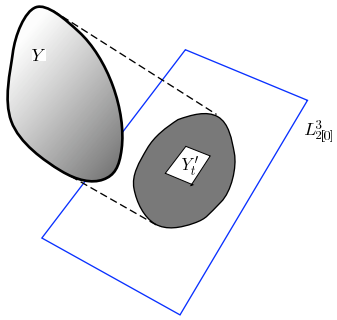
$$\beta_1 = \int_{L^2_1} I((\partial Y \cap L^3_{2[0]}) \cap L^2_1)\, \rho dL^2_1 = 2 \cdot \text{Area}(H_{Y\cap L^3_{2[0]}}).$$

De esta forma obtenemos una equivalencia entre fórmulas integrales mediante proyecciones de Y (fórmulas de Cauchy) y fórmulas mediante secciones de Y a través de planos $L^3_{2[0]}$ (fórmulas pivotales).



$S(\partial Y) = 4 \cdot \mathbb{E}\,\text{area}\left(H_{Y\cap L^3_{2[0]}}\right)$

PIVOTAL SECTION FORMULA



$S(\partial Y) = 4 \cdot \mathbb{E}\,\text{area}(Y'_t)$

CAUCHY'S PROJECTION FORMULA

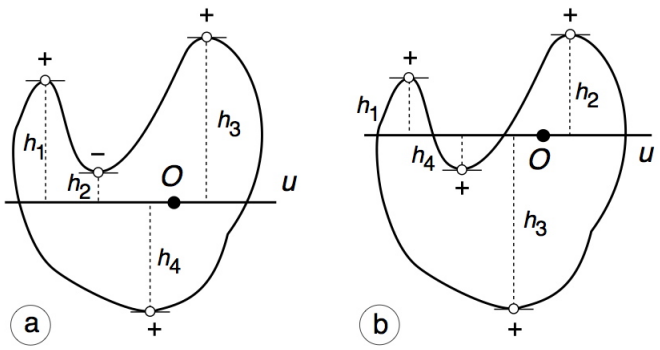
Para la segunda interpretación, consideraremos una representación de Morse de β_1 a partir de los puntos críticos de la función altura en $L^3_{2[0]}$ a partir del punto O; para ello utilizamos coordenadas polares para obtener la siguiente expresión:

$$\rho dL^2_1 = \rho\, d\rho\, d\phi.$$

Así, como fijado un ángulo ϕ, el valor de ρ con respecto a la función altura sólo cambia al atravesar un punto crítico obtenemos:

$$\begin{aligned} \beta_1 &= \int_0^\pi \left(\int_0^\infty I((\partial Y \cap L^3_{2[0]}) \cap L^2_1)\, \rho d\rho \right) d\phi \\ &= \int_0^\pi \left(\sum_0^m \epsilon_k \frac{\rho_k^2}{2} \right) d\phi, \end{aligned} \tag{2.20}$$

donde ρ_k es la distancia al $k-$ésimo punto crítico de la función distancia y ϵ_k es $+1$ o -1 dependiendo de si el punto crítico es máximo o mínimo.



En el Capítulo 8 de la tesis obtenemos fórmulas rotacionales para el área y para las integrales de curvatura media de la superficie frontera de un dominio compacto en un espacio de curvatura constante λ, M^n_λ (si $\lambda = 0$, $M^n_\lambda = \mathbb{R}^n$, si $\lambda > 0$, M^n_λ, es la esfera y si $\lambda < 0$, M^n_λ es el espacio hiperbólico).

Capítulo 3

Cuerpos caracterizados por campos vectoriales

El objetivo de este capítulo es describir cómo podemos caracterizar cuerpos geométricos mediante campos vectoriales utilizando los *currents*. Este procedimiento es fundamental en esta memoria, ya que, es trasversal a las tres primeras aportaciones que se detallan en la Parte II (Capítulo 5, 6 y 7). Las tres parten de cuerpos geométricos expresados como *currents* y desarrollan un modelo teórico para aplicar técnicas estadísticas a una base de datos de cuerpos. El procedimiento que veremos en este capítulo permite transformar los cuerpos en funciones vectoriales, de tal modo es posible generalizar algunas técnicas estadísticas al caso funcional utilizando la teoría de FDA.

3.1. Currents representados por campos vectoriales de un RKHS.

En esta sección estudiamos la extensión de los espacios *scalar-valued RKHS* (vista en la sección 2.4) al caso de funciones vectoriales. A continuación, a partir de la representación de Riesz y utilizando las propiedades del espacio, podremos expresar cada cuerpo, representado por un *current*, como un campo vectorial en un *vector-valued RKHS*.

3.1.1. Espacio test de campos vectoriales en el que integrar los currents.

Para caracterizar hipersuperficies, principalmente curvas en $\mathbb{R}^2$ y superficies en $\mathbb{R}^3$, medimos cómo varían las integrales que determinan los *currents*

(Ec. (2.3), (2.4) y (2.5)) cuando varían los campos vectoriales ω. Sin embargo, en vez de considerar todos los campos vectoriales, definiremos un espacio test de cuadrado integrable de donde tomaremos ω. En particular, escogeremos como espacio test un *vector-valued Reproducing Kernel Hilbert Space*.

Es importante notar que como consecuencia de la restricción del dominio de integración, $S \longrightarrow C_S$ deja de ser inyectiva, es decir, el mismo *current* geométrico C_S, como una aplicación definida en un RKHS, podría representar dos hipersuperficies diferentes. Por tanto, la elección de un RKHS como espacio test apropiado dependerá de la aplicación propuesta, pues en general, nos interesará diferenciar con claridad los *currents* que proceden de hipersuperficies distintas.

3.1.2. Kernels operator-valued y RKHS de campos vectoriales.

En este apartado, estudiamos la extensión de la teoría descrita en la sección 2.4 a espacios de funciones vectoriales, es decir, espacios de campos vectoriales de $\mathbb{R}^n$ a $\mathbb{R}^n$ que han sido ampliamente estudiados (ver Carmeli *et al.* (2006), Micchelli and Pontil (2005) and Caponnetto *et al.* (2008)).

Sea $\mathcal{L}(\mathbb{R}^n) := \{f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ campos vectoriales acotados$\}$ que es un espacio de Banach.

Definición 3.1.1. *Denotamos por W el espacio de Hilbert formado por campos vectoriales de $\mathbb{R}^n$ a $\mathbb{R}^n$. Definimos el operador K,*

$$\begin{aligned} K :&\mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathcal{L}(\mathbb{R}^n) \\ &(x, y) \longmapsto K(x, y) : \mathbb{R}^n \longrightarrow \mathbb{R}^n, \end{aligned}$$

donde $K(x, y)$ es un campo vectorial que actúa sobre $\alpha \in \mathbb{R}^n$, dando como resultado un elemento $K(x, y)(\alpha) \in \mathbb{R}^n$.
Diremos que K es un operator-valued reproducing kernel asociado a W si:

1. *$\forall x, \alpha \in \mathbb{R}^n$, $K(\cdot, x)(\alpha) \in W$ (con $K(\cdot, x)(\alpha)(y) = K(x, y)(\alpha) \in \mathbb{R}^n$, $\forall y \in \mathbb{R}^n$) y,*

2. *K satisface la "reproducing property"; es decir, $\forall \omega \in W$ y $x, \alpha \in \mathbb{R}^n$*

$$\omega(x) \cdot \alpha = \langle K(\cdot, x)(\alpha), \omega \rangle_W.$$

Definición 3.1.2. *Sea W el espacio de Hilbert de campos vectoriales de $\mathbb{R}^n$ a $\mathbb{R}^n$. W es un vector-valued RKHS si existe un operator-valued reproducing kernel (rk) asociado a W.*

Además,

Teorema 3.1.1. *El rk de un RKHS es un operador único, simétrico y definido positivo.*

A continuación, se define formalmente el concepto de *operator-valued kernel*:

Definición 3.1.3. *Un operador $K : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathcal{L}(\mathbb{R}^n)$ es definido positivo y autoadjunto si para cada par $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$, $K(x, y) \in \mathcal{L}(\mathbb{R}^n)$ es un operador autoadjunto y*

$$\sum_{i,j=1}^{N} \alpha_i \cdot K(x_i, x_j)(\alpha_j) \geq 0,$$

para cualquier conjunto finito de puntos $\{x_i\}_{i=1}^N$ en $\mathbb{R}^n$ y $\{\alpha_i\}_{i=1}^N$ en $\mathbb{R}^n$. En tal caso, diremos que K es un operator-valued kernel.

El siguiente teorema expresa la relación entre los *operator-valued reproducing kernel* y los *operator-valued kernel*, extensión del Teorema (2.4.2) dado en los *scalar-valued RKHS*.

Teorema 3.1.2. *Si K es un operator-valued kernel, entonces hay un único RKHS, W, tal que K es el operator-valued reproducing kernel (rk) asociado a W.*

La prueba del teorema se basa en la construcción de W a través de la completación de $H_0 := span\{K(\cdot, x)(\alpha) \ / \ x, \alpha \in \mathbb{R}^n\} \subset W$.

Si $\omega_1 = \sum_{i=1}^{N_1} K(\cdot, x_i)(\alpha_i)$, $\omega_2 = \sum_{j=1}^{N_2} K(\cdot, x_j)(\beta_j) \in H_0$, se define

$$\langle \omega_1, \omega_2 \rangle_{H_0} := \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \alpha_i \cdot K(x_i, y_j)(\beta_j). \tag{3.1}$$

El *vector-valued RKHS* W asociado al kernel K, que, desde ahora, se denotará H_K, es la clausura de H_0, es decir, el espacio generado por campos vectoriales de la forma $K(x, \cdot)(\alpha)$ para todo $x \in \mathbb{R}^n$ y $\alpha \in \mathbb{R}^n$ es denso en

H_K. Por esta razón, el producto interior en H_K ($\langle \cdot, \cdot \rangle_{H_K}$) entre $\omega_1, \omega_2 \in \overline{H_0}$ es el límite de la expresión Ec. (3.1) cuando N_1, N_2 tiende a infinito. Por tanto, la idea es construir el espacio vectorial como combinación lineal de los campos vectoriales de la forma $K(x, .)(\alpha)$ y completar este espacio añadiendo el límite de todas las sucesiones de Cauchy. Esta construcción hace posible procesar mallas discretas de superfícies y superfícies continuas (límite de una combinación finita) en el mismo espacio.

Un vez establecido el espacio test H_K, el dominio de un *current* queda restringido a H_K. Por tanto, el espacio de *currents* considerado es H_K^*. El espacio de *currents* H_K^* es un espacio vectorial con las operaciones suma ($+$) y producto ($\cdot$) como en un espacio estándard de funciones (Durrleman, 2010).

3.1.3. Curvas y superficies como elementos de un vector-valued RKHS

Si utilizamos las propiedades del RKHS, H_K, podemos reescribir los *currents* geométricos asociados a curvas en $\mathbb{R}^2$ y superficies en $\mathbb{R}^3$.

Sea ω un campo vectorial en $H_K(\mathbb{R}^2, \mathbb{R}^2)$. Entonces, usando la "reproducing property" y la Ec. (2.4), el *current* geométrico asociado a la curva α es:

$$C_\alpha^K(\omega) = \int_a^b \omega(\alpha(t)) \cdot \alpha'(t)\, dt = \langle \int_a^b K(\alpha(t), \cdot)(\alpha'(t)) dt, \omega \rangle_{H_K},$$

y por Ec. (2.5), el *current* geométrico asociado con la superficie parametrizada $S = r(U)$ es:

$$C_S^K(\omega) = \int_U \bar{\omega}(r(x)) \cdot \eta(x)\, dx = \langle \int_U K(r(x), \cdot)(\eta(x))\, dx, \omega \rangle_{H_K},$$

donde $x = (u, v)$, $dx = du\, dv$, $\eta(x) = r_u \wedge r_v$ es el vector ortogonal a la superficie S en el punto $r(x)$, y ω y $\int_S K(r(x), \cdot)(\eta(x)) dx \in H_K(\mathbb{R}^3, \mathbb{R}^3)$.

Hasta ahora, cada hipersuperficie ha sido asociada a un elemento en el espacio vectorial H_K^*. Sin embargo, el Teorema de Riesz Frèchet ((Conway, 2013)) establece que hay una aplicación isométrica, lineal y biyectiva $\mathcal{L}_{H_K} : H_K \to H_K^*$, definida por $\mathcal{L}_{H_K}(\omega)(\omega') := \langle \omega, \omega' \rangle_{H_K}$, $\forall \omega, \omega' \in H_K$. Como consecuencia, el espacio H_K^* es isométrico a H_K, y entonces cada hipersuperficie puede ser identificada con un campo vectorial en H_K.

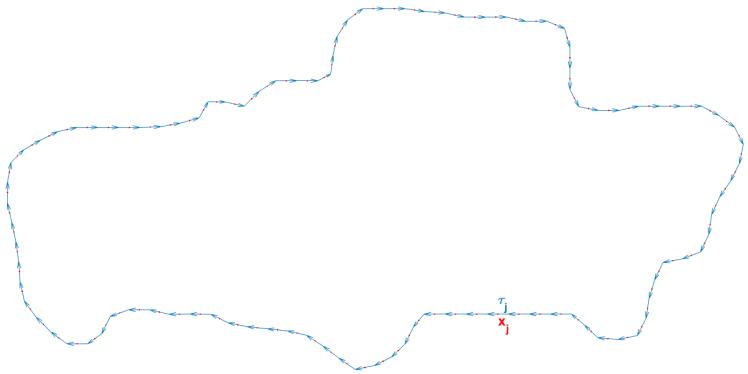


Figura 3.1: Curva en $\mathbb{R}^2$ con sus elementos

Por tanto, gracias al Teorema de Riesz-Frèchet, una curva parametrizada por α definida en $[a, b]$, se puede representar por un elemento en $H_K(\mathbb{R}^2, \mathbb{R}^2)$; es decir,

$$\alpha \longrightarrow C_\alpha^K(\omega) = \langle \int_a^b K(\alpha(t), \cdot)(\alpha'(t))dt, \omega \rangle_{H_K} \cong \int_a^b K(\alpha(t), \cdot)(\alpha'(t))dt,$$

donde $\cong$ denota el elemento isométrico que proporciona el Teorema de Riesz-Frèchet, y para representar una superficie parametrizada por S,

$$\begin{aligned} S \longrightarrow C_S^K(\omega) = \int_U \bar{\omega}(r(x)) \cdot \eta(x)\, dx &= \langle \int_U K(r(x), \cdot)(\eta(x))\, dx, \omega \rangle_{H_K} \\ &\cong \int_U K(r(x), \cdot)(\eta(x))\, dx. \end{aligned}$$

Consideremos ahora que la curva α solo es conocida en un número finito p, de puntos $\{t_1 < t_2 < \ldots, < t_p\}$, que son una partición del intervalo $[a, b]$. Sea $y_j = \alpha(t_j) \in \mathbb{R}^2\ \forall j = 1, \cdots, p$, y denotamos por $x_j \in \mathbb{R}^2$ el centro de los segmentos $[y_j, y_{j+1}]$ y por $\tau_j = y_{j+1} - y_j$ una aproximación del vector tangente (cuanto más fina es la partición, mejor es la aproximación). Entonces,

$$\alpha \longrightarrow C_\alpha^K(\omega) \cong \int_a^b K(\alpha(t), \cdot)(\alpha'(t))dt \sim \sum_{j=1}^{p} K(x_j, \cdot)(\tau_j).$$

En aplicaciones prácticas, cada curva α será representada por una suma finita que aproxima al campo vectorial.

Si suponemos que tenemos una triangularización de S donde cada triángulo $y_j y_{j+1} y_{j+2}$ está representado por el campo vectorial $K(x_j, \cdot)(\tau_j)$ donde

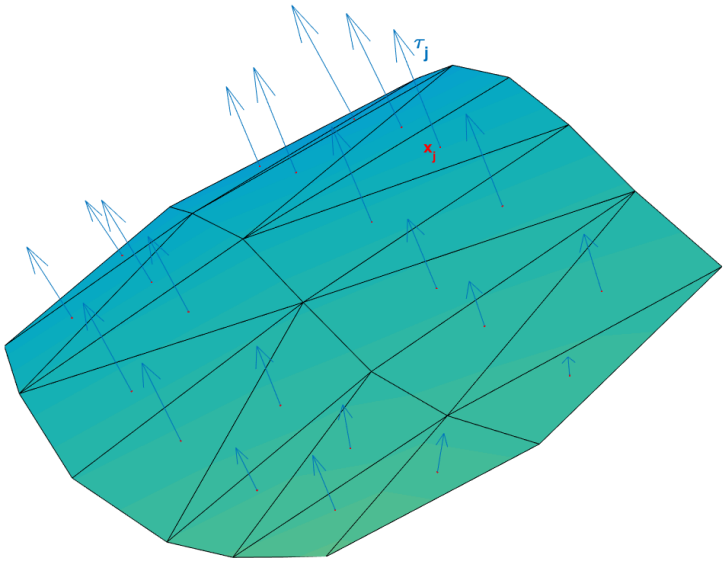


Figura 3.2: Superficie triangularizada en $\mathbb{R}^3$ con sus elementos

$x_j = \frac{1}{3}(y_j + y_{j+1} + y_{j+2})$ y $\tau_j = \frac{1}{2}(y_{j+1} - y_j) \wedge (y_{j+2} - y_j)$ (τ_j es el vector normal al triángulo, cuya norma codifica el área del triángulo); entonces,

$$S \longrightarrow C_S^K(\omega) \cong \int_U K(r(x), \cdot)(\eta(x))\, dx \sim \sum_{j=1}^{p} K(x_j, \cdot)(\tau_j).$$

La suma finita tiende hacia la integral conforme la malla se va haciendo más fina. Esta suma finita es una aproximación del campo vectorial $\int_S K(r(x), \cdot)(\eta(x))dx$, y será la representación que usaremos por su simplicidad computacional.

Como consecuencia, nosotros trabajamos con curvas y superficies como campos vectoriales en H_K. Entonces, la distancia entre dos superficies (o curvas) se define como la distancia entre los correspondientes elementos en H_K; es decir, si φ_1 y φ_2 son dos elementos en el RKHS asociados a las dos superficies S_1 y S_2; entonces,

$$\begin{aligned} d(S_1, S_2)^2 = d(\varphi_1, \varphi_2)^2 &= \langle \varphi_1 - \varphi_2, \varphi_1 - \varphi_2 \rangle_{H_K} \\ &= \langle \varphi_1, \varphi_1 \rangle_{H_K} - 2\langle \varphi_1, \varphi_2 \rangle_{H_K} + \langle \varphi_2, \varphi_2 \rangle_{H_K}. \end{aligned}$$

Usando las aproximaciones finitas a $\varphi_1 = \sum_{j=1}^{p} K(x_j^1, \cdot)(\tau_j^1)$ y $\varphi_2 = \sum_{j=1}^{p} K(x_j^2, \cdot)(\tau_j^2)$

tenemos

$$\langle \varphi_1, \varphi_2 \rangle_{H_K} = \sum_{i=1}^{p} \sum_{j=1}^{p} \tau_i^1 \cdot K(x_i^1, x_j^2)(\tau_j^2). \tag{3.2}$$

Además, la media muestral en el vector-valued RKHS H_K se calcula de la misma manera que en un espacio Euclídeo (Hsing and Eubank, 2015). Dada la muestra $\varphi_1, ..., \varphi_m$, la media muestral es:

$$\bar{\varphi} = \frac{\sum_{l=1}^{m} \varphi_l}{m} = \frac{\left(\sum_{l=1}^{m} \sum_{j=1}^{p} K(x_j^l, \cdot)(\tau_j^l) \right)}{m}, \tag{3.3}$$

$\bar{\varphi} \in H_K$; sin embargo, en general $\bar{\varphi}$ no es un *current* geométrico asociado a la superfície (detallado en sección 2.2.2).

La distancia entre superficies, obtenidas desde la distancia entre los correspondientes *currents* geométricos, da una estimación global de la diferencia de forma entre objetos. Esta distancia se usará, entre otras cosas, para la definición de un sistema de tallaje para la población infantil, se necesita una medida que marque las diferencias de forma globales de los cuerpos y no resalte las diferencias locales.

3.2. Elección de un operator-valued reproducing kernel

La elección del kernel determina el *vector-valued RKHS* y especialmente su métrica. La elección de esta métrica es crucial y se debe adaptar a toda aplicación particular. Basándonos en esto, usaremos un tipo particular de *operator-valued kernel* K que se define como sigue:

Definición 3.2.1. *Sea $D \subseteq \mathbb{R}^n$ un subconjunto no vacío, y $K : D \times D \to \mathcal{L}(\mathbb{R}^n)$ una aplicación. Entonces, para cada $x, y \in D$, el operador $K(x, y)$ se define como*

$$\begin{aligned} K(x,y): \quad & \mathbb{R}^n \longrightarrow \mathbb{R}^n \\ & \alpha \longrightarrow K(x,y)(\alpha) := k(x,y) \cdot \alpha, \end{aligned} \tag{3.4}$$

donde $k : D \times D \to \mathbb{R}$ es una función simétrica y definida positiva (función kernel).

Proposición 3.2.1. *El operator-valued K establecido en la definición anterior está bien definido, es simétrico y definido positivo, por lo que existe un único vector-valued RKHS $H_K(D, \mathbb{R}^n) \subset W := \{\omega : D \longrightarrow \mathbb{R}^n\}$ (o simplemente H_K) con K como su rk.*

Proof. Dado $(x, y) \in D \times D$, el operador $K(x, y)$ es obviamente lineal. También, $K(x, y)$ está acotado porque su norma está acotada por $|k(x, y)|$:

$$\|K(x,y)\| = \sup\{\|K(x,y)(\alpha)\|_{\mathbb{R}^n} : \alpha \in \mathbb{R}^n, \|\alpha\|_{\mathbb{R}^n} \leq 1\} \leq |k(x,y)|.$$

Además, como k es simétrico y definido positivo, K tiene inmediatamente esas propiedades. Entonces, K es un *operator-valued reproducing kernel* y hay un único *vector-valued RKHS* $H_K(D, \mathbb{R}^n)$ con K como su rk. □

A pesar de que no se conoce cómo elegir el "mejor" kernel para una aplicación dada, frecuentemente se usan kernels isotrópicos invariantes bajo traslaciones de la forma $k(x, y) = k(\|x-y\|_{\mathbb{R}^n})$. En particular, la función Gausiana (también llamado Kernel Gaussiano)

$$k(x,y) := e^{\frac{-\|x-y\|^2_{\mathbb{R}^n}}{\lambda^2}}. \tag{3.5}$$

donde $\lambda > 0$ es el parámetro escala (*bandwidth*), define un *operator-valued* $K : D \times D \to \mathcal{L}(\mathbb{R}^n)$ particularmente importante en la literatura, llamado *vector-valued Gaussian Kernel*, que fija un *vector-valued RKHS* $H_K(D, \mathbb{R}^n)$ con K como su rk. Este *vector-valued RKHS* tiene la siguiente expresión (Quang *et al.* (2010))

$$H_K(D,\mathbb{R}^n) = \{f \in C_0(D,\mathbb{R}^n) \cap L^2(D,\mathbb{R}^n) : \int_{\mathbb{R}^n} e^{\frac{\lambda^2\|\xi\|^2_{\mathbb{R}^n}}{4}} \|\widehat{f}(\xi)\|^2_{\mathbb{R}^n} d\xi < \infty\}, \tag{3.6}$$

donde $\widehat{f}$ es la transformada de Fourier de f.

Proposición 3.2.2. *Dado $D \subseteq \mathbb{R}^n$ un subconjunto no vacío; $K_\lambda : D \times D \to \mathcal{L}(\mathbb{R}^n)$ el operador definido como en Ec.(3.4) usando el Kernel Gausiano k con parámetro λ (Ec. (3.5)), y H_{K_λ} el RKHS establecido por K_λ, si $\lambda_1, \lambda_2 \in \mathbb{R} : 0 < \lambda_1 \leq \lambda_2$, entonces $H_{K_{\lambda_2}} \subseteq H_{K_{\lambda_1}}$.*

Proof. Si f estuviera en $H_{K_{\lambda_2}}$, usando la expresión del espacio Ec. (3.6), la integral con λ_2 sería finita y más grande que la integral con el parámetro λ_1,

$$\int_{\mathbb{R}^n} e^{\frac{\lambda_1^2\|\xi\|^2}{4}} \|\widehat{f}(\xi)\|^2 d\xi \leq \int_{\mathbb{R}^n} e^{\frac{\lambda_2^2\|\xi\|^2}{4}} \|\widehat{f}(\xi)\|^2 d\xi < \infty.$$

Entonces, $f \in H_{K_{\lambda_1}}$. En conclusión, cuando más pequeño es el valor de $\lambda > 0$, más grande es el espacio H_{K_λ} establecido. □

Nota 1. ***Elección del parámetro*** λ

λ es un parámetro que tiene un papel muy importante en los resultados y que debe ser escogido de acuerdo a los datos muestrales. Hay que tener en cuenta que por la Proposición (3.2.2), $0 < \lambda_1 \leq \lambda_2$ implica que $H_{K_{\lambda_2}} \subseteq H_{K_{\lambda_1}}$. Por tanto, cuanto menor es el valor de λ, más grande es el espacio H_{K_λ} establecido. Así, con un valor pequeño de λ será más probable identificar diferentes valores que toma el *current* en los campos vectoriales y entonces mejor se caracterizarán los datos geométricos. Pero, por esta misma razón, si λ es demasiado pequeño, la distancia en el espacio RKH detecta detalles geométricos minúsculos y se puede capturar demasiado ruido. En conclusión, es esencial elegir un parámetro adecuado haciendo balance entre las dos ideas previas.

Para ello en Durrleman *et al.* (2009) se propone tomar λ como la escala típica en la que los campos vectoriales $\omega \in H_K(D, \mathbb{R}^n)$ varían espacialmente. Esta idea junto con un método de validación cruzada será usada en las aplicaciones.

Capítulo 4

Estudio experimental

En este capítulo se describe la aplicación que motivó el inicio de nuestra investigación, así como la base de datos de cuerpos de niños y niñas escaneados. Los algoritmos desarrollados en las tres primeras aportaciones se llevan a cabo sobre esta base de datos, proponiendo aplicaciones innovadoras en los sistemas de tallaje y en la venta de ropa infantil online.

Además, en este capítulo se presentan las bases de datos experimentales que se utilizarán para comprobar la bondad de los procedimientos aportados teóricamente en la Parte II.

4.1. Proyecto: prendas de vestir para niños y niñas

Como se menciona en la introducción, el presente trabajo forma parte de un proyecto del Ministerio de Economía y Competitividad, que tiene como objetivo desarrollar herramientas para la selección de tallas de prendas de ropa infantil.

En el proceso actual de compra de ropa infantil online los consumidores basan la selección de talla en su experiencia previa o en la tabla de tallas que normalmente se incluye en la propia tienda online. Pero estos métodos son poco fiables y el resultado final es un elevado porcentaje de devoluciones de la ropa infantil vendida en tiendas online y la reticencia de muchos consumidores a comprar ropa infantil a través de este canal.

Con el fin de solucionar estos problemas, algunos sistemas comerciales ofrecen probadores virtuales (Olaru *et al.* (2012)), que mediante un avatar paramétrico (Fig. (4.1)) que representa al usuario, realizan simulaciones físicas para predecir la tensión del tejido de las prendas al ajustarlas sobre

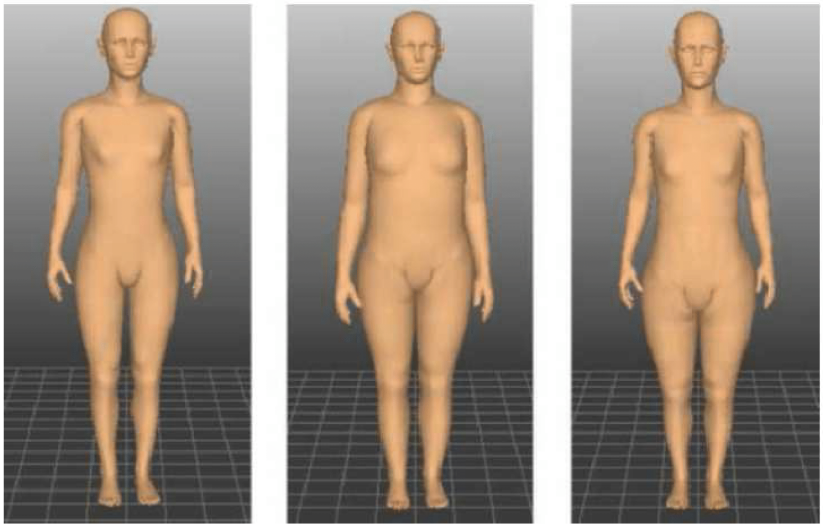


Figura 4.1: Configuración del avatar paramétrico.

la forma $3D$ del avatar. Este mapa de tensiones describe las zonas donde la prenda proporciona un ajuste holgado y las zonas donde se produce compresión.

Estas herramientas son muy sofisticadas pero resultan inservibles para los consumidores, en especial por su precisión. Por un lado, la configuración del avatar $3D$ que representa al usuario suele tener muy baja precisión, pues se ajusta con apenas tres o cuatro medidas del cuerpo tomadas en casa por el usuario, y por otro, la simulación física requiere la caracterización de los materiales textiles. Además, la configuración de la prenda en $3D$ que viste el avatar se alinea y ajusta al cuerpo a partir de patrones $2D$.

La selección en base a la antropometría del niño parece la aproximación más adecuada para predecir la talla y el ajuste de la prenda en la población infantil. Sin embargo, los sistemas de adquisición de la antropometría $3D$ disponibles en la actualidad presentan diversos inconvenientes. Los escáneres corporales 3D son demasiado caros para su uso doméstico o en puntos de venta y su uso se ha visto limitado a investigación. Los avatares paramétricos configurados a partir de medidas manuales del usuario presenta tres fuentes importantes de imprecisión: no está basado en estadística de poblaciones reales sino en modelos y proporciones que priman la estética, el número de

medidas introducido no depende del tipo de prenda seleccionada ni de las medidas críticas para el ajuste asociadas, y la toma de medidas la realiza un usuario no entrenado, con el error que ello puede conllevar. Los sistemas de bajo coste que utilizan tecnología doméstica para capturar medidas corporales todavía no han conseguido la precisión suficiente para la asignación de talla o la predicción del ajuste.

Por todos estos motivos, uno de los objetivos del proyecto que abordó el grupo del IBV, consistió en desarrollar un sistema de captura de la morfometría $3D$ del cuerpo que sea preciso, fácil de usar y que pueda realizarse en el hogar. El sistema reconstruye el cuerpo del niño o de la niña en $3D$ a partir de dos o tres fotografías tomadas con tecnología doméstica (smartphone, tablet o cámara digital), como se muestra en la Fig. (4.2). Se utilizan modelos representativos de la morfometría infantil Europea como base de reconstrucción. Pero, la reconstrucción $3D$ del cuerpo completo tiene un reto añadido: hay que eliminar la variabilidad de la postura de brazos y piernas en las bases de datos $3D$ que se utilizan para la reconstrucción y controlar o eliminar la postura en el proceso de captura. El IBV desarrolló metodologías para la generación de bases de datos $3D$ homólogas del cuerpo y la corrección de la postura mediante la integración de un maniquí cinemático. La aplicación de esta metodología requiere de una base de datos $3D$ representativa de la población infantil. El IBV dispone de bases de datos antropométricas $3D$ de población infantil Europea, que permiten llevar a cabo la reconstrucción.

Aunque el estudio antropométrico $3D$ del cuerpo humano ha experimentado un desarrollo científico- tecnológico importantísimo en los últimos años, principalmente de la mano del sector textil que ha impulsado la realización de estudios antropométricos nacionales en muchos países, la transferencia de este conocimiento a la industria ha sido mínima.

Otro de los objetivos del proyecto consistió en desarrollar nuevas metodologías estadísticas que permitan la predicción del ajuste y la asignación de talla óptima que maximice el confort y la vida útil de la prenda a partir de métodos de aprendizaje en grandes conjuntos de datos. Estas metodologías deben estar basadas en la forma $3D$ del cuerpo del niño o de la niña, los gustos/hábitos del comprador y en los modelos estadísticos de la forma del cuerpo infantil y de su crecimiento.

En el presente trabajo afrontamos el problema de asignación de talla y el de la predicción de ajuste en los Capítulos 6 y 7 respectivamente.

Además, observamos que los sistemas de tallaje actuales no ofrecen un buen ajuste a la mayoría de la población infantil. Los procesos de diseño ergonómico y la definición de tamaño de la ropa para niños y niñas tienen varias diferencias con respecto a la vestimenta para adultos. En primer lugar,

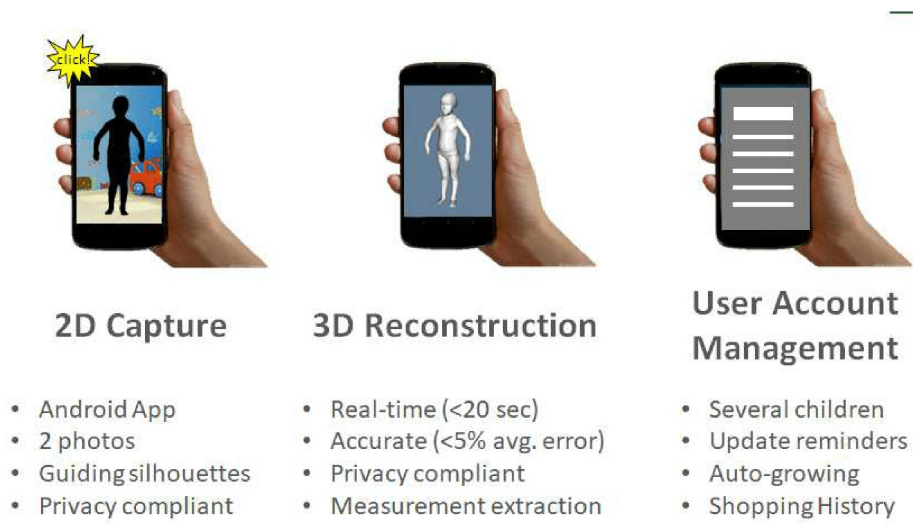


Figura 4.2: Sistema de captura de la morfometría $3D$ desde el hogar.

la designación del tamaño de la ropa infantil generalmente se etiqueta en años, que no es una medida del cuerpo, por lo que generalmente se relaciona con una altura corporal concreta por edad, que no necesariamente tiene que ser cercana a un niño de esa edad, debido a la alta variabilidad de altura por edad en niños. De acuerdo con la norma europea UNE-EN 13402-3, el rango de edades entre 3 y 12 años tiene 10 tamaños diferentes asociados (950-1010 mm, 1010-1070 mm, 1070-1130 mm, 1130-1190 mm, 11901250 mm, 1250-1310 mm, 1310-1370 mm, 1370-1430 mm, 1430-1490 mm, 1490-1550 mm).

Cuando se quiere comprar una camiseta a un niño o a una niña, se ha de comprar la talla asociada a su altura. Es importante notar que esta talla se ha diseñado para una forma específica de cuerpo, sin embargo hay una gran variabilidad de formas de cuerpos de niños con una misma altura. Por ejemplo, si al niño asociado a una altura no le ajusta una talla por la forma de su cuerpo, tendrá que comprar la talla anterior o la siguiente de camiseta, la cual será probablemente demasiado corta o demasiado larga para él. En conclusión, es esencial crear un nuevo sistema de tallaje que tenga en cuenta la forma y tamaño del cuerpo humano. Resolvemos este problema en la aportación que aparece desarrollada en el Capítulo 5.

Como comprar ropa en Internet es un problema tanto para el cliente como

para la industria del sector textil, en los últimos años, tanto las administraciones nacionales como los grupos industriales de este sector, han fomentado las encuestas antropométricas nacionales en diferentes países. En este sentido y para conseguir los objetivos anteriormente comentados, el Instituto de Biomecánica de Valencia realizó en 2004 un estudio antropométrico $3D$ de la población infantil en España. Este estudio se describe en la siguiente sección.

4.2. Base de datos de escáneres de niños y niñas

En esta sección se detalla la base de datos del proyecto. En el estudio que se menciona en el apartado anterior, 739 niños y niñas españoles de entre 3 y 12 años fueron elegidos aleatoriamente y escaneados usando el escaner corporal *Vitus Smart 3D de Human Solutions*, un sistema de láser no intrusivo formado por cuatro columnas que albergan el sistema óptico, el cual se mueve desde la cabeza hasta los pies en segundos, realizando un barrido de todo el cuerpo.

Cada niño se situó de pie mirando hacia adelante; y se almacenó la forma del cuerpo como un conjunto de 3075 $3D$-*landmarks* (puntos homólogos en su superficie). A los niños se les pidió que vistieran unas prendas estándares para estandarizar las medidas tomadas (ver Fig. (4.3)).

Utilizando una malla $3D$, se calcularon muchas medidas antropométricas semi-automáticamente. Para cada niño o niña, los *landmarks* observados fueron suficientes para definir una superficie triangularizada, suave y orientada en $\mathbb{R}^3$, con un total de 4668 triángulos.

La Fig. (4.4) muestra la superficie triangularizada que define el escáner del cuerpo de un niño. En esta imagen no aparecen manos, pies ni cabeza, pues en nuestro estudio se han eliminado estas partes del cuerpo por no ser relevantes en referencia a tallas de prendas de vestir.

En un estudio realizado posteriormente un experto probó a una submuestra de 76 niños y niñas una prenda de tres tamaños consecutivos de un mismo modelo de camiseta: la talla supuestamente correcta, la siguiente y la anterior, codificando la bondad de ajuste de la prenda en una variable ordinal: pequeña, buen ajuste o grande. Esta será la submuestra utilizada en la tercera aportación de la tesis (Capítulo 7). Puesto que en este estudio se hace referencia a la talla de camiseta, se descarta también la parte del escáner correspondiente a las piernas (Fig. (4.5)), en este caso se obtiene una superficie triangularizada con 2766 triángulos.

De este modo, en cualquiera de los casos, el contorno del cuerpo de cada

Figura 4.3: Niño siendo escaneado por el escáner corporal *Vitus Smart 3D de Human Solutions*.

Figura 4.4: Superficie triangularizada a partir del escáner de un niño de la base de datos.

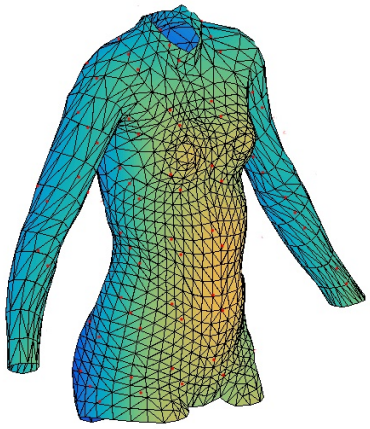


Figura 4.5: Superficie triangularizada a partir del escáner de un niño de la base de datos (parte superior del cuerpo).

niño o niña se representa por una superficie suave triangularizada. Para el $j-$ésimo triángulo de la superficie del $k-$ésimo niño de la base de datos, se calcula su baricentro x_j^k y el vector normal a la superficie en x_j^k, τ_j^k. Entonces, según la teoría desarrollada en el Capítulo 3, el contorno del cuerpo de un niño queda asociado al campo vectorial $\varphi_k = \sum_{j=1}^{p} K(x_j^k, \cdot)(\tau_j^k)$ en H_K, donde p es el número de triángulos de la superficie ($p = 4668$ o 2766 dependiendo del caso considerado).

4.3. Bases de datos sintéticas

4.3.1. Base 2D

En esta sección, describimos la base de datos de figuras sintéticas llamada MPEG7CEShape-1 PartB. Está formada por figuras binarias agrupadas en categorías como coches, caras, relojes, caballos y pájaros con imágenes correspondientes al mismo grupo, pero presentando formas diferentes.

Para nuestro trabajo, consideramos tres clases de figuras sintéticas de la base de datos: coches, caras y relojes. Cada clase contiene 20 elementos (imágenes binarias) excepto la clase de los relojes en la que descartamos dos de ellos (reloj-2 es un elemento atípico porque es muy grande y reloj-8 está girado y nuestra estructura teórica considera forma y tamaño). Las figuras

se centraron y el contorno α_k de cada una de ellas define una curva suave orientada la cual se discretizó mediante 100 puntos $\{y_j^k\}_{j=1}^{100}$ ($y_1^k = y_{100}^k$) para $k = 1, \ldots, 58$. Además, las figuras con forma de cara se giraron 90 grados, para mantener la orientación horizontal común a todas las figuras sintéticas de la base de datos (estableciendo la correspondencia con la base de datos de los niños, donde todos los elementos son registrados y tienen la misma posición). Para cada $k \in \{1, \ldots, 58\}$, de $\{y_j^k\}_{j=1}^{100}$, definimos $x_j^k = \frac{1}{2}(y_j^k + y_{j+1}^k)$ los centros de los segmentos y los vectores $\tau_j^k = y_{j+1}^k - y_j^k$, $\forall j = 1, \ldots, 99$, que definen los campos vectoriales $\sum_{j=1}^{99} K(x_j^k, \cdot)(\tau_j^k)$ in $H_K(D, \mathbb{R}^2)$.

En este estudio experimental, estamos interesados en analizar situaciones similares a las que aparecerán en nuestra aplicación real (descrita en sección 4.2). Por lo que consideramos dos escenarios. En el primero, las 58 figuras sintéticas se expandieron o contrajeron para que tuvieran la misma longitud en el eje X, estableciendo la semejanza entre esta longitud y la altura de un niño.

La Fig. (4.6) muestra un ejemplo de un objeto de cada clase en esta primera situación. Los puntos $\{x_j^k\}_{j=1}^{99}$ están marcados en negro y los vectores $\{\tau_j^k\}_{j=1}^{99}$ de cada curva aparecen pintados en diferentes colores.

En el segundo escenario, la mitad de las figuras sintéticas de cada categoría se agrandaron por el factor escala 1.5. En este caso, hay dos "alturas" diferentes para cada clase de figuras. Además, cada figura de la muestra se multiplicó por un coeficiente aleatorio entre 1 y 1.1, para cambiar un poco la "altura" de las figuras

La Fig. (4.7) muestra un ejemplo de un objeto de cada grupo en la segunda situación, en la que hay dos "alturas" para cada clase de formas.

4.3.2. Base 3D

En esta sección se describe con detalle la base experimental de figuras $3D$, utilizada en nuestras aportaciones. Con ella se quieren probar las mismas situaciones que con la base $2D$. Ahora consideramos tres clases de objetos $3D$: elipsoides, esferas y peras, que fueron generados por MATLAB. Cada clase contiene 10 elementos centrados en el origen y el contorno de cada uno de ellos está definido por una superficie triangularizada suave y orientada en $\mathbb{R}^3$. Cada superficie k está definida por 10000 triángulos con baricentro x_j^k y vectores normales τ_j^k, $j = 1, \ldots, 10000$, siendo $\sum_{j=1}^{10000} K(x_j^k, \cdot)(\tau_j^k)$ en $H_K(D, \mathbb{R}^3)$

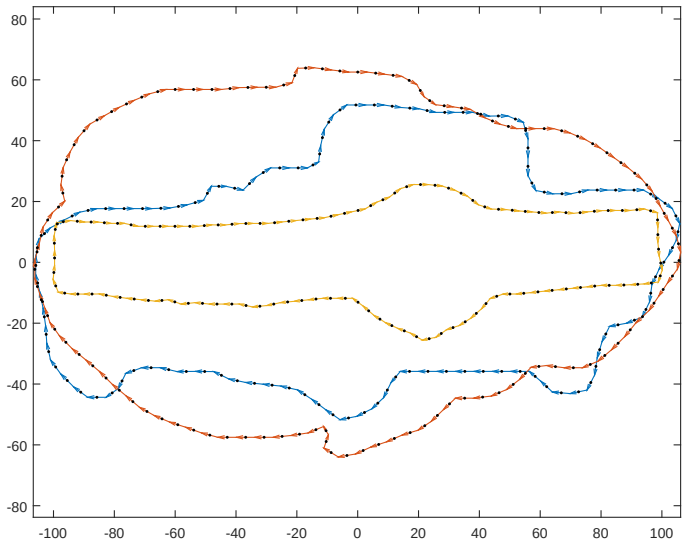


Figura 4.6: Un objeto de cada clase de la base de datos experimental $2D$ (escenario 1).

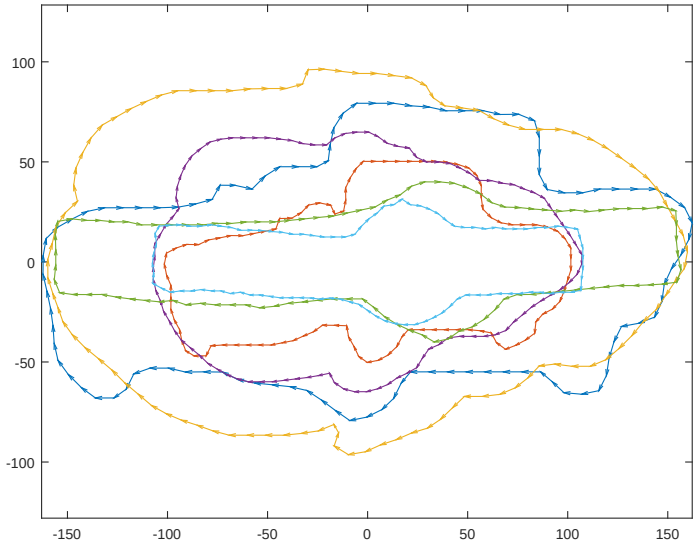


Figura 4.7: Un objeto de cada clase de la base de datos experimental $2D$ (escenario 2).

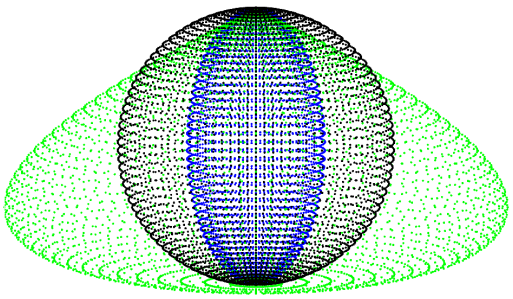


Figura 4.8: Un objeto de cada clase de la base de datos experimental $3D$ (escenario 1).

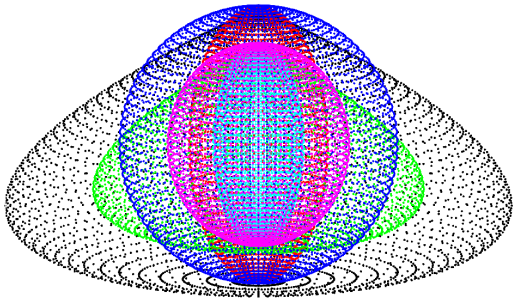


Figura 4.9: Un objeto de cada clase de la base de datos experimental $3D$ (escenario 2).

el campo vectorial asociado a cada superficie.

Consideramos los mismos escenarios que en la base de datos $2D$. En el primero, Fig. (4.8), todas las figuras tienen aproximadamente la misma longitud en el eje Y (misma "altura").

La Fig. (4.9) muestra el segundo escenario, en el cual hay dos "alturas" para cada clase de objetos $3D$.

De este modo, el primer escenario nos permite experimentar con una muestra de cuerpos geométricos que tienen la misma "altura". En ella centramos la atención en las distintas formas de los objetos. En cambio, en el segundo escenario no se seleccionan objetos de la misma "altura", sino que se trabaja con cuerpos con distinta forma y "altura". Estos escenarios los encontramos en los Capítulos 5 y 6 al trabajar con la base de datos real de niños y niñas escaneados.

Parte II

Aportaciones

Capítulo 5

Clasificación no supervisada: propuesta de nuevos sistemas de tallaje

Unsupervised classification of children's bodies using currents. **Sonia Barahona**, Ximo Gual Arnau, M. Victoria Ibáñez y Amelia Simó. Publicado en *Advances in Data Analysis and Classification*, páginas 1-33, año 2017. DOI: 10.1007/s 11634-017-0283-0.

El objetivo de esta aportación es proponer una nueva metodología para clasificar en grupos homogéneos cuerpos geométricos según su forma y tamaño. En particular, se aplicará a la clasificación de las superficies de los cuerpos de los niños de nuestra base de datos (sección 4.2) con el fin de proponer nuevos sistemas de tallaje infantiles que proporcionen un buen ajuste. Para ello, caracterizamos la superficie de cada niño mediante un *current* geométrico (Ec. (2.2)) y utilizando el desarrollo detallado en el Capítulo 3 conseguimos expresar cada cuerpo como un campo vectorial en un *vector-valued RKHS*.

En nuestro caso, recordemos que se usa el *operator-valued* kernel K, que se define a continuación.

Sea $D \subseteq \mathbb{R}^3$ un conjunto no vacío y $K : D \times D \to \mathcal{L}(\mathbb{R}^3)$ una aplicación, donde $\mathcal{L}(\mathbb{R}^3) := \{f : \mathbb{R}^3 \longrightarrow \mathbb{R}^3 \text{ campos vectoriales acotados}\}$. Entonces, para cada $x, y \in D$, el operador $K(x, y)$ se define

$$\begin{aligned} K(x,y): \quad & \mathbb{R}^3 \longrightarrow \mathbb{R}^3 \\ & \alpha \longrightarrow K(x,y)(\alpha) := k(x,y) \cdot \alpha \end{aligned}$$

donde $k : D \times D \to \mathbb{R}$ es una función simétrica y semidefinida positiva, (es

decir, $k(x,y) = k(y,x)\ \forall x, y \in D$ y $\sum_{i,j} a_i k(x_i, x_j)(a_j) \geq 0$ para conjuntos finitos $\{a_i\} \in \mathbb{R}$, $\{x_i\} \in D$).

En la aplicación práctica, se escoge como k la función Gaussiana

$$k_G(s,t) := e^{\frac{-\|s-t\|^2_{\mathbb{R}^n}}{\lambda^2}},$$

donde $\lambda > 0$ es un parámetro de escala, de modo que usando el Teorema (3.1.2) queda fijado el espacio *vector-valued RKH* en el que trabajaremos: $H_K(D, \mathbb{R}^3)$.

En este espacio, si consideramos la superficie del k-ésimo niño de la base de datos triangularizada, recordemos que el campo vectorial que la representa es

$$\varphi_k = \sum_{j=1}^{4668} K(x_j^k, \cdot)(\tau_j^k) \in H_K(D, \mathbb{R}^3), \tag{5.1}$$

donde x_j^k es el baricentro y τ_j^k es el vector normal al triángulo en x_j^k.

El procedimiento por el que podemos expresar cuerpos geométricos como campos vectoriales en un *vector-valued RKHS*, es parte del desarrollo teórico de esta aportación, pero aparece detallado en el apartado de Fundamentos (Capítulos 2 y 3) por su importancia en esta memoria, ya que las dos siguientes aportaciones también utilizan este proceso.

Este es uno de los aspectos novedosos de este trabajo, ya que, los espacios *vector-valued RKH* no han sido trabajados tan ampliamente como los espacios RKH escalares. Con la caracterización desarrollada, en el espacio RKH podemos calcular la media de cuerpos (Ec. (3.3)) y la distancia entre ellos (Ec. (3.2)), por lo que es posible llevar a cabo una clasificación no supervisada de datos geométricos en función de su forma y tamaño. Para ello, utilizamos el algoritmo k-medias. Es sencillo generalizar el algoritmo k-medias del caso multivariante (sección 2.6.1) al caso de campos vectoriales en un espacio de Hilbert.

Sea la muestra $\varphi_1, \varphi_2, \ldots, \varphi_m$ de campos vectoriales en $H_K(D, \mathbb{R}^3)$ de la forma de la Ec. (5.1) y queremos dividirla en k grupos de modo que, por un lado, los elementos pertenecientes a un mismo grupo sean muy similares entre sí, y por otro, los objetos pertenecientes a grupos diferentes sean lo más distintos posibles. La similaridad en este contexto se mide a través de la distancia d del espacio $H_K(D, \mathbb{R}^3)$ (Ec. (3.2)). Es decir, partiendo de un

número pre-fijado de *clusters* o grupos k, se asigna cada objeto a uno y solo a un grupo y se busca una asignación que minimice la suma de las distancias entre todos los pares de elementos de cada grupo.

Dada una k-partición $\mathcal{C} = (C_1, \ldots, C_k)$ del conjunto $\mathcal{O} = \{1, \ldots, m\}$, sea:

$$W(\mathcal{C}) = \sum_{i=1}^{k} \sum_{l \in C_i} d(\varphi_l - \bar{\varphi}_i)^2,$$

donde $\bar{\varphi}_i$ denota el centroide de los C_i y se calcula mediante la Ec. (3.3).

Este problema de optimización uniparamétrico en el que se pretende encontrar una k-partición de la muestra que alcance el mínimo valor de $W(\mathcal{C})$, es equivalente al problema de optimización con dos parámetros:

$$W(\mathcal{C}, \phi_1, \ldots, \phi_k) = \sum_{i=1}^{k} \sum_{l \in C_i} d(\varphi_l - \phi_i)^2, \tag{5.2}$$

donde la minimización es también respecto a todos los vectores $Z = (\phi_1, \ldots, \phi_k)$ de k puntos $\phi_1, \ldots, \phi_k$ de $H_K(D, \mathbb{R}^3)$ (clases representativas).

Los pasos a iterar del algoritmo son los mismos que en el caso multivariante (sección 2.6.1).

- **PASO 1.** Dada una partición inicial $\mathcal{C}(0)$, obtenemos el vector centroide $Z(0) = (\bar{\varphi}_1^0, \ldots, \bar{\varphi}_k^0)$. Sea $i = 1$.
- **PASO 2.** Dado el vector centroide $Z(i-1)$, obtenemos $\mathcal{C}(i)$ minimizando Ec. (5.2) con respecto a $\mathcal{C}$, asignando cada punto a una clase cuyo centroide tiene la distancia Euclídea mínima a ello.
- **PASO 3.** Dado $\mathcal{C}(i)$, minimizamos Ec.(5.2) con respecto a Z, obteniendo el nuevo vector centroide $Z(i) = (\bar{\varphi}_1^i, \ldots, \bar{\varphi}_k^i)$, la media muestral.
- **PASO 4.** Sea $i = i + 1$ y volvemos al PASO 2 hasta que se alcance la convergencia.

La sucesión de particiones con valores decrecientes de la función objetivo Ec. (5.2) converge hacia el valor mínimo.

Del mismo modo que en el caso multivariante, el nuevo vector centroide que se obtiene en el PASO 3 de este algoritmo hace decrecer el valor de la función objetivo porque la media muestral minimiza la distancia d_{H_K} de cualquier elemento en el *cluster*.

Una vez adaptado el algoritmo de clasificación al espacio de Hilbert, se hace un estudio experimental utilizando la base sintética *2D*, descrita en la sección 4.3.1. En el caso en el que los cuerpos a caracterizar son curvas, la adaptación es muy similar y se expresan como elementos en $H_K(D, \mathbb{R}^2)$ como sigue:

$$\sum_{j=1}^{p} K(x_j, \cdot)(\tau_j),$$

donde los puntos y_j, $j = 1, \cdots, p$ son una partición de la curva; x_j el centro de los segmentos $[y_j, y_{j+1}]$ y $\tau_j = y_{j+1} - y_j$ una aproximación del vector tangente.

En la experimentación, partimos de una base de datos en la que se conoce cuáles son los grupos, ya que, quedan determinados por las formas de las curvas: caras, coches y relojes y pretendemos observar qué errores de clasificación se obtienen cuando aplicamos el algoritmo k-medias propuesto a los dos escenarios planteados en la sección 4.3.1. Además, queremos ver cómo afecta a la métrica la variación del parámetro λ del Kernel Gaussiano, incluso el cambio del kernel por otro kernel isotrópico de traslación invariante: Kernel Epanechnikov. En la aportación, mostramos los resultados obtenidos tras aplicar el k-medias calculando los porcentajes de errores de clasificación. Se concluye que el criterio teórico que se había descrito para la elección del parámetro λ del Kernel Gaussiano (nota 1) es coherente con lo observado experimentalmente, y se consiguen clasificaciones con 0 % de error para estos valores. Además, si utilizamos el Kernel de Epanechnikov también se determina un intervalo en el que si tomamos el parámetro de este kernel, la clasficación que hace el k-medias es correcta. Para visualizar con mayor inmediatez los grupos obtenidos tras aplicar el algoritmo k-medias, se utiliza la técnica del *Multidimensional Scaling (MDS)* sobre las matrices de distancias. A cada clase de figuras se le asigna un color y nos interesan los parámetros que hagan que los indicadores de un mismo color queden lo más agrupados posible entre sí y lo más alejados de otros grupos.

Se hace de nuevo un estudio experimental, pero en este caso utilizando la base sintética 3D (descrita en la sección 4.3.2). Se observa que el método desarrollado es poco sensible a las diferentes densidades de triángulos que pueden presentar las superficies frontera de los cuerpos geométricos a clasificar: peras, esferas y elipsoides.

Finalmente, se aplica el desarrollo teórico a la base de datos real de cuerpos de niños y niñas escaneados, descrita en la sección 4.2. Como hay distintos

sistemas de tallaje para cada sexo, elegimos un subconjunto de la muestra formado por 195 niñas mayores de 6 años, con el fin de mostrar nuestro procedimiento. De acuerdo con los estándares europeos UNE-EN 13402-3, este rango de edades tiene asociadas 4 tallas distintas, que quedan determinadas únicamente por los siguientes intervalos de alturas: $1190 - 1250$ mm, $1250 - 1310$ mm, $1310 - 1370$ mm, $1370 - 1430$ mm.

Por ello, se proponen dos sistemas de tallaje como mejora al ajuste de las prendas de vestir infantiles en la actualidad, utilizando el algoritmo k-medias, que tendrá en cuenta tanto forma como altura de la niña. Cada uno de los sistemas de tallaje propuestos corresponde a uno de los escenarios planteados con las bases de datos experimentales (sección 4.3).

El primero de los sistemas de tallaje, se obtiene al dividir en dos grupos (según la forma) cada uno de los rangos de altura que sugiere como talla la norma UNE-EN 13402-3. Para describir las tallas de este sistema de tallaje, se calcula la media de las medidas antropométricas (altura, pecho, cintura, cadera) de cada grupo como se muestra en la Tabla (5.1).

Talla	Altura	Contorno de pecho	Contorno de cintura	Contorno de cadera	Tamaño del grupo
T1 chica pequeña (1190-1250)	1209	604	538	656	29
T2 chica grande (1190-1250)	1227	670	610	739	15
T3 chica pequeña (1250-1310)	1273	643	563	688	31
T4 chica grande (1250-1310)	1282	696.5	643.5	767	20
T5 chica pequeña (1310-1370)	1331	644	564.5	701	32
T6 chica grande (1310-1370)	1346	733	669	807	23
T7 chica pequeña (1370-1430)	1393	677	586	750	31
T8 chica grande (1370-1430)	1410.5	797.5	722	856.5	14

Tabla 5.1: Primer sistema de tallaje propuesto para alturas comprendidas entre 1190 y 1430 mm.

A pesar de que el sistema de tallaje propuesto ajusta a un alto porcentaje de la población, puede que no interese a las empresas textiles, porque cuanto mayor sea el número de tallas, mayor es el coste. Por ello, tratando de reducir el número de tallas, proponemos un segundo sistema de tallaje, que se consigue tras aplicar el algoritmo k-medias directamente a la muestra de 195 niñas agrupando según forma y tamaño (sin hacer previamente la criba por alturas). Para decidir el número de grupos (tallas) en que queda dividida la muestra se utilizan las técnicas *silueta* y el *criterio del codo* que determinan $k = 5$. Así, se obtiene el segundo sistema de tallaje, que se describe en la Tabla (5.2) usando las medias de las medidas antropométricas de cada grupo:

Talla	Altura	Contorno de pecho	Contorno de cintura	Contorno de cadera	Tamaño del grupo
T1 niña	1241	610	540	661	57
T2 niña	1259	678	622	754	39
T3 niña	1362	660	571.5	723.5	56
T4 niña	1361.5	747.5	673	814	34
T5 niña	1417	832	767	903	9

Tabla 5.2: Segundo sistema de tallaje propuesto para alturas comprendidas entre 1190 y 1430 mm.

En el trabajo que aparece a continuación también mostramos de manera más visual los resultados, utilizando *MDS* y diagramas de caja de las medidas antropométricas para cada grupo, en cada uno de los sistemas de tallaje propuestos.

Capítulo 6

Clasificación supervisada: asignación de talla

Classification of geometrical objects by integrating currents and functional data analysis. An application to online sales of children's wear. **Sonia Barahona**, Pablo Centella, Ximo Gual Arnau, M. Victoria Ibáñez y Amelia Simó. Sometido a *Applied Stochastic models in business and industry.* arXiv: 1707.02147v1 [stat.ME] 7 Julio 2017.

En esta aportación se propone adaptar y generalizar las metodologías del análisis estadístico de datos funcionales (sección 2.5) a nuestro tipo particular de datos en un *vector valued RKHS*. En particular, en este trabajo se adecúa el algoritmo de análisis discriminante (sección 2.6.2) a un conjunto muestral de objetos geométricos (cuerpos) caracterizados mediante *currents*.

Como ejemplo de aplicación, usaremos esta metodología para asignar a cada niño o niña la talla correspondiente de una determinada prenda de vestir a partir de las formas *3D* de sus cuerpos completos en la base de datos de la sección 4.2. Una vez definido un sistema de tallaje que ajusta adecuadamente a una gran parte de la población infantil, surge la necesidad de proporcionar herramientras que ayuden a la selección de talla en las compras online de ropa para niños y niñas.

Gracias al desarrollo teórico del Capítulo 3, la forma y tamaño del cuerpo de un niño quedan representados a través de un campo vectorial en un espacio *vector-valued RKHS*, donde el *operator-valued* kernel K se define como sigue. Sea $D \subseteq \mathbb{R}^3$ un conjunto compacto no vacío y $(x, y) \in D \times D$,

$$K(x,y) := k(x,y)\,\mathbb{I}_{3\times 3} = e^{-\frac{\|x-y\|^2_{\mathbb{R}^3}}{\lambda^2}}\,\mathbb{I}_{3\times 3},$$

con $\mathbb{I}_{3\times 3}$ la matriz identidad y $\lambda > 0$ es el parámetro escala.

En este espacio, $H_K(D,\mathbb{R}^3)$, si consideramos la superficie del k-ésimo niño de la base de datos triangularizada, recordemos que el campo vectorial que la representa es

$$\varphi_k = \sum_{j=1}^{4668} K(x_j^k, \cdot)(\tau_j^k) \in H_K(D,\mathbb{R}^3), \tag{6.1}$$

donde x_j^k es el baricentro y τ_j^k es el vector normal al triángulo en x_j^k.

Para poder utilizar la teoría de Análisis de Datos Funcionales y adaptar el algoritmo del análisis discriminante funcional a este espacio, surgen algunas dificultades que han de ser superadas previamente y que exponemos a continuación.

En primer lugar, utilizamos la teoría de regularización detallada en sección 2.5.1, para expresar todos nuestros datos muestrales $\varphi_k \in H_K(D,\mathbb{R}^3)$, $k = 1,\dots,m$ de manera similar. Es importante recordar que, aunque la regularización suele utilizarse para transformar un conjunto de observaciones puntuales en una función suave, nuestras observaciones muestrales ya son funciones en $H_K(D,\mathbb{R}^3)$. Por tanto, en nuestro caso, partimos de un campo vectorial $\varphi_k \in H_K(D,\mathbb{R}^3)$ expresado como en Ec. (6.1) con puntos y vectores de la superficie de un niño y consideramos la malla $\{a_i\}_{i=1}^N \subset D$. Si denotamos $b_{k,i} := \varphi_k(a_i)$, $b_{k,i} = (b_{k,i}^1, b_{k,i}^2, b_{k,i}^3) \in \mathbb{R}^3$, podemos aplicar el Teorema de Representación, de manera que existe un único campo vectorial $\overline{\varphi_k} \in H_K(D,\mathbb{R}^3)$ tal que:

$$\overline{\varphi_k} = \underset{g\in H_K(D,\mathbb{R}^3)}{\arg\min} \frac{1}{N}\sum_{i=1}^{N} \|g(a_i) - b_i\|_{\mathbb{R}}^2 + \gamma\|g\|_{H_K}^2, \tag{6.2}$$

donde $\gamma > 0$ es un parámetro de regularización. De este modo, podemos aproximar φ_k por un campo vectorial $\overline{\varphi}_k \in H_K(D,\mathbb{R}^3)$, tan suave como queramos (dependiendo de la elección de γ) que ajusta los datos, es decir, $\overline{\varphi_k}(a_i)$ es cercano a $\varphi_k(a_i)$.

Además, la única solución $\overline{\varphi_k}$ a la Ec. (6.2) tiene la expresión

$$\overline{\varphi_k}(\cdot) = \sum_{i=1}^{N} K(a_i, \cdot)(\beta_{k,i}),$$

donde los vectores $\beta_{k,i} \in \mathbb{R}^3$ se obtienen resulviendo este sistema matricial:

$$(\gamma\, N\, \mathbb{I}_{N\times N} + K|_a)\beta_k = b_k,$$

con $K|_a$ la matriz definida como $K|_a(i,j) = k(a_i, a_j) \in \mathbb{R}$, $i,j = 1,\ldots,N$, y β_k, b_k son las siguientes matrices $N \times 3$

$$\beta_k = \begin{pmatrix} \beta_{k,1}^1 & \beta_{k,1}^2 & \beta_{k,1}^3 \\ \beta_{k,2}^1 & \beta_{k,2}^2 & \beta_{k,2}^3 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \beta_{k,N}^1 & \beta_{k,N}^2 & \beta_{k,N}^3 \end{pmatrix};\ b_k = \begin{pmatrix} b_{k,1}^1 & b_{k,1}^2 & b_{k,1}^3 \\ b_{k,2}^1 & b_{k,2}^2 & b_{k,2}^3 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ b_{k,N}^1 & b_{k,N}^2 & b_{k,N}^3 \end{pmatrix}.$$

Tras aplicar este teorema a todos los campos vectoriales muestrales $\varphi_k \in H_K(D,\mathbb{R}^3)$ que caracterizan a los niños de la base de datos (detallada en la sección 4.2), conseguimos que todos tengan una misma expresión:

$$\overline{\varphi}_k(\cdot) = \sum_{i=1}^{N} K(a_i,\cdot)\beta_{i,k} \in H_K(D,\mathbb{R}^3),\ k = 1,\ldots,m.$$

A continuación, buscamos una base ortonormal del espacio *vector-valued RKH* $H_K(D,\mathbb{R}^3)$ donde proyectar nuestros datos. Así, truncaremos estas proyecciones de manera óptima para obtener una aproximación finita que nos permitirá usar el método del análisis discriminante multivariante (sección 2.6.2).

Consideramos el espacio de Hilbert,

$$L^2(D,\mathbb{R}^3) := \{f = (f_1,f_2,f_3)\colon D \longrightarrow \mathbb{R}^3 \,/\, \|f\|^2_{L^2(D,\mathbb{R}^3)} = \sum_{i=1}^{3} \int_D |f_i(x)|^2\, dx < \infty\}$$

y el operador integral, $L_{K,D}\colon L^2(D,\mathbb{R}^3) \longrightarrow L^2(D,\mathbb{R}^3)$, definido por

$$L_{K,D}f(x) = \int_D K(x,y)(f(y))\, dy = \left(\int_D k(x,y)f_i(y)\, dy\right)_{i=1}^{3} = (L_k f_i(x))_{i=1}^3.$$

Sean $\{\lambda_l\}_{l=1}^{\infty}$ los valores propios del operador integral L_k y las correspondientes funciones propias $\{\phi_l\}_{l=1}^{\infty}$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ y $\lim_{l\to\infty} \lambda_l = 0$, por el Teorema Espectral (2.3.8).

Si ϕ_l es una función propia de L_k con valor propio asociado λ_l, y denotamos $\psi_l^1 = (\phi_l,0,0)$, $\psi_l^2 = (0,\phi_l,0)$, $\psi_l^3 = (0,0,\phi_l)$, entonces ψ_l^j es una función propia de $L_{K,D}$ asociada al mismo valor propio ($j = 1,2,3$).

Teorema 6.0.1. *Sea* $\overline{\varphi_k}(\cdot) = \sum_{i=1}^{N} K(\cdot, a_i)(\beta_{k,i})$ *un campo vectorial que representa una superficie* S_k *en* $\mathbb{R}^3$. *Entonces,* $\overline{\varphi_k}$ *se puede expresar como*

$$\overline{\varphi_k}(\cdot) = \sum_{j=1}^{3} \sum_{l=1}^{\infty} \mu_{l,k}^{j} \left(\sqrt{\lambda_l} \psi_l^j(\cdot) \right), \tag{6.3}$$

donde $\{\{\sqrt{\lambda_l}\psi_l^j\}_{l=1}^{\infty}\}_{j=1}^{3}$ *es una base ortonormal para* $H_K(D, \mathbb{R}^3)$.

Además, los primeros $d = \operatorname{rank}(K|_a)$ *coeficientes* $\mu_{l,k}^j$ *se pueden aproximar por*

$$\widehat{\mu_{l,k}^j} = \sqrt{\ell_l}(v_l \cdot \beta_k^j)$$

para $j = 1, 2, 3$, *donde* $v_l \in \mathbb{R}^N$ *son los vectores propios de* $K|_a$, ℓ_l *son los valores propios de* $K|_a$ *y* $\beta_k^j = (\beta_{k,1}^j, \beta_{k,2}^j, \ldots, \beta_{k,N}^j)$.

Una vez expresado cada campo vectorial con respecto a la base ortonormal dada por la descomposición de funciones propias del kernel K, es importante señalar que los elementos de la base están ordenados siguiendo un criterio de optimalidad.

Es decir, si fijamos un número entero $r > 0$:

$$\min_{f_1,\ldots,f_r \in H_K(D,\mathbb{R}^3)} \int \int_{D\times D} \left(K(y,x) - \sum_{j=1}^{3} \sum_{l=1}^{\infty} f_l^j(y) f_l^j(x) \right)^2 dydx = \sum_{j=r+1}^{\infty} \lambda_j^2,$$

donde el mínimo se alcanza por $\sum_{l=1}^{r} \psi_l^j(y)\psi_l^j(x)$. Esto significa que como nuestros datos funcionales son de la forma $\sum_{j=1}^{3}\sum_{l=1}^{\infty} \mu_{l,k}^j \left(\sqrt{\lambda_l}\psi_l^j(x)\right)$, y la descomposición truncada de valores y vectores propios son la mejor aproximación de K, al truncar esta representación se reduce la dimensión de manera óptima.

Entonces, si truncamos el segundo sumatorio en la Ec. (6.3) a $d = \operatorname{rank}(K|_a)$, cada superficie S_k para $k = 1, \ldots, m$, viene dada por los coeficientes $\mu_{l,k}^j$ para $j = 1, 2, 3$ y $l = 1, \ldots, d$ (estimados por $\widehat{\mu_{l,k}^j}$), en la base ortonormal $\{\{\sqrt{\lambda_l}\psi_l^j\}_{l=1}^{\infty}\}_{j=1}^{3}$. Por tanto, S_k se puede expresar como un vector $(3 \cdot d)$-dimensional

$$\mu_k = (\widehat{\mu_{1,k}^1}, \widehat{\mu_{1,k}^2}, \widehat{\mu_{1,k}^3}, \widehat{\mu_{2,k}^1}, \widehat{\mu_{2,k}^2}, \widehat{\mu_{2,k}^3}, \ldots, \widehat{\mu_{d,k}^1}, \widehat{\mu_{d,k}^2}, \widehat{\mu_{d,k}^3}). \tag{6.4}$$

Esta expresión reduce de manera óptima el problema infinito-dimensional a uno finito-dimensional y así podemos aplicar el método clásico multivariante del análisis discriminante lineal (sección 2.6.2) a los vectores μ_k,

$k = 1, \ldots, m$ de la Ec. (6.4). En nuestro caso utilizaremos el algoritmo implementado en MATLAB (2015).

El procedimiento descrito que permite asociar la superficie de un niño o una niña con un vector, está desarrollado en la aportación para una hipersuperficie genérica en $\mathbb{R}^n$. Es decir, podemos aplicar este método a curvas en $\mathbb{R}^2$ ($n = 2$) y también a superficies en $\mathbb{R}^3$ ($n = 3$), como ya ha sido detallado.

Para analizar la robustez del método desarrollado, llevamos a cabo una validación cruzada sobre las bases sintéticas $2D$ y $3D$ (descritas en las secciones 4.3.1 y 4.3.2 respectivamente). Lo que pretendemos observar con estas bases de datos que tienen grupos claramente diferenciados según la forma y el tamaño, es que si dejamos un elemento de la base fuera y le pedimos al método que le asigne el grupo correspondiente lo hace de manera adecuada. Por tanto, aplicando la validación cruzada (*leave-one-out cross-validation)*, podemos evaluar el método mediante el porcentaje de error. Llevamos a cabo el proceso de validación cruzada con MATLAB (2015), considerando los distintos escenarios descritos en las secciones 4.3.1 y 4.3.2. En todas las situaciones planteadas, obtenemos un error del 0 % tomando los parámetros adecuados (λ del Kernel Gaussiano escogido según el criterio del Capítulo anterior y γ parámetro de regularización). De este estudio experimental también se deduce que el parámetro γ tiene poca influencia en los resultados de clasificación.

Finalmente, se aplica la clasificación supervisada a la base de datos de niños y niñas del proyecto (descrita en la sección 4.2). En particular, se trabaja con la misma muestra de 195 niñas que se utiliza en el Capítulo 5, ya que se consideran los grupos (tallas) definidos por los sistemas de tallaje propuestos en ese Capítulo. Así, se estudia a través de la validación cruzada si a un nuevo individuo se le asignaría la talla adecuada. Los resultados son bastante prometedores. Los errores son menores del 10 % en cada rango de altura del primer sistema de tallaje propuesto, ya que, en este hay un filtro previo según la altura. En el segundo sistema de tallaje obtenemos un error del 16,92 %.

Capítulo 7

Regresión ordinal: predicción de ajuste de una prenda

Generalized linear models for geometrical currents predictors. An application to predict the garment fit. **Sonia Barahona**, Pablo Centella, Ximo Gual Arnau, M. Victoria Ibáñez y Amelia Simó. Sometido a *Annals of Applied Statistics*. arXiv: 1803.03019v1 [stat.AP] 8 Marzo 2018.

En esta aportación se construye un modelo teórico y práctico que podrá utilizarse para la predicción del ajuste de una prenda a un niño o una niña. Esto permitiría implementar una aplicación web que haga posible seleccionar la talla correcta de una prenda sin la necesidad de probársela.

Como en las aportaciones anteriores, caracterizamos el contorno del cuerpo del niño o niña mediante un *current* y usamos el desarrollo del Capítulo 3 para trabajar en un espacio *vector-valued RKH*. Así, el contorno del cuerpo de un niño queda asociado al campo vectorial:

$$\varphi_k = \sum_{j=1}^{2766} K(x_j^k, \cdot)(\tau_j^k),$$

donde los puntos x_j^k y los vectores τ_j^k son diferentes de una superficie a otra. Como se describe en el Capítulo 6, todos estos campos vectoriales se representan en el mismo grid de puntos $\{a_i\}_{i=1}^N$ elegidos en un conjunto compacto $D \subset \mathbb{R}^3$, por lo que cada φ_k se aproxima por una función suave en $H_K(D, \mathbb{R}^3)$:

$$\overline{\varphi}_k = \sum_{j=1}^{N} K(a_j, \cdot)(\beta_j^k),$$

evaluada en este grid común. Esta aproximación permite aplicar a este tipo de datos la metodología del Análisis de Datos Funcionales (FDA).

El objetivo particular es obtener un modelo para predecir el ajuste de una prenda a un niño como "demasiado pequeña" ($Y = -1$), "buen ajuste" ($Y = 0$) o "demasiado grande" ($Y = 1$), en función de la talla de la prenda, la edad del niño, el sexo y la superficie del cuerpo del niño caracterizada por $\overline{\varphi}_k$. Para ello, usaremos técnicas de Análisis de Datos Funcionales para adaptar a nuestro contexto un Modelo Lineal Generalizado de Regresión Ordinal (sección 2.6.3).

Los GLM (sección 2.6.3) asumen que los predictores tienen dimensión finita, James (2002) amplió los GLM a modelos lineales generalizados funcionales (FGLM), sustituyendo la suma del espacio de dimensión finita (Ec. (2.15)) por una integral en L^2. En el caso particular de un modelo de respuesta ordinal Y con J categorías basado en *logits acumulados*, la Ec. (2.16) quedaría:

$$logit[P(Y \leq j)] = \alpha_j + \int \sigma(t)X(t)\,dt, \forall j \in \{1, \ldots, J-1\},$$

donde $\sigma(\cdot)$ es el análogo funcional a β_i.

En nuestro caso, al estar trabajando con predictores funcionales en $H_K(D, \mathbb{R}^3)$, el producto interior en L^2 se sustiuye por el producto interior en $H_K(D, \mathbb{R}^3)$ y el modelo quedaría:

$$logit[P(Y \leq j)] = \alpha_j + \langle \sigma, \overline{\varphi} \rangle_{H_K}, \forall j \in 1, \{1, \ldots, J-1\}. \tag{7.1}$$

Como no es posible estimar directamente las funciones paramétricas, la solución más utilizada en FDA consiste en considerar que las observaciones funcionales pertenecen al espacio generado por una base de funciones (Silverman and Ramsay, 2005). En la literatura se han utilizado diferentes bases: funciones splines, funciones trigonométricas o funciones wavelets. Un enfoque un poco distinto pero muy popular consiste en usar la base ortonormal de funciones propias de la función covarianza (FPCA). En la práctica, los desarrollos en la base quedan truncados y se aproximan por sumas finitas, con lo cual el modelo se convierte en un GLM múltiple y podemos aplicar las técnicas clásicas.

En nuestro caso, $\sigma(\cdot)$ y $\overline{\varphi}(\cdot)$ pertenecen a $H_K(D, \mathbb{R}^3)$, por eso, si $\{\phi_l(\cdot)\}_{l=1}^{\infty}$ es una base del espacio:

$$\overline{\varphi}(x) \cong \sum_{l=1}^{r} c_l \phi_l(x); \qquad \sigma(x) \cong \sum_{l=1}^{r} b_l \phi_l(x). \tag{7.2}$$

De las ecuaciones Eqs. (7.1) y (7.2), obtenemos:

$$logit[P(Y \leq j)] = \alpha_j + \sum_{p=1}^{r}\sum_{l=1}^{r} b_p c_l \langle \phi_p, \phi_l \rangle_{H_K}, \forall j \in 1, \{1, \ldots, J-1\}$$

donde $\{\alpha_j\}_{j=1}^{J-1}$ y $\{b_l\}_{l=1}^{r}$ son los parámetros a estimar.

Como nuestros datos funcionales son campos vectoriales en un RKHS, proponemos tres bases alternativas a las habituales para expresar los campos vectoriales $\varphi(\cdot)$ respecto a ellas.

En primer lugar proponemos usar la base ortonormal dada por el operador integral $L_K \colon L^2 \longrightarrow L^2$ definido por el kernel reproductivo K de $H_K(D, \mathbb{R}^3)$:

$$L_K f(x) := \int_D K(x, y)(f(y))\, dy.$$

Así, si $\{\lambda_l\}_{l=1}^{\infty}$ son los valores propios del operador y $\{\psi_l\}_{l=1}^{\infty}$ las correspondientes funciones propias, usando el Teorema de Mercer (2.3.9), $\rho_l = \sqrt{\lambda_l}\psi_l$, $\{\rho_l\}_{l=1}^{\infty}$ es una base ortonormal de H_K y el campo vectorial $\overline{\varphi}_k$ como elemento de H_K se puede expresar como sigue (Hsing and Eubank, 2015):

$$\overline{\varphi}_k = \sum_{l=1}^{\infty} \langle \overline{\varphi_k}, \rho_l \rangle_{H_K} \rho_l = \sum_{l=1}^{\infty} \mu_l^k \rho_l. \tag{7.3}$$

Como $\overline{\varphi}_k(x) = \sum_{i=1}^{N} K(a_i, x)(\beta_i^k)$, la generalizacíón del Teorema (2.3.10) al caso vector-valued asegura que el truncamiento de Ec. (7.3) reduce la dimensión de manera óptima. En la práctica, los coeficientes respecto de esta base se estiman usando la matriz aproximación al kernel (procedimiento visto en el Capítulo 6).

La segunda propuesta consiste en adaptar a nuestro escenario, una estructura más habitual: la propia del operador integral definido usando las funciones covarianza. El hecho de que un elemento aleatorio de $H_K(D, \mathbb{R}^3)$ sea un proceso estocástico (sección 2.4.3), permite considerar las funciones covarianza $\gamma_{ij}(x, y) := Cov(\Phi_i(x), \Phi_j(y)), \forall i, j = 1, 2, 3$. Sea $\gamma(x, y)$ la (3×3)-matriz cuyos elementos son $\gamma_{ij}(x, y)$. Entonces, $\Gamma(x, y)(\alpha) := \gamma(x, y)\alpha$ es una función *vector-valued* simétrica y definida no negativa. Consideramos el operador integral L_Γ:

$$L_\Gamma f(x) := \int_D \Gamma(x, y)(f(y))\, dy.$$

Usando el Teorema Espectral (2.3.8), obtenemos las funciones propias de este operador, $\{v_l\}_{l=1}^{\infty}$ asociadas a los valores propios $\{\ell_l\}_{l=1}^{\infty}$. $\{v_l\}_{l=1}^{\infty}$ forma una base ortonormal para L^2.

$$\overline{\varphi}_k = \sum_{l=1}^{\infty} \langle \overline{\varphi}_k, v_l \rangle_{L^2} v_l = \sum_{l=1}^{\infty} \varsigma_l^k v_l. \tag{7.4}$$

Para poder usar este resultado necesitamos demostrar que estos vectores forman también una base de H_K y obtener los productos (en H_K) de los vectores de la base. Se demuestra facilmente que es una base puesto que:

- $H_K \subset L^2$
- Usando de nuevo el Teorema de Mercer, $\{\sqrt{\ell_l} v_l\}_{l=1}^{\infty}$ forman una base ortonormal de H_Γ y asumiendo que Γ es continuo entonces $H_\Gamma \subset H_K$, (extensión al caso vectorial de la Ec.(2.9)) y por lo tanto $v_j \in H_K$.

En cuanto a los productos, se demuestra que:

$$\langle v_i, v_j \rangle_{H_K} = \sum_{k=1}^{\infty} \lambda_k^{-1} \langle v_i, \psi_k \rangle_{L^2} \langle v_j, \psi_k \rangle_{L^2}. \tag{7.5}$$

De nuevo en la práctica los coeficientes ς_l^k se estiman usando la matriz que aproxima la función covarianza.

El truncamiento de la expresión en la Ec. (7.4), también es óptimo, ya que, existen variables $I_\Phi(v_l)$ con media cero, varianzas decrecientes e incorreladas tal que (Hsing and Eubank, 2015):

$$\lim_{n \to \infty} \sup_{t \in D} E[\|\Phi(t) - \sum_{l=1}^{n} I_\Phi(v_l) v_l(t)\|] = 0.$$

Así pues, tenemos hasta ahora dos bases para expresar nuestros datos funcionales: la determinada por el kernel reproductor que depende del espacio de funciones y la determinada por las funciones de covarianza que depende de los datos muestrales. Cada una de estas bases permite reducir la dimensionalidad de una manera óptima en algun sentido y cada una de ellas puede tener fortalezas o debilidades según la naturaleza de los datos y del problema a resolver. Por ejemplo, cuando se utiliza PCA en problemas de clasificación para reducir la dimensionalidad del análisis, tenemos que ser conscientes de que no hay garantía de que la separación entre grupos estará en la dirección

de los primeros PCs; la separación entre grupos puede estar en las direcciones de los últimos. Por este motivo sería interesante buscar otra base de funciones que represente un compromiso entre ambas.

La tercera base propuesta surge de una relación entre ambos operadores. Definimos el operador lineal $G := L_K^{\frac{1}{2}} \circ L_\Gamma \circ L_K^{\frac{1}{2}}$ y denotamos $\{\eta_j, w_j\}_{j=1}^{\infty}$ los valores y funciones propias de G. A partir de la base $\{w_l\}_{l=1}^{\infty}$, en el Teorema (7.0.1), definimos un nuevo conjunto de funciones en H_K, $\{u_l\}_{l=1}^{\infty}$, que es sistema generador de H_K y del que obtenemos una relación ("diagonalización simultánea") entre los operadores L_K y L_Γ.

Teorema 7.0.1. *Sean $\{\eta_j, w_j\}_{j=1}^{\infty}$ los valores y funciones propias de G. Definimos $u_j := L_K^{\frac{1}{2}}(w_j)$, $\forall j$.*

Entonces, $\forall f \in H_K$, $f = \sum_{j=1}^{\infty} \xi_j \, u_j$, *donde* $\xi_j = \langle f \ , \ L_K^{-1} \, u_j \rangle_{L^2}$,

$$L_\Gamma \, u_j = \eta_j \, L_K^{-1} \, u_j$$

y

$$\langle u_i \ , \ L_\Gamma \, u_j \rangle_{L^2} = \eta_j \, \delta_{ij}.$$

Los productos de vectores u_j se pueden calcular usando una ecuación similar a la Ec. (7.5).

Finalmente, aplicamos todo este desarrollo teórico a la base de datos del proyecto de los niños y niñas para predecir la bondad del ajuste de una talla de camiseta. Se definen tres modelos diferentes dependiendo de la base usada con el fin de experimentar cuál da mejores resultados a esta aplicación práctica. Denotamos la base utilizada cada vez por $\{\phi_l\}_{l=1}^{\infty}$, y los correspondientes coeficientes a la k-ésima superficie con $\{c_l^k\}_{l=1}^{\infty}$. $\forall l = 1, \ldots, \infty$ consideraremos *'caso 1'*, donde $\phi_l = \rho_l$ y $c_l^k = \langle \overline{\varphi_k}, \rho_l \rangle_{H_K}$ (como en Ec. (7.3)), *'caso 2'*, donde $\phi_l = v_l$ y $c_l^k = \langle \overline{\varphi_k}, v_l \rangle_{L^2}$ (como en Ec. (7.4)), y *'caso 3'*, donde $\phi_l = u_l$ y $c_l^k = \langle \overline{\varphi_k}, \ L_K^{-1} \, u_l \rangle_{L^2}$ (como en Teorema (7.0.1)).

Dado que gran parte de la información inherente a los datos originales se captura por las primeras componentes funcionales y los coeficientes asociados, estas bases se truncan con un número bajo r de términos. En el *'caso 1'*, $\overline{\varphi(x)} \cong \sum_{l=1}^{r} c_l \phi_l(x)$ se trunca considerando $r = 7$ elementos de la base, en el *'caso 2'*, $r = 8$ elementos y en el *'caso 3'*, $r = 7$ elementos.

Recordemos que (ver sección 4.2) nuestro conjunto de datos para este estudio contiene ahora los escáneres $3D$ de la parte superior de los cuerpos de 76 niños y niñas y la bondad del ajuste de tres tallas distintas codificadas

como -1 (pequeño), 0 (correcto) y 1 (grande). El modelo asume observaciones independientes, pero en nuestro caso hay muchas medidas tomadas en cada niño o niña, por eso transformamos el modelo como un *logit* acumulativo mixto (Agresti, 2010):

$$\begin{aligned} logit[P(Y_{k,i} \leq j)] =& \alpha_j + \beta_1 shirt.size_i + \beta_2 sex_k + \beta_3 age_k + \\ &+ \sum_{p=1}^{r}\sum_{l=1}^{r} b_p^k c_l^k \langle \phi_p, \phi_l \rangle_{H_K} + u(k), \\ &\forall i = 1, \cdots, n_k; j \in \{-1, 0\}; k = 1, \cdots, 78, \end{aligned}$$

donde $n_k \in \{1, 2, 3\}$ es el número de observaciones tomadas en el k-ésimo niño y los efectos del niño se asumen que son aleatorios, independientes e identicamente distribuidos siguiendo una distribución Gaussiana, es decir, $u(k) \sim N(0, \sigma_k^2)$.

Usamos la función *clmm* del paquete ordinal de R (Christensen, 2015) para ajustar el modelo en los tres casos. Además, llevamos a cabo un proceso de validación cruzada (*leave-one-out cross validation*) para comprobar el poder predictivo del modelo en las diferentes bases. Para cada base, se estima el modelo iterativamente teniendo en cuenta todos los datos excepto las observaciones disponibles del niño o niña en turno.

Por último, calculamos un porcentaje de acierto entre las predicciones y los datos reales en cada uno de los tres modelos (ver Tabla (7.1)). A partir de los resultados concluimos que a pesar de que la base de funciones propias de la covarianza es uno de los enfoques más utilizados, hemos encontrado dos bases adicionales con un poder predictivo similar para esta aplicación. A pesar de que las diferencias son bastante pequeñas, los resultados son ligeramente mejores con la base mixta, como esperábamos inicialmente.

		Caso 1			Caso 2			Caso 3		
		Predicción			Predicción			Predicción		
		-1	0	1	-1	0	1	-1	0	1
Decisión del experto	-1	47	16	1	51	12	1	51	12	1
	0	10	41	10	11	38	12	10	39	12
	1	1	12	54	0	13	54	1	11	55
% de aciertos		73,95 %			74,48 %			75,52 %		

Tabla 7.1: Resultados del procedimiento de validación cruzada para el modelo mixto de regresión ordinal estimado con las diferentes bases.

Además se evalúa experimentalmente la influencia del número de puntos del grid en el método, concluyendo que el procedimiento es bastante robusto y resaltando el hecho de que los mejores resultados se obtienen en un equilibrio entre un número grande y pequeño de puntos en el grid. Finalmente, comparamos el procedimiento desarrollado con otros métodos tradicionales y los resultados de clasificación obtenidos son ligeramente peores que los obtenidos con la metodología desarrollada en este trabajo.

Capítulo 8

Fórmula de Gauss-Bonet e integrales rotacionales en espacios de curvatura constante

Gauss-Bonnet formulae and rotational integrals in constant curvature spaces. **Sonia Barahona** y Ximo Gual Arnau. Publicado en *Differential Geometry and its Applications*, volumen 50, páginas 116-125, año 2017. DOI: 10.1016/j.difgeo.2016.11.005.

En este capítulo, detallamos el trabajo desarrollado dentro del ámbito de la Estereología, en el que obtenemos fórmulas rotacionales para el área y para las integrales de curvatura media de la superficie frontera de un dominio compacto en un espacio de curvatura constante λ, M^n_λ.

Antes de enunciar los resultados más relevantes demostrados en esta aportación, es conveniente fijar la siguiente notación. Sea L^n_r un $r-$plano, $(r \leq n)$, al que llamamos subvariedad totalmente geodésica de dimensión r en M^n_λ, y $\mathrm{d}L^n_r$ la correspondiente densidad que es invariante bajo el grupo de movimientos, Euclídeos o no Euclídeos. Denotamos por $L^n_{r[0]}$ un $r-$plano que pasa por un punto fijo O en M^n_λ, y su densidad invariante $\mathrm{d}L^n_{r[0]}$.

Sea $Q \subset M^n_\lambda$ un dominio compacto con frontera suave $S = \partial Q$, $\chi(Q)$ la característica de Euler-Poincaré de Q y M_i la $i-$ésima integral de curvatura media de S. Denotamos por $O_k = \mathrm{vol}(\mathbb{S}^k)$ el área de la superficie de la esfera unidad k-dimensional.

El trabajo tiene una estructura semejante a los pasos seguidos en el ejemplo de la sección 2.7. Por tanto, en primer lugar se obtiene una fórmula integral, del tipo de la Ec. (2.17), en la que aparecen diversas integrales de curvatura media de S.

Teorema 8.0.1. *Para n y r impar, o n y r par, se cumple la siguiente relación,*

$$\frac{1}{2}O_n\chi(Q) - c_{n-1}M_{n-1} - \lambda c_{n-3}M_{n-3} - \cdots - \lambda^{\frac{n-r-2}{2}}c_{r+1}M_{r+1} = \lambda^{\frac{n-r}{2}}\frac{O_r\ldots O_1}{O_{n-1}\ldots O_{n-r}}\int_{\mathcal{L}_r}\chi(Q\cap L_r^n)\mathrm{d}L_r^n,$$

donde,

$$c_h = \begin{pmatrix} n-1 \\ h \end{pmatrix}\frac{O_n}{O_hO_{n-1-h}}.$$

El resultado enunciado supone una generalización del trabajo Solanes (2006), en su caso lo demuestra para $r = n-2$ siguiendo un procedimiento distinto. Además, como consecuencia del teorema anterior, podemos deducir el siguiente corolario, que permite expresar cada curvatura integral a partir de fórmulas integrales.

Corolario 1.

$$\begin{aligned} M_r = {} & \frac{(n-r-1)O_r\ldots O_0}{O_{n-2}\ldots O_{n-r-2}}\int_{\mathcal{L}_{r+1}}\chi(Q\cap L_{r+1}^n)\mathrm{d}L_{r+1}^n \\ & - \lambda\frac{rO_{r-2}\ldots O_0}{O_{n-2}\ldots O_{n-r}}\int_{\mathcal{L}_{r-1}}\chi(Q\cap L_{r-1}^n)\mathrm{d}L_{r-1}^n. \end{aligned} \tag{8.1}$$

A continuación, obtenemos fórmulas rotacionales para las integrales de curvatura media de las superficies en M_λ^n, similares a la Ec. (2.19). Éstas permiten expresar las integrales de curvatura media a partir de secciones del objeto que pasan por el punto O.

En este contexto, a partir de la Ec. (8.1), encontramos funciones α_r definidas en $L_{r+2[0]}^n\cap Q$ con media rotacional igual a M_r, es decir,

$$M_r = \int_{L_{r+2[0]}^n\cap Q\neq\emptyset}\alpha_r(L_{r+2[0]}^n\cap Q)\mathrm{d}L_{r+2[0]}^n.$$

Teorema 8.0.2. *Sea $Q\subset M_\lambda^n$ un dominio compacto con frontera suave $S=\partial Q$. Las funciones medida α_r correspondientes a la $r-$ésima integral de*

curvatura media de S, M_r, se pueden expresar como

$$\alpha_r(L^n_{r+2[0]} \cap Q) = \frac{O_{r-2}\dots O_0}{O_{n-2}\dots O_{n-r-2}}$$
$$\left[(n-r-1)O_rO_{r-1}\beta_{r+1} - \lambda r O_1 O_0 \int\limits_{(Q\cap L^n_{r+2[0]})\cap L^{r+2}_{r[0]}\neq\emptyset} \beta_{r-1}\mathrm{d}L^{r+2}_{r[0]}\right],$$

donde

$$\beta_r = \int\limits_{(L^n_{r+1[0]}\cap Q)\cap L^{r+1}_r\neq\emptyset} \chi((L^n_{r+1[0]}\cap Q)\cap L^{r+1}_r) s^{n-r-1}_\lambda(\rho)\mathrm{d}L^{r+1}_r,$$

ρ es la distancia desde O hasta el plano L^{r+1}_r; y

$$s_\lambda(\rho) = \begin{cases} \lambda^{-1/2}\sin(\rho\sqrt{\lambda}), & \lambda > 0 \\ \rho, & \lambda = 0 \\ |\lambda|^{-1/2}\sinh(\rho\sqrt{|\lambda|}), & \lambda < 0 \end{cases}.$$

Finalmente, utilizamos resultados clásicos de la teoria de Morse para hallar una representación de las fórmulas rotacionales equivalente a la de la Eq. (2.20). Así, el integrando de la curvatura media se puede interpretar geométricamente utilizando los puntos críticos de funciones altura.

Introducimos a continuación la notación necesaria para el último resultado obtenido. Sea u_r un vector unitario en $\mathbb{S}^r \subset T_O L^n_{r+1[0]}$, la geodésica $\gamma_{u_r} : \mathbb{R} \longrightarrow L^n_{r+1[0]}$, con $\gamma_{u_r}(0) = O$ y $\gamma'(0) = u_r$ viene dada por $\gamma_{u_r}(t) = c_\lambda(t)O + s_\lambda(t)u_r$, donde $c_\lambda(t) = \frac{\mathrm{d}}{\mathrm{d}t}s_\lambda(t)$. Dado u_r, sea $h_{u_r} : L^n_{r+1[0]} \longrightarrow \mathbb{R}$ la función altura cuyas hipersuperficies de nivel son los $r-$planos L^{r+1}_r perpendiculares a la geodésica $\gamma_{u_r}(t)$. Notar que en el caso Euclídeo ($\lambda = 0$) esta función altura coincide con la función estándar.

Para $r < n \in \{1, 2, \dots\}$, definimos:

$$I_{n-r-1,r}(\rho) = \int s^{n-r-1}_\lambda(|\rho|)\, c^r_\lambda(\rho)\,\mathrm{d}\rho.$$

Teorema 8.0.3. *Sea O un punto en M^n_λ y $Q \subset M^n_\lambda$ un dominio compacto contenido en el hemisferio de M^n_λ con polo O cuando $\lambda > 0$. Sea $L^n_{r+1[0]}$ un*

plano genérico, por tanto, $Q_{r+1} = Q \cap L^n_{r+1[0]}$ *es un dominio con frontera suave en* $L^n_{r+1[0]}$*. Entonces, para* $r \in \{0, 1, \ldots, n-2\}$*,*

$$\beta_r = \frac{1}{2} \int_{\mathbb{S}^r} \left(\sum_{k=1}^{m} \epsilon_k \, I_{n-r-1,r}(\rho_k) \right) \mathrm{d}u_r,$$

donde $\rho_1, < \rho_2 < \ldots < \rho_m$ *representan los valores que toma la función altura en los puntos críticos* $\mathrm{Crit}(h_{u_r}|_{\partial Q_{r+1}})$ *correspondientes a la dirección* u_r *y* ϵ_k *es +1 o -1 dependiendo de si el punto crítico es máximo o mínimo.*

Cabe resaltar que si $\lambda \neq 0$ las rectas y planos pasan a ser subvariedades totalmente geodésicas de M^n_λ, y la función altura ya no es tan intuitiva.

Parte III

Conclusiones y trabajo futuro

Capítulo 9

Conclusiones

El objetivo de la tesis ha sido doble. Por una parte, proponer un tipo de objeto matemático llamado *current*, que permite caracterizar cuerpos geométricos. Tras el estudio del espacio donde quedan embebidos estos objetos geométricos y el uso de sus prácticas propiedades, se desarrollan nuevas metodologías estadísticas (clasificación supervisada, clasificación no supervisada y regresión ordinal) para trabajar con una muestra de cuerpos caracterizados mediante estos *currents*. Por otra parte, obtener fórmulas rotacionales para el área y para las integrales de curvatura media de la superficie frontera de un dominio compacto en un espacio de curvatura constante λ, M^n_λ.

En los tres primeros trabajos se ha utilizado la caracterización de la forma y el tamaño de cuerpos geométricos mediante *currents*, permitiendo asociar a cada cuerpo geométrico un elemento de un espacio *vector-valued RKH*. En estas aportaciones también se han realizado estudios de simulación para poder validar los resultados obtenidos y se ha aplicado la metodología propuesta en cada caso, a problemas reales que plantea el proyecto en el que se enmarca el trabajo.

En el primer trabajo se ha propuesto un método de clasificación no supervisada para cuerpos geométricos y se ha aplicado para la definición de sistemas de tallaje infantiles más eficientes. Tras la transformación de los datos, se ha adaptado el conocido algoritmo k-medias a este contexto y se han propuesto dos sistemas de tallas infantiles para conseguir un mejor ajuste de las prendas. El primero de ellos se ha definido segmentando previamente los datos según su altura y a continuación, se ha aplicado el algoritmo k-medias para establecer dos tallas en cada grupo de alturas. De esta manera, la altura nos sirve de filtro inicial, y las tallas obtenidas optimizan el ajuste de la prenda a la forma del cuerpo. Para reducir el número de tallas, se ha propuesto un segundo sistema de tallaje aplicando a todos los miembros de

la muestra el algoritmo k-medias. El número de grupos a generar se ha establecido usando criterios de optimalidad. En los dos casos, se han descrito las tallas propuestas según las medidas antropométricas del grupo.

En el segundo trabajo se ha desarrollado una metodología para la clasificación supervisada de cuerpos geométricos. Utilizando la teoría de FDA se ha adaptado el método de análisis discriminante para poder emplearlo en este contexto. Este desarrollo se ha aplicado para la construcción de un modelo que asigna a cada niño o niña su talla en cada uno de los sistemas de tallaje definidos en el primer trabajo. Se han obtenido resultados prometedores al emplear este método con la base de datos de niños y niñas escaneados que aporta el proyecto.

En el tercer trabajo se ha propuesto una nueva metodología para modelizar una variable de respuesta ordinal en función de objetos geométricos $3D$. Una vez se han transformado los datos geométricos en elementos en un *vector-valued RKHS*, se han expresado en tres bases de funciones del espacio diferentes. En primer lugar, los predictores se han expresado en la base dada por el kernel del RKHS. En segundo lugar, se ha usado una base que se obtiene de manera similar a la base habitual dada por el kernel covarianza en el caso escalar. En tercer lugar, se ha buscado una base de funciones que conecta los beneficions de las dos bases anteriores mediante la "diagonalización simultánea" de operadores. Se han estimado los coeficientes de cada predictor geométrico expresado en las tres bases de funciones. Por último, este método se ha aplicado para predecir si la talla de una prenda ajusta a un cliente o si es grande o pequeña para él/ella. Los resultados obtenidos con las diferentes bases fueron semejantes y bastante prometedores.

El cuarto trabajo se encuentra ubicado en el ámbito de la Estereología. En él se han obtenido generalizaciones del principal resultado de Solanes (2006) y se han dado interpretaciones geométricas de combinaciones lineales de las integrales de curvatura media que aparecen en la fórmula de Gauss-Bonnet para hipersuperficies en espacios M^n_λ. Entonces, se han combinado estos resultados con la clásica teoría de Morse para obtener nuevas fórmulas rotacionales para la k-ésima integral de curvatura media de una hipersuperficie en M^n_λ.

Capítulo 10

Líneas de investigación futuras

A continuación, se describen algunas posibles líneas de investigación futuras. Es importante notar que en Estadística no se había trabajado hasta el momento con elementos de un espacio *vector-valued RKH*. Por ello, se han tenido que generalizar las propiedades que usa el Análisis de Datos Funcionales del caso escalar al vectorial, abriendo un nuevo campo de posibilidades y de problemas abiertos. En esta línea, es fundamental señalar que el operador integral definido a partir de las funciones covarianza L_Γ (Capítulo 7), en general, no es equivalente al operador covarianza $\mathcal{K}$ de χ (Ec. (2.8)) considerado como un elemento aleatorio en H_K. Por ello, como trabajo futuro nos planteamos estudiar en profundidad su relación (Ec. (2.10)): en qué situaciones son operadores equivalentes, la correspondientes descomposiciones en valores y funciones propias, etc.

Por otra parte, en este trabajo de tesis doctoral se han caracterizado los cuerpos geométricos (curvas en $2D$ y superficies en $3D$) a través de funciones, *currents*. Pero, esta representación geométrica del conjunto de formas, que se engloba dentro del Análisis de Forma y Tamaño, se puede hacer desde otros puntos de vista. Una idea natural cuando se trabaja con formas en $2D$ o en $3D$, es la de considerar curvas o superficies parametrizadas que representan la frontera del dominio de interés. Se puede trabajar con parametrizaciones de dos maneras diferentes. Se puede fijar una parametrización que sea equivalente en todas las curvas (o superficies) y mantener esta parametrización fija en todo el análisis de formas. Este tipo de método se llama análisis de formas *no elástico* o análisis de formas "bending-only". También se puede trabajar con parametrizaciones arbitrarias y tener en cuenta las variaciones que se producen al cambiar de parametrización. Este método se llama análisis de formas de tipo *elástico*.

Tanto esta nueva aproximación, como la ya desarrollada en el presente

trabajo se pueden aplicar a una base de datos de escáneres y fotografías de pies, con el fin de resolver el problema de asignación de calzado. En la tesis se ha trabajado tomando como base un sistema que, a partir de 2 o 3 fotografías tomadas con tecnología doméstica (smartphone, tablet o cámara web/digital) extrae las medidas del cuerpo entero de niños y niñas y también una aproximación a la reconstrucción $3D$ del cuerpo entero. Este mismo sistema se ha utilizado con éxito para obtener una reconstrucción $3D$ del pie. Por tanto, la información que tenemos para desarrollar y aplicar técnicas estadísticas vuelven a ser formas en $2D$ y en $3D$.

En el caso $2D$ (trabajando directamente con las fotografías tomadas con tecnología doméstica) se puede considerar cada contorno (fotografía) como una curva plana cerrada. El conjunto de todas las curvas parametrizadas planas cerradas tiene una estructura de variedad diferenciable de dimensión infinita que, con una métrica elástica adecuada, nos permite calcular distancias entre curvas y también la curva media.

En el caso de $3D$ (cuando disponemos de una malla triangular que represennta la forma $3D$ del pie tras ser escaneado) proponemos trabajar con parametrizaciones que obtendremos a partir de la malla, desde los dos puntos de vista. Es decir, utilizando análisis de formas de tipo elástico y no elástico.

Como la reconstrucción $3D$ de todos los pies, a partir de las fotografías $2D$ se hace de manera semejante, supondremos que hay una correspondencia entre todas las parametrizaciones de los pies y de esta manera fijaremos una única parametrización para todas las formas $3D$ (análisis de formas no elástico). Se propone comparar los resultados obtenidos mediante el análisis de formas no elástico, con otro basado en el análisis de formas elástico. Por tanto, con la idea de considerar una métrica en el espacio de formas, que no dependa de parametrizaciones, definiremos una función asociada a cada superficie: la función factor multiplicación de áreas o la función raíz cuadrada del campo normal.

Bibliografía

Agresti A (2010). Analysis of ordinal categorical data, vol. 656. John Wiley & Sons.

Aronszajn N (1950). Theory of reproducing kernels. Transactions of the American mathematical society :337–404.

Auneau J, Jensen EBV (2010). Expressing intrinsic volumes as rotational integrals. Advances in Applied Mathematics 45:1–11.

Bauer M, Harms P, Michor PW (2011). Sobolev metrics on shape space of surfaces. Journal of Geometric Mechanics 3:389–438.

Bock HH (2007). Clustering methods: a history of k-means algorithms. In: Brito P, Bertrand P, Cucumel G, de Carvalho F, eds., Selected Contributions in Data Analysis and Classification. Springer Berlin Heidelberg, 161–72.

Caponnetto A, Micchelli CA, Pontil M, Ying Y (2008). Universal multi-task kernels. The Journal of Machine Learning Research 9:1615–46.

Carmeli C, De Vito E, Toigo A (2006). Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. Analysis and Applications 4:377–408.

Conway JB (2013). A course in functional analysis, vol. 96. Springer Science & Business Media.

Cruz-Orive LM, Gual-Arnau X (2015). The invariator design: an update. Image Analysis Stereology 34:147–59.

Cucker F, Smale S (2001). On the mathematical foundations of learning. American Mathematical Society 39:1–49.

Cuevas A (2014). A partial overview of the theory of statistics with functional data. Journal of Statistical Planning and Inference 147:1–23.

Do Carmo MP (2012). Differential forms and applications. Springer Science & Business Media.

Durrleman S (2010). Statistical models of currents for measuring the variability of anatomical curves, surfaces and their evolution. Ph.D. thesis, Université Nice Sophia Antipolis.

Durrleman S, Pennec X, Trouvé A, Ayache N (2009). Statistical models of sets of curves and surfaces based on currents. Medical image analysis 13:793–808.

Ferraty F, Vieu P (2006). Nonparametric functional data analysis: theory and practice. Springer Science & Business Media.

Gardner RJ (1995). Geometric tomography, vol. 1. Cambridge University Press Cambridge.

Gual-Arnau X, Cruz-Orive LM (2009). A new expression for the density of totally geodesic submanifolds in space forms, with stereological applications. Differential Geometry and its Applications 27:124–8.

Gual-Arnau X, Cruz-Orive LM (2016). New rotational integrals in space forms, with an application to surface area estimation. Applications of Mathematics 61:489–501.

Gual-Arnau X, Cruz-Orive LM, Nuno-Ballesteros J (2010). A new rotational integral formula for intrinsic volumes in space forms. Advances in Applied Mathematics 44:298–308.

Gual-Arnau X, Herold-García S, Simó A (2015). Geometric analysis of planar shapes with applications to cell deformations. Image Analysis Stereology 34.

Hsing T, Eubank R (2015). Theoretical foundations of functional data analysis, with an introduction to linear operators. John Wiley & Sons.

James GM (2002). Generalized linear models with functional predictors. Journal of the Royal Statistical Society Series B Statistical Methodology 64:411–32.

Jensen, EBV (1998). Local Stereology. 1998.

Keerthi SS, Lin CJ (2003). Asymptotic behaviors of support vector machines with gaussian kernel. Neural computation 15:1667–89.

Kendall M, Stuart A (1977). The advanced theory of statistics. vol. 1: Distribution theory. London Griffin 1977 4th ed .

Lukić M, Beder J (2001). Stochastic processes with sample paths in reproducing kernel hilbert spaces. Transactions of the American Mathematical Society 353:3945–69.

MATLAB (2015). version 8.6.0 (R2015b). Natick, Massachusetts: The MathWorks Inc.

Micchelli C, Pontil M (2005). On learning vector-valued functions. Neural computation 17:177–204.

Munoz A, Moguerza JM (2006). Estimation of high-density regions using one-class neighbor machines. IEEE Transactions on Pattern Analysis and Machine Intelligence 28:476–80.

Olaru S, Filipescu E, Filipescu E, Niculescu C, Salistean A (2012). 3d fit garment simulation based on 3d body scanner anthropometric data. In: 8th International DAAAM Baltic Conference on Industrial Engineering, vol. 146.

Pennec X (2006). Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. J Math Imaging Vis 25:127–54.

Quang MH, Kang SH, Le TM (2010). Image and video colorization using vector-valued reproducing kernel hilbert spaces. Journal of Mathematical Imaging and Vision 37:49–65.

Santaló LA (2004). Integral geometry and geometric probability. Cambridge university press.

Silverman B, Ramsay J (2005). Functional Data Analysis. Springer.

Solanes G (2006). Integral geometry and the gauss-bonnet theorem in constant curvature spaces. Transactions of the American Mathematical Society 358:1105–15.

Thórisdóttir Ó, Kiderlen M (2014). The invariator principle in convex geometry. Advances in Applied Mathematics 58:63–87.

Thórisdóttir Ó, Rafati AH, Kiderlen M (2014). Estimating the surface area of nonconvex particles from central planar sections. Journal of microscopy 255:49–64.