USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Tests de independencia entre obxectos estatísticos

Carlos García Meixide

2020/2021

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Tests de independencia entre obxectos estatísticos

Carlos García Meixide

Xullo 2021

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Probabilidade e Estatística
Título: Tests de independencia entre obxectos estatísticos
Breve descrición do contido
Preténdese expoñer un enfoque moderno do problema clásico da independencia estatística entre variables aleatorias asociadas a diversos espazos de probabilidade. Para concluír, ilústranse as ferramentas teóricas desenvoltas cunha aplicación no ámbito da física de materiais.
Recomendacións
Outras observacións

A miña nai, Isabel

Índice xeral

Resumo

Examínase cunha linguaxe marcada pola análise funcional e a teoría da medida o problema matemático da independencia estatística dende unha perspectiva contemporánea. Introdúcese o marco conceptual preciso de cara a acadar tal obxectivo para posteriormente desenvolver de xeito rigoroso os dous formalismos que historicamente guiaron a evolución do estudo da independencia, así como o sorprendente paralelismo que entre eles gardan. Transpasando a fronteira entre o mundo poboacional e o empírico, preséntase de xeito conciso un corpus de resultados relativos a U-estatísticos, abstracción fundamental da teoría da estimación. A finalidade deste capítulo é proporcionar a maquinaria precisa para dar paso á formulación empírica do contraste de independencia de dúas variables aleatorias dada unha mostra. Finalmente, condénsanse as contribucións orixinais nos dous derradeiros capítulos proporcionando unha ponte entre o exposto até este punto e a teoría de machine learning máis actual, para así dar paso á aplicación do presente traballo no eido da física de materiais e a novedosa representación da súa estrutura en forma de grafos.

Abstract

The mathematical problem of statistical independence is examined from a contemporary perspective with a language marked by functional analysis and measure theory. We introduce the necessary conceptual framework in order to achieve this objective and then rigorously develop the two formalisms that historically guided the evolution of the study of independence, as well as the surprising parallelism between them. Crossing the border between the population and empirical worlds, we present in a concise way a corpus of results related to U-statistics, a fundamental abstraction of estimation theory. The purpose of this chapter is to provide the machinery needed to give way to the empirical formulation of testing given a sample whether two random variables are independent or not. Finally, the original contributions are condensed in the two following chapters, providing a bridge between what has been exposed up to this point and the most current machine learning theory, in order to show the applicability of this work in the field of materials physics and the novel representation of their structure in the form of graphs.

Introdución

Toda disciplina matemática
pasa por tres períodos de desenvolvemento:
o inxenuo, o formal e o crítico.
David Hilbert

A independencia é un concepto fundamental na matemática actual. Sobre ela cae a responsabilidade de que a aleatoriedade non sexa unha mera particularización da teoría da medida, así como de articular o pensamento estatístico da comunidade científica á hora de modelar un problema real.

Neste sentido, é importante determinar se supoñer *independencia* nun determinado contexto é realista ou non cara a dilucidar se se pode esperar que se garantan certas propiedades teóricas dos algoritmos empregados ou avaliar a bondade de axuste dun modelo estatístico.

Os enfoques clásicos do test de independencia centrábanse na configuración máis simple-variables Euclidianas unidimensionais- e adoitaban ter potencia considerable soamente contra clases restrinxidas de alternativas. Nesta etapa foron deseñados test baseados na correlación de Pearson (por exemplo, [Pearson, 1920]), o coeficiente de correlación de rangos de Spearman ([Spearman, 1904]), o tau de Kendall ([Kendall, 1938]) ou o D de Hoeffding ([Hoeffding, 1948b]).

Non obstante, a eclosión do machine learning marcou un ritmo acelerado en canto á xeneralización do espazo no que viven os datos dun problema concreto. Isto, xunto co desexo de gañar potencia contra clases máis amplas de hipóteses alternativas, desencadeou no renacemento da interese polos tests de independencia nos últimos anos.

Para comprender a fauna de tests de independencia que figura a día de hoxe na lite-

ratura é importante ter en mente que **en escenarios típicos de interese non existe o test uniformemente máis potente** [Berrett et al., 2020]. Como consecuencia deste sonoro resultado xurdiron varias perspectivas diferentes e novas, entre as que destacan dúas liñas temáticas: os test baseados no criterio de independencia de Hilbert – Schmidt, da metodoloxía *núcleo* ([Gretton et al., 2005]); e os test baseados en distance covariance, da metodoloxía *energy statistics* ([Szekely, 2000]. O concepto de *distance covariance* foi introducido por [Székely et al., 2007] e pode expresarse como unha norma pesada de L^2 da diferenza entre a función característica da distribución conxunta e o produto das funcións características das marxinais. En [Sejdinovic et al., 2013] demostrouse que os tests baseados en distance covariance son equivalentes a tests con RKHS cunha escolla específica do núcleo. Estes test foron amplamente estudados pola comunidade machine learning, cunha comprensión serodia do tema en [Bach and Jordan, 2002] e coa proposta do test de independencia Hilbert–Schmidt en [Gretton et al., 2005]. Estes tests están baseados en facer embeddings da distribución conxunta e do produto das marxinais nun espazo de Hilbert e considerar a norma da súa diferenza neses espazo. Como desvantaxe, o rendemento destes métodos podería estar fortemente afectada pola elección do *núcleo*.

Esta é precisamente a motivación da parte máis orixinal deste traballo: explorar o impacto das propiedades teóricas derivadas de construír directamente un *embedding* no canto de seguir o procedemento paradigmático do *machine learning* (escoller directamente un núcleo obviando as propiedades da súa aplicación canónica). En consoancia con esta perspectiva, empregamos o traballo duns investigadores do MIT [Xie and Grossman, 2018a] para deseñar un embedding para o conxunto de estruturas cristalinas que un material pode amosar que nos leva a un espazo con boas propiedades topolóxicas de cara a realizar un contraste de independencia. A correspondencia explorada no capítulo 2 serve de guía matemática para facer da elección do núcleo a resolución dun problema de optimización convexa e ser capaces de inferir conclusión válidas na sección de Resultados. Todos os códigos implementados, que abranguen maioritariamente as linguaxes de programación `R`, `Python` e `MATLAB`, poden atoparse en `https://github.com/carmeiga/statindep`

O soño de Hilbert non consistía simplemente en demostrar que a matemática pura é consistente, senón que esta xamais pode conducir a conclusións disonantes co mundo real. Nun intento de emular esta conduta, os oito capítulos deste documento foron articulados e ordenados de tal xeito que nunca se perdan de vista o coñecemento racional e o coñecemento empírico entre eles, mantendo viva a oscilación que a praxe matemática debe sempre experimentar.

Capítulo 1

Aspectos preliminares de teoría da probabilidade e análise funcional

1.1. Espazos semimétricos de tipo negativo

O concepto de tipo negativo é antigo pero está vivindo un rexurdimento grazas á súa versatilidade en ciencias da computación teórica [Naor, 2010].

Definición 1.1 (*Semimétrica*)**.** Sexa $\mathcal{Z}$ un conxunto non baleiro e $\rho : \mathcal{Z} \times \mathcal{Z} \to [0, \infty)$ unha función tal que

$\forall z, z' \in \mathcal{Z}$

1. $\rho(z, z') = 0$ se e só se $z = z'$,
2. $\rho(z, z') = \rho(z', z)$

Entón dise que $(\mathcal{Z}, \rho)$ é un espazo semimétrico e que ρ é unha semimétrica en $\mathcal{Z}$.

Definición 1.2. Dise do espazo semimétrico $(\mathcal{Z}, \rho)$ que é de *tipo negativo* se $\forall n \geq 2, z_1, \ldots, z_n \in \mathcal{Z}$, e $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, con $\sum_{i=1}^{n} \alpha_i = 0$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \rho(z_i, z_j) \leq 0$$

Proposición 1.3. *Sexa $(\mathcal{Z}, \rho)$ un espazo semimétrico de tipo negativo. Entón*

1. *Se ρ é de tipo negativo entón ρ^q tamén, para $0 < q < 1$.*
2. *ρ é unha semimétrica de tipo negativo se e só se existe un espazo de Hilbert H e unha aplicación inxectiva $\phi : \mathcal{Z} \to H$, tal que*

$$\rho\left(z, z'\right) = \left\|\phi(z) - \phi\left(z'\right)\right\|_{H'}^2$$

1.2. Teoría da medida e martingalas

1.2.1. Fundamentos

Lembramos que unha *σ -álxebra* nun conxunto $\mathcal{Z}$ dado é unha clase de subconxuntos de $\mathcal{Z}$ á que pertence $\mathcal{Z}$ e pechada respecto a tomar complementarios, e unións numerables e interseccióńns. A clase máis pequena con esta propiedade é $\mathcal{Z}$ xunto co conxunto baleiro; a máis grande é o conxunto partes $2^{\mathcal{Z}}$ de todos os subconxuntos de $\mathcal{Z}$. Ter introducido esta noción vén motivado polo feito de precisarmos σ -álxebras como dominios de definición de medidas.
Toda clase $\mathcal{S}$ de subconxuntos de $\mathcal{Z}$ leva asociada a σ -álxebra máis pequena que contén a $\mathcal{S}$. Denótase polo símbolo $\sigma(\mathcal{S})$ e chámaselle *σ-álxebra xerada pola clase* $\mathcal{S}$. Formalmente a súa construción é ben simple: tan só tomamos a intersección de todas as σ -álxebras en $\mathcal{Z}$ que conteñen a S. De todos xeitos, é raro que se poidan describir de xeito construtivo os elementos de $\sigma(\mathcal{S})$. Neste exemplo sinxelo si que se podería: a σ -álxebra xerada por todos os conxuntos finitos en $\mathcal{Z}$ está formada por conxuntos como moito numerables e os seus complementarios.

Unha función $\mu : \mathscr{F} \to \mathbb{R}$ dunha σ -álxebra $\mathscr{F}$ nun conxunto $\mathcal{Z}$ chámase *medida con signo estendida* se para toda colección numerable ou finita de conxuntos disxuntos $A_n \in \mathscr{F}$, temos

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu\left(A_n\right)$$

Tal medida vén representada de xeito único na forma da chamada descomposición de Hahn-Jordan

$$\mu = \mu^+ - \mu^-,$$

onde μ^+ e μ^- son medidas non negativas en $\mathscr{F}$ concentradas en conxuntos disxuntos $\mathcal{Z}^+$ e $\mathcal{Z}^-$ (dando a descomposición de Hahn do espazo $\mathcal{Z}$) e chamadas as partes positivas e negativas da medida μ . Se μ sempre toma valores finitos daquela chámaselle a μ *medida con signo finita* e se $\mu \geqslant 0$ e $\mu(\mathcal{Z}) = 1$, entón chámaselle a μ *medida de probabilidade*.

Definimos o espazo de *medidas positivas finitas* en $\mathcal{Z}$

$$\mathcal{M}_+(\mathcal{Z}) = \{\mu : \mu = \mu^+; \quad \mu(\mathcal{Z}) < \infty\}$$

Cando $\mathcal{Z} = \mathbb{R}$ aparece o importante concepto de integral de Fourier dunha medida.

Proposición 1.4. *Para $\mu \in \mathcal{M}_+(\mathbb{R})$, definimos a transformada de Fourier como*

$$\mathcal{F}(\mu) = \hat{\mu}(\xi) = \int_{\mathbb{R}} e^{-ix\xi} d\mu(x)$$

Entón

$$\mathcal{F} : \mathcal{M}_+(\mathbb{R}) \to C_b(\mathbb{R})$$

onde $C_b(\mathbb{R})$ é o espazo de funcións continuas e acotadas sobre $\mathbb{R}$.

Ademais a transformada de Fourier está ben definida. Temos

$$|\hat{\mu}(\xi)| \leq \left| \int_{\mathbb{R}} e^{-ix\xi} d\mu(x) \right| \leq \int_{\mathbb{R}} \left| e^{-ix\xi} \right| d\mu(x) = \mu(\mathbb{R}) < \infty$$

Dado que $\mu(\mathbb{R}) < \infty$, a función $g(x) \equiv 1$ é integrable e serve como maiorante, $\left|e^{-ix\xi}\right| \leq 1 = g(x)$. Polo *Teorema da converxencia dominada de Lebesgue*, $\hat{\mu}(\xi)$ é continua.

Definición 1.5. A función característica f_X dunha variable aleatoria X defínese como

$$f_X(t) = \mathbb{E}\left(e^{itX}\right) = \int_{-\infty}^{\infty} e^{itx} d\mu_X(x)$$

En xeral, salvo por un signo, a función característica dunha variable aleatoria X é a transformada de Fourier da distribución de X

$$\chi_X(t) = \widehat{\mu_X}(-t)$$

Se $\mathcal{Z}$ é un espazo topolóxico, con $\mathcal{B}(\mathcal{Z})$ denotamos a chamada σ -álxebra de *conxuntos de Borel en $\mathcal{Z}$* definida como a σ -álxebra máis pequena que conteña todos os conxuntos abertos da topoloxía de $\mathcal{Z}$, i.e., a intersección de todas esas σ -álxebras. Asemade, está claro que esta é a σ -álxebra máis pequena que contén todos os pechados. Nótese que esta definición de σ -álxebra de Borel en $\mathcal{Z}$ semella simple mais en tódols casos non triviais os conxuntos de Borel non admiten descrición construtiva.

Definición 1.6. Toda medida en $\mathcal{B}(\mathcal{Z})$ recibe o nome de *medida de Borel*.

A non ser que se indique o contrario, asumiremos que $(\mathcal{Z}, \rho)$ é calquera espazo semimétrico de tipo negativo no que se poden definir medidas de Borel (nótese que non todo espazo semimétrico é un espazo topolóxico para poder ter σ-álxebra de Borel aínda que todo espazo métrico o sexa, véxase o Teorema 1.2 en [Sims, 1962]). Denotaremos por $\mathcal{M}(\mathcal{Z})$ o conxunto de todas as medidas de Borel con signo finitas en $\mathcal{Z}$, e por $\mathcal{M}_+^1(\mathcal{Z})$ o conxunto de todas as medidas de probabilidade de Borel en $\mathcal{Z}$.

O obxecto primario da teoría da probabilidade na axiomática de Kolmogorov é un *espazo de probabilidade* $(\Omega, \mathscr{F}, P)$, onde Ω é un conxunto non baleiro (*espazo amostral*) cunha σ -álxebra $\mathscr{F}$ de subconxuntos de Ω e unha medida de probabilidade $P : \mathscr{F} \to [0,1]$ en $\mathscr{F}$. Clasicamente, as funcións $X : \Omega \to \mathbb{R}$ medibles respecto a $\mathscr{F}$, o que significa que $X^{-1}(B) \in \mathscr{F}$ para todo conxunto de Borel B da recta real, chámanse *variables aleatorias* e a súa integral respecto da medida P en caso de existir chámase *esperanza*. Para toda variable aleatoria X a medida inducida $P \circ X^{-1}$ na recta real chámase *distribución* da variable aleatoria X.

Aquí atopamos o primeiro obstáculo cara a xeneralizar o codominio da nosa variable aleatoria X a un espazo semimétrico de tipo negativo: simplemente definir unha noción de esperanza. En [Lyons, 2013] xeneralízase este concepto a espazos métricos de tipo negativo e en [Sejdinovic et al., 2013] dáse un paso alén e establécese en espazos *semi*métricos de tipo negativo.

Definición 1.7. Para $\theta > 0$, dise que $v \in \mathcal{M}(\mathcal{Z})$ ten *θ -momento finito con respecto á semimétrica* ρ de tipo negativo se se existe $z_0 \in \mathcal{Z}$, tal que $\int \rho^\theta (z, z_0)\, d|v|(z) < \infty$. Denotamos

$\mathcal{M}_\rho^\theta(\mathcal{Z}) = \left\{ v \in \mathcal{M}(\mathcal{Z}) : \exists z_0 \in \mathcal{Z} \text{ tal que } \int \rho^\theta (z, z_0)\, d|v|(z) < \infty \right\}$

Un concepto importante é a *independencia* de variables aleatorias $X_1, \dots, X_n$ nun espazo de probabilidade xeral $(\Omega, \mathscr{F}, P)$ entendida como a igualdade

$$P(X_1 \in B_1, \dots, X_n \in B_n) = P(X_1 \in B_1) \cdots P(X_n \in B_n)$$

para todos os conxuntos de Borel B_i. A independencia dunha colección infinita de variables aleatorias enténdese como independencia de cada subcolección finita nela.

Non debe confundirse a noción de independencia coa de independencia dous-a-dous, pois a primeira implica a segunda pero a afirmación recíproca non é certa. Supoñamos que tres persoas lanzan cada unha a súa moeda A, B e C. Considérense as tres variables aleatorias Bernoulli que valen un cando (A,B) coinciden, (B,C) coinciden, e (A,C) coinciden respectivamente. Claramente, estas variables aleatorias son independentes dous a dous. Pero calquera par delas determina a terceira.

1.2.2. Converxencia

Definimos a continuación os tres tipos de converxencia que empregaremos posteriormente:

Definición 1.8. Considérense variables aleatorias $X_1, X_2, ...$ e X en $(\Omega, \mathscr{F}, P)$. Dise que X_n *converxe con probabilidade 1* (*ou de xeito forte, ou en case todo punto, ou case seguro*) a X se

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1$$

Escribimos $X_n \xrightarrow{\text{wp1}} X, n \to \infty$.

Observación 1.9.

$$\left\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\} = \bigcap_{\varepsilon>0} \bigcup_{n=1}^{\infty} \{\omega : |X_m(\omega) - X(\omega)| < \varepsilon, \text{ all } m \geq n\}$$

Isto permite afirmar que o conxunto $\{\omega : X_n(\omega) \to X(\omega)\}$ pertence verdadeiramente a $\mathscr{F}$, tal é como precisamos para que a converxencia con probabilidade 1 estea ben definida.

Definición 1.10. Considérense funcións de distribución $F_1(\cdot), F_2(\cdot), ...$ e $F(\cdot)$. Sexan $X_1, X_2, ...$ e X variables aleatorias (non necesariamente definidas nun espazo de probabilidade común) distribuíndose consonte estas leis respectivamente. Dicimos que X_n *converxe en distribución* a X se

$$\lim_{n\to\infty} F_n(t) = F(t), \text{en case todo punto de continuidade } t \text{ de } F.$$

Escribimos $X_n \xrightarrow{d} X$, or $d - \lim_{n\to\infty} X_n = X$.

Proposición 1.11 (Desigualdade de Jensen)**.** *Se $g(\cdot)$ é unha función convexa en $\mathbb{R}$, e X e $g(X)$ son tales que a súa media existe entón*

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)]$$

Definición 1.12. Considérense variables aleatorias $X_1, X_2, ...$ e X en $(\Omega, \mathscr{F}, P)$. Para $r > 0$, dicimos que X_n *converxe en r-media* a X se

$$\lim_{n\to\infty} \mathbb{E}\,|X_n - X|^r = 0$$

Isto escríbese $X_n \xrightarrow{\text{r}} X$ ou $L^r - \lim_{n\to\infty} X_n = X$. Canto máis alto sexa o valor de r, máis forte é a condición en virtude da desigualdade de Jensen.

$$X_n \xrightarrow{r} X \Rightarrow X_n \xrightarrow{\text{s}} X, 0 < s < r$$

1.2.3. Esperanza condicional

Definición 1.13 (*σ-álxebra xerada por unha variable aleatoria*)**.** Para X variable aleatoria, definimos $\mathscr{F}_X = \left\{X^{-1}(B); B \text{ Boreliano}\right\}$

Pode probarse que $\mathscr{F}_X$ é a σ-álxebra máis pequena $\mathscr{F}$ que fai de X unha función medible.

Definición 1.14 (*σ-álxebra xerada por variables aleatorias*)**.** A σ -álxebra xerada pola colección de variables aleatorias $(X_\gamma)_{\gamma\in\Gamma}$ é a σ -álxebra máis pequena $\mathscr{F}_\Gamma$ en Ω máis pequena tal que para cada $\gamma \in \Gamma$ a variable aleatoria X_γ é $\mathscr{F}_\Gamma$ -medible. Denotamos a σ -álxebra xerada pola colección por $\mathscr{F}_\Gamma = \sigma\left((X_\gamma)_{\gamma\in\Gamma}\right)$

A σ -álxebra que cumpre a anterior definición existe e é única: é a intersección de todas as σ -álxebras que satisfagan a propiedade enunciada na definición.

Teorema 1.15. *Se X e Y son variables aleatorias, entón Y pode escribirse como función Borel-medible de X, i.e. $Y = f(X)$ para algunha f Borel-medible se e só se*

$$\mathscr{F}_Y \subset \mathscr{F}_X$$

Definición 1.16. Sexa $\mathcal{G}$ unha sub-sigma álxebra de $\mathscr{F}$. Se X é unha variable aleatoria cuxa media existe, entón a *esperanza condicional* de X dado $\mathcal{G}$ é calquera variable aleatoria Z que cumpra as seguintes dúas propiedades:

(i) Z é $\mathcal{G}$ -medible

(ii) se $\Lambda \in \mathcal{G}$, entón

$$\int_\Lambda Z dP = \int_\Lambda X dP$$

Denotamos Z por $\mathbb{E}[X \mid \mathcal{G}]$. Vai implícito en (ii) que Z debe ser integrable.

Esta definición revela a interpretación das σ-álxebras como coñecemento porque canto máis fina é $\mathcal{G}$, máis esperanzas condicionais de X podemos calcular (máis *sucesos* haberá no noso espazo de probabilidade)

Proposición 1.17. *Se Z e Z' son dúas variables aleatorias satisfacendo (i) e (ii), entón $Z = Z'$ en case todo punto.*

A esperanza condicional no sentido *clásico* $\mathbb{E}[X \mid Y]$ concíbese como función de Y que ten a mesma integral ca X sobre conxuntos da forma $\{\omega : Y(\omega) = y\}$. Entón, en virtude de 1.13 e 1.15 temos que

$$\begin{aligned} &\mathbb{E}[X \mid Y] \text{ é} \mathscr{F}_Y \text{ -medible.} \\ &\int_\Lambda \mathbb{E}[X \mid Y] dP = \int_\Lambda X dP, \quad \forall \Lambda \in \mathscr{F}_Y \end{aligned} \tag{1.1}$$

Corolario 1.18.

$$\mathbb{E}[X \mid Y] = \mathbb{E}\left[X \mid \mathscr{F}_Y\right]$$

Acabamos de ver que a esperanza condicional $g(y) = \mathbb{E}(X \mid Y = y), y \in \mathcal{Z}$, pode transferirse a Ω formando $h(\omega) = g(Y(\omega))$ e entón $h = \mathbb{E}\left[X \mid \mathscr{F}_Y\right]$. Reciprocamente, calquera esperanza condicional $\mathbb{E}(X \mid \mathscr{G})$, con $\mathscr{G}$ unha sub σ-álxebra e $\mathscr{F}$, emerxe dun obxecto aleatorio Y desta maneira. Simplemente tomemos $Y : (\Omega, \mathscr{F}) \to (\Omega, \mathscr{G})$ a aplicación identidade: $Y(\omega) = \omega, \omega \in \Omega$. Entón $Y^{-1}(\mathscr{G}) = \mathscr{G}$, daquela se $g(y) = \mathbb{E}(X \mid Y = y)$ temos $h = \mathbb{E}(X \mid \mathscr{F}_Y) = \mathbb{E}(X \mid \mathscr{G})$ Agora intuitivamente, $\mathbb{E}(X \mid \mathscr{G}) = \mathbb{E}(X \mid Y)$ é a media de X dado un valor de Y coñecido. Pero que significa "saber"$Y : (\Omega, \mathscr{F}) \to (\Omega, \mathscr{G})$? Os eventos que involucran a Y son conxuntos da forma $\{Y \in G, G \in \mathscr{G}\}$, e xa que Y é a aplicación identidade, $\{Y \in G\} = G$.Como os eventos en realidade son preguntas que teñen unha resposta de *si* ou *non*, $\mathbb{E}(X \mid \mathscr{G})$ pódese interpretar como a media de $X(\omega)$, sabendo para cada $G \in \mathscr{G}$, se $\omega \in G$ ou non [Ash, 2014]

Por último, a probabilidade condicional é tan só un caso especial das esperanzas condicionais. Se B é un evento e I_B a súa función indicadora, entón I_B é unha variable aleatoria con $P\left\{I_B = 1 \mid A\right\} = P\{B \mid A\}$, e $P\left\{I_B = 0 \mid A\right\} = 1 - P\{B \mid A\}$ de tal xeito que

$$P\{B \mid A\} = \mathbb{E}\left\{I_B \mid A\right\}$$

1.2.4. Martingalas

Definición 1.19 (*Martingala regresiva*, [Serfling, 2009])**.** Consideremos unha sucesión de σ -álxebras $\{\mathscr{F}_n\}$ contida en $\mathscr{F}$, tal que Y_n é $\mathscr{F}_n$ -medible e $\mathbb{E}\,|Y_n| < \infty$. Entón a sucesión $\{Y_n, \mathscr{F}_n\}$ chámase martingala regresiva se

(i) $\mathscr{F}_1 \supset \mathscr{F}_2 \supset \ldots$

(ii) $\mathbb{E}\left[Y_n \mid \mathscr{F}_{n+1}\right] = Y_{n+1}$ con prob. 1, para todo n

Teorema 1.20 (Converxencia de submartingalas, [Loeve, 1978])**.** *Sexan variables aleatorias X_n que forman unha submartingala ou submartingala regresiva*

(i) Se $\sup \mathbb{E}X_n^+ < \infty$, *entón* $X_n \xrightarrow{c.s.} X < \infty$ *con* $\mathbb{E}X \leqq \sup \mathbb{E}X_n^+$ *e* $\mathbb{E}|X| \leqq \sup \mathbb{E}\,|X_n|$.

(ii) $X_n \xrightarrow{r} X$ *onde* $r \geqq 1$ *se é so se* $|X_n|^r$ *son uniformemente integrables [Takaoka, 1999], e daquela* $X_n \xrightarrow{c.s.} X$.

Observación 1.21. As martingalas son en particular submartingalas que saturan a desigualdade da súa definición, logo este resultado é válido para martingalas

1.3. Operadores normais en espazos de Hilbert

Empezamos definindo os conceptos de operador Hilbert-Schmidt e de clase traza. Sexa H un espazo de Hilbert arbitrario

Lema 1.22 ([Conway, 2019]). *Se $\{e_i\}$ e $\{f_j\}$ son dúas bases ortonormais de H e $A \in \mathscr{B}(H)$, entón*

1. $\sum_i \|Ae_i\|^2 = \sum_j \|Af_j\|^2 = \sum_i \sum_j |\langle Ae_i, f_j\rangle|^2$
2. *Se $A \in \mathscr{B}(H)$ e $\{e_i\}$ é unha base para H, definimos*

$$\|A\|_2 = \left[\sum_i \|Ae_i\|^2\right]^{1/2}$$

   Por 1. $\|A\|_2$ é independente da base escollida e entón está ben definido. Se $\|A\|_2 < \infty$, A chámase operador Hilbert-Schmidt e $\mathscr{B}_2 = \mathscr{B}_2(H)$ denota o conxunto de todos os operadores Hilbert-Schmidt.

3. *$\|A\| \leqslant \|A\|_2$ para todo A en $\mathscr{B}(H)$ e $\|\cdot\|_2$ é unha norma en $\mathscr{B}_2$.*

Sexa $\mathscr{B}_1(H) = \{AB : A \text{ e } B \in \mathscr{B}_2(H)\}$. Os operadores de $\mathscr{B}_1(H)$ chámanse de clase traza e $\mathscr{B}_1(H) = \mathscr{B}_1$ chámase clase traza.

Lema 1.23 ([Conway, 2019]). *Se $A \in \mathscr{B}_1(H)$ e $\{e_i\}$ é unha base entón $\sum |\langle Ae_i, e_i\rangle| < \infty$. Ademais, a suma $\sum \langle Ae_i, e_i\rangle$ é independente da elección de base.*

Definición 1.24. Se $\{e_i\}$ é unha base de H, definimos tr: $\mathscr{B}_1 \to \mathbb{C}$ por

$$\operatorname{tr}(A) = \sum_i \langle Ae_i, e_i\rangle$$

Por 1.23 a definición de $\operatorname{tr}(A)$ non depende da elección dunha base; $\operatorname{tr}(A)$ chámase a traza de A. Se $\dim H < \infty$, entón $\operatorname{tr}(A)$ é precisamente a suma dos termos da diagonal de calquera representación de A

1.4. Operador de covarianzas (cruzadas)

Sexa H_1 (resp., H_2) un espazo de Hilbert real e separable con produto escalar $\langle\cdot,\cdot\rangle_1$ (resp., $\langle\cdot,\cdot\rangle_1$) e σ -álxebra de Borel $\mathscr{F}_1$ (resp., H_2). Sexa $\mathscr{F}_1 \times \mathscr{F}_2$ a σ -álxebra xerada polos rectángulos medibles $A \times B, A \in \mathscr{F}_1, B \in \mathscr{F}_2$. Definimos $H_1 \times H_2 = \{(\mathbf{u}, \mathbf{v}) : \mathbf{u} \text{ en } H_1, \mathbf{v} \text{ en}$

$H_2\}$. $H_1 \times H_2$ é un $\mathbb{R}$-espazo vectorial con suma e multiplicación por escalares definidas por $(\mathbf{u},\mathbf{v})+(\mathbf{z},\mathbf{y}) = (\mathbf{u}+\mathbf{z},\mathbf{v}+\mathbf{y})$ e $k(\mathbf{u},\mathbf{v}) = (ku, kv)$. $H_1 \times H_2$ é un espazo de Hilbert separable co produto escalar $[\cdot,\cdot]$ definido por $[(u,v),(t,z)] = \langle u,t\rangle_1 + \langle v,z\rangle_2$; ademais, os abertos da topoloxía inducida por este produto escalar xeran $\mathscr{F}_1 \times \mathscr{F}_2$ [Parthasarathy, 2005]. Sexa $\|\cdot\|_1$ (resp., $\|\cdot\|_2$) a norma en H_1 (resp., H_2) inducida polo produto escalar, e sexa $\|\cdot\|$ a norma en $H_1 \times H_2$ inducida por $[\cdot,\cdot]$. Unha medida de probabilidade en $(H_1 \times H_2, \mathscr{F}_1 \times \mathscr{F}_2)$ chámase *medida conxunta*. Unha medida de probabilidade μ_i on $(H_i, \mathscr{F}_i)$ $(i = 1$ or $2)$ que satisfai

$$\int_{H_i} \|\mathbf{x}\|_i^2 d\mu_i(\mathbf{x}) < \infty$$

define un operador $\mathbf{R}_i$ en H_i e un elemento *media* $\mathbf{m}_i$ de H_i por

$$\langle \mathbf{m}_i, \mathbf{u}\rangle_i = \int_{H_i} \langle \mathbf{x}, \mathbf{u}\rangle_i d\mu_i(\mathbf{x})$$

e

$$\langle \mathbf{R}_i\mathbf{u}, \mathbf{v}\rangle_i = \int_{H_i} \langle \mathbf{x}-\mathbf{m}_i, \mathbf{u}\rangle_i \langle \mathbf{x}-\mathbf{m}_i, \mathbf{v}\rangle_i \, d\mu_i(\mathbf{x})$$

Dise que $\mathbf{R}_i$ é un *operador de covarianzas*; i.e., é linear, acotado, non negativo, autoadxunto, e de clase traza [Baker, 1973]

Agora supoñamos que μ_{XY} é unha medida conxunta satisfacendo

$$\int_{H_1\times H_2} \|(\mathbf{x},\mathbf{y})\|^2 d\mu_{XY}(\mathbf{x},\mathbf{y}) < \infty$$

Definimos un funcional G en $H_1 \times H_2$ por

$$G(\mathbf{u},\mathbf{v}) \equiv \int_{H_1\times H_2} \langle \mathbf{x}-\mathbf{m}_X, \mathbf{u}\rangle_1 \langle \mathbf{y}-\mathbf{m}_Y, \mathbf{v}\rangle_2 \, d\mu_{XY}(\mathbf{x},\mathbf{y})$$

Fixando $\mathbf{u}$ (resp $\mathbf{v}$), G é un funcional linear en H_2 (resp. H_1). Ademais, $|G(\mathbf{u},\mathbf{v})|^2 \leq \left\|\mathbf{R}_X^{1/}\mathbf{u}\right\|_1^2 \left\|\mathbf{R}_Y^{1/2}\mathbf{v}\right\|_2^2$, onde $\mathbf{R}_X$ e $\mathbf{R}_Y$ son os operadores de covarianzas de μ_X e μ_Y. Daquela, para $\mathbf{u}$ fixado, existe polo teorema de representación de Riesz un único elemento $\mathbf{q}_\mathrm{u}$ en H_2 tal que $G(\mathbf{u},\mathbf{v}) = \langle \mathbf{q}_\mathrm{u}, \mathbf{v}\rangle_2$ para todo $\mathbf{v}$ en H_2. De xeito semellante, para $\mathbf{v} \in H_2$ fixado existe un único elemento $\mathbf{g}_v \in H_1$ tal que $G(\mathbf{u},\mathbf{v}) = \langle \mathbf{g}_v, \mathbf{u}\rangle_1$ para todo $\mathbf{u} \in H_1$. Definimos a aplicación $\mathbf{R}_{XY} : H_2 \to H_1$ como $\mathbf{R}_{XY}\mathbf{v} = \mathbf{g}_v \cdot \mathbf{R}_{XY}$, que está ben definida polo feito de que $\mathbf{g}_\mathrm{v}$ é único. $\mathbf{R}_{XY}$ está definida en todo H_2, é claramente linear e é acotada porque

$$\begin{aligned}\|\mathbf{R}_{XY}\mathbf{v}\|_1^2 = \|\mathbf{g}_\mathbf{v}\|_1^2 &= \sup_{\mathbf{u}\epsilon H_1} \frac{\langle \mathbf{g}_\mathbf{v}, \mathbf{u}\rangle_1^2}{\|\mathbf{u}\|_1^2} = \sup_{\mathbf{u}\in H_1} \frac{|\mathbf{G}(\mathbf{v},\mathbf{u})|^2}{\|\mathbf{u}\|_1^2} \\ &\leq \sup_{\mathbf{u}\in H_1} \frac{\left\|\mathbf{R}_X^{1/2}\mathbf{u}\right\|_1^2}{\|\mathbf{u}\|_1^2} \left\|\mathbf{R}_Y^{1/2}\mathbf{v}\right\|_2^2 \leq \|\mathbf{R}_X\|_1 \|\mathbf{R}_Y\|_2 \|\mathbf{v}\|_2^2\end{aligned}$$

Claramente $\mathbf{R}_{XY}^{*} : H_1 \to H_2$ vén definida por $\mathbf{R}_{XY}^{*}\mathbf{u} = \mathbf{q}_{\mathrm{u}}$. Entón $G(\mathbf{u}, \mathbf{v}) = \langle \mathbf{R}_{XY}\mathbf{v}, \mathbf{u} \rangle_1 = \langle \mathbf{v}, \mathbf{R}_{XY}^{*}\mathbf{u} \rangle_2$ para todo $\mathbf{u}$ en H_1 e $\mathbf{v}$ en H_2. Definimos $\mathbf{R}_{XY}^{*} = \mathbf{R}_{YX}$. O operador $\mathbf{R}_{XY}$ chámase *operador de covarianzas cruzadas* de μ_{XY}.

Capítulo 2

O dualismo distancia e núcleo

Malia a súa flagrante similitude, o enlace teórico entre as filosofías coñecidas como *energy statistics* e *aprendizaxe núcleo* foron un problema aberto até a publicación de [Sejdinovic et al., 2013]. *Distance covariance* é unha metodoloxía baseada na noción de distancia entre observacións que se pode empregar en contrastes de independencia condicional, screening, clustering, análise de series de tempo e dependencia en grafos, entre outros moitos. Á parte das aplicacións discutidas, paga a pena mencionar algunha contribución metodolóxica da distance covariance para enfrontarse a outros mecanismos de mostraxe como o *missing data* para examinar a relación entre biomarcadores da diabetes e cambios no volume de glicosa en sangue a longo prazo [Matabuena et al., 2021].

Doutra banda, o criterio de independencia de Hilbert-Schmidt é un método baseado en núcleos para contrastar independencia e igualmente popular en tarefas relacionadas.

Estes dous métodos fundacionais comparten formulacións similares e moitas propiedades comúns tal e como veremos neste capítulo.

2.1. Distance covariance

Introducíase en [Székely et al., 2007] un revolucionario enfoque para contrastar independencia entre dous vectores aleatorios Euclidianos. O novo marco teórico dos *estatísticos enerxía* nacera o derradeiro ano do século pasado no contexto de contrastar igualdade de distribucións [Szekely, 2000] e recibe ese nome para recalcar a súa natureza: igual que a enerxía potencial gravitatoria [electromagnética] dun sistema formado por un número de finito de masas [cargas], a versión empírica dos estatísticos enerxía son funcións da distancia euclidiana entre pares de observacións [Székely and Rizzo, 2017]. De feito, a observación clave foi decatarse de que a distancia L^2 entre funcións de distribución [Cramér, 1928] pode escribirse do seguinte xeito:

$$\int_{-\infty}^{\infty} (F(x) - G(x))^2 \ \mathrm{d}x = 2E|\mathrm{X} - Y| - E\left|X - X'\right| - E\left|Y - Y'\right|$$

Neste contexto, [Székely et al., 2007] ataca o problema de contrastar independencia entre X e Y, onde X é un vector aleatorio de $\mathbb{R}^p$ e Y é un vector aleatorio de $\mathbb{R}^q$, a través dunha distancia L_2 pesada entre as funcións características da distribución conxunta de X e Y e o produto das súas marxinais Se f_X e f_Y denotan as funcións características de X e Y, respectivamente, e a súa función característica conxunta é $f_{X,Y}$, entón X e Y son independentes se e só se $f_{XY} = f_X f_Y$. A definición orixinal de distance covariance de X e Y con media finita é o número non negativo $\mathcal{V}(X,Y)$ definido por

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \|f_{X,Y}(t,s) - f_X(t)f_Y(s)\|^2 \\ &= \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2}{|t|_p^{1+p}|s|_q^{1+q}} dtds\end{aligned}$$

onde $c_d = \frac{\pi^{(1+1)/2}}{\Gamma\left(\frac{d+1}{2}\right)}$

En virtude do Teorema de Fubini e do Lema 5.2 temos

$$\begin{aligned}\mathcal{V}^2(X,Y) = \mathbb{E}\left[\left|X - X'\right|_p \left|Y - Y'\right|_q\right] + \mathbb{E}|X - X'|_p \mathbb{E}|Y - Y''|_q \\ - 2\mathbb{E}\left[\left|X - X'\right|_p \left|Y - Y''\right|_q\right]\end{aligned}$$

onde $(X,Y), (X',Y)$, and (X'',Y'') son copias i.i.d. En [Lyons, 2013] deuse o paso cara a xeneralización da distance covariance a espazos métricos de tipo negativo, sendo estes un conepto clave de machine learning [Vapnik, 1995]. Máis tarde, a noción estendeuse a espazos *semi*métricos de tipo negativo [Sejdinovic et al., 2013].

Definición 2.1 (Distance covariance en espazos semimétricos de tipo negativo)**.** Sexan $(\mathcal{X}, \rho_{\mathcal{X}})$ e $(\mathcal{Y}, \rho_{\mathcal{Y}})$ espazos semimétricos de tipo negativo e sexan $X \sim P_X \in \mathcal{M}^2_{\rho_{\mathcal{X}}}(\mathcal{X})$ e $Y \sim P_Y \in \mathcal{M}^2_{\rho_{\mathcal{Y}}}(\mathcal{Y})$, con distribución conxunta P_{XY}. A distance covariance xeneralizada de X e Y é

$$\begin{aligned}\mathcal{V}^2_{\rho_{\mathcal{X}},\rho_{\mathcal{Y}}}(X,Y) = \mathbb{E}_{XY}\mathbb{E}_{X'Y'}\rho_{\mathcal{X}}\left(X,X'\right)\rho_{\mathcal{Y}}\left(Y,Y'\right) + \mathbb{E}_X\mathbb{E}_{X'}\rho_{\mathcal{X}}\left(X,X'\right)\mathbb{E}_Y\mathbb{E}_{Y'}\rho_{\mathcal{Y}}\left(Y,Y'\right) \\ - 2\mathbb{E}_{XY}\left[\mathbb{E}_{X'}\rho_{\mathcal{X}}\left(X,X'\right)\mathbb{E}_{Y'}\rho_{\mathcal{Y}}\left(Y,Y'\right)\right]\end{aligned}$$

A condición de que as variables aleatorias teñan media finita vai ser suficiente para que estas esperanzas existan. Precisaremos afondar na teoría máis a fondo para dar unha proba rigorosa desta afirmación en 2.2.2.

Para que distance covariance caracterice independencia nun espazo métrico [i.e., $\mathcal{V}^2_{\rho_{\mathcal{X}},\rho_{\mathcal{Y}}}(X,Y) = 0$ se e só se X e Y son independentes] precisamos que as métricas $\rho_{\mathcal{X}}$ e $\rho_{\mathcal{Y}}$ cumpran unha

propiedade adicional chamada *tipo negativo forte*. Veremos a relación que gardan cunha clase moi especial de núcleos (característicos, ver Sección 2.2) no final do documento.

2.2. Espazo de Hilbert con núcleo reprodutor

Definición 2.2 (Espazo de Hilbert con núcleo reprodutor)**.** Sexa H un espazo de Hilbert de funcións reais definidas en $\mathcal{Z}$. Unha función $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ é un *núcleo reprodutor* de H se satisfai as seguintes condicións:

1. $h_z \equiv k(\cdot, z) \in H$ para todo $z \in \mathcal{Z}$;
2. $\langle h_z, f \rangle_H = f(z)$ para todo $z \in \mathcal{Z}$ e $f \in H$.

Se tal función núcleo k existe, entón H chámase RKHS (do inglés, *reproducing kernel Hilbert space*).

Observación 2.3. O *funcional avaliación* que actúa sobre H é un funcional linear que avalía cada función de H nun punto fixado $z \in \mathcal{Z}$, $L_z : f \mapsto f(z)$, $f \in H$. A definición orixinal de RKHS é un espazo de Hilbert de funcións reais sobre $\mathcal{Z}$ tal que para todo z en $\mathcal{Z}, L_z$ é continua en cada f de H ou, equivalentemente, se L_x é un operador acotado en H; i.e., existe $M_z > 0$ tal que

$$|L_z(f)| := |f(z)| \leq M_z \|f\|_H \quad \forall f \in H$$

Aínda que se asume $M_z < \infty$ para todo $z \in Z$, podería ser o caso de que $\sup_z M_z = \infty$. Liñas arriba, presentamos unha definición de RKHS máis intuitiva obtida a través de observar que esta última definición permite representar o funcional avaliación tomando o produto escalar de f cunha función k_z de H. Isto próbase facilmente utilizando dúas veces seguidas o teorema de representación de Riesz.

Definición 2.4. Sexa $\mathcal{Z}$ un conxunto non baleiro. Unha función $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ chámase *definida positiva* (e **non** *semi*definida positiva) se

$$\sum_{j,k=1}^{n} c_j c_k k\left(z_j, z_k\right) \geqq 0$$

para todo $n \in \mathbb{N}, \{z_1, \ldots, z_n\} \subseteq \mathcal{Z}$ e $\{c_1, \ldots, c_n\} \subseteq \mathbb{R}$.

A posibilidade de representar un núcleo $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ como $k(x, y) = \langle \varphi(x), \varphi(y) \rangle$ para certa función φ sobre $\mathcal{Z}$ que tome valores nun certo espazo de Hilbert sóubose por suposto que era certa no caso de $\mathcal{Z}$ finito. Para $\mathcal{Z}$ numerable, o resultado probouno Kolmogorov en

1941, e anos máis tarde Aronszajn plantexou a pregunta plenamente no primeiro dos seus- importantísimos- traballos [Aronszajn, 1943] e [Aronszajn, 1950], dando no segundo deles o primeiro tratamento sistemático da teoría de núcleos reprodutores. Dende entón, estas innovacións atoparon moitas aplicacións dentro da análise matemática, como na teoría de funcións complexas [Hille, 1972], ou na análise de series de tempo [Parzen, 1974].

Teorema 2.5 (Moore-Aronszajn)**.** *Para toda función simétrica e definida positiva $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, existe un RKHS asociado, H_k, núcleo reprodutor k. A aplicación $\varphi : z \mapsto h_z \in H_k$ chámase aplicación canónica ou de Aronszajn de k.*

O teorema de Moore-Aronszajn dá pé á seguinte definición:

Definición 2.6. Chamamos *núcleo* a toda función simétrica e definida positiva $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ e dicimos que k é *non dexenerado* se a aplicación canónica é inxectiva.

A noción de aplicación canónica pode estenderse a embeddings de medidas de $\mathcal{M}(\mathcal{Z})$ do seguinte xeito [Smola et al., 2007]:

Definición 2.7 (*Kernel embedding*)**.** Sexa k un núcleo sobre $\mathcal{Z}$, e $v \in \mathcal{M}(\mathcal{Z})$. O kernel embedding de v no RKHS H_k, se existe, é $\mu_k(v) \in H_k$ tal que $\int f(z)dv(z) = \langle f, \mu_k(v) \rangle_{H_k}$ para todo $f \in H_k$

Precisamos restrinxir a nosa atención a unha clase particular de medidas para as que existe o kernel embedding (veremos máis adiante que isto está relacionado con que as variables aleatorias consideradas no distance covariance deben ter media finita en 2.2.2). Sexa k un núcleo medible en $\mathcal{Z}$, e denotamos para $\theta > 0$

$$\mathcal{M}_k^\theta(\mathcal{Z}) = \left\{ v \in \mathcal{M}(\mathcal{Z}) : \int k^\theta(z,z) d|v|(z) < \infty \right\}$$

Claramente, $\theta_1 \leq \theta_2 \quad \Rightarrow \quad \mathcal{M}_k^{\theta_2}(\mathcal{Z}) \subseteq \mathcal{M}_k^{\theta_1}(\mathcal{Z})$. Nótese que o kernel embedding $\mu_k(v)$ está ben definido $\forall v \in \mathcal{M}_k^{1/2}(\mathcal{Z})$, polo teorema de representación de Riesz.

Tal e como vimos, os kernel embeddings de medidas de probabilidade de Borel en $\mathcal{M}_+^1(\mathcal{Z}) \cap \mathcal{M}_k^{1/2}(\mathcal{Z})$ existen, e podemos introducir a noción de distancia entre medidas de probabilidade de Borel neste conxunto usando a distancia inducida polo produto escalar do espazo de Hilbert no que viven os seus kernel embeddings.

Definición 2.8 (*Maximum mean discrepancy (MMD)*)**.** Sexa k un núcleo en $\mathcal{Z}$, e sexan $P, Q \in \mathcal{M}_+^1(\mathcal{Z}) \cap \mathcal{M}_k^{1/2}(\mathcal{Z})$. O MMD γ_k entre P e Q defínese como

$$\gamma_k(P,Q) = \|\mu_k(P) - \mu_k(Q)\|_{H_k}$$

A seguinte representación alternativa do cadrado do MMD vai ser útil [Gretton et al., 2012]

$$\gamma_k^2(P,Q) = \mathbb{E}_{ZZ'}k\left(Z,Z'\right) + \mathbb{E}_{WW'}k\left(W,W'\right) - 2\mathbb{E}_{ZW}k(Z,W)$$

onde Z, Z' i.i.d. P e W, W' i.i.d. Q. Se a restrición de μ_k a certo $\mathcal{P}(\mathcal{Z}) \subseteq \mathcal{M}_+^1(\mathcal{Z})$ está ben definida e é inxectiva, entón dise que k é *característico respecto de* $\mathcal{P}(\mathcal{Z})$ é dise que é *característico* (sen ningún adxectivo máis) se é característico respecto de $\mathcal{M}_+^1(\mathcal{Z})$. Cando k é característico, γ_k é unha métrica en todo $\mathcal{M}_+^1(\mathcal{Z})$, e en particular $\gamma_k(P,Q) = 0$ se e só se $P = Q, \forall P, Q \in \mathcal{M}_+^1(\mathcal{Z})$

A MMD pode empregarse para medir dependencia estatística entre variables aleatorias [Gretton et al., 2005]. Sexan $\mathcal{X}$ e $\mathcal{Y}$ dous espazos topolóxicos non baleiros e sexan $k_\mathcal{X}$ e $k_\mathcal{Y}$ núcleos sobre $\mathcal{X}$ ae $\mathcal{Y}$, cos seus respectivos RKHSs $H_{k_\mathcal{X}}$ e H_{k_y}. Entón podemos aplicar o seguinte

Lema 2.9 (Produto de núcleos, [Steinwart and Christmann, 2008]). *Sexa k_1 un núcleo en X_1 e k_2 núcleo en X_2. Entón $k_1 \cdot k_2$ é un núcleo en $X_1 \times X_2$. En particular, se $X_1 = X_2$, logo $k\left(x,x'\right) := k_1\left(x,x'\right) k_2\left(x,x'\right), x, x' \in X$, define un núcleo en X.*

para deducir que

$$k\left((x,y),\left(x',y'\right)\right) := k_\mathcal{X}\left(x,x'\right) k_\mathcal{Y}\left(y,y'\right)$$

é un núcleo no espazo produto $\mathcal{X} \times \mathcal{Y}$ con RKHS H_k isometricamente isomorfo ao produto tensorial $H_{k_\mathcal{X}} \otimes H_{k_\mathcal{Y}}$.

Definición 2.10 (*Criterio de independencia Hilbert-Schmidt (HSIC)*). Sexan $X \sim P_X$ e $Y \sim P_Y$ variables aleatorias sobre $\mathcal{X}$ e $\mathcal{Y}$, respectivamente, con distribución conxunta P_{XY}. Ademais, sexa k o núcleo en $\mathcal{X} \times \mathcal{Y}$ construído como na conclusión do Lema anterior. O HSIC de X e Y é a MMD γ_k entre a distribución conxunta P_{XY} e o produto das marxinais $P_X P_Y$

$$\gamma_k^2\left(P_{XY}, P_X P_Y\right)$$

Definiamos o operador de covarianzas cruzadas $C_{xy} : H_{k_\mathcal{X}} \to H_{k_\mathcal{Y}}$ como aquel tal que para todo $f \in H_{k_\mathcal{Y}}$ e $g \in H_{k_\mathcal{X}}$

$$\langle f, C_{xy} g\rangle_{H_{k_\mathcal{Y}}} = \mathbb{E}_{xy}\left(\left[f(x) - \mathbb{E}_x(f(x))\right]\left[g(y) - \mathbb{E}_y(g(y))\right]\right).$$

De acordo con [Gretton et al., 2005], o operador de covarianzas cruzadas pode escribirse como

$$C_{xy} := \mathbb{E}_{xy}\left[(\phi(x) - \mu_x) \otimes (\psi(y) - \mu_y)\right]$$

onde $\mu_x := \mathbb{E}_x\phi(x), \mu_y := \mathbb{E}_y\phi(y)$, ϕ e ψ son as aplicacións de Aronsajn de cada RKHS e $\otimes$ é o produto tensorial. Podemos empregar a norma Hilbert-Schmidt do operador de covarianzas (que ven a ser a suma dos valores singulares ao cadrado), que ten o seguinte valor poboacional

$$\|C_{xy}\|^2_{H_{k\mathcal{X}}\otimes H_{k\mathcal{Y}}} = \|\mathbb{E}_{XY}\left[k_{\mathcal{X}}(\cdot,X)\otimes k_{\mathcal{Y}}(\cdot,Y)\right] - \mathbb{E}_X k_{\mathcal{X}}(\cdot,X)\otimes\mathbb{E}_Y k_{\mathcal{Y}}(\cdot,Y)\|^2_{H_{k\mathcal{X}}\otimes H_{k\mathcal{Y}}}$$

$$\begin{aligned}&= \mathbb{E}_{XYX'Y'}k_{\mathcal{X}}\left(X,X'\right)k_{\mathcal{Y}}\left(Y,Y'\right) + \left(\mathbb{E}_X\mathbb{E}_{X'}k_{\mathcal{X}}\left(X,X'\right)\right)\left(\mathbb{E}_Y\mathbb{E}_{Y'}k_{\mathcal{Y}}\left(Y,Y'\right)\right)\\ &\quad -2\mathbb{E}_{X'Y'}\left[\mathbb{E}_X k_{\mathcal{X}}\left(X,X'\right)\mathbb{E}_Y k_{\mathcal{Y}}\left(Y,Y'\right)\right]\end{aligned}$$

Pode demostrarse que esta cantidade é igual ao HSIC entre X e Y [Gretton et al., 2005], e será esta expresión a partir da cal escribamos un estimador empírico de HSIC. $\gamma_k^2\left(P_{XY},P_XP_Y\right)$ está ben definido sempre e cando $P_X\in\mathcal{M}^1_{k_{\mathcal{X}}}(\mathcal{X})$ e $P_Y\in\mathcal{M}^1_{k_Y}(\mathcal{Y})$. De feito, isto é suficiente para que $\mu_k\left(P_{XY}\right)$ exista, xa que isto implica que $P_{XY}\in\mathcal{M}_k^{1/2}(\mathcal{X}\times\mathcal{Y})$, o que se pode deducir da desigualdade de Cauchy-Schwartz,

$$\begin{aligned}&\int k^{1/2}((x,y),(x,y))dP_{XY}(x,y)\\ &\quad= \int k_{\mathcal{X}}^{1/2}(x,x)k_{\mathcal{Y}}^{1/2}(y,y)dP_{XY}(x,y)\\ &\quad\leq \left(\int k_{\mathcal{X}}(x,x)dP_X(x)\int k_{\mathcal{Y}}(y,y)dP_Y(y)\right)^{1/2}\end{aligned}$$

Ademais, o embedding do produto das marxinais $\mu_k\left(P_XP_Y\right)$ tamén existe porque se pode identificar co produto tensorial $\mu_{k_{\mathcal{X}}}\left(P_X\right)\otimes\mu_{k_{\mathcal{Y}}}\left(P_Y\right)$, onde $\mu_{k_{\mathcal{X}}}\left(P_X\right)$ existe porque $P_X\in\mathcal{M}^1_{k_{\mathcal{X}}}(\mathcal{X})\subset\mathcal{M}^{1/2}_{k_{\mathcal{X}}}(\mathcal{X})$, e $\mu_{k_{\mathcal{Y}}}\left(P_Y\right)$ existe porque $P_Y\in\mathcal{M}^1_{k\mathcal{Y}}(\mathcal{Y})\subset\mathcal{M}^{1/2}_{k\mathcal{Y}}(\mathcal{Y})$

2.2.1. Correspondencia entre núcleos e semimétricas

Nesta sección desenvolvemos a correspondencia entre semimétricas de tipo negativo e a teoría dos RKHS; i.e., núcleos simétricos e definidos positivos. Esta correspondencia será chave para probar a equivalencia entre a distance covariance e HSIC.

Lema 2.11 ([Berg et al., 1984])**.** *Sexa $\mathcal{Z}$ un conxunto non baleiro, e $\rho:\mathcal{Z}\times\mathcal{Z}\to\mathbb{R}$ unha semimétrica en $\mathcal{Z}$. Sexa $z_0\in\mathcal{Z}$, e denotemos $k\left(z,z'\right)=\rho\left(z,z_0\right)+\rho\left(z',z_0\right)-\rho\left(z,z'\right)$. Entón k é definido positivo se e só se o espazo semimétrico $(\mathcal{Z},\rho)$ é de tipo negativo*

Definición 2.12 (Núcleo inducido por unha distancia)**.** . Sexa ρ unha semimétrica de tipo negativo en $\mathcal{Z}$ e sexa $z_0\in\mathcal{Z}$. O núcleo

$$k\left(z,z'\right)=\frac{1}{2}\left[\rho\left(z,z_0\right)+\rho\left(z',z_0\right)-\rho\left(z,z'\right)\right]$$

o núcleo inducido por ρ e centrado en z_0.

Por brevidade, omitimos o cualificativo *inducido*, e diremos que k é simplemente o *núcleo distancia*, abusando un pouco da terminoloxía.

Variando o punto z_0, obtemos unha familia

$$\mathcal{K}_\rho = \left\{ \frac{1}{2} \left[\rho\left(z, z_0\right) + \rho\left(z', z_0\right) - \rho\left(z, z'\right) \right] \right\}_{z_0 \in \mathcal{Z}}$$

de núcleos distancia inducidos por ρ. A seguinte proposición conclúese da definición de $\mathcal{K}_\rho$ e amosa que sempre se pode expresar o apartado 2 da Proposición 1.3 empregando a aplicación canónica do RKHS H_k.

Proposición 2.13. *Sexa $(\mathcal{Z}, \rho)$ un espazo semimétrico de tipo negativo, e $k \in \mathcal{K}_\rho$. Entón:*

1. $\rho\left(z, z'\right) = k(z, z) + k\left(z', z'\right) - 2k\left(z, z'\right) = \left\| k(\cdot, z) - k\left(\cdot, z'\right) \right\|_{H_k}^2$
2. *k é non dexenerado; i.e., a aplicación canónica $z \mapsto k(\cdot, z)$ é inxectiva.*

Agora damos un páso alén no vínculo entre as semimétricas de tipo negativo e os núcleos. Empezamos cun corolario da Proposición 1.3.

Corolario 2.14. *Sexa k un núcleo non dexenerado en $\mathcal{Z}$. Daquela,*

$$\rho\left(z, z'\right) = k(z, z) + k\left(z', z'\right) - 2k\left(z, z'\right)$$

define unha semimétrica de tipo negativo válida ρ en $\mathcal{Z}$.

Definición 2.15 (Núcleos equivalentes)**.** Sempre que o núcleo k e a semimétrica ρ satisfagan 2.14, diremos que k xera ρ. Se dous núcleos xeran a mesma semimétrica diremos que son *núcleos equivalentes*.

Parece claro que cada núcleo distancia $\tilde{k} \in \mathcal{K}_\rho$ inducido por ρ, tamén xera ρ. Non obstante, hai moitos outros núcleos que xeran ρ. A seguinte proposición dá unha condición baixo a cal dous núcleos son equivalentes

Proposición 2.16. *Sexan k e $\tilde{k}$ dous núcleos sobre $\mathcal{Z}$ .Entón k e $\tilde{k}$ son equivalentes se e só se $\tilde{k}\left(z, z'\right) = k\left(z, z'\right) + f(z) + f\left(z'\right)$, para algunha función shift $f : \mathcal{Z} \to \mathbb{R}$.*

Non toda elección de función shift f na Proposición 2.16 vai ser válida, xa que ambos k e $\tilde{k}$ teñen que ser definidos positivos. Unha clase importante de funcións shift pode derivarse empregando as funcións que viven no RKHS. De feito, sexa k un núcleo sobre $\mathcal{Z}$ e sexa $f \in H_k$,. Definimos o núcleo

$$\begin{aligned} \tilde{k}_f\left(z, z'\right) &= \left\langle k(\cdot, z) - f, k\left(\cdot, z'\right) - f \right\rangle_{H_k} \\ &= k\left(z, z'\right) - f(z) - f\left(z'\right) + \|f\|_{H_k}^2 \end{aligned}$$

Como é representable como un produto escalar nun espazo de Hilbert, $\tilde{k}_f$ é un núcleo válido que é equivalente a k pola Proposición 2.16.Como caso especial, se $f = \mu_k(P)$ para algunha $P \in \mathcal{M}_+^1(\mathcal{Z})$, obtemos o *núcleo centrado na medida de probabilidade* P :

$$\tilde{k}_P\left(z, z'\right) := k\left(z, z'\right) + \mathbb{E}_{WW'} k\left(W, W'\right) - \mathbb{E}_W k(z, W) - \mathbb{E}_W k\left(z', W\right) \tag{2.1}$$

con W, W' i.i.d. consonte P. Nótese que $\mathbb{E}_{ZZ' \text{ i.i.d. } P} \tilde{k}_P(Z, Z') = 0$, e entón $\mu_{\tilde{k}_P}(P) = 0$. Os núcleos da forma 2.1 que están centrados nas masas puntuais $P = \delta_{z_0}$ son precisamente os núcleos distancia equivalentes a k.

2.2.2. Existencia do kernel embedding a través da semimétrica que xera

Despois da Definición 2.7, vimos que unha condición suficiente para que exista o kernel embedding $\mu_k(v)$ de $v \in \mathcal{M}(\mathcal{Z})$ era que $v \in \mathcal{M}_k^{1/2}(\mathcal{Z})$. Interpretaremos agora esta condición en termos da semimétrica ρ xerada por k, relacionando $\mathcal{M}_k^\theta(\mathcal{Z})$ co espazo $\mathcal{M}_\rho^\theta(\mathcal{Z})$ de medidas con θ -momento finito respecto a ρ.

Proposición 2.17. *Sexa k un núcleo que xera a semimétrica ρ, e sexa $n \in \mathbb{N}$. Entón $\mathcal{M}_k^{n/2}(\mathcal{Z}) = \mathcal{M}_\rho^{n/2}(\mathcal{Z})$. En particular, se k_1 e k_2 xeran a mesma semimétrica ρ, entón $\mathcal{M}_{k_1}^{n/2}(\mathcal{Z}) = \mathcal{M}_{k_2}^{n/2}(\mathcal{Z})$.*

Observación 2.18. Agora estamos en condicións de demostrar que $P, Q \in \mathcal{M}_\rho^1(\mathcal{Z})$ é suficiente para demostrar a validez da definición de distance covariance para semimétricas de tipo negativo en xeral ρ. En efecto, sexa k calquera núcleo que xera a ρ e $P, Q \in \mathcal{M}_k^1(\mathcal{Z})$. Entón,

$$\mathbb{E}_{ZW} \rho(Z, W) = \mathbb{E}_Z k(Z, Z) + \mathbb{E}_W k(W, W) - 2\mathbb{E}_{ZW} k(Z, W) < \infty$$

onde o primeiro termo é finito porque $P \in \mathcal{M}_k^1(\mathcal{Z})$, o segundo termo é finito porque $Q \in \mathcal{M}_k^1(\mathcal{Z})$, e o terceiro termo é finito porque $|k(z, w)| \leq k^{1/2}(z, z) \times k^{1/2}(w, w)$ e $P, Q \in \mathcal{M}_k^1(\mathcal{Z}) \subset \mathcal{M}_k^{1/2}(\mathcal{Z})$.

A Proposición 2.17 dá unha interpretación natural das condicións que lle pedimos ás medidas de probabilidade en termos de momentos con respecto a ρ. En efecto, o kernel embedding $\mu_k(P)$, onde o núcleo k xera a semimétrica ρ, existe para toda P con momento $1/2$ finito respecto a ρ e entón a MMD $\gamma_k(P, Q)$ entre P e Q está ben definida sempre e cando P e Q teñan momentos $1/2$ finitos con respecto a ρ. Ademais, o HSIC entre variables aleatorias X e Y está ben definido sempre e cando as marxinais P_X e P_Y momentos de primeira orde finitos con respecto á semimétrica $\rho_{\mathcal{X}}$ e $\rho_{\mathcal{Y}}$ xerada por núcleos $k_{\mathcal{X}}$ e $k_{\mathcal{Y}}$ nos seus respectivos dominios $\mathcal{X}$ e $\mathcal{Y}$.

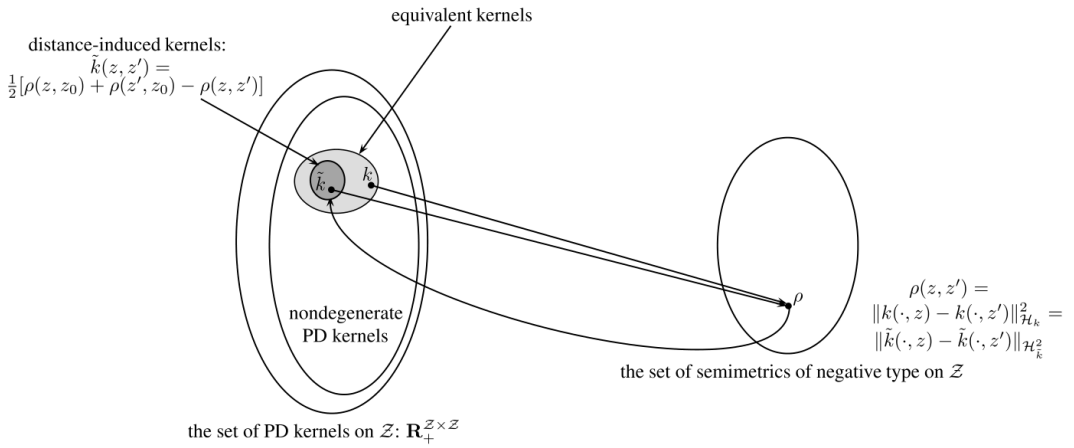


Figura 2.1: A relación entre núcleos e semimétricas. Un conxunto de núcleos definidos positivos non dexenerados asígnase a unha única semimétrica de tipo negativo, e os núcleos distancia inducidos por esa semimétrica en cuestión forman soamente un subconxunto desa clase. Diagrama tomado de [Sejdinovic et al., 2013]

2.3. Equivalencia entre HSIC e distance covariance

Teorema 2.19 (Equivalencia entre HSIC e distance covariance)**.** *Sexan* $(\mathcal{X}, \rho_{\mathcal{X}})$ *e* $(\mathcal{Y}, \rho_{\mathcal{Y}})$ *espazos semimétricos de tipo negativo e sexan* $X \sim P_X \in \mathcal{M}^2_{\rho_{\mathcal{X}}}(\mathcal{X})$ *e* $Y \sim P_Y \in \mathcal{M}^2_{\rho_{\mathcal{Y}}}(\mathcal{Y})$*, con distribución conxunta* P_{XY}*. Sexa* $k_{\mathcal{X}}$ *e* k_y *dous núcleos calquera sobre* $\mathcal{X}$ *e* $\mathcal{Y}$ *que xeran a* $\rho_{\mathcal{X}}$ *e a* $\rho_{\mathcal{Y}}$*, respectivamente e denotemos* $k\left((x,y),(x',y')\right) = k_{\mathcal{X}}\left(x,x'\right) k_{\mathcal{Y}}\left(y,y'\right)$ *Logo*

$$\mathcal{V}^2_{\rho_{\mathcal{X}},\rho_{\mathcal{Y}}}(X,Y) = 4\gamma_k^2\left(P_{XY}, P_X P_Y\right)$$

Demostración. Definimos $v = P_{XY} - P_X P_Y$. Entón

$$\begin{aligned}
&\mathcal{V}^2_{\rho_{\mathcal{X}},\rho_{\mathcal{Y}}}(X,Y) = \\
&= \iint \rho_{\mathcal{X}}\left(x,x'\right) \rho_{\mathcal{Y}}\left(y,y'\right) dv(x,y) dv\left(x',y'\right) \\
&= \iint \left(k_{\mathcal{X}}(x,x) + k_{\mathcal{X}}\left(x',x'\right) - 2k_{\mathcal{X}}\left(x,x'\right)\right) \times \left(k_{\mathcal{Y}}(y,y) + k_{\mathcal{Y}}\left(y',y'\right) - 2k_{\mathcal{Y}}\left(y,y'\right)\right) dv(x,y) dv\left(x',y'\right) \\
&= 4 \iint k_{\mathcal{X}}\left(x,x'\right) k_{\mathcal{Y}}\left(y,y'\right) dv(x,y) dv\left(x',y'\right) \\
&= 4\gamma_k^2\left(P_{XY}, P_X P_Y\right)
\end{aligned}$$

onde empregamos a descomposición do HSIC derivada da súa identificación coa norma do operador de covarianzas cruzadas e que $v(\mathcal{X}\times\mathcal{Y}) = 0$,, que

$$\int g\left(x,y,x',y'\right) dv(x,y) dv\left(x',y'\right) = 0$$

cando g non depende de un ou máis dos seus argumentos, xa que v tamén tén medidas marxinais nulas. A converxencia das integrais da forma $\int k_{\mathcal{X}}(x,x)\times\ k_{\mathcal{Y}}(y,y)dv(x,y)$ está asegurada polas condicións que impuxemos aos momentos das marxinais.

□

Observación 2.20. Para asegurar a existencia da distance covariance, impuxemos unha condición máis forte nas marxinais: $P_X \in \mathcal{M}^2_{k_{\mathcal{X}}}(\mathcal{X})$ e $P_Y \in \mathcal{M}^2_{k_y}(\mathcal{Y})$, mentres que $P_X \in \mathcal{M}^1_{k_{\mathcal{X}}}(\mathcal{X})$ e $P_Y \in \mathcal{M}^1_{k_Y}(\mathcal{Y})$ son suficientes para a existencia do criteiro de independecia de Hilbert-Schmidt.

Capítulo 3

U-estatísticos

Os obxectos denominados U-estatísticos [Hoeffding, 1948a] xeneralizan de xeito elegante e útil a noción de media amostral e son un dos obxectos máis universais da estatística matemática. Alxebraicamente, son máis complicados que simples sumas de variables independentes polos padróns combinatorios que os caracterizan. Estatisticamente, están estruturados por complexas relacións de dependencia que eles mesmos amosan nas súas propiedades de martingala, que tamén explorarmos brevemente.

3.1. Fundamentos

Sexan $X_1, \ldots, X_k$ observacións independentes e identicamente distribuídas que toman valores nun espazo topolóxico e medible coa súa σ-álxebra de Borel dada pola topoloxía. Supoñamos tamén que $X_1, \ldots, X_k$ se distribúen consonte a función de distribución F. Considérese un funcional θ ben definido nun conxunto de funcións de distribución $\mathcal{F}$ que contén a F (por simplicidade só pediremos isto, de momento):

$$\theta : \mathcal{F} \to \mathbb{R}$$

Supoñamos que $F \in \mathcal{F}$ é tal que $\theta(F)$ admite un estimador insesgado. Temos entón que

$$\theta(F) = E_F \left\{ h\left(X_1, \ldots, X_k\right) \right\} = \int \cdots \int h\left(x_1, \ldots, x_k\right) dF\left(x_1\right) \cdots dF\left(x_k\right)$$

para certa función $h = h\left(x_1, \ldots, x_k\right)$. O funcional θ recibe o nome de *funcional estatístico regular* de grao k, e a función h recibe o nome de *núcleo do funcional*.

Contamos co seguinte resultado concibido nun período histórico da estatística marcado pola busca frenética de estimadores insesgados de varianza mínima (MVUE).

Teorema 3.1 ([Fraser, 1954]). *Sexa θ un funcional estatístico regular de orde k con núcleo h definido nun conxunto $\mathcal{F}$ de funcións de distribución que conteña a todas as absolutamente continuas. Entón*

(i)

$$h^{[n]}(X_1, \dots, X_n) = \frac{(n-k)!}{n!} \sum h(X_{i_1}, \dots, X_{i_k}) \tag{3.1}$$

é o ***único*** *estimador simétrico e insesgado de θ, onde a suma percorre todos as $n!/(n-k)!$ permutacións $(i_1, \dots, i_k)$ de enteiros distintos*

(ii) O estimador $h^{[n]}$ é de mínima varianza na clase de todos os estimadores insesgados de θ.

O resultado tamén é válido para familias $\mathcal{F}$ que conteñen todas as funcións de distribución de soporte finito [Lee, 2019].

Observación 3.2. Unha función simétrica é aquela invariante baixo permutacións dos índices

Utilizando a propiedade asociativa, podemos agrupar os termos da suma que involucren avaliacións de h sobre permutacións de observacións correspondentes ao mesmo subconxunto da mostra. Escribimos

$$h^{[n]}(X_1, \dots, X_n) = \frac{(n-k)!k!}{n!} \sum_I \frac{1}{k!} \sum_{\sigma \in S_k} \psi\left(X_{i_{\sigma(1)}}, \dots, X_{i_{\sigma(k)}}\right) \tag{3.2}$$

Na maioría de problemas podemos asumir que h é simétrica e deste xeito resulta que todos os $k!$ sumandos son os mesmos en cada sumatorio interior. Entón os únicos estimadores insesgados e simétricos de θ son da forma

$$U_n = \binom{n}{k}^{-1} \sum_{1 \le i_1 < \dots < i_k \le n} h(X_{i_1}, \dots, X_{i_k})$$

onde a suma se extende ás *combinacións* de k elementos da mostra e chámanse *U-estatísticos*.

Exemplo 3.3 (Varianza amostral). Sexa $\mathcal{F}$ o conxunto de todas as distribucións con momento de segunda orde finito

$$\mathcal{F} = \left\{ F : \int |x|^2 dF(x) < \infty \right\}$$

Entón podemos definir o *funcional varianza* sobre $\mathcal{F}$ con

$$\operatorname{Var} F = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2} (x_1 - x_2)^2 \, dF(x_1) \, dF(x_2)$$

que estimamos coa varianza amostral $s_n^2 = \begin{pmatrix} n \\ 2 \end{pmatrix}^{-1} \sum_{1 \leq i < j \leq n} \frac{1}{2} (X_i - X_j)^2$.

O U-estatístico empregado para estimar $\theta(F) = E_F\{h\}$, leva asociado o seu estatístico de von Mises (*V-estatístico*)

$$V_n = \frac{1}{n^k} \sum_{i_1=1}^{n} \cdots \sum_{i_k=1}^{n} h\left(X_{i_1}, \ldots, X_{i_k}\right)$$

Definición 3.4. Sexan $\{X_i\}$ i.i.d. con función de distribucón F. Para cada mostra de tamaño $n, \{X_1, \ldots, X_n\}$, a *función de distribución empírica* F_n constrúese como

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I\left(X_i \leq x\right), \quad -\infty < x < \infty$$

Para cada mostra $\{X_1, \ldots, X_n\}, F_n(\cdot)$ é unha función de distribución, considerada como función de x. Por outro lado, para cada valor fixado de $x, F_n(x)$ é unha variable aleatoria, considerada como función da mostra. Dende un punto de vista que compatibilice ambas perspectivas, $F_n(\cdot)$ é unha *función de distribución aleatoria*. En definitiva, a función de distribución empírica é unha lei de probabilidade con soporte discreto e masa de probabilidade $\frac{1}{n}$ en cada punto observado. Agora podemos escribir V_n na súa forma integral

$$V_n = \int \ldots \int h\left(x_1, \ldots, x_k\right) dF_n\left(x_1\right) \ldots dF_n\left(x_k\right)$$

En consecuencia, $V_n = \theta\left(F_n\right)$ é o estimador que máis natural parece para θ. Pero é fundamental ter en conta que V_n non é un estimador insesgado de θ

3.2. Algunhas propiedades asintóticas da teoría de U-estatísticos

Con todo, baixo condicións apropiadas dos momentos, o U-estatístico e V -estatístico asociado a h están intimamente relacionados en comportamento, tal e como amosa o seguinte resultado.

Lema 3.5 ([Serfling, 2009]). *Sexa r un enteiro positivo. Supoñamos que* $\mathbb{E}_{\mathrm{F}} \left|h\left(X_{i_1}, \ldots, \mathrm{X}_{i_k}\right)\right|^r < \infty$, *para todo* $i \leq i_1, \ldots, i_k \leq$ *Entón*

$$\mathbb{E}\left|\mathrm{U_n} - \mathrm{V_n}\right| = \mathcal{O}\left(\mathrm{n}^{-r}\right)$$

Corolario 3.6. *Tomando* $r = 2$ *dedúcese que*

$$n^{1/2}\left(U_n - V_n\right) \xrightarrow{p} 0$$

En consecuencia $n^{1/2}\left(U_n - \theta\right)$ e $n^{1/2}\left(V_n - \theta\right)$ teñen a mesma distribución límite.

3.2.1. A estrutura martingala dos U-estatísticos

As propiedades asintóticas dos U-estatísticos están determinadas pola estrutura de martingala subxacente a eles.

Un U-estatístico é unha martingala regresiva respecto do fluxo non decrecente das σ -álxebras

$$B_n = \sigma\{\omega : U_n, U_{n+1}, ...\}, \quad n \geq 1.$$

xeradas pola sucesión de U-estatísticos $U_n, U_{n+1}, \ldots$, $n \geq k$. É obvio que $B_n \supseteq B_{n+1}$ para todo$n = k, k+1, \ldots$, e U_n é unha variable aleatoria B_n-medible

Lema 3.7 ([V. S. Koroljuk, 1994]). *A sucesión* $(U_n, B_n), n \geq k$, *forma unha martingala regresiva.*

Supoñamos que $E|h| < \infty$ e

$$\begin{aligned} h_c(x_1, \ldots, x_c) &= \mathbb{E}h(x_1, \ldots, x_c, X_{c+1}, \ldots, X_m) \\ &= E(h(X_1, \ldots, X_m) \mid X_1 = x_1, \ldots, X_c = x_c), \quad c = 0, 1, \ldots, m \end{aligned}$$

Sexa $h_0 = \theta(F), h_m = h$; pode verse que

$$h_c(x_1, \ldots, x_c) = \mathbb{E}h_{c+1}(x_1, \ldots, x_C, X_{c+1})$$

para $1 \leq c \leq m-1$. Por comodidade definimos

$$\tilde{h} = h - \theta, \quad \tilde{h}_c = h_c - \theta, \quad 1 \leq c \leq m$$

A función h_1 está intimamente relacionadas coas denominadas *proxeccións de Hájek* de U-estatísticos coa finalidade de aproximar U_n por unha suma de variables aleatorias **independentes** que mellor o explique e enunciar resultados asintóticos.

Definición 3.8. Sexa H un espazo de Hilbert e sexa $\mathcal{S} \subseteq H$ un subespazo pechado de H. Para calquera $v \in H$ definimos a *proxección* de v sobre $\mathcal{S}$ como

$$\pi_{\mathcal{S}}(v) := \underset{s \in \mathcal{S}}{\operatorname{argmin}} \left\{ \|s - v\|_2^2 \right\}$$

Teorema 3.9 ([Van der Vaart, 2000]). *Sexa* h *un núcleo simétrico de orde* r *tal que* $\mathbb{E}\left[h^2\right] < \infty$ *e sexa ,* U_n *o U-estatístico asociado* $\theta = \mathbb{E}[U_n] = \mathbb{E}[h(X_1, \ldots, X_n)]$. *Definindo con* $\hat{U}_n$ *a proxección de* $U_n - \theta$ *en* $\mathcal{S}_n = \{\sum_{i=1}^n g_i(X_i) : g_i(X_i) \in L_2(F)\}$ *daquela*

$$\hat{U}_n = \sum_{i=1}^n \mathbb{E}[U_n - \theta \mid X_i] = \frac{r}{n} \sum_{i=1}^n h_1(X_i)$$

Seguidamente definimos [Efron and Stein, 1981]

$$\begin{aligned}
g_1(x_1) &= \tilde{h}_1(x_1) \\
g_2(x_1,x_2) &= \tilde{h}_2(x_1,x_2) - g_1(x_1) - g_1(x_2) \\
g_3(x_1,x_2,x_3) &= \tilde{h}_3(x_1,x_2,x_3) - \sum_{i=1}^{3} g_1(x_i) - \sum_{1\le i<j\le 3} g_2(x_i,x_j) \\
&\dots \\
g_m(x_1,\dots,x_m) &= \tilde{h}_m(x_1,\dots,x_m) - \sum_{i=1}^{m} g_1(x_i) \\
&\quad - \sum_{1\le i_1<i_2\le m} g_2(x_{i_1},x_{i_2}) - \dots - \sum_{1\le i_1<\dots<i_{m-1}\le m} g_{m-1}\left(x_{i_1},\dots,x_{i_{m-1}}\right)
\end{aligned}$$

Definición 3.10 (Rango dun U-estatístico)**.** . Sexa $r \ge 1$ o enteiro máis pequeno para o que se cumpre

$$g_1 = \dots = g_{r-1} = 0, \quad g_r \neq 0$$

É obvio que r toma valores en $1,\dots,k$. O enteiro r que satisfai esta identidade chámase *rango do U-estatístico* (ou do núcleo h). Se $r = 1$, un U-estatístico (ou un núcleo h) chámase *non dexenerado*. Se $r \ge 2$, entón dicimos que o U-estatístico (ou n núcleo h) é *dexenerado*, e r chámase a *orde de dexeneración*. Se $r = k$, dicimos que o núcleo h é *completamente dexenerado*.

3.2.2. Converxencia débil

Considérese a situación xeral dada por un núcleo arbitrario h de rango r

Teorema 3.11 (Distribución asintótica de U-estatísticos de rango 1 e 2 [V. S. Koroljuk, 1994])**.** *Supoñamos que o núcleo h ten rango r e satisfai as condicións*

$$E\left|g_c(X_1,\dots,X_c)\right|^{2c/(2c-r)} < \infty$$

para todo $c = r, r+1, \dots, k$. Entón

1. *Se $r = 1$, entón*
$$n^{1/2}U_n \xrightarrow{d} N\left(0, m^2\mathbb{E}g_1^2\right)$$
2. *Se $r = 2$, entón*
$$nU_n \xrightarrow{d} \begin{pmatrix} m \\ 2 \end{pmatrix} \sum_{i=1}^{\infty} \lambda_i\left(\tau_i^2 - 1\right),$$

onde $\{\lambda_i\}$ *son os autovalores do operador*

$$S : f \to E\left(g_2\left(X_1, X_2\right) f\left(X_2\right) \mid X_1\right)$$

actuando de L^2 *a* L^2*, e* $\sum_{i=1}^{\infty} \lambda_i^2 = \mathbb{E}g_2^2$ *e* τ_i *son variables aleatorias normais estandarizadas independentes.*

De feito, este operador integral

$$Sf(x) = \int g_2(x, y) f(y) dP(y)$$

é un operador compacto e autoadxunto. Entón, polo teorema espectral para operadores compactos e autoadxuntos [Conway, 2019], S posúe unha sucesión de autovectores φ_i con autovalores λ_i, e por riba, $\{\varphi_i\}$ é unha base ortonormal.

3.2.3. Lei forte dos grandes números

Teorema 3.12 (Lei forte dos grandes números para U-estatísticos)**.** *Supoñamos que o núcleo* h *é tal que* $E\|h\| < \infty$*. Entón*

$$U_n \to \theta \text{ (a.s.)}$$

cando $n \to \infty$*; a converxencia tamén ten lugar en media* (L^1)*, i. e.*

$$E\left\|U_n - \theta\right\| \to 0$$

Sexa agora o V-estatístico asociado ao núcleo h

Teorema 3.13. *Supoñamos agora que o núcleo* h *satisfai a condición*

$$\max\left\{E\left|h\left(X_{i_1}, \ldots, X_{i_m}\right)\right| : 1 \leq i_1, \ldots, i_k \leq m\right\} < \infty$$

Entón

$$V_n \to \theta$$

cando $n \to \infty$ *con probabilidade 1 e en* L^1*.*

Capítulo 4

Estimadores empíricos e contraste de hipóteses

Estas coleccións de enfoques orixina cuestións teóricas naturais sobre os retos estatísticos de carácter fundamental que ofrece o test de independencia. Un deles é o feito de que a distribución asintótica do estatístico do test baixo a hipótese nula (independencia) depende de características descoñecidas das distribucións marxinais relevantes, dificultando a obtención apropiada dun p-valor.

Unha estratexia atractiva, polo tanto, é empregar un test de permutacións, que emprega as mesmas para imitar o comportamento do estatístico do test baixo a hipótese nula. Aínda que o principio de funcionamento é coñecido dende hai case cun século ([Fisher, 1935], capítulo 21), os test de permutacións son cada vez máis populares en machine learning debido á súa facilidade de uso e o seu control garantizado do erro tipo I con mostras finitas en todo o espazo de parámetros da hipótese nula, soamente asumindo a intercambiabilidade dos datos baixo a nula.

Ao resolver un problema de aprendizaxe estatística, moitas veces existen varios procedementos a escoller para atacalo. Isto leva á seguinte pregunta: como podemos saber se un algoritmo de aprendizaxe estatística se comporta mellor que outro? A teoría minimax, que é un conxunto de técnicas para minimizar as posibles perdas no peor dos escenarios, aporta respostas a esta pregunta. En particular, resulta de grande utilidade para comprender mellor as propiedades da potencia dos diversos test de permutacións no contexto dos test de independencia non paramétricos.

4.1. Descrición do test

Describimos a continuación un test de independencia de dúas variables aleatorias baseado no estatístico $\mathrm{HSIC}_b(Z)$, $Z = (X, Y)$ que definiremos a continuación inspirado polo valor poboacional de HSIC. Comezamos cunha introdución formal á terminoloxía dos contrastes de hipóteses. Dada a mostra $Z := (X, Y) = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ i.i.d., o test, $\mathcal{T}(Z)$: $(X \times Y)^n \mapsto \{0, 1\}$ emprégase para distinguir a *hipótese nula* $H_0 : P_{XY} = P_X P_Y$ e a *hipótese alternativa* $H_1 : P_{XY} \neq P_X P_Y$. Isto conséguese comparando o estatístico, no noso caso $\mathrm{HSIC}_b(Z)$, cun valor limiar: se se excede, entón o test rexeita a hipótese nula (tendo en conta que un valor de 0 para o HSIC poboacional é equivalente a $P_{XY} = P_X P_Y$ se o núcleo co que se constrúe $\mathrm{HSIC}_b(Z)$ é *característico*). A *rexión de aceptación* do test defínese entón como o conxunto de números reais por debaixo do limiar. Como o test está baseado nunha mostra finita, é posible que se retorne unha resposta incorrecta: o *erro de tipo I* defínese como a probabilidade de rexeitar H_0 baseándonos na mostra malia seren X e Y independentes. Analogamente, o *erro tipo II* é a probabilidade de aceptar $P_{XY} = P_X P_Y$ cando as variables son dependentes. O *nivel* α é unha cota superior que se impón manualmente ao erro de tipo I, sendo deste xeito un parámetro de deseño do test empregado para fixar o limiar. Un test *consistente* acada un nivel α e un erro tipo II cero cando $n \longrightarrow \infty$.

Como fixamos entón o limiar do test dado α? Tentaremos deducir a distribución asintótica do estimador $\mathrm{HSIC}_b(Z)$ baixo H_0 e H_1; a derradeira tamén a necesitamos para probar a consistencia do test. Veremos que a distribución baixo a nula ten unha forma complicada e non se pode avaliar directamente. Dito isto, describiremos como aproximar o cuantil $1-\alpha$ desta distribución.

4.2. Estimando HSIC en base á mostra.

Un estimador empírico insesgado de HSIC é a suma de tres U-estatísticos seguindo a ecuación 2.2:

$$\mathrm{HSIC}(Z) = \frac{1}{(n)_2} \sum_{(i,j)\in \mathbf{i}_2^n} k_{ij} l_{ij} + \frac{1}{(n)_4} \sum_{(i,j,q,r)\in \mathbf{i}_4^n} k_{ij} l_{qr} - 2\frac{1}{(n)_3} \sum_{(i,j,q)\in \mathbf{i}_3^n} k_{ij} l_{iq}$$

onde $(n)_m := \frac{n!}{(n-m)!}$, o conxunto de índices $\mathbf{i}_r^n$ denota o conxunto de todas as r -plas extraídas sen reemprazamento do conxunto $\{1, \ldots, n\}$, $k_{ij} := k_{\mathcal{X}}(X_i, X_j)$, and $l_{ij} := k_{\mathcal{Y}}(Y_i, Y_j)$. Para contrastar independencia veremos que é máis sinxelo empregar o curmán sesgado

deste estimador reemprazando o U-estatístico polo seu correspondente V-estatístico

$$\mathrm{HSIC}_b(Z) = \frac{1}{n^2}\sum_{i,j}^{m} k_{ij}l_{ij} + \frac{1}{n^4}\sum_{i,j,q,r}^{n} k_{ij}l_{qr} - 2\frac{1}{n^3}\sum_{i,j,q}^{n} k_{ij}l_{iq} = \frac{1}{n^2}\,\mathrm{trace}(\mathbf{KHLH})$$

onde os índices dos sumatorios agora denotan *todas* as r -plas elixidas con reemprazamento de $\{1,\ldots,n\}$ ficando o número de índices baixo o símbolo do sumatorio, $\mathbf{K}$ é a matriz $n\times n$ con entradas $k_{ij}, \mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, e $\mathbf{1}$ é un vector $n\times 1$ repleto de 1 (o custo de calcular este estatístico é de $\mathrm{O}\left(n^2\right)$).

4.2.1. Converxencia débil

Teorema 4.1. *Sexa*

$$h_{ijqr} = \frac{1}{4!}\sum_{(t,u,v,w)}^{(i,j,\sigma,r)} k_{tu}l_{tu} + k_{tu}l_{vw} - 2k_{tu}l_{tv}$$

onde a suma percorre todas as cuádruplas ordenadas (t,u,v,w) extraídas sen reemprazamento de (i,j,q,r), e asumimos $\mathbb{E}\left(h^2\right) < \infty$. baixo $H_1, \mathrm{HSIC}_b(Z)$ converxe en distribción a unha Gaussiana dacordo con

$$n^{\frac{1}{2}}\left(\mathrm{HSIC}_b(Z) - \mathrm{HSIC}\right) \xrightarrow{D} \mathcal{N}\left(0,\sigma_u^2\right)$$

Á varianza é $\sigma_u^2 = 16\left(\mathbb{E}_i\left(\mathbb{E}_{j,q,r}h_{ijqr}\right)^2 - \mathrm{HSIC}\left(P_{XY},\mathcal{F},\mathcal{G}\right)\right)$, onde $\mathbb{E}_{j,q,r} := \mathbb{E}_{z_j,z_q,z_n}$

Demostración. Escribimos,

$$\mathrm{HSIC}_b(Z) = \frac{1}{n^4}\sum_{i,j.q,r}^{m} h_{ijqr}$$

como un único V-estatístico, e decatámonos de que h_{ijqr} é invariante baixo permutacións dos índies. O U-estatístico asociado HSIC (Z) converxe en distribución a $\mathcal{N}\left(0,\sigma_u^2\right)$ en virtude do Teorema 3.11 con varianza σ_u^2. Como a diferenza entre $\mathrm{HSIC}_b(Z)$ e $\mathrm{HSIC}(Z)$ cae con $1/n$ [Gretton et al., 2005], $\mathrm{HSIC}_b(Z)$ converxe á mesma distribución en virtude do Lema 3.5. □

Corolario 4.2. *Un test construído con $\mathrm{HSIC}_b(Z)$ é consistente; i.e., o erro tipo II converxe a cero cando $n\to\infty$*

Demostración. Baixo H_1 a varianza asintótica de $\mathrm{HSIC}_b(Z)$ decrece con n^{-1} □

O segundo teorema é valido baixo H_0. O núcleo $\tilde{k}_P$ centrado en P definido en 2.1 xoga un papel crucial en caracterizar a distribución baixo a nula of do V-estatístico, que sae dexenerado. Vexamos por que.

Teorema 4.3 (Operadores integrais núcleo I, [Steinwart and Christmann, 2008]). *Sexa $\mathcal{Z}$ un espazo medible, μ unha medida σ -finita sobre $\mathcal{Z}$, e H un RKHS separable sobre X with measurable núcleo $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$. Supoñamos que existe $p \in [1, \infty)$ tal que*

$$\|k\|_{L_p(\mu)} := \left(\int_X k^{p/2}(x,x) d\mu(x) \right)^{1/p} < \infty$$

Entón H consiste en funcións p -integrables e a inclusión id : $H \to L_p(\mu)$ é continua con $\|\mathrm{id} : H \to L_p(\mu)\| \leq \|k\|_{L_p(\mu)}$. Ademais, o adxunto desta inclusión é o operador $S_k : L_{p'}(\mu) \to H$ definido por

$$S_k g(x) := \int_X k\left(x, x'\right) g\left(x'\right) d\mu\left(x'\right), \quad g \in L_{p'}(\mu), x \in X$$

onde p' vén definido por $\frac{1}{p} + \frac{1}{p'} = 1$.

Teorema 4.4 (Operadores integrais núcleo II, [Steinwart and Christmann, 2008]). *Sexa $\mathcal{Z}$ un espazo medible con medida σ -finite μ e H un RKHS separable sobre $\mathcal{Z}$ con núcleo medible $k : X \times X \to \mathbb{R}$ satisfacendo $\|k\|_{L_2(\mu)} < \infty$. Entón $S_k : L_2(\mu) \to H$ definida polo teorema anterior é un operador de Hilbert-Schmidt tal que*

$$\|S_k\|_{\mathrm{HS}} = \|k\|_{L_2(\mu)}$$

A $\tilde{k}_P$ asociámoslle o operador integral núcleo $S_{\tilde{k}_P} : L_P^2(\mathcal{Z}) \to L_P^2(\mathcal{Z})$ dado por

$$S_{\tilde{k}_P} g(z) = \int_{\mathcal{Z}} \tilde{k}_P(z,w) g(w) dP(w)$$

A condición $P \in \mathcal{M}_k^1(\mathcal{Z})$, e, en consecuencia, que $\tilde{k}_P \in L_{P\times P}^2(\mathcal{Z} \times \mathcal{Z})$ está intimamente relacionada coas propiedades que nos gustaría que tivese o operador integral.

Teorema 4.5 ([Reed, 2012]). *Sexa $\langle M, \mu \rangle$ un espazo de medida e $H = L^2(M, d\mu)$. Entón $A \in \mathscr{L}(\mathscr{H})$ é Hilbert-Schmidt se e só se existe unha función*

$$K \in L^2(M \times M, d\mu \otimes d\mu)$$

con

$$(Af)(x) = \int K(x,y) f(y) d\mu(y)$$

Ademais,

$$\|A\|_2^2 = \int |K(x,y)|^2 d\mu(x) d\mu(y)$$

En consecuencia, $S_{\tilde{k}_P}$ é un operador Hilbert-Schmidt e daquela clase traza.

Agora estamos preparados para probar que a distribución baixo a nula de HSIC ten a forma dunha combinación linear de χ^2s, con coeficientes correspondendo ao produto dos autovalores dos operadores integrais $S_{\tilde{k}_{P_X}} : L^2_{P_X}(\mathcal{X}) \to L^2_{P_X}(\mathcal{X})$ e $S_{\tilde{k}_{P_Y}} : L^2_{P_Y}(\mathcal{Y}) \to L^2_{P_Y}(\mathcal{Y})$. Pediremos que $P_X \in \mathcal{M}^1_{k_\mathcal{X}}(\mathcal{X})$ e $P_Y \in \mathcal{M}^1_{k_Y}(\mathcal{Y})$, implicando que os operadores integrais $S_{\tilde{i}}$ e $S_{\tilde{j}}$ son operadores de clase traza

Teorema 4.6 (from [Zhang et al., 2012]). *Sexa* $\mathbf{Z} = \{(X_i, Y_i)\}_{i=1}^n$ *unha mostra i.i.d. de* $P_{XY} = P_X P_Y$*, que toma valores en* $\mathcal{X} \times \mathcal{Y}$*, tal que* $P_X \in \mathcal{M}^1_{k_\mathcal{X}}(\mathcal{X})$ *e* $P_Y \in \mathcal{M}^1_{k_Y}(\mathcal{Y})$*. Baixo* H_0*, o U-estatístico* $\mathrm{HSIC}(Z)$ *correspondente ao* V *-estatístico en 4.1 é dexenerado, i.e.* $\mathbb{E}_i h_{ijqr} = 0$*. Entón*

$$n\,\mathrm{HSIC}\left(\mathbf{Z}; k_\mathcal{X}, k_y\right) \xrightarrow{D} \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \lambda_i \eta_j N_{i,j}^2$$

onde $N_{i,j} \sim \mathcal{N}(0,1), i, j \in \mathbb{N}$ *son independentes e* $\{\lambda_i\}_{i=1}^\infty$ *e* $\{\eta_j\}_{j=1}^\infty$ *son os autovalores dos operadores* $S_{\tilde{k}_{P_X}}$ *e* $S_{\tilde{k}_{P_Y}}$*, respectivamente.*

Demostración. Isto é consecuencia do Teorema 3.11 tendo en conta que estamos traballando cun V-estatístico (de aí que os termos da distribución limite non estean centrados: no caso dun U-estatístico, a suma sería sobre os termos $\lambda_l\left(z_l^2 - 1\right)$). □

4.3. Aproximando o cuantil $1-\alpha$ da distribución baixo a nula

Un test de hipóteses que empregue $\mathrm{HSIC}_b(Z)$ podería derivarse do teorema anterior calculando o cuantil $(1-\alpha)$ da distribución límite baixo $\mathcal{H}_0$ dada polo teorema anterior. Pero a forma analítica desta é complexa e daquela preguntámonos como aproximar de xeito eficiente e preciso os seus cuantiles.

A estratexia que adoptemos será empregar remostraxe Monte Carlo: permutamos a mostra de Y mantendo fixada a de X, e o cuantil $1-\alpha$ obténse de xeito directo da distribución resultante dos valores simulados de HSIC_b. Isto pode ser moi custoso computacionalmente para mostras non excesivamente grande porque un conxunto de n observacións pode permutarse de $n!$ xeitos distintos. En caso de existir problemas computacionais por esta razón, un segundo enfoque consiste en aproximar a distribución baixo a nula cunha distribución Gamma. Véxase [Gretton et al., 2007] para máis información sobre os parámetros da mesma.

Capítulo 5

Machine learning e embeddings en espazos de Hilbert separables

Este capítulo constitúe a transición cara a aplicación do traballo aos contrastes de independencia en espazos con pouca estrutura.

A partir de agora denominaremos *aplicación obxecto* e *espazo de obxectos* a calquera parella ϕ e F nas hipóteses da mesma pedindo a maiores que F sexa separable para simplificar as garantías teóricas e porque o espazo que nos interesa para a aplicación nun problema real é separable tal e como veremos.

Comezamos destacando que toda aplicación obxecto $\phi : \mathcal{Z} \longrightarrow F$, con F espazo de Hilbert separable, define un núcleo $k_\phi(x, y) := \langle \phi(x), \phi(y) \rangle_F$ Claramente k é simétrica e definida positiva en virtude das propiedades do produto escalar de F. Esta representación implica dende un punto de vista computacional que as entradas da matriz k_{ij} son produtos escalares de elementos no espazo de obxectos. Daquela, usando núcleos podemos construír algoritmos con fronteiras de decisión non lineares a partir de un no que si o son sen modificar o fluxo do algortimo, véxase a Figura 5.

Esta observación, coñecida popularmente como *kernel trick* abriu a porta a que moitos procederes da aprendizaxe estatística fosen *kernelizados*, sendo o exemplo máis trivial o *kernel ridge regression* [Vovk, 2013].

Unha segunda vantaxe do *kernel trick* é que o espazo de entrada $\mathcal{Z}$ non ten que ser un subconxunto de $\mathbb{R}^d$ xa que todos os cálculos a nivel computacional se fan no espazo de obxectos. Pódense atopar na literatura núcleos definidos en datos como texto, histogramas, cadeas de ADN ou árbores.

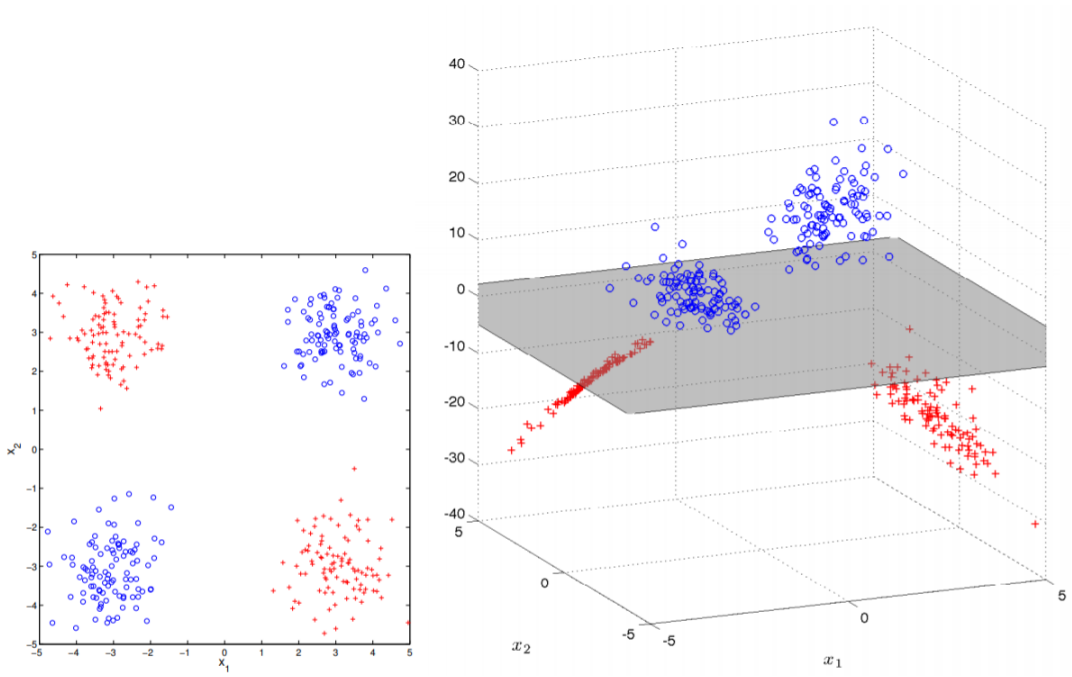


Figura 5.1: O *support vector machine* (SVM) [Steinwart and Christmann, 2008] é un algoritmo de machine learning supervisado que clasifica datos axustando un hiperplano que os separe. Á esquerda, representación dos datos no espazo orixinal. Ningún clasificador linear sería capaz de separar os puntos vermellos dos azuis. Mergullando os datos nun espazo de obxectos de dimensión superior a través de $\phi(x) = [x_1 \quad x_2 \quad x_1 x_2] \in \mathbb{R}^3$, obtemos subconxuntos linearmente separables. Unha posible fronteira de decisión amósase co plano gris [Gretton, 2019].

Aproveitamos a Proposición 1.3 para probar os seguintes resultados

Proposición 5.1. $\mathbb{R}$ *coa distancia euclidiana é de tipo negativo*

Demostración. Definimos

$$\begin{aligned}\phi\colon \mathbb{R} &\longrightarrow L^2(\mathbb{R},\lambda)\\ r &\longmapsto \phi(r) := \mathbf{1}_{[0,\infty)} - \mathbf{1}_{[x,\infty)} \in L^2(\mathbb{R},\lambda)\end{aligned}$$

onde λ é a medida de Lebesgue. Tense que $\|\varphi(r) - \varphi(r')\|_{L^2}^2 = |r - r'|$.
Para $n \geq 2$, definimos

$$f_x(s) := \|x - s\|^{-(n-1)/2}, \quad g_x := f_x - f_0$$

para $x \in \mathbb{R}^n$. Pode probarse que $g_x \in L^2(\mathbb{R}^n, \lambda^n)$ e que existe $c \in \mathbb{R}$ tal que $\|g_x - g_{x'}\|_2 = \|g_{x-x'}\|_2 = c\|x - x'\|^{1/2}$ [Lyons, 2013], de tal xeito que $\phi(x) := g_x/c$ cumpre as hipóteses que buscamos e denominámola *embedding de Riesz.* □

Lema 5.2 ([Szekely et al., 2005])**.** *Se* $0 < \alpha < 2$*, para todo* $x \in \mathbb{R}^d$

$$\int_{\mathbb{R}^d} \frac{1 - \cos(t,x)}{\|t\|^{d+\alpha}} dt = C(d,\alpha)\|x\|^\alpha$$

onde $t \in \mathbb{R}^d$*, e* $C(d,\alpha) > 0$ *é unha constante que depende soamente de* d *e* α*. (As integrais en* 0 *e* ∞ *son no sentido de valor principal:* $\lim_{\varepsilon\to 0}\int_{\mathbb{R}^d\setminus\{\varepsilon B + \varepsilon^{-1}\bar{B}\}}$*, onde* B *é a bola unidade* $\mathbb{R}^d$ *e* $\bar{B}$ *é o complementario de* B*.)*

Outro embedding Φ para $\mathbb{R}^n$ é o que segue:

$$\begin{aligned}\Phi\colon \mathbb{R}^n &\longrightarrow L^2(\mathbb{R}^n, F\lambda^n)\\ x &\longmapsto \Phi(x) := c\left(1 - e^{-i\langle\cdot,x\rangle}\right) \in L^2(\mathbb{R}^n, F\lambda^n)\end{aligned}$$

onde $F(s) := \|s\|^{-(n+1)}$ e $c \in \mathbb{R}$

Pode demostrarse que o é empregando o Lema 5.2 xunto con que

$$\left\|\varphi(x) - \varphi\left(x'\right)\right\|_F = \left\|-e^{-i\langle\cdot,x\rangle} + e^{-i\langle\cdot,x'\rangle}\right\|_F$$

e tamén

$$(-e^{-i\langle\cdot,x\rangle} + e^{-i\langle\cdot,x'\rangle})\overline{(-e^{-i\langle\cdot,x\rangle} + e^{-i\langle\cdot,x'\rangle})} = 2(1 - \cos\langle\cdot, x - x'\rangle)$$

.

Φ é a composición do embedding de Riesz coa isometría de Fourier, e entón referiremonos a Φ como o *embedding de Fourier*

Resulta natural preguntarse cal é a relación entre F e o RKHS asociado a ϕ. Con todo cómpre indicar que no que segue aparecerán aplicacións do estilo de $\beta_\phi : \mu \mapsto \int \phi(x) d\mu(x)$, con μ unha medida de probabilidade que xunto con ϕ terá que cumprir certas condicións para que poidamos escribir esta integral. Non esquezamos que $\phi(x) \in F$, con F espazo de Hilbert separable. Daquela precisamos entender o que significa integrar funcións que toman valores nun espazo de Hilbert. A integral de Bochner é a xeneralización natural da integral de Lebesgue a funcións cuxo codominio é un espazo de Banach. Desenvolver chegados a este punto unha teoría da integración en espazos de Banach levaríanos demasiado lonxe, polo que enunciamos simplemente os resultados e condicións suficientes que precisamos para que β_ϕ estea ben definida, é dicir, sexa un elemento de F. Partimos da seguinte definición e tendo en conta que no que segue $(\mathcal{Z}, \Sigma, \mu)$ é un espazo de medida e E un espazo de Banach.

Definición 5.3. Unha función $\Phi : \mathcal{Z} \to E$ é *fortemente μ -medible* se existe unha sucesión $(f_n)_{n \geqslant 1}$ de funcións μ -simples converxente a f en μ-case todo $z \in \mathcal{Z}$.

Proposición 5.4. *Unha función $f : \mathcal{Z} \to E$ fortemente μ -medible é Bochner integrable se e só se*

$$\int_Z \|f\| d\mu < \infty$$

En virtude do ben coñecido *teorema de medibilidade de Pettis* [Pettis, 1938] **as nocións de medibilidade forte e débil coinciden en espazos de Banach separables**. Daquela soamente temos que probar que f é *debilmente medible*, isto é, que para todo funcional linear e acotado $g : E \to \mathbb{R}$, a función

$$g \circ f : \mathcal{Z} \to \mathbb{R}$$

é medible respecto a Σ e á σ -álxebra de Borel en $\mathbb{R}$.

Proposición 5.5. *Se $f : \mathcal{Z} \to E$ é μ -Bochner integrable e T é un operador linear e limitado entón Tf é μ-Bochner-integrable e ademais*

$$\int_{\mathcal{X}} Tf d\mu = T \int_{\mathcal{Z}} f d\mu$$

A continuación enunciamos un resultado crucial para comprender que o RKHS dunha aplicación obxecto é o espazo de obxectos "máis pequeno" no sentido de que sempre vai existir unha isometría sobrexectiva que nos leve de calquera outro espazo obxecto ao RKHS, podéndose considerar este último entón como o máis natural de todos.

Teorema 5.6 ([Steinwart and Christmann, 2008]). *Sexa $X \neq \emptyset$ e sexan $\phi_0 : X \to H_0$. unha aplicación obxecto co seu espazo de obxectos $\mathcal{H}_0$. Sexa $k_{\phi_0}(x, y) := \langle \phi_0(x), \phi_0(x') \rangle_{\mathcal{H}_0}$. Entón $H := \{f : X \to \mathbb{K} \mid \exists w \in H_0 \text{ con } f(x) = \langle w, \phi_0(x) \rangle_{H_0} \text{ para todo } x \in X\}$ equipado coa norma*

$$\|f\|_H := \text{ínf} \left\{ \|w\|_{H_0} : w \in H_0 \text{ con } f = \langle w, \phi_0(\cdot) \rangle_{H_0} \right\}$$

*é o **único** RKHS para o que k é núcleo reprodutor. Ademais, o operador $V : H_0 \to H$ definido por*

$$Vw := \langle w, \phi_0(\cdot) \rangle_{H_0}, \quad w \in H_0$$

é tal que existe un subespazo $\hat{H}$ de $\mathcal{H}_0$, $\hat{H} = (\ker(V))^{\perp}$, tal que $V_{|\hat{H}} : \hat{H} \to H$. é un isomorfismo isométrico de espazos de Hilbert.

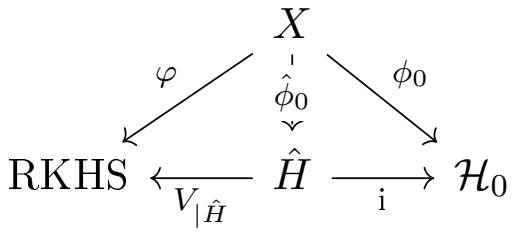
Tense pola construción de $\hat{H}$ que $\phi_0(x) \in \hat{H}$ para todo $x \in X$. Daquela podemos restrinxir o codominio de ϕ_0 e definir

$$\begin{aligned} \hat{\phi}_0 \colon X &\longrightarrow \hat{H} \\ x &\longmapsto \hat{\phi}_0(x) := \phi_0(x) \in \hat{H} \end{aligned}$$

Resulta que compoñendo con V temos que

$$\varphi := (V \circ \hat{\phi}_0)(x) = V(\hat{\phi}_0(x)) = \left\langle \hat{\phi}_0(x), \phi_0(\cdot) \right\rangle_{H_0} = \langle \phi_0(x), \phi_0(\cdot) \rangle_{H_0} = k_{\phi_0}(x, \cdot)$$

é a plicación de Aronszajn do RKHS inducido por k_{ϕ_0}! Isto permite afirmar que o seguinte diagrama é conmutativo en cada unha das súas metades



Chegamos deste xeito a un- impresionante- resultado froito da axuda que aportaron comunicacións persoais con Russell Lyons:

Proposición 5.7. *Sexan $\mathcal{X} = \mathbb{R}^p$, $\mathcal{Y} = \mathbb{R}^q$, cada un co seu embedding de Fourier Φ_p e Φ_q que inducen senllos núcleos f_p e f_q. Sexan X e Y dúas variables aleatorias que toman valores en $\mathcal{X}$ e $\mathcal{Y}$ con medida conxunta $\theta \in \mathcal{M}_+^1(\mathbb{R}^p \times \mathbb{R}^q) \cap \mathcal{M}_f^{1/2}(\mathbb{R}^p \times \mathbb{R}^q)$ e marxinais μ e ν, respectivamente. Daquela o valor poboacional do estatístico HSIC de X e Y construído con $f_p f_q$ coincide coa definición orixinal [Székely et al., 2007] de distance covariance.*

Demostración. En virtude do 2.9 temos que $f\left((x,y),(x',y')\right) := f_p\left(x,x'\right) f_q\left(y,y'\right)$ é un núcleo en $\mathcal{X}\times\mathcal{Y}$ con RKHS H_f isometricamente isomorfo ao produto tensorial $H_{f_p}\otimes H_{f_q}$ onde H_{f_p} e H_{f_q} son os RKHS's de f_1 e f_2 respectivamente. Sexa $\Phi(x,y)=\Phi_p(x)\Phi_q(x)\in L^2_F := L^2(G_pG_q\lambda^{p+q})$

Sexa agora a aplicación

$$\begin{aligned}\beta_\Phi\colon \mathcal{M}_+^1(\mathbb{R}^p\times\mathbb{R}^q)\cap\mathcal{M}_f^{1/2}(\mathbb{R}^p\times\mathbb{R}^q) &\longrightarrow L^2(\mathbb{R}^p\times\mathbb{R}^q)\\ \eta &\longmapsto \int\Phi(x,y)d\eta(x,y)\end{aligned}$$

β_Φ está ben definida como integral de Bochner en virtude da Proposición 5.4 xa que

- $\eta\in\mathcal{M}_+^1(\mathbb{R}^p\times\mathbb{R}^q)$ e entón η é compatible coa topoloxía de $\mathbb{R}^p\times\mathbb{R}^q$, polo tanto ao ser Φ continua (por ser isometría no sentido de espazos métricos) tamén será debilmente medible.
- $\eta\in\mathcal{M}_f^{1/2}(\mathbb{R}^p\times\mathbb{R}^q)$ e daquela $\int_{\mathbb{R}^p\times\mathbb{R}^q}\|\Phi(x,y)\|d\eta(x,y)<\infty$

Por unha banda

$$\|\beta_\Phi(\theta-\mu\times\nu)\|_{L^2_F}=\|V\beta_\Phi(\theta-\mu\times\nu)\|_{H_{f_p}\otimes H_{f_q}}=\|\left\langle\Phi(\cdot,\cdot),\int\Phi(x,y)d(\theta-\mu\times\nu)(x,y)\right\rangle\|_{H_{f_p}\otimes H_{f_q}}=$$

$$=\|\int\langle\Phi(\cdot,\cdot),\Phi(x,y)\rangle\,d(\theta-\mu\times\nu)(x,y)\|_{H_{f_p}\otimes H_{f_q}}=\|\int f(x,y,\cdot,\cdot)d(\theta-\mu\times\nu)(x,y)\|_{H_{f_p}\otimes H_{f_q}}=$$

$$=\|\mu_f(\theta-\mu\times\nu)\|_{H_{H_{f_p}\otimes H_{f_q}}}=\gamma_f(\theta,\mu\times\nu)$$

Por outra banda,

$$\|\beta_\Phi(\theta-\mu\times\nu)\|_{L^2_F}=\int_{\mathbb{R}^{p+q}}\frac{|\beta_\Phi(\theta-\mu\times\nu)(t,s)|^2}{\|t\|_p^{1+p}\|s\|_q^{1+q}}dtds$$

Ademais,

$$\beta_\Phi(\theta-\mu\times\nu)(t,s)=\left(\int\Phi(x,y)d\theta(x,y)\right)(t,s)-\left(\int\Phi(x,y)d(\mu\times\nu)(x,y)\right)(t,s)$$

En virtude do teorema de Fubini, o sustraendo vale $(1-\hat\mu(t))(1-\hat\nu(s))$, onde $\hat\mu$ e $\hat\nu$ son senllas transformadas de Fourier das medidas μ e ν.
Empregando que μ e ν son as marxinais de θ dedúcese que o minuendo vale $1-\hat\mu(t)-\hat\nu(s)+\widehat{\mu\times\nu}(t,s)$. Facendo a resta e tendo en conta que a función característica é o conxugado da transformada de Fourier e que ao facer o módulo ao cadrado desaparecen os conxugados, chegamos ao resultado buscado.

□

Capítulo 6

Aplicación aos contrastes de independencia na física de materiais

A regularidade microscópica subxacente á materia cristalina foi por moito tempo a hipótese que daba explicación á xeometría macroscópica de certos cristais, dos que se observara que as súas caras planas soamente forman determinados ángulos unhas coas outras. Isto verificouse experimentalmente de xeito directo en 1913 co traballo de W. e L. Bragg, os fundadores da Cristalografía con raios X e pioneiros na investigación de como se ordenan os átomos nun sólido. Dende entón, a *física do estado sólido* (base teórica da física de materiais) viu como se desenvolvían no seu marco a través da irrupción da mecánica cuántica novas ideas- como a teoría de bandas- e tecnoloxías- como os semiconductores, o transistor e os circuítos integrados- que fixeron que o mundo no que vivimos fose un sitio inconcibible no caso de non teren aparecido.

En particular, as *perovskitas* son nanomateriais cun tremendo potencial no deseño de sensores de glicosa, neurotransmisores para tratar o Alzheimer, tecnoloxías de catálise, celas de combustible e sensores electroquímicos. Por exemplo a desvantaxe obvia das placas solares de silicio é o alto custo monetario que supón a produción de enerxía eléctrica con elas, de aí que recentemente se puxese atención en *células fotovoltaicas* sintetizadas con determinados tipos de perovskitas en virtude das extraordinarias propiedades de estabilidade e eficiencia que amosan neste tipo de aplicacións. Outro caso práctico interesante é o das *pilas de combustible*, que se comezaron a contemplar como alternativas eficientes aos motores de combustión dado o seu potencial de minimización do impaco medioambiental que supón a queima de combustibles fósiles. Unha pila de combustible emprega certo tipo de fonte enerxética química, a que transforma directamente en enerxía eléctrica. Son atrac-

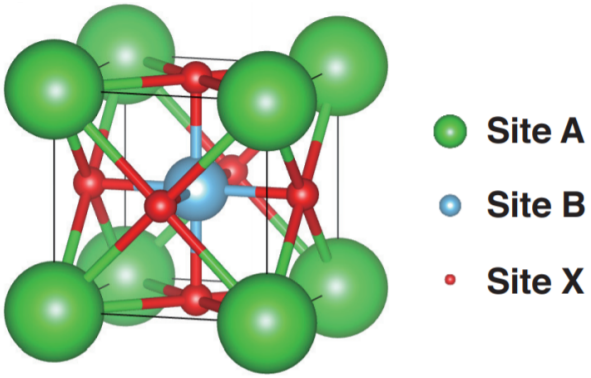
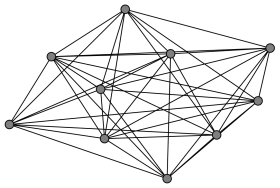
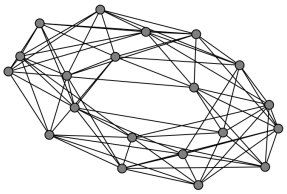
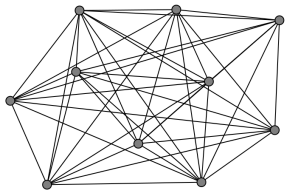
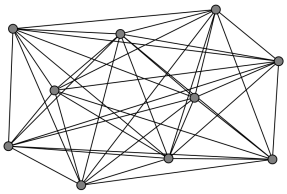
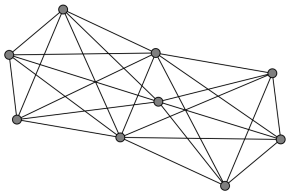


Figura 6.1: A forma da estrutura cristalina das perovskitas é do tipo ABX_3, onde os átomos tipo A se sitúan nas esquinas, os tipo B en posicións centradas no corpo e os tipo X en posicións centradas na cara. Non obstante, a razón de que as perovskitas sexan un hot-topic na ciencia de materiais é que manifestan excelentes propiedades termoeléctricas, dieléctricas; de condutividade electrónica, actividade estrutural, mobilidade dos ións a través da rede cristalina e supermagnetismo, entre moitas outras.

tivas pola súa eficiencia, natureza, baixas emisións, contaminación acústica nula e papel na economía do hidróxeno. Resulta que moitas compoñentes das pilas de combustible como os electrolitos, electrodos e interconexións xa están no punto de mira como beneficiarios potenciais das perovskitas

Semella interesante entón dilucidar se existe algunha relación entre a xeometría cristalina das perovskitas e os seus observables macroscópicos (*bulk properties*). Contestar a esta pregunta podería aportar novas ideas que orienten a investigación física e guiar o deseño de novos materiais.

Para iso, antes precisamos modelizar de xeito matemático a disposición microscópica cristalina. O traballo pioneiro de [Xie and Grossman, 2018a] desenvolto no MIT está baseado na creación dun algoritmo que transforma a estrtura cristalina de calquera material nun multigrafo (grafos aos que se lles permite ter múltiplas arestas entre o mesmo par de nodos terminais) e que tomaremos a nosa aplicación obxecto. Os nodos e arestas representan átomos e enlaces químicos respectivamente. Unha pseudodescrición do fluxo deste algoritmo podería ser a seguinte: para cada átomo búscanse veciños nun raio de 6 Åe considéranse conectados por un vértice se comparten unha cara no diagrama de Voronoi [Blatov, 2004] e teñen distancia interatómica menor que a suma das lonxitudes de enlace

Material e id	Grafo cristalino (X)	Enerxía de formación por átomo (Y)
$CsFeBr_3$ `mp-1079847`		-1.248 eV
$MgGeO_3$ `mp-1078588`		-2.403 eV
$BaTaSe_3$ `mp-1078588`		-1.564 eV
$BaZrSe_3$ `mp-1079300`		-1.871 eV
$BaEuFe_2O_5$ `mp-656144`		-2.370 eV
...	...	...

Cadro 6.1: Submostra exemplificativa das $n = 67$ perovskitas que empregaremos como mostra extraídas da base de datos [Jain et al., 2013]

covalente [Cordero et al., 2008] con tolerancia de 0,25 Å. Entón só se consideran enlaces fortes na creación do grafo cristalino.

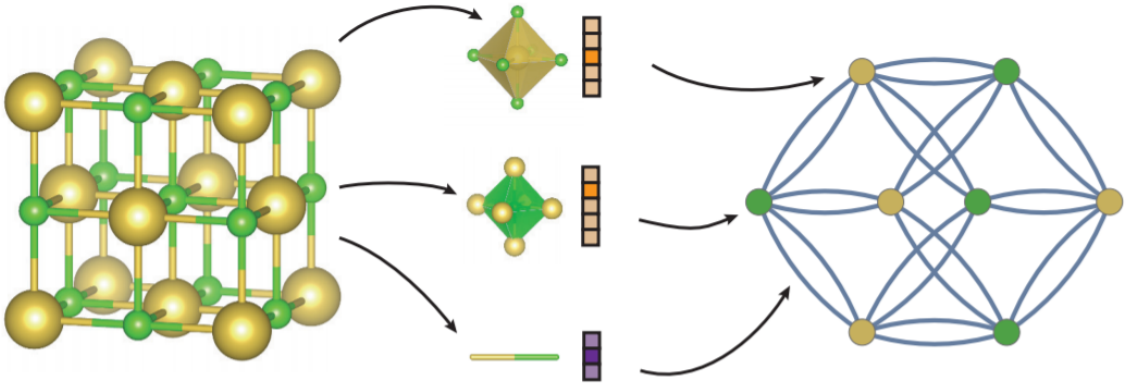


Figura 6.2: Amósase a construción do grafo cristalino do sal común (NaCl) consonte a metodoloxía desenvolta en [Xie and Grossman, 2018b]. A rede cristalina coa súa base atómica convértese nun multigrafo con nodos representando átomos na cela unitaria e arestas codificando veciñanza. Na nosa aproximación, se hai máis dunha aresta conectando dous vértices quedamos con unha soa.

No seu traballo, empregan o output deste algortitmo para deseñar unha rede neuronal profunda convolucionada co fin de extraer representacións óptimas cara á predición de propiedades, despois de adestrala con información precalculada con DFT (*density-functional theory*). Nós, non obstante, aproveitaremos os multigrafos que devolve o algoritmo de [Xie and Grossman, 2018a] para formar un espazo métrico de tipo negativo forte e poder empregalo con vistas a contrastar independencia estatística con outra variable real: a enerxía de formación.

6.1. Construción do espazo métrico de tipo negativo forte

6.1.1. A distancia de Gromov–Hausdorff

A distancia de Gromov–Hausdorff é unha ferramenta utilizada no estudo das propiedades topolóxicas de familias de espazos métricos. Gromov introduciu por primeira vez a noción de distancia de Gromov–Hausdorff (GH) en Helsinki no ano 1979 co obxectivo de estudar todas as estruturas métricas Riemannianas, no sentido de darlle estrutura a este espazo e estudiar a súa completitude, converxencia, familias compactas e outros conceptos relacionados.

Dado $(X, d_X), (Y, d_Y) \in \mathcal{M}$, onde $\mathcal{M}$ denota o conxunto de todos os espazos métricos

compactos, a distancia de GH mide canto lle falta a dous espazos métricos para ser isométricos. Considera un terceiro espazo métrico o *suficientemente rico* como para conter copias isométricas de X e Y e calcular a distancia de Hausdorff entre estas copias. O obxectivo é minimizar facendo unha boa elección das copias isométricas é de Z how far the two metric spaces are from being isometric. Formalmente,

$$d_{\mathcal{GH}}(X,Y) \stackrel{\text{def}}{=} \inf_{Z,\phi_X,\phi_Y} d_{\mathcal{H}}^{Z}\left(\phi_X(X),\phi_Y(Y)\right)$$

onde $\phi_X : X \to Z$ e $\phi_Y : Y \to Z$ son embeddings isométricos de X e Y e Z, e $d_{\mathcal{H}}^{Z}$ é a distancia de Hausdorff en Z :

$$d_{\mathcal{H}}^{Z}(S,T) \stackrel{\text{def}}{=} \max\left\{\sup_{s\in S}\inf_{t\in T} d_Z(s,t), \sup_{t\in T}\inf_{s\in S} d_Z(s,t)\right\} \quad \forall S,T\subseteq Z \ .$$

Gromov probou en [Gromov, 1981] que $d_{\mathcal{GH}}$ é unha métrica no conxunto de clases de isometría de $\mathcal{M}$, constituíndo un *espazo de Gromov-Hausdorff*. A partir de agora sexa $\mathcal{G}$ a colección de todos os espazos métricos compactos.

Teorema 6.1 ([Tuzhilin, 2020])**.** *O espazo $(\mathcal{G}, d_{\mathcal{GH}})$ é separable e completo, é dicir, un espazo Polish.*

6.1.2. Algoritmo para estimar $\widehat{d}_{\mathcal{GH}}(X,Y)$ entre grafos sen pesos non dirixidos

Dende a súa concepción hai catro décadas, a distancia GH empregouse principalmente dende unha perspectiva teórica e o seu cálculo explícito plantexa un problema combinatorio NP-completo [Oles et al., 2019].

Aínda que a definición escrita liñas arriba exhibe de xeito claro o concepto de distancia de GH e ten moi boas propiedades teóricas, non é moi práctica dende o punto de vista computacional. En [Mémoli, 2012] introdúcese a distancia de Gromov–Hausdorff modificada (mGH), unha relaxación da distancia GH que preserva a propiedade de ser unha métrica nas clases de isometría de espazos métricos compactos.

Malia que calcular a distancia mGH é de menor complexidade en comparación coa distancia GH estándar, tamén require resolver un problema de optimización NP-completo.

O algoritmo para estimar mGH entre espazos métricos compactos X e Y consiste en chamar a dúas rutinas: `FINDLOWERBOUND` e `FINDUPPERBOUND` de xeito que $\widehat{d}_{\mathcal{GH}}(X,Y)$ calcúlase de xeito exacto polo algoritmo cando os outputs das dúas chamadas coinciden.

Para obter resultados máis prácticos, consideramos un caso especial de espazos métricos inducidos por grafos sen pesos non dirixidos. Estimar mGH entre estes espazos métricos ten en moitas aplicacións complexidade $O\left(N^3 \log N\right)$

Sexa $G = (V_G, E_G)$ un grafo non dirixido. Para cada par de vértices $v, v' \in V_G$, definimos $d_G(v, v')$ como o camiño máis curto de v a v'. Se os pesos das arestas in E_G son positivos, a función con eles construída $d_G : V_G \times V_G \to [0, \infty]$ é unha métrica en V_G. Dicimos que o espazo métrico (V_G, d_G) vén inducido polo grafo G. Por convención, o camiño máis curto entre vértices de distintas compoñentes conexas dun grafo defínese ∞- casuística que non se dará na nosa aplicación- e entón (V_G, d_G) é compacto se e só se G é conexo [Oles et al., 2019].

Por brevidade, usamos a notación $G \overset{\text{def}}{=} (V_G, d_G)$, asumindo que a distinción entre grafos G e espazos métricos inducidos por dito grafo está clara do contexto. En particular, referímonos á distancia mGH entre espazos métricos compactos inducidos por grafos conexos non dirixidos G e H como $\widehat{d}_{\mathcal{GH}}(G, H)$, e abusámos lixeiramente da nomenclatura chamándoa "distancia mGH entre os grafos conexos sen pesos non dirixidos G e H".

6.1.3. Implementación

O algoritmo para estimar mGH implementouse en Python 3,7 como parte do paquete scikit-tda package [Saul and Tralie, 2019] `https://github.com/scikit-tda`. A implementación toma matrices de adxacencia de grafos sen pesos non dirixidos.

6.2. Elección matemática do núcleo

Sexan b_L e b_U as cotas inferior e superior da matriz de distancias, ambas calculadas co algoritmo. Orientándonos a partir do lema 2.11, para cada z_i da mostra construímos n funcións simétricas

$$k_L^i\left(z, z'\right) = \frac{1}{2}\left[b_L\left(z, z_i\right) + b_L\left(z', z_i\right) - b_L\left(z, z'\right)\right], \quad 1 \leq i \leq n$$

e outras tantas

$$k_U^i\left(z, z'\right) = \frac{1}{2}\left[b_U\left(z, z_i\right) + b_U\left(z', z_i\right) - b_U\left(z, z'\right)\right], \quad 1 \leq i \leq n$$

.

Para cada $1 \leq i \leq n$ queremos resolver o seguinte problema de programación

$$\begin{array}{ll} \min & \|X - k_L^i\|_2 + \|X - k_U^i\|_2 \\ \text{s.a} & X \succeq 0 \end{array}$$

onde $X \in \mathcal{S}^n$ é a variable de optimización, $X \succeq 0$ denota X semidefinida positiva e $\mathcal{S}^n$ é o conxunto de matrices simétricas n x n. Como $\|A\|_2 = \text{máx}_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}$ por definición, temos

para todo $s \geq 0$

$$\begin{aligned}
\|A\|_2 \leq s &\iff \frac{\langle Ax, Ax\rangle}{\langle x, x\rangle} = \frac{\|Ax\|_2^2}{\|x\|_2^2} \leq s^2 \quad \forall x \neq 0 \\
&\iff \langle Ax, Ax\rangle \leq s^2 \langle x, x\rangle \quad \forall x \neq 0 \\
&\iff \left\langle A^T Ax, x\right\rangle - s^2\langle x, x\rangle \leq 0 \quad \forall x \neq 0 \\
&\iff \left\langle \left(A^T A - s^2 I\right) x, x\right\rangle \leq 0 \quad \forall x \neq 0 \\
&\iff A^T A \preceq s^2 I
\end{aligned}$$

Daquela podemos escribir o problema anterior na forma dun problema de optimización convexa cunha función obxectivo linear sobre unha intersección (convexa) de conxuntos matriciais convexos:

$$\begin{aligned}
\min \quad & s + t \\
\text{s.a} \quad & X \succeq 0 \\
& (X - k_L^i)^T (X - k_L^i) \preceq s^2 I \\
& (X - k_U^i)^T (X - k_U^i) \preceq t^2 I
\end{aligned}$$

Finalmente, supoñamos que k é o índice do problema de optimización que *menos sufriu*. Escollemos como núcleo K a solución X^k, aquela que menor valor do óptimo presenta entre todos os n problemas de optimización resoltos.

Podemos empregar o Corolario 2.14 para obter a semimétrica de tipo negativo

$$\rho\left(z, z'\right) = k(z, z) + k\left(z', z'\right) - 2k\left(z, z'\right)$$

Podendo empregar finalmente o resultado de [Lyons, 2013]

Corolario 6.2 ([Lyons, 2013]). *Se $(\mathscr{X}, d)$ é un espazo métrico separable de tipo negativo, entón $\left(\mathscr{X}, d^{\frac{1}{2}}\right)$ é de tipo negativo **forte**.*

Empregamos de novo 2.11 para recuperar unha familia de kernels, todos eles **característicos** en virtude do seguinte resultado:

Proposición 6.3 ([Sejdinovic et al., 2013]). *Sexa un núcleo k que xera ρ. Entón $(\mathcal{Z}, \rho)$ é de tipo negativo forte se e só se k é característico respecto de $\mathcal{M}_+^1(\mathcal{Z}) \cap \mathcal{M}_k^1(\mathcal{Z})$.*

Eliximos un elemento da mostra de xeito aleatorio, digamos j, e entón

$$k_j\left(z, z'\right) = \frac{1}{2}\left[d^{\frac{1}{2}}\left(z, z_i\right) + d^{\frac{1}{2}}\left(z', z_i\right) - d^{\frac{1}{2}}\left(z, z'\right)\right]$$

é o núcleo $k_{\mathcal{X}}$ que empregaremos finalmente no test.

Canto menor sexa a enerxía de formación dun cristal (en valor absoluto), máis estable será experimentalmente a súa sintetización.

6.3. Resultados

Empregamos a información que temos ata agora:

- Matriz k_{ij} que representa o kernel de $\mathcal{X}$
- Variable resposta real, neste caso enerxía de formación por átomo en eV.

Modificamos a función `hsicTestBoot` escrita en `MATLAB` [Gretton et al., 2005] (`http://www.kyb.mpg.de/bs/people/arthur/indep.htm`) para que admita núcleos precalculados nunha das variables. Como Y é real, empregamos o implementado por defecto no código dos autores: a *función base radial*

$$K\left(y_i, y_j\right) = \exp\left(-\frac{|y_i - y_j|^2}{2(\text{Mediana}\left\{|y_p - y_q|, 1 \leq p < q \leq n\right\})^2}\right)$$

Para concluír, amosamos na Figura 6.3 unha representación gráfica dos resultados computacionais finais á vista dos cales parece razoable afirmar que contamos con probas significativas de que a xeometría cristalina e o valor da enerxía de formación dun material están profundamente conectados.

Os resultados son perfectamente compatibles co coñecemento experto proveniente da física do estado sólido. Por exemplo, no caso do enlace iónico a ecuación de Born-Landé [Quane, 1970] permite calcular a enerxía de formación dun material a partir da contribución marcada polo potencial de Coulomb da rede e un termo repulsivo. A xeometría do material vén codificada nesta ecuación por medio principalmente da *constante de Madelung*, puramente dependente do material e determinada empregando técnicas experimentais.

En conclusión, o razoamento importante a ter en conta é que as propiedades macroscópicas dun cristal resultan da colocación regular dos átomos formando unha rede e das forzas atractivas de natureza electrostática, covalente ou derivadas da estrutura metálica se fose o caso.

Acabamos de probar que a estrutura microscópica dun cristal e polo tanto a configuración das forzas que manteñen unidos os seus átomos ten un efecto dramático, polo menos, nunha das súas propiedades: a enerxía de formación. A metodoloxía desenvolta neste traballo podería estenderse á regresión non-paramétrica con vistas á predición de numerosas propiedades interesantes na ciencia de materiais, por nomear algunhas: condutividade, resistencia á corrosión, gap de enerxía, densidade, ductilidade, elasticidade ou plasticidade.

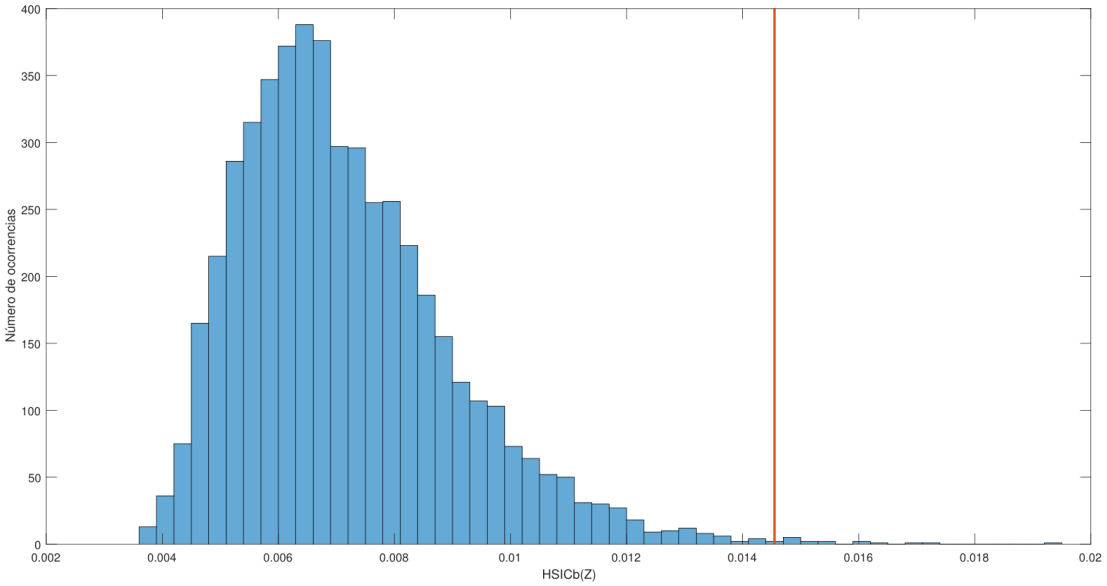


Figura 6.3: En azul, a distribución baixo a nula simulada con $B = 5000$ permutacións. En vermello, valor de $HSIC_b(Z)$ avaliado na nosa mostra. O p-valor estimado é 0.003 e polo tanto contamos con evidencia a un nivel de significación do 0.3 % de que a **enerxía de formación dun material depende da súa estruturación microscópica**, como era de esperar.

Agradecementos

Ao profesor Wenceslao González pola súa implicación e por todas as interesantes conversacións e bos momentos que xurdiron durante a elaboración deste traballo. Ao profesor Russell Lyons, da Universidade de Indiana, pola súas aclaracións á hora de elaborar o capítulo 5. Finalmente, a Marcos Matabuena pola súa inestimable axuda, a súa paciencia e amizade, e por descubrirme este fascinante tema que cambiou para sempre a miña percepción do pensamento científico.

Bibliografía

[Aronszajn, 1950] Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404.

[Aronszajn, 1943] Aronszajn, P.Ñ. (1943). La théorie des noyaux reproduisants et ses applications première partie. *Mathematical Proceedings of the Cambridge Philosophical Society*, 39(3):133–153.

[Ash, 2014] Ash, R. B. (2014). *Real Analysis and Probability: Probability and Mathematical Statistics: a Series of Monographs and Textbooks*. Academic press.

[Bach and Jordan, 2002] Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48.

[Baker, 1973] Baker, C. R. (1973). Joint measures and cross-covariance operators. *Transactions of the American Mathematical Society*, 186:273–289.

[Berg et al., 1984] Berg, C., Christensen, J. P. R., and Ressel, P. (1984). *Harmonic analysis on semigroups: theory of positive definite and related functions*, volume 100. Springer.

[Berrett et al., 2020] Berrett, T. B., Kontoyiannis, I., and Samworth, R. J. (2020). Optimal rates for independence testing via u-statistic permutation tests. *arXiv preprint arXiv:2001.05513*.

[Blatov, 2004] Blatov, V. A. (2004). Voronoi–dirichlet polyhedra in crystal chemistry: theory and applications. *Crystallography Reviews*, 10(4):249–318.

[Conway, 2019] Conway, J. B. (2019). *A course in functional analysis*, volume 96. Springer.

[Cordero et al., 2008] Cordero, B., Gómez, V., Platero-Prats, A. E., Revés, M., Echeverría, J., Cremades, E., Barragán, F., and Alvarez, S. (2008). Covalent radii revisited. *Dalton Trans.*, pages 2832–2838.

[Cramér, 1928] Cramér, H. (1928). On the composition of elementary errors. *Scandinavian Actuarial Journal*, 1928(1):141–180.

[Efron and Stein, 1981] Efron, B. and Stein, C. (1981). The Jackknife Estimate of Variance. *The Annals of Statistics*, 9(3):586 – 596.

[Fisher, 1935] Fisher, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

[Fraser, 1954] Fraser, D. (1954). Completeness of order statistics. *Canadian Journal of Mathematics*, 6:42–45.

[Gretton, 2019] Gretton, A. (2019). Introduction to rkhs, and some simple kernel algorithms. `http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/lecture4_introToRKHS.pdf`.

[Gretton et al., 2012] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

[Gretton et al., 2005] Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer.

[Gretton et al., 2007] Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., Smola, A. J., et al. (2007). A kernel statistical test of independence. In *Nips*, volume 20, pages 585–592. Citeseer.

[Gromov, 1981] Gromov, M. (1981). Groups of polynomial growth and expanding maps. *Publications Mathématiques de l'Institut des Hautes Études Scientifiques*, 53(1):53–78.

[Hille, 1972] Hille, E. (1972). Introduction to general theory of reproducing kernels. *The Rocky Mountain Journal of Mathematics*, 2(3):321–368.

[Hoeffding, 1948a] Hoeffding, W. (1948a). A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 19(3):293 – 325.

[Hoeffding, 1948b] Hoeffding, W. (1948b). A non-parametric test of independence. *The annals of mathematical statistics*, pages 546–557.

[Jain et al., 2013] Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., and Persson, K. a. (2013). The

Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002.

[Kendall, 1938] Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

[Lee, 2019] Lee, A. J. (2019). *U-statistics: Theory and Practice*. Routledge.

[Loeve, 1978] Loeve, M. (1978). *Probability theory II*. Graduate Texts in Mathematics. Springer, 4th edition.

[Lyons, 2013] Lyons, R. (2013). Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3284 – 3305.

[Matabuena et al., 2021] Matabuena, M., Félix, P., García-Meixide, C., and Gude, F. (2021). Five-year prediction of glucose changes with missing data in a reproducing kernel hilbert space.

[Mémoli, 2012] Mémoli, F. (2012). Some properties of gromov–hausdorff distances. *Discrete & Computational Geometry*, 48(2):416–440.

[Naor, 2010] Naor, A. (2010). L1 embeddings of the heisenberg group and fast estimation of graph isoperimetry. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1549–1575. World Scientific.

[Oles et al., 2019] Oles, V., Lemons, N., and Panchenko, A. (2019). Efficient estimation of a gromov–hausdorff distance between unweighted graphs. *arXiv preprint arXiv:1909.09772*.

[Parthasarathy, 2005] Parthasarathy, K. R. (2005). *Probability measures on metric spaces*, volume 352. American Mathematical Soc.

[Parzen, 1974] Parzen, E. (1974). Some recent advances in time series modeling. *IEEE Transactions on Automatic Control*, 19(6):723–730.

[Pearson, 1920] Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, 13(1):25–45.

[Pettis, 1938] Pettis, B. J. (1938). On integration in vector spaces. *Transactions of the American Mathematical Society*, 44(2):277–304.

[Quane, 1970] Quane, D. (1970). Crystal lattice energy and the madelung constant. *Journal of Chemical Education*, 47(5):396.

[Reed, 2012] Reed, M. (2012). *Methods of modern mathematical physics: Functional analysis*. Elsevier.

[Saul and Tralie, 2019] Saul, N. and Tralie, C. (2019). Scikit-tda: topological data analysis for python. *URL https://doi. org/10.5281/zenodo*, 2533369.

[Sejdinovic et al., 2013] Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263 – 2291.

[Serfling, 2009] Serfling, R. J. (2009). *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons.

[Sims, 1962] Sims, B. T. (1962). Some properties and generalizations of semi-metric spaces.

[Smola et al., 2007] Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer.

[Spearman, 1904] Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

[Steinwart and Christmann, 2008] Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

[Szekely, 2000] Szekely, G. (2000). Technical report 03-05: E-statistics: energy of statistical samples. *Department of Mathematics and Statistics, Bowling Green State University*.

[Székely et al., 2007] Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794.

[Szekely et al., 2005] Szekely, G. J., Rizzo, M. L., et al. (2005). Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of classification*, 22(2):151–184.

[Székely and Rizzo, 2017] Székely, G. J. and Rizzo, M. L. (2017). The energy of data. *Annual Review of Statistics and Its Application*, 4(1):447–479.

[Takaoka, 1999] Takaoka, K. (1999). Some remarks on the uniform integrability of continuous martingales. In *Séminaire de Probabilités XXXIII*, pages 327–333. Springer.

[Tuzhilin, 2020] Tuzhilin, A. A. (2020). Lectures on hausdorff and gromov-hausdorff distance geometry.

[V. S. Koroljuk, 1994] V. S. Koroljuk, Y. V. B. a. (1994). *Theory of U-Statistics*. Mathematics and Its Applications 273. Springer Netherlands, 1 edition.

[Van der Vaart, 2000] Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

[Vapnik, 1995] Vapnik, V.Ñ. (1995). The nature of statistical learning. *Theory*.

[Vovk, 2013] Vovk, V. (2013). Kernel ridge regression. In *Empirical inference*, pages 105–116. Springer.

[Xie and Grossman, 2018a] Xie, T. and Grossman, J. C. (2018a). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301.

[Xie and Grossman, 2018b] Xie, T. and Grossman, J. C. (2018b). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120:145301.

[Zhang et al., 2012] Zhang, K., Peters, J., Janzing, D., and Schoelkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery.