USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

UNIVERSIDADE DA CORUÑA

Universidade Vigo

Traballo de Fin de Mestrado

Contrastes de independencia para datos xenómicos de alta dimensión

Fernando Castro Prado

Mestrado en Técnicas Estatísticas

Curso 2017–2018

Universidade de Santiago de Compostela

Traballo de Fin de Mestrado

Contrastes de independencia para datos xenómicos de alta dimensión

Fernando Castro Prado

Mestrado en Técnicas Estatísticas

Setembro de 2018

Proposta de Traballo de Fin de Mestrado

Título en galego: *Contrastes de independencia para datos xenómicos de alta dimensión.*

Título en castelán: *Contrastes de independencia para datos genómicos de alta dimensión.*

Título en inglés: *Independence testing for high-dimensional genomic data.*

Modalidade: *A* (traballos teóricos ou aplicados).

Autor: Fernando Castro Prado (Universidade de Santiago de Compostela).

Directores:

- Wenceslao González Manteiga (Universidade de Santiago de Compostela),
- Javier Costas Costas (Instituto de Investigación Sanitaria de Santiago).

Breve resumo do traballo:

O obxectivo deste traballo é estudar o uso da *correlación de distancias* para a detección de epistase (asociación xenética) a partir de datos caso-control relativos a polimorfismos dun único nucleótido. Esta análise levarase a cabo desde o punto de vista teórico (da estatística matemática) e aplicado (simulacións e datos reais).

Recomendacións:

Malia estar motivado por un exemplo aplicado, este traballo posúe unha notable compoñente metodolóxica. En consecuencia, é aconsellable unha sólida base de coñecementos estatísticos. Máis precisamente, prevese utilizar aqueles das materias Teoría da Probabilidade, Estatística Matemática, Contrastes de Especificación e Técnicas de Remostraxe. Tamén é aconsellable un bo nivel de programación nas linguaxes *R* e *C*.

Outras observacións: Proposta de TFM presentada polo alumno Fernando Castro Prado, co visto e prace dos directores arriba indicados. Wenceslao González Manteiga (USC) será o titor de referencia en canto á metodoloxía estatística e Javier Costas Costas (IDIS), no tocante ao eido de aplicación, a xenética, que motiva o traballo.

Agradecementos

Antes de nada, quero recoñecer a abnegación do investigador posdoutoral da USC David Rodríguez Penas á hora de implementar as distintas técnicas estatísticas no CESGA.

Ademais, debo mencionar que os labores de investigación previos á redacción do presente documento foron financiados parcialmente polas bolsas do Instituto de Matemáticas da USC (IMAT), nas modalidades de grao (curso 2016–2017) e mestrado (curso 2017–2018).

Tampouco podo esquecer que a xénese deste traballo tivo lugar durante as miñas prácticas externas no IDIS-Fundación Ramón Domínguez, durante o verán de 2016, no marco do Grao en Matemáticas da USC. Quero aproveitar este espazo para expresar a miña gratitude cara os que fixeron posible o convenio entre estas dúas entidades.

Por último, pero non por iso menos importante, grazas aos meus codirectores por todo o que teñen feito por min, tanto durante a realización deste TFM coma antes. Todos os debates, consellos, reunións de traballo, etc. fixéronme aprender moito.

Índice

Índice

Prefacio

Este traballo versa sobre a análise estatística de bases de datos xenómicos para a detección de *epistase* (interacción entre mutacións puntuais) en relación con trastornos complexos. No capítulo 1 introdúcese este problema, de indubidable interese práctico, que no entanto carece dunha solución clara desde o punto de vista da metodoloxía estatística, a causa das dificultades matemáticas e informáticas que entraña a alta dimensión. Un enfoque habitual para afrontalas é o baseado na correlación de Pearson, que non é válido cando as variables aleatorias consideradas son discretas, como no caso dos datos xenéticos que aquí se manexan.

Porén, si que está xustificado o uso da *correlación de distancias*, que caracteriza a independencia con xeneralidade, e non só a linear. No capítulo 2, desenvólvese o marco teórico desta medida de asociación, unificando resultados publicados moi recentemente (algúns, hai poucos meses) e resolvendo varias contradicións e ambigüidades atopadas na bibliografía.

A continuación (capítulo 3), analízase a validez desta metodoloxía no contexto do problema xenético de partida, propoñendo un método de contraste *ad hoc*. Con posterioridade, o comportamento desta técnica avalíase mediante simulacións e criterios biolóxicos (capítulos 4– 5), obtendo resultados bastante satisfactorios.

O último capítulo, o 6, expón as principais conclusións que se poden extraer do traballo feito ata agora e os avances que sería interesante levar a cabo nun futuro.

Capítulo 1

Motivación do traballo

Vanse explicar as motivacións do traballo, que xustifican a elección da metodoloxía que se expoñerá no capítulo 2. O punto de partida é un problema xenómico (§ 1.1), do que se analizan a súa importancia e interese. Así mesmo, resúmese o estado do coñecemento estatístico neste campo (§ 1.2); dando especial relevancia a unha das técnicas máis recentes (§ 1.3), pola solidez dos resultados teóricos en que se basea. Todo isto permitirá formular varias hipóteses, recompiladas en forma de preguntas retóricas na § 1.4.

1.1. Epistase en trastornos complexos

A maioría dos trastornos psiquiátricos son caracteres multifactoriais complexos, cuxa aparición é debida á combinación de factores xenéticos e ambientais; ningún dos cales é, porén, necesario nin suficiente. Máis aínda, o efecto individual de cada un deles é, xeralmente, moi pequeno (Biernacka *et al.*, 2013; Sullivan *et al.*, 2017).

En concreto, o xenoma pode explicar ata o 80 % da susceptibilidade a padecer algunha destas doenzas, tal e como é o caso da esquizofrenia (Sullivan *et al.*, 2017, Fig. 3C). Pola súa parte, as adiccións, caracterizadas pola dependencia compulsiva e incontrolada dunha substancia ou actividade nociva, téñense considerado tradicionalmente como enfermidades psiquiátricas con tendencia moderada ou incluso alta a ser herdadas (Goldman *et al.*, 2005). Na actualidade, estímase que esta proporción é

do 50 % (Sullivan *et al.*, 2017). De feito, hai polo menos tres décadas que se considera «ben establecida» a pertinencia de estudos xenéticos sobre o abuso de substancias, en particular para o alcoholismo (Devor e Cloninger, 1989).

A predisposición xenética a padecer un trastorno psiquiátrico repártese entre numerosas variantes ao longo do xenoma. Aínda que habitualmente véñense propoñendo modelos meramente aditivos (Purcell *et al.*, 2009), o coñecemento biolóxico suxire que a interacción xene-xene (chamada *epistase*) podería ser un dos factores que expliquen a «herdabilidade perdida» (*missing heritability*), a cal se traduce na ineficiencia dos *estudos de asociación de xenoma completo* (GWAS) á hora de explicar a causalidade de enfermidades (Manolio *et al.*, 2009; Gusareva e Van Steen, 2014).

O Grupo de Xenética Psiquiátrica do Instituto de Investigación Sanitaria de Santiago (IDIS) dispón de datos de estudos de asociación caso-control ao longo do xenoma (noutras palabras, do xenotipo de «enfermos» e «sans») para alcoholismo e esquizofrenia, aínda que neste traballo só se estudarán estes últimos (polos argumentos expostos en 5.1). O desafío estatístico que desde alí se lanza é usar esta información para detectar pares de alelos de susceptibilidade ou protección fronte aos correspondentes trastornos psiquiátricos, para finalmente corroborar tales achados mediante criterios biolóxicos: confluencia nunha mesma ruta metabólica, unión de proteínas en complexos, coexpresión xénica no espazo (na mesma rexión do cerebro) e no tempo...

Os datos dispoñibles (que se describirán en 5.1) refírense a *polimorfismos dun único nucleótido* (SNP) exónicos, que son variantes nunha das bases das rexións do ADN que se transcriben a ARN. Máis concretamente, considéranse unicamente variantes autosómicas, polo que cada individuo pode contar con 0, 1 ou 2 copias do alelo *menor* (o menos frecuente dos dous posibles) no seu xenoma diploide.

Todo isto supón que debe levarse a cabo inferencia estatística nun contexto de alta dimensión en relación ao tamaño mostral (HDLSS), sendo as variables explicativas *ternarias* (discretas con soporte de cardinal 3).

1.2. Ferramentas para a detección de epistase

O recente desenvolvemento das *-ómicas* transcorreu en paralelo ao de ferramentas bioinformáticas para o procesamento da inxente cantidade de datos que producen. Tanto é así que, na actualidade, hai tal variedade deste tipo de *software* que ata xurdiu a necesidade de crear *metaferramentas* que indexen as distintas técnicas

existentes. Un bo exemplo disto constitúeo *OMICtools* (Henry *et al.*, 2014), cuxo directorio contén máis de 12 000. Entre estas, 300 están dedicadas a GWAS e 100, á detección de epistase.

A existencia dun espectro tan amplo de solucións propostas para este problema tan concreto responde á sorprendente diversidade dos métodos estatísticos válidos para abordar este problema: modelos lineares (estándar e xeneralizados), regresión loxística, contrastes de correlacións, tests de permutacións, inferencia non paramétrica bayesiana, *random forests*, cadeas de Markov, índices de coinformación, teoría de grafos, modelos de probabilidade con máxima entropía, criterios de independencia da estatística máis recente...

Outras causas da existencia de tantas alternativas poden encontrarse en que algunhas das ferramentas están enfocadas a subproblemas concretos (interaccións xene-xene fronte ás de orden superior, variable resposta binaria fronte a continua, herdanza entre dúas xeracións, «pedigrís» [familias], poboacións estratificadas, estudo do desequilibrio de ligamento...) e nos distintos «trucos» informáticos que empregan para poder ofrecer resultados nun tempo razoable (uso de coñecemento biolóxico, paralelización, tarxetas gráficas [GPU], operacións booleanas, aprendizaxe automática, colonias de formigas [ACO]...).

Na Táboa 1.1 ilústranse algúns dos métodos existentes, incluíndo os revisados por Gusareva e Van Steen (2014) e Niel *et al.* (2015), e engadindo algúns dos artigos que se estudaron durante as pescudas previas á redacción do presente TFM.

En calquera caso, non resulta práctico estudar pormenorizadamente o centenar de métodos. En lugar diso, realizouse primeiramente unha revisión superficial de todos eles, en busca de algún criterio que permitise reducir a lista a aqueles máis representativos e eficaces. Porén, esta estratexia resultou escasamente satisfactoria porque, curiosamente, os creadores da inmensa maioría deles defenden nos seus artigos de presentación que melloran o *software* preexistente. En termos xerais, esta circunstancia non se debe a que cada poucos meses se estea desenvolvendo unha ferramenta que desbanque as anteriores con rotundidade; senón á falta dunha forma estándar de comparar detectores de epistase.

En primeiro lugar, hai diferenzas entre a importancia relativa outorgada á rapidez e á exactitude. E á marxe disto, mentres que a velocidade é unha magnitude que se pode cuantificar obxectivamente, a maneira de medir a efectividade varía substancialmente duns autores a outros.

Ferramenta	Metodoloxía estatística	«Trucos» informáticos	Ano
BEAM	MCMC bayesiano	Ningún	2007
BOOST	Regresión loxística	Operacións booleanas. Paralelización	2010
BiForce	Regresión linear	Operacións booleanas. Paralelización	2012
CES	Algoritmos evolutivos	Intelixencia artificial	2015
EpiGPU	Regresión linear	Arquitecturas GPU	2011
EpiACO	Teoría da información	Heurísticas tipo ACO	2017
EpiBlaster	Correlacións de Pearson	Arquitecturas GPU	2011
GLIDE	Regresión linear	Arquitecturas GPU	2012
GWIS	Análise de curvas ROC	Arquitecturas GPU	2013
IndOR	Regresión loxística	Filtro previo	2012
MDR	Combinatoria. Remostraxe	Filtro previo	2001
Random Jungle	*Random forests*	Paralelización	2010

Táboa 1.1: *Cadro sinóptico dalgunhas ferramentas salientables para a detección da epistase en GWAS.*

1.3. Test de correlacións a grande escala (LCT)

Varias ferramentas (a de Kam-Thong *et al.* [2011], por exemplo) fundamentan a súa detección de interaccións xene-xene na procura de comportamentos distintos das correlacións (de Pearson) entre casos e controis. Isto non é estraño, xa que son varios os autores (De la Fuente, 2010; Camacho *et al.*, 2005; D'Haeseleer *et al.*, 2000) que argumentan a conveniencia do test de igualdade de correlacións para este propósito cando os datos son continuos (expresión xénica, metabolómica...). Porén, para SNP isto non é así (§ 1.3.5).

Ademais, o correcto funcionamento de técnicas como a de Kam-Thong *et al.* (2011) depende do cumprimento de hipóteses de normalidade, claramente demasiado restritivas. Por isto, resulta interesante o procedemento introducido por Cai e Liu (2016); quen, partindo do mesmo enfoque, establecen resultados teóricos moito máis fortes e de notable utilidade práctica. Esta metodoloxía é sumamente recente, ao igual que o resto da teoría de contrastes sobre estruturas de covarianzas en alta dimensión, que se desenvolveu case desde cero durante os últimos 5 anos (Cai, 2017).

1.3.1. Introdución aos LCT

Dados $L \in \mathbb{Z}^+$ SNP, considéranse os correspondentes vectores aleatorios $X = (X_j)_{j=1}^L$, para os enfermos ou «casos», e $Y = (Y_j)_{j=1}^L$, para os sans ou «controis». As compoñentes destes vectores virán codificadas, como anteriormente, mediante os valores $\{0, 1, 2\}$. Malia o seu marcado carácter ordinal, tratalas como variables absolutamente continuas podería ser unha simplificación aceptable en algúns contextos, segundo autores como os xa mencionados Kam-Thong *et al.* (2011). Porén, en 1.3.5 comprobarase que este non é un deses casos.

Vanse denotar as respectivas matrices de correlacións dos vectores aleatorios X e Y mediante:

$$(\rho_{ij1})_{i,j} := \mathrm{Cor}(X, X) \in \mathcal{M}_L(\mathbb{R}) \quad \text{e} \quad (\rho_{ij2})_{i,j} := \mathrm{Cor}(Y, Y) \in \mathcal{M}_L(\mathbb{R}).$$

Nesta situación, preténdese contrastar:

$$\begin{cases} H_{0ij} : \rho_{ij1} = \rho_{ij2} \\ H_{1ij} : \rho_{ij1} \neq \rho_{ij2} \end{cases}$$

para cada par $(i, j) \in ([1, L] \cap \mathbb{Z})^2$ tal que $i < j$; en base ás mostras aleatorias simples $\{X_k\}_{k=1}^{n_1}$ i.i.d. X e $\{Y_k\}_{k=1}^{n_2}$ i.i.d. Y, supostas independentes entre si.

1.3.2. LCT: enfoque clásico

Un enfoque pouco innovador do problema é a estabilización da varianza dos coeficientes de correlación mostrais mediante a transformación Z de Fisher:

$$\hat{Z} := \mathrm{atanh}(\hat{\rho}) = \frac{1}{2} \log\left(\frac{1 + \hat{\rho}}{1 - \hat{\rho}}\right).$$

Cando as distribucións son gaussianas, o estatístico

$$F_{ij} := \frac{\sqrt{n_1 n_2}}{2\sqrt{n_1 + n_2}} \left[\log\left(\frac{1 + \hat{\rho}_{ij1}}{1 - \hat{\rho}_{ij1}}\right) - \log\left(\frac{1 + \hat{\rho}_{ij2}}{1 - \hat{\rho}_{ij2}}\right)\right]$$

aproximase a unha $\mathrm{N}(0, 1)$ a medida que $n_1, n_2 \to \infty$ (Anderson, 2003).

Podería pensarse en engadir a isto un procedemento de control da *proporción de falsos descubrimentos* (FDR), coma os de Benjamini e Hochberg (1995) ou Benjamini e Yekutieli (2001), para establecer o test de correlacións a grande escala (LCT)

buscado. O inconveniente desta idea é que, cando a normalidade non está garantida, o comportamento de F_{ij} apártase ostensiblemente do anteriormente descrito. Simulando (Cai e Liu, 2016), compróbase que os resultados aos que conduce este método apártanse notablemente do desexable (tanto con BH como BY), máxime se se compara cos LCT das §§ 1.3.3–1.3.4.

1.3.3. LCT con aproximación normal (LCT-N)

O primeiro contraste proposto por Cai e Liu (2016), o LCT-N, está baseado no estatístico

$$T_{ij} := \frac{\hat{\rho}_{ij1} - \hat{\rho}_{ij2}}{\sqrt{\frac{\hat{\kappa}_1}{3n_1}\left(1-\tilde{\rho}_{ij}^2\right)^2 + \frac{\hat{\kappa}_2}{3n_2}\left(1-\tilde{\rho}_{ij}^2\right)^2}},$$

sendo $\hat{\kappa}_1$ e $\hat{\kappa}_2$ as respectivas curtoses de X e Y, e $\tilde{\rho}_{ijl}$ unha versión «recortada» da correlación mostral $\hat{\rho}_{ijl}$, para $l \in \{1,2\}$; con $\tilde{\rho}_{ij}^2 := \text{máx}\{\tilde{\rho}_{ij1}^2, \tilde{\rho}_{ij2}^2\}$.

Rexeitarase H_{0ij} cando $|T_{ij}|$ sexa maior que un certo limiar (*threshold*) $\hat{t}_\alpha \in \mathbb{R}^+$ (Cai e Liu, 2016), dependente do nivel nominal $\alpha \in]0,1[$ por debaixo do cal se desexe manter a FDR. O cálculo deste $\hat{t}_\alpha$ asume que as distribucións de partida son gaussianas ou non distan moito de selo (contornos elípticos). Por tanto, o LCT-N desaconséllase en multitude de contextos.

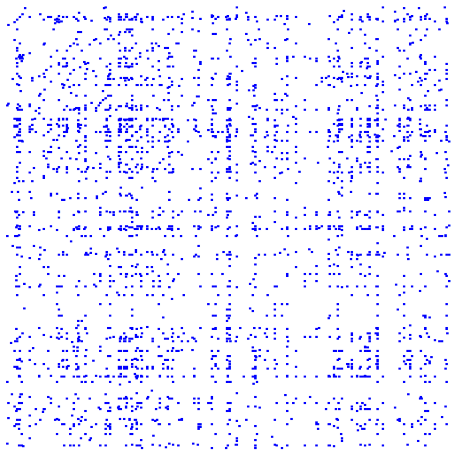
1.3.4. LCT con *bootstrap* (LCT-B)

Se as distribucións de X e Y son totalmente descoñecidas, o razoable é usar técnicas de remostraxe para aproximar a cola da distribución de T_{ij}, da cal depende $\hat{t}_\alpha$. En concreto, o esquema *bootstrap* proposto por Cai e Liu (2016) é consistente e permite obter un limiar $\hat{t}_\alpha^*$ que dá lugar ao LCT-B. A bondade deste contraste ven avalada por varias propiedades teóricas desexables, todas elas demostradas polos devanditos autores.

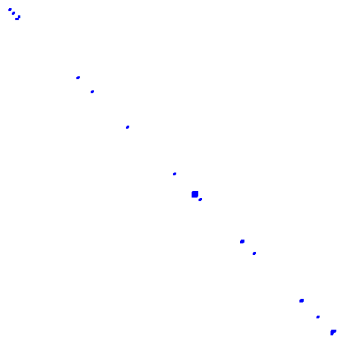
1.3.5. Inadecuación dos LCT aos datos de SNP

Dado que o código executable dos LCT non está dispoñible, programouse desde cero. Para comprobar o seu funcionamento, reproduciuse paso por paso o exemplo do

artigo orixinal, obtendo unha matriz de adxacencia da rede epistática (Figura 1.1a) idéntica á de Cai e Liu (2016).



a) *Datos de expresión xénica para cancro de próstata (Broad Institute).*



b) *Datos de SNP para esquizofrenia (IDIS).*

Figura 1.1: *Matriz de adxacencia da presunta rede epistática detectada por o LCT-B.*

En concreto, estudáronse os datos de expresión xénica de Singh *et al.* (2002), previa redución da dimensión a 500 mediante o test de Welch–Satterthwaite (problema de Behrens–Fisher). Como se trata de variables continuas, o LCT-B ofrece resultados «cribles»; no senso de que a matriz resultante é *sparse*, pero non demasiado. Tendo en conta que Cai e Liu (2016) non empregaron ningún criterio biolóxico para verificar os achados realizados, estes carecen de utilidade práctica e, o que é máis, son bastante cuestionables.

Por outra parte, se os datos estudados son os de SNP de esquizofrenia (descritos en 5.1), marcadamente discretos, a matriz de adxacencia adopta un aspecto moi distinto (Figura 1.1b) á do caso anterior (Figura 1.1a). As únicas entradas non nulas son as próximas á diagonal, o que indica que unicamente se están detectando pares de posición cromosómica moi próxima, que se encontran en *desequilibrio de ligamento* (fenómeno consistente en que a frecuencia dun par de alelos difire significativamente da esperable a partir das correspondentes frecuencias marxinais). Tales achados carecen de utilidade práctica, pois non reflicten unha asociación vinculada ao trastorno complexo estudado, senón que se producen de maneira independente del.

Con maior xeneralidade, é fácil comprobar que, para diversos modelos de variable

ternaria, a potencia dos LCT é próxima a cero. O feito de que a técnica máis robusta para detectar a epistase baseada en correlacións teña este comportamento orienta o desenvolvemento posterior do traballo e, en particular as hipóteses de partida.

1.4. Cuestións que motivan o traballo

Á vista do anterior, está xustificado preguntarse:

1. Que medidas de asociación caracterizan a independencia de variables ternarias?
2. Como se poden xeneralizar os LCT de Cai e Liu (2016) de maneira que sexan aplicables a datos de SNP?
3. Poderá lograrse, ademais, un procedemento adaptable ao tipo de interacción buscada?
4. Gozará semellante método dun comportamento satisfactorio, en termos de calibración do nivel de significación e de potencia?
5. Terán sentido desde o punto de vista biolóxico as interaccións achadas?

Capítulo 2

Correlación de distancias en espazos métricos xerais

A *enerxía dos datos* (Székely e Rizzo, 2017) e todos os resultados da estatística matemática que dela derivan, entre os que se contan a caracterización da independencia en espazos euclidianos (§ 2.1) e moitos outros (Székely e Rizzo, 2010, 2009, 2013), están xeralmente moi ben establecidos e gozan dunha notable solidez teórica desde o seu comezo (Bakirov *et al.*, 2006; Székely *et al.*, 2007; Székely e Rizzo, 2017).

No entanto, o artigo (Lyons, 2013) en que se presenta a correlación de distancias en espazos métricos «deixa unha sorprendentemente grande cantidade de detalles para o lector» (Jakobsen, 2017, páx. 2). A omisión de tantos pasos intermedios supuxo que, durante anos, pasase desapercibida a falsidade de parte importante dos resultados teóricos do artigo (Lyons, 2018), a causa do elevado grao de abstracción de moitos dos argumentos e tamén do prestixio da revista en que foi publicado.

Durante o traballo previo á realización deste TFM, en 2016, detectáronse xa algunhas destas deficiencias. Pero non sería ata a publicación dunha «tese» de fin de mestrado na Universidade de Copenhague sobre o tema (Jakobsen, 2017), cando se desvelaría a magnitude destes erros. O devandito traballo danés dedica 150 páxinas a explicalos e a corrixir boa parte deles, en contraste coas 10 que ocupa Lyons (2013).

Así, o obxectivo do presente capítulo é presentar unha versión corrixida da teoría de Lyons (2013). Para iso, incluiranse de maneira resumida os achados de Jakobsen (2017), engadindo algunhas pequenas demostracións orixinais, e incorporaranse os

contidos da retractación da práctica totalidade das demostracións do artigo orixinal publicada o pasado mes de xuño (Lyons, 2018).

2.1. Correlación de distancias en espazos euclidianos

Cando dúas variables aleatorias (ou vectores) X e Y toman valores en espazos euclidianos ($\mathbb{R}^L$ e $\mathbb{R}^M$, para $L, M \in \mathbb{Z}^+$), é posible definir unha medida que caracterice a súa independencia, denominada *correlación de distancias* (Székely *et al.*, 2007). Máis concretamente, a covarianza de distancias poboacional é unha norma da diferenza da función característica conxunta e o produto das marxinais:

$$\mathrm{dCov}(X,Y) := \|\varphi_{X,Y} - \varphi_X\varphi_Y\|_w \equiv \sqrt{\int_{\mathbb{R}^L\times\mathbb{R}^M} |\varphi_{X,Y}(t,s) - \varphi_X(t)\varphi_Y(s)|^2 w(t,s)\,\mathrm{d}t\,\mathrm{d}s};$$

onde w é unha función «peso» dependente da dimensión dos espazos euclidianos considerados (gozando de unicidade (Székely e Rizzo, 2012)):

$$w(t,s) := \frac{\Gamma\left(\frac{L+1}{2}\right)}{\left(\|t\|\sqrt{\pi}\right)^{L+1}} \frac{\Gamma\left(\frac{M+1}{2}\right)}{\left(\|s\|\sqrt{\pi}\right)^{M+1}},\ (t,s) \in \mathbb{R}^L \times \mathbb{R}^M.$$

E, como de costume,

$$\varphi_X(t) := \mathrm{E}\left[e^{i\langle t,X\rangle}\right],\ t \in \mathbb{R}^L; \qquad \varphi_Y(s) := \mathrm{E}\left[e^{i\langle s,Y\rangle}\right],\ s \in \mathbb{R}^M.$$

Como é lóxico, a correlación de distancias defínese como o cociente da covarianza e as desviacións típicas, carecendo tamén de signo:

$$\mathrm{dCor}(X,Y) := \frac{\mathrm{dCov}(X,Y)}{\sqrt{\mathrm{dCov}(X,X)\,\mathrm{dCov}(Y,Y)}}.$$

A correlación de distancias xeneraliza o cadrado da correlación linear, xa que:

- Toma valores en $[0,1]$. Isto non é estraño: en $\mathbb{R}$, a ordenación total supón que ten sentido que a correlación teña signo (un só pode moverse a «esquerda» e «dereita»). Porén, nun espazo euclidiano de dimensión arbitraria non se mantén esta noción.

- É nula se e só se X e Y son independentes (de aquí o seu interese).

Malia a complexidade da definición da correlación de distancias poboacional, a súa versión mostral é moi facilmente manexable. Dada unha mostra aleatoria simple

$$(X_1, Y_1), \dots, (X_n, Y_n) \text{ i.i.d. } (X, Y);$$

defínese $a_{ij} := d(X_i, X_j)$ para $i, j \in [1, n] \cap \mathbb{Z}$. Con esta notación, as distancias dobremente centradas son:

$$A_{ij} := a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{j\cdot} + \bar{a}_{\cdot\cdot}$$

Se $\{b_{ij}\}_{i,j}$ e $\{B_{ij}\}_{i,j}$ se definen analogamente para $\{Y_i\}_i$, a covarianza de distancias mostral non é máis que o número real positivo cuxo cadrado é:

$$\widehat{\mathrm{dCov}}_n(X, Y)^2 := \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij} B_{ij}$$

A distribución asintótica do estatístico anterior multiplicado por $\frac{n}{\bar{a}_{\cdot\cdot}\bar{b}_{\cdot\cdot}}$ é unha forma cuadrática:

$$\sum_{j=1}^{\infty} \lambda_j Z_j^2,$$

dada por $\{Z_j\}_j$ i.i.d. $\mathrm{N}(0, 1)$ e $\{\lambda_j\}_j \subset \mathbb{R}$; todo isto sempre e cando $\{X, Y\}$ sexan independentes. Desafortunadamente, este resultado carece de utilidade práctica á hora de contrastar a nulidade da covarianza de distancias.

Aquí entran en xogo as técnicas de remostraxe. O máis sensato para aproximar a distribución nula do estatístico de contraste é incorporar ao deseño da remostraxe a información que achega H_0, isto é, a independencia, a través dos *tests de permutacións*.

2.2. Contexto e notacións

2.2.1. Formulación xeral do problema non paramétrico de independencia

Sexan $(\mathscr{X}, d_{\mathscr{X}})$ e $(\mathscr{Y}, d_{\mathscr{Y}})$ dous espazos métricos separables arbitrarios (a restrición da separabilidade explícase en 2.2.2). Dado o elemento aleatorio $Z = (X, Y)$, definido

sobre o espazo de probabilidade $(\Omega, \mathcal{F}, \mathrm{P})$ e con valores en $\mathscr{X} \times \mathscr{Y}$, considérase a súa distribución

$$\theta : \mathcal{B}(\mathscr{X} \times \mathscr{Y}) \longrightarrow [0, 1].$$

As distribucións marxinais veñen dadas por:

- $X \sim \mu := \theta \circ \pi_1^{-1}$, marxinal en $\mathscr{X}$; onde $\pi_1 : (x, y) \in \mathscr{X} \times \mathscr{Y} \mapsto x \in \mathscr{X}$.
- $Y \sim \nu := \theta \circ \pi_2^{-1}$, marxinal en $\mathscr{Y}$; onde $\pi_2 : (x, y) \in \mathscr{X} \times \mathscr{Y} \mapsto y \in \mathscr{Y}$.

Con esta notación, o test non paramétrico de independencia para X e Y consiste no contraste de $H_0 : \theta = \mu \times \nu$ fronte á alternativa $H_1 : \theta \neq \mu \times \nu$. Cómpre matizar que o produto $\mu \times \nu$ está definido do xeito habitual: é a única medida en $\mathcal{B}(\mathscr{X}) \otimes \mathcal{B}(\mathscr{Y})$ para a cal

$$(\mu \times \nu)(A \times B) := \mu(A)\nu(B);\ A \in \mathcal{B}(\mathscr{X}),\ B \in \mathcal{B}(\mathscr{Y}).$$

2.2.2. Separabilidade dos espazos marxinais

A primeira vantaxe de considerar espazos $\mathscr{X}$ e $\mathscr{Y}$ separables é que deste xeito, a $\sigma-$álxebra de Borel xerada polo seu produto topolóxico redúcese á $\sigma-$álxebra produto:

$$\mathcal{B}(\mathscr{X} \times \mathscr{Y}) = \mathcal{B}(\mathscr{X}) \otimes \mathcal{B}(\mathscr{Y}) := \sigma\{A \times B : A \in \mathcal{B}(\mathscr{X}), B \in \mathcal{B}(\mathscr{Y})\}.$$

Esta igualdade é útil *per se* (por exemplo, é crucial na demostración do lema 3.10 de Jakobsen [2017]), pero o seu maior interese é que garante que as métricas sexan «conxuntamente medibles» (*jointly measurable*): $d_{\mathscr{Z}}$ é $\mathcal{B}(\mathscr{Z}) \otimes \mathcal{B}(\mathscr{Z}) / \mathcal{B}(\mathbb{R})-$medible, para $\mathscr{Z} \in \{\mathscr{X}, \mathscr{Y}\}$. A relevancia disto reside en que, noutro caso, as integrais de Lebesgue involucradas na definición da covarianza de distancias (§ 2.3) non estarían definidas. Un exemplo de tal espazo sería $\mathscr{X} := \mathbb{R}^{\mathbb{R}}$, equipado coa métrica discreta. Isto é un caso particular do fenómeno denominado *patoloxía de Nedoma* (véxanse Schechter [1996, proposición 21.8] e Bogachev [2007, exemplo 6.4.3] para máis detalles).

Finalmente, a separabilidade úsase explicitamente nas demostracións dalgunhas propiedades importantes da covarianza de distancias (Jakobsen, 2017, teorema 4.4 e lema 5.8), o que indica que non é unha hipótese «desmesurada». Ademais, no caso da detección da epistase, os espazos son finitos e, en consecuencia, separables; polo que estas consideracións, aínda que cruciais no plano teórico, pódense obviar na práctica.

O artigo orixinal de presentación da correlación de distancias en espazos métricos (Lyons, 2013) non tivo en conta estes aspectos (cf. Lyons [2018]).

2.2.3. Medidas con signo

Dise que $\mu : \mathcal{B}(\mathscr{X}) \longrightarrow \mathbb{R}$ é unha medida (de Borel) con signo e finita, e denótase $\mu \in M(\mathscr{X})$, se e só se $|\mu|$ é unha medida finita. Para toda $\mu \in M(\mathscr{X})$, existe a denominada *descomposición de Hahn–Jordan* e é esencialmente única (páxs. 125–126 de Rudin, 1921; teorema 32.1 de Billingsley, 1995), é dicir, que é posible atopar medidas non negativas $\mu^{\pm} \in M(\mathscr{X})$ tales que

$$\mu = \mu^{+} - \mu^{-}$$

e unha partición do espazo $\mathscr{X} = \mathscr{X}^{+} \sqcup \mathscr{X}^{-}$ que verifique:

$$\mu^{+}(\mathscr{X}^{-}) = 0 = \mu^{-}(\mathscr{X}^{+});$$

o que significa que μ^{+} e μ^{-} son *mutuamente singulares* («ortogonais»).

Isto permite definir de maneira natural as integrais (de Lebesgue) con respecto a medidas con signo. Así, dada $f : \mathscr{X} \longrightarrow \mathbb{R}$ medible,

$$\int_{\mathscr{X}} f \,\mathrm{d}\mu := \int_{\mathscr{X}} f \,\mathrm{d}\mu^{+} - \int_{\mathscr{X}} f \,\mathrm{d}\mu^{-};$$

que estará ben definida cando f sexa integrable respecto de $\mu^{\pm}$, posto que $|\mu| = \mu^{+} + \mu^{-}$.

Por outra banda, tamén vai ser preciso integrar respecto de medidas produto. Para comezar, considérese $\nu \in M(\mathscr{Y})$, con descomposición de Hahn–Jordan dada por $(\mathscr{Y}^{\pm}, \nu^{\pm})$. Entón:

- $\mu^{+} \times \nu^{+} + \mu^{-} \times \nu^{-}$ é unha medida (non negativa) con soporte $(\mathscr{X}^{+} \times \mathscr{Y}^{+}) \sqcup (\mathscr{X}^{-} \times \mathscr{Y}^{-})$
- $\mu^{+} \times \nu^{-} + \mu^{-} \times \nu^{+}$ é unha medida (non negativa) con soporte $(\mathscr{X}^{+} \times \mathscr{Y}^{-}) \sqcup (\mathscr{X}^{-} \times \mathscr{Y}^{+})$

Polo carácter disxunto dos seus soportes, as medidas anteriores son mutuamente singulares e, en consecuencia (Rudin, 1921, corolario do teorema 6.14), determinan a descomposición de Hahn–Jordan de $\mu \times \nu$:

$$\mu \times \nu = (\mu^{+} \times \nu^{+} + \mu^{-} \times \nu^{-}) - (\mu^{+} \times \nu^{-} + \mu^{-} \times \nu^{+}).$$

Deste xeito, a integral dunha función Borel-medible $h : \mathscr{X} \times \mathscr{Y} \longrightarrow \mathbb{R}$ respecto de $\mu \times \nu$ será:

$$\int h \, \mathrm{d}\mu \times \nu := \int h \, \mathrm{d}\mu^+ \times \nu^+ + \int h \, \mathrm{d}\mu^- \times \nu^- - \int h \, \mathrm{d}\mu^+ \times \nu^- - \int h \, \mathrm{d}\mu^- \times \nu^+;$$

o que supón que $\mathcal{L}^1(\mu \times \nu)$ é a intersección dos catro espazos $\mathcal{L}^1(\mu^\pm \times \nu^\pm)$. Na ecuación anterior omitiuse o conxunto de integración, por sobreentenderse que se trata do máis grande posible (neste caso, $\mathscr{X} \times \mathscr{Y}$). Este abuso de notación, tomado de Lyons (2013), será dos poucos que se empregarán neste capítulo, mentres que se emendarán aqueles que causaron erros no artigo orixinal e mesmo no seu *corrigendum* (Lyons, 2018).

A última observación pertinente sobre a integración con respecto ao produto de medidas con signo é que admiten un teorema de Fubini–Tonelli xeneralizado (Bogachev, 2007, § 3.3):

$$\forall \, h \in \mathcal{L}^1(\mu \times \nu), \; \int h \, \mathrm{d}\mu \times \nu = \iint h \, \mathrm{d}\mu \, \mathrm{d}\nu = \iint h \, \mathrm{d}\nu \, \mathrm{d}\mu.$$

2.2.4. Regularidade dunha medida

Antes de proseguir, cómpre enunciar e demostrar a denominada *desigualdade* «c_r» (c_r *inequality*, en inglés). Dados $\alpha, \beta, r \in \mathbb{R}^+$, tense que: $(\alpha+\beta)^r \leq c_r(\alpha^r + \beta^r)$, onde

$$c_r = \begin{cases} 1, & r < 1 \\ 2^{r-1}, & r \geq 1 \end{cases}.$$

Demostración. (1) Se $\boxed{\mathrm{r<1}}$, trátase de probar que

$$(t+1)^r \leq t^r + 1, \; t := \frac{\alpha}{\beta}$$

ou, noutras palabras:

$$f(t) := t^r + 1 - (t+1)^r \geq 0.$$

E isto último verifícase por ser $r - 1 < 0$:

$$\forall t \in \mathbb{R}^+, \; f'(t) = r(t^{r-1} - (t+1)^{r-1}) > 0 \Rightarrow \forall t \in \mathbb{R}^+, \; f(t) \geq f(0) = 1 - 1^r = 0.$$

(2) Para $\boxed{r \geq 1}$, a función $g(x) := x^r$ é convexa en todo $x \in \mathbb{R}^+$:

$$g''(x) = r(r-1)x^{r-2} > 0, \; x \in \mathbb{R}^+.$$

Xeometricamente isto significa que:

$$g\left(\frac{\alpha+\beta}{2}\right) \leq \frac{g(\alpha)+g(\beta)}{2} \Leftrightarrow (\alpha+\beta)^r \leq 2^{r-1}(\alpha^r+\beta^r).$$

□

Neste punto, xa é posible estender o concepto de «regularidade» dunha distribución a medidas con signo. Dise que $\mu \in M(\mathscr{X})$ ten momentos de orde r finitos, e escríbese $\mu \in M^r(\mathscr{X})$, se e só se:

$$\exists\, o \in \mathscr{X},\ \int d_{\mathscr{X}}(o,x)^r \,\mathrm{d}|\mu|(x) < +\infty.$$

A desigualdade «c_r» permite concluír que, de cumprirse a condición anterior, será válida para calquera orixe:

$$\mu \in M^r(\mathscr{X}) \Leftrightarrow \forall o \in \mathscr{X},\ \int d_{\mathscr{X}}(o,x)^r \,\mathrm{d}|\mu|(x) < +\infty.$$

Ademais, dirase que unha medida con signo $\theta \in M(\mathscr{X}\times\mathscr{Y})$ pertence a $M^{r,r}(\mathscr{X}\times\mathscr{Y})$ se as súas dúas marxinais teñen $r-$ésimos momentos finitos. Por último, empregarase o subíndice «1» para indicar as medidas de probabilidade:

$$M_1(\mathscr{X}) := \{\mu \in M(\mathscr{X}) :\ \mu \geq 0,\ \mu(\mathscr{X}) = 1\};$$

$$M_1^r(\mathscr{X}) := M^r(\mathscr{X}) \cap M_1(\mathscr{X}); \quad M_1^{r,r}(\mathscr{X}\times\mathscr{Y}) := M^{r,r}(\mathscr{X}\times\mathscr{Y}) \cap M_1(\mathscr{X}\times\mathscr{Y}).$$

2.3. Definición formal de *dcov*

Na sección anterior introduciuse o marco teórico no que ten sentido falar de covarianza de distancias, solventando así algunhas das inconsistencias de Lyons (2013), o que permitirá agora definir o operador *dcov* de xeito rigoroso, ampliando e corrixindo a construción efectuada por Jakobsen (2017).

2.3.1. Integrabilidade da métrica

Para poder definir *dcov*, é importante ter en conta que:

$$\forall\, \mu_1, \mu_2 \in M^1(\mathscr{X}) :\ d_{\mathscr{X}} \in \mathcal{L}_1(\mu_1 \times \mu_2).$$

Isto é consecuencia de Fubini e da desigualdade triangular:

$$\int d_{\mathscr{X}}\,\mathrm{d}|\mu_1|\times|\mu_2| \leq \int d_{\mathscr{X}}(x,o)\,\mathrm{d}|\mu_1|\times|\mu_2|(x,x') + \int d_{\mathscr{X}}(o,x')\,\mathrm{d}|\mu_1|\times|\mu_2|(x,x') =$$

$$= |\mu_2|(\mathscr{X})\int d_{\mathscr{X}}(x,o)\,\mathrm{d}|\mu_1|(x) + |\mu_1|(\mathscr{X})\int d_{\mathscr{X}}(x,o)\,\mathrm{d}|\mu_2|(x) < +\infty.$$

2.3.2. Distancias esperadas e algunhas desigualdades

A covarianza de distancias calcúlase a partir de distancias dobremente centradas (§ 2.3.3), pero para iso antes hai que comprobar que as esperanzas involucradas estean ben definidas. En primeiro lugar, sexa $\mu \in M^1(\mathscr{X})$. A seguinte función dá o valor esperado da distancia dun elemento aleatorio $X \sim \mu$ a cada punto:

$$\begin{aligned} a_\mu : \mathscr{X} &\longrightarrow \mathbb{R} \\ x &\longmapsto \int d_{\mathscr{X}}(x,x')\,\mathrm{d}\mu(x') \end{aligned}$$

Evidentemente, está ben definida. Ademais é $|\mu|(\mathscr{X})-$lipschitziana (e, polo tanto, continua):

$$\forall x,x' \in \mathscr{X}:\ |a_\mu(x) - a_\mu(x')| \leq \int |d_{\mathscr{X}}(x,z) - d_{\mathscr{X}}(x',z)|\,\mathrm{d}|\mu|(z) \leq$$

$$\leq \int d_{\mathscr{X}}(x,x')\,\mathrm{d}|\mu|(z) = |\mu|(\mathscr{X})d_{\mathscr{X}}(x,x').$$

Por outra banda, o visto en 2.3.1 supón que a integral $D(\mu)$ é sempre un número real:

$$D(\mu) := \int a_\mu\,\mathrm{d}\mu = \int d_{\mathscr{X}}\,\mathrm{d}\mu\times\mu.$$

As seguintes desigualdades serán moi útiles no sucesivo e son certas sempre que $\mu \in M_1^1(\mathscr{X})$ e $x,y \in \mathscr{X}$:

1. $D(\mu) \leq 2a_\mu(x)$;
2. $D(\mu) \leq a_\mu(x) + a_\mu(y)$;
3. $d_{\mathscr{X}}(x,y) \leq a_\mu(x) + a_\mu(y)$;
4. $a_\mu(x) \leq d_{\mathscr{X}}(x,y) + a_\mu(y)$.

Demostración. (1) $D(\mu) = \int d_{\mathscr{X}}(x', x'')\,\mathrm{d}\mu^2(x', x'') \leq \mu(\mathscr{X}) \int d_{\mathscr{X}}(x', x)\,\mathrm{d}\mu(x') + \mu(\mathscr{X}) \int d_{\mathscr{X}}(x, x'')\,\mathrm{d}\mu(x'') = 2a_\mu(x)$.

(2) Aplicando *(1)* a y e sumando membro a membro, chégase a que: $2D(\mu) \leq 2a_\mu(x) + 2a_\mu(y)$.

(3) Obtense integrando respecto de z os dous membros de: $d_{\mathscr{X}}(x, y) \leq d_{\mathscr{X}}(x, z) + d_{\mathscr{X}}(y, z)$.

(4) Ídem ao anterior: $d_{\mathscr{X}}(x, z) \leq d_{\mathscr{X}}(x, y) + d_{\mathscr{X}}(y, z)$.

2.3.3. Distancias dobremente centradas

Dada $\mu \in M^1(\mathscr{X})$, a distancia $d_{\mathscr{X}}$ dobremente $\mu-$centrada é:

$$\begin{aligned} d_\mu : \mathscr{X} \times \mathscr{X} &\longrightarrow \mathbb{R} \\ (x_1, x_2) &\mapsto d_{\mathscr{X}}(x_1, x_2) - a_\mu(x_1) - a_\mu(x_2) + D(\mu) \end{aligned}$$

A modificación anterior de $d_{\mathscr{X}}$ non é, en xeral, unha métrica; aínda que si que é sempre continua (por selo $d_{\mathscr{X}}$, a_μ, π_1 e π_2) e, en particular, Borel-medible. Ademais, é importante observar que a notación «d_μ» non fai referencia ao espazo métrico sobre o que está definida, pero este pequeno abuso alixeira notablemente a notación e non ocasiona ningún malentendido. Este non é o caso doutras abreviacións de Lyons, como por exemplo a consistente en tomar $d := d_{\mathscr{X}}$, o cal dá a entender que os espazos marxinais $\mathscr{X}$ e $\mathscr{Y}$ deben compartir a mesma estrutura métrica, hipótese totalmente superflua para o desenvolvemento posterior.

A última propiedade salientable de d_μ é que:

$$\forall \mu, \mu_1, \mu_2 \in M_1^1(\mathscr{X}) :\ d_\mu \in \mathcal{L}^2(\mu_1 \times \mu_2).$$

Demostración. En primeira instancia, é conveniente probar que, para $(x, y) \in \mathscr{X}^2$,

$$|d_\mu(x, y)| \leq 2a_\mu(y).$$

Para isto, hai que considerar dous casos:

- Se $\boxed{d_\mu(x, y) \geq 0}$, abonda con aplicar as desigualdades vistas en 2.3.2:

$$|d_\mu(x, y)| = d_\mu(x, y) \overset{(3)}{\leq} D(\mu) \overset{(1)}{\leq} 2a_\mu(y).$$

- Se $\boxed{d_\mu(x,y) < 0}$, os argumentos de Jakobsen (2017, páx. 10) fan uso de hipóteses demasiado «fortes». No seu lugar, pode argüírse que:

$$\forall z,t \in \mathscr{X} : \; d_{\mathscr{X}}(x,z) \leq d_{\mathscr{X}}(x,y) + d_{\mathscr{X}}(y,t) + d_{\mathscr{X}}(t,z) \Rightarrow$$

$$\Rightarrow a_\mu(x) \leq d_{\mathscr{X}}(x,y) + a_\mu(y) + D(\mu);$$

  o que equivale a que $|d_\mu(x,y)| \leq 2a_\mu(y)$.

Utilizando o anterior, probar que $d_\mu \in \mathcal{L}^2(\mu_1 \times \mu_2)$ resulta moi sinxelo:

$$\int d_\mu(x,y)^2 \,\mathrm{d}\mu_1 \times \mu_2(x,y) \leq 4 \int a_\mu(x) a_\mu(y) \,\mathrm{d}\mu_1 \times \mu_2(x,y) \overset{\text{Fubini}}{=}$$

$$= 4 \int d_{\mathscr{X}}(x,z) \,\mathrm{d}\mu_1 \times \mu(x,z) \int d_{\mathscr{X}}(y,z) \,\mathrm{d}\mu_2 \times \mu(y,z) \overset{\boxed{d_{\mathscr{X}} \in \mathcal{L}^1}}{<} +\infty.$$

□

2.3.4. A medida de asociación *dcov*

A covarianza de distancias xeneralizada defínese como:

$$\mathrm{dcov}(\theta) := \int_{(\mathscr{X}\times\mathscr{Y})^2} d_\mu(x,x') d_\nu(y,y') \,\mathrm{d}\theta^2 \left((x,y),(x',y')\right), \; \theta \in M_1^{1,1}(\mathscr{X} \times \mathscr{Y});$$

onde, como anteriormente, $\mu := \theta \circ \pi_1^{-1}$ e $\nu := \theta \circ \pi_2^{-1}$.

Para ver que *dcov* está ben definida, obsérvese que a integral do produto de dúas funcións respecto dunha medida (non negativa) é sempre un produto escalar (bilinear, semidefinido positivo) e, polo tanto, verifica a desigualdade de Cauchy–Bunyakovsky –Schwarz. Tamén é sinxelo demostrar este caso particular da desigualdade de Hölder «directamente»:

$$0 \leq \int [d_\mu(v) d_\nu(w) - d_\mu(w) d_\nu(v)]^2 \,\mathrm{d}\theta^2(v,w) = 2 \int d_\mu^2 \,\mathrm{d}\theta^2 \int d_\nu^2 \,\mathrm{d}\theta^2 - 2 \left(\int d_\mu d_\nu \,\mathrm{d}\theta^2 \right)^2 \Rightarrow$$

$$\overset{\boxed{d_\mu, d_\nu \in \mathcal{L}^2}}{\Rightarrow} |\,\mathrm{dcov}(\theta)| \leq \sqrt{\int d_\mu^2 \,\mathrm{d}\theta^2 \int d_\nu^2 \,\mathrm{d}\theta^2} < +\infty.$$

Outra forma de razoalo sería ver que:

$$(d_\mu \pm d_\nu)^2 \geq 0 \Leftrightarrow \frac{d_\mu^2 + d_\nu^2}{2} \geq \pm d_\mu d_\nu \Leftrightarrow \frac{d_\mu^2 + d_\nu^2}{2} \geq |d_\mu d_\nu|,$$

que non é máis que un caso particular da desigualdade «media aritmética-xeométrica» (AM-GM) e da de Young.

Agora só falta aclarar por que son finitas as dúas integrais do membro dereito. Por exemplo, para d_μ, tense que:

$$\int d_\mu(x,x')^2 \,\mathrm{d}\theta^2((x,y)(x',y')) \overset{\text{Fubini}}{=} \iint d_\mu(x,x') \,\mathrm{d}\theta(x,y) \,\mathrm{d}\theta(x',y') \overset{\text{ACOV}}{=}$$

$$= \int d_\mu(x,x') \,\mathrm{d}\mu^2(x,x') \overset{d_\mu \in \mathcal{L}^2(\mu\times\mu)}{<} +\infty.$$

As siglas «ACOV» designan o teorema de cambio de variable abstracto (*abstract change of variables*, en inglés), particularizado para o caso en que a función do cambio de variable é unha proxección. Máis formalmente, sexa f unha función medible no seguinte esquema:

$$(\mathscr{X}\times\mathscr{Y}, \mathcal{B}(\mathscr{X})\otimes\mathcal{B}(\mathscr{Y}), \theta) \xrightarrow{\pi_1} (\mathscr{X}, \mathcal{B}(\mathscr{X})) \xrightarrow{f} (\mathbb{R}, \mathcal{B}(\mathbb{R})).$$

Cando $h \in \mathcal{L}^1(\theta\circ\pi_1^{-1})$, está garantido que:

$$\int_{\pi_1(\mathscr{X}\times\mathscr{Y})} f \,\mathrm{d}(\theta\circ\pi_1^{-1}) = \int_{\mathscr{X}\times\mathscr{Y}} (f\circ\pi_1)\,\mathrm{d}\theta$$

ou, tomando $\mu \overset{\text{def.}}{=} \theta\circ\pi_1^{-1}$:

$$\int_{\mathscr{X}} f(x)\,\mathrm{d}\mu(x) = \int_{\mathscr{X}\times\mathscr{Y}} f(x)\,\mathrm{d}\theta(x,y).$$

□

As diversas comprobacións de integrabilidade feitas ata agora permiten escribir *dcov* en termos de esperanzas. Para iso, se $X \sim \mu \in M_1^1(\mathscr{X})$ e $Y \sim \nu \in M_1^1(\mathscr{Y})$, con distribución conxunta $\theta := \mathrm{P}\circ\binom{X}{Y}^{-1}$, a súa covarianza de distancias vén dada por:

$$\mathrm{dcov}(X,Y) \overset{\text{Abuso}}{:=} \mathrm{dcov}(\theta) = \mathrm{E}[d_\mu(X,X')d_\nu(Y,Y')] =$$

$$= \mathrm{E}\left[\Big(d_{\mathscr{X}}(X,X') - \mathrm{E}[d_{\mathscr{X}}(X,X')|X] - \mathrm{E}[d_{\mathscr{X}}(X,X')|X'] + \mathrm{E}[d_{\mathscr{X}}(X,X')]\Big)\cdot\right.$$

$$\left.\cdot\Big(d_{\mathscr{Y}}(Y,Y') - \mathrm{E}[d_{\mathscr{Y}}(Y,Y')|Y] - \mathrm{E}[d_{\mathscr{Y}}(Y,Y')|Y'] + \mathrm{E}[d_{\mathscr{Y}}(Y,Y')]\Big)\right];$$

onde, os apóstrofos indican copias independentes e identicamente distribuídas do elemento aleatorio correspondente.

Para concluír este apartado, nótese que *dcov* sempre é unha medida de asociación, no sentido de que se anula baixo independencia:

$$\mathrm{dcov}(\mu\times\nu) = \int d_\mu d_\nu \,\mathrm{d}(\mu\times\nu)^2 \overset{\text{Fubini}}{=}$$

$$= \left(\int d_{\mathscr{X}} \, \mathrm{d}\mu^2 - 2 \int a_\mu \, \mathrm{d}\mu^2 + \int D(\mu) \, \mathrm{d}\mu^2 \right) \left(\int d_{\mathscr{Y}} \, \mathrm{d}\nu^2 - 2 \int a_\nu \, \mathrm{d}\nu^2 + \int D(\nu) \, \mathrm{d}\nu^2 \right) =$$
$$= [D(\mu) - 2D(\mu) + D(\mu)][D(\nu) - 2D(\nu) + D(\nu)] = 0.$$

Ademais, baixo certas condicións, *dcov* é non negativa e pode reescalarse no intervalo $[0, 1]$ (véxase 2.5.1), sendo así unha medida de asociación «normalizada» (Bishop *et al.*, 1975, páxs. 375–276).

2.4. A covarianza de distancias en espazos de tipo negativo

Posto que xa se viu que:

$$\theta = \mu \times \nu \Rightarrow \mathrm{dcov}(\theta) = 0,$$

é natural preguntarse en que espazos é certa a implicación en sentido contrario. Estes espazos son os de tipo negativo forte, xa que neles $\mathrm{dcov}(\theta)$ é unha función inxectiva de $\theta - \mu \times \nu$.

Con este propósito, comezarase por introducir intuitivamente os espazos de tipo negativo (§ 2.4.1), que son aqueles nos que *dcov* admite a devandita representación alternativa (aínda que a inxectividade non estea garantida). Logo introducirase a versión «forte» desta condición (§ 2.4.3) e concluirase cun resultado moi relevante: o tipo negativo forte non só é unha condición necesaria para que *dcov* caracterice a independencia, senón que é suficiente (cunha pequena excepción, nada restritiva).

2.4.1. Espazos métricos de tipo negativo

O concepto de espazo métrico de tipo negativo é moi antigo (Wilson, 1935) e nos últimos tempos está a gozar dunha «segunda xuventude»: en primeira instancia, polo seu papel na informática algorítmica (Deza e Laurent, 1997, § 6.1.; Naor, 2010) e, máis recentemente, da man da *enerxía dos datos* (Székely e Rizzo, 2017).

O espazo métrico $(\mathscr{X}, d_{\mathscr{X}})$ dise de tipo negativo se e só se:

$$\forall n \in \mathbb{N};\ \forall x, y \in \mathscr{X}^n :\ 2 \sum_{i,j=1}^{n} d_{\mathscr{X}}(x_i, y_j) \geq \sum_{i,j=1}^{n} [d_{\mathscr{X}}(x_i, x_j) + d_{\mathscr{X}}(y_i, y_j)].$$

A expresión analítica anterior correspóndese coa seguinte interpretación xeométrica: dados n puntos vermellos e outros tantos azuis, a suma das distancias entre os $2n^2$ pares ordenados da mesma cor non é inferior que a correspondente suma para distinta cor. Ademais, esta condición pódese formular dun xeito «aparentemente máis xeral» (Székely e Rizzo, 2017), dado polo carácter *condicionalmente definido negativo* da métrica. Porén, en realidade ambas son equivalentes tal e como se comproba tomando repeticións dos puntos e lembrando que $\mathbb{Q}$ é denso en $\mathbb{R}$:

$$\forall n \in \mathbb{N};\ \forall x \in \mathscr{X}^n;\ \forall \alpha \in \mathbb{R}^n, \sum_{i=1}^{n} \alpha_i = 0 :\ \sum_{i,j=1}^{n} \alpha_i \alpha_j d_{\mathscr{X}}(x_i, x_j) \leq 0.$$

O anterior quere dicir que un espazo métrico de tipo negativo é aquel no que a métrica actúa coma un núcleo definido negativo (coma os estudados pormenorizadamente en: Klebanov [2005] e Berg *et al.* [1984]).

O anterior tamén se pode estudar en termos de núcleos definidos positivos, xa que un espazo $(\mathscr{X}, d_{\mathscr{X}})$ é de tipo negativo se e só se a seguinte *diverxencia antipodal absoluta*:

$$d_o(x, y) := d_{\mathscr{X}}(x, o) + d_{\mathscr{X}}(y, o) - d_{\mathscr{X}}(x, y),\ (x, y) \in \mathscr{X}^2$$

é definida positiva para algún $o \in \mathscr{X}$.

Teñen tipo negativo, por exemplo, os espazos euclidianos e, con maior xeneralidade, todos os espazos de Hilbert, tal e como se verá en 2.4.2.

2.4.2. Representación en espazos de Hilbert

A continuación vanse presentar algúns resultados que involucran espazos de Hilbert. Por simplicidade, vaise supoñer que en todos eles o corpo escalar é $\mathbb{R}$; pero, en xeral, todas as afirmacións que se van facer son certas de maneira equivalente para $\mathbb{C}$. Isto demóstrase «realificando» ou «complexificando» (páxs. 132–135 de Jakobsen, 2017), segundo o caso.

Tamén vai ser preciso integrar funcións $f : \mathscr{X} \longrightarrow \mathcal{H}$ con rango un espazo de Hilbert. Se non se estivese esixindo a separabilidade de $\mathscr{X}$ (cf. § 2.2.2), como en Lyons (2013), os espazos $\mathcal{H}$ que xorden tampouco terían por que ser separables. En tal caso, unicamente sería posible definir as integrais no sentido «débil» ou de Pettis (1938), e non no «forte» ou de Bochner. Dada $\mu \in M(\mathscr{X})$, se f é integrable segundo Pettis (ou, máis concretamente, «escalarmente $\mu-$integrable»), a integral queda definida de maneira inequívoca pola súa conmutatividade respecto de calquera

aplicación do espazo dual $\mathcal{H}^*$:

$$I = \int_{\mathscr{X}} f \,\mathrm{d}\mu \overset{I \in \mathcal{H}}{\Leftrightarrow} \forall h^* : \mathcal{H} \longrightarrow \mathbb{R} \text{ linear e continua, } h^*(I) = \int_{\mathscr{X}} (h^* \circ f) \,\mathrm{d}\mu.$$

Porén, no marco teórico deste traballo, todos os espazos de Hilbert van ser separables, o que supón que as integrais de Pettis se reducen ás de Bochner.

Tras estas observacións técnicas, xa se pode enunciar o *teorema de Schoenberg* (Schoenberg, 1937 e 1938), que caracteriza os espazos de tipo negativo $(\mathscr{X}, d_{\mathscr{X}})$ como aqueles tales que $\left(\mathscr{X}, \sqrt{d_{\mathscr{X}}}\right)$ admite un «mergullo» isométrico nalgún espazo de Hilbert:

$$\exists\, \mathcal{H} \text{ esp. Hilbert; } \exists\, \varphi : \mathscr{X} \longrightarrow \mathcal{H};\ \forall x, y \in \mathscr{X} :\ \|\varphi(x) - \varphi(y)\|_{\mathcal{H}}^2 = d_{\mathscr{X}}(x, y).$$

Para unha demostración sinxela desta propiedade, que involucra a diverxencia antipodal absoluta (introducida en 2.4.1), véxase Jakobsen (2017, teorema 3.7), que corrixe a de Lyons (2013). Con independencia disto, este teorema supón que, ao manexar soamente espazos métricos separables (§ 2.2.2), os espazos de Hilbert que xorden tamén o serán.

O último «ingrediente» que falta para poder enunciar a representación de *dcov* en espazos de Hilbert é a definición do *operador baricentro*: dadas unha aplicación monótona $\varphi : \left(\mathscr{X}, \sqrt{d_{\mathscr{X}}}\right) \longrightarrow \mathcal{H}_1$ como a do teorema anterior e mais $\mu \in M^1(\mathscr{X})$, está garantida a existencia da integral de Pettis

$$\beta_\varphi(\mu) := \int_{\mathscr{X}} \varphi \,\mathrm{d}\mu \in \mathcal{H}_1$$

e denomínase *baricentro*, xa que é a media dun campo de vectores de $\mathcal{H}_1$ sobre o espazo $\mathscr{X}$ conforme á distribución dada por μ (recordando así a idea xeométrica dun centro de gravidade). De feito, se $X \sim \mu \in M_1^1(\mathscr{X})$,

$$\beta_\varphi(\mu) = \mathrm{E}[\varphi(X)].$$

Por outra banda, se $\psi : \left(\mathscr{Y}, \sqrt{d_{\mathscr{Y}}}\right) \longrightarrow \mathcal{H}_2$ tamén é isométrica, para cada $\theta \in M^{1,1}(\mathscr{X} \times \mathscr{Y})$, tamén está definido o baricentro do produto tensorial $\varphi \times \psi$:

$$\beta_{\varphi\otimes\psi}(\theta) := \int_{\mathscr{X}\times\mathscr{Y}} (\varphi \otimes \psi) \,\mathrm{d}\theta \in \mathcal{H}_1 \otimes \mathcal{H}_2.$$

Pero o máis interesante é que se, ademais, $\theta \in M_1^{1,1}(\mathscr{X} \times \mathscr{Y})$ ten como marxinais (μ, ν):

$$\mathrm{dcov}(\theta) = 4\,\|\beta_{\phi\otimes\psi}(\theta - \mu \times \nu)\| - \mathcal{H}_1 \otimes \mathcal{H}_2{}^2.$$

En conclusión, *dcov* caracterizará a independencia naqueles espazos nos que o devandito baricentro sexa inxectivo, que se estudarán no apartado seguinte.

2.4.3. Espazos de tipo negativo forte

Se $(\mathscr{X}, d_{\mathscr{X}})$ é de tipo negativo, verifícase a seguinte propiedade (cuxa demostración é sorprendentemente complexa [Jakobsen, 2017, lema 3.16]):

$$\forall \mu_1, \mu_2 \in M_1^1(\mathscr{X}) : \ D(\mu_1 - \mu_2) \leq 0.$$

Ademais, o espazo $(\mathscr{X}, d_{\mathscr{X}})$ dise de tipo negativo *forte* se, nel, o operador D é quen de «separar» as medidas de probabilidade (con primeiros momentos finitos):

$$D(\mu_1 - \mu_2) = 0 \Leftrightarrow \mu_1 = \mu_2.$$

A extensión do teorema de Schoenberg caracteriza estes espazos como aqueles para os que existe $\varphi : (\mathscr{X}, \sqrt{d_{\mathscr{X}}}) \longrightarrow \mathcal{H}_1$ isométrica e ademais verificando que β_φ é inxectiva. Máis aínda, se se consideran dous espazos de tipo negativo forte $\mathscr{X}$ e $\mathscr{Y}$, sempre se poden atopar aplicacións isométricas $\varphi : (\mathscr{X}, \sqrt{d_{\mathscr{X}}}) \longrightarrow \mathcal{H}_1$ e $\psi : (\mathscr{Y}, \sqrt{d_{\mathscr{Y}}}) \longrightarrow \mathcal{H}_2$ tales que o baricentro $\beta_{\varphi\otimes\psi} : M^{1,1}(\mathscr{X} \times \mathscr{Y}) \longrightarrow \mathcal{H}_1 \otimes \mathcal{H}_2$ é inxectivo. Polo tanto, se os espazos $\mathscr{X}$ e $\mathscr{Y}$ son de tipo negativo forte,

$$\operatorname{dcov}(X, Y) = 0 \Leftrightarrow X, Y \text{ independentes},$$

para calquera elemento aleatorio $Z = (X, Y) : \Omega \longrightarrow \mathscr{X} \times \mathscr{Y}$.

Polo tanto, o tipo negativo forte é unha condición *suficiente* dos espazos marxinais para que se verifique a equivalencia anterior, pero será tamén *necesaria*? A resposta é un *si* cun pequeno matiz: se $(\mathscr{X}, d_{\mathscr{X}})$ non é de tipo negativo forte (e simetricamente para $\mathscr{Y}$), é posible achar $\theta \in M_1^{1,1}(\mathscr{X} \times \mathscr{Y})$ tal que:

$$\operatorname{dcov}(\theta) = 0 \text{ e, no entanto, } \theta \neq (\theta \circ \pi_1^{-1}) \times (\theta \circ \pi_2^{-1});$$

sempre e cando $\min\{\#\mathscr{X}, \#\mathscr{Y}\} > 1$. De feito, esta construción é moi sinxela: abonda definir

$$\theta := \frac{\mu_1 \times \delta_{y_1} + \mu_2 \times \delta_{y_2}}{2};$$

onde μ_1, μ_2 son dúas medidas diferentes de $M_1^1(\mathscr{X})$ tales que $D(\mu_1 - \mu_2) = 0$ e $y_1, y_2 \in \mathscr{Y}$ tamén son distintos entre si.

Finalmente, vaise demostrar que a exclusión dos casos en que algún espazo marxinal é unitario non é restritiva; posto que, nestas situacións «dexeneradas», claramente se ten que dcov $\equiv 0$ (xa que $d_\nu \equiv 0$ cando $\#\mathscr{Y} = 1$) e que toda $\theta \in M_1^{1,1}(\mathscr{X} \times \mathscr{Y})$ é produto das súas marxinais. Para ver isto último, primeiro nótese que:

$$\mathscr{Y} = \{y\} \Rightarrow \mathcal{B}(\mathscr{Y}) = \{\emptyset, \{y\}\} = \{\emptyset, \mathscr{Y}\}.$$

Entón, se $B \in \mathcal{B}(\mathscr{Y})$,

$$\forall A \in \mathcal{B}(\mathscr{X}),\ \theta(A \times B) = \begin{cases} \theta(A \times \emptyset) = \theta(\emptyset) = 0 = \mu(A)\nu(\emptyset) \\ \theta(A \times \mathscr{Y}) = \theta\pi_1^{-1}(A) \equiv \mu(A) = \mu(A)\nu(\mathscr{Y}) \end{cases};$$

do que se conclúe que $\theta = \mu \times \nu$. Este resultado analítico simplemente quere dicir que, como é lóxico, se un elemento aleatorio Y toma sempre o mesmo valor, as observacións de calquera outro X aleatorio van ser independentes das de Y.

Tras o desenvolvemento teórico anterior, queda claro o interese de identificar os espazos métricos de tipo negativo forte. A estes efectos, é de grande utilidade práctica saber que entre eles están todos os espazos de Hilbert separables. Aínda que este resultado é pouco sorprendente, a súa demostración é notablemente dificultosa (Jakobsen, 2017, páxs. 49–60).

2.5. Correlación de distancias en espazos métricos

2.5.1. A medida de asociación *dcor*

Como anteriormente, sexa $(X, Y) \sim \theta \in M_1^{1,1}(\mathscr{X} \times \mathscr{Y})$ con marxinais (μ, ν), onde $(\mathscr{X}, d_{\mathscr{X}})$ e $(\mathscr{Y}, d_{\mathscr{Y}})$ son dous espazos métricos (separables) calquera. Entón, verifícanse as seguintes desigualdades:

$$|\operatorname{dcov}(X, Y)| \leq \sqrt{\operatorname{dvar}(X)\operatorname{dvar}(Y)} \leq D(\mu)D(\nu);$$

onde, loxicamente, $\operatorname{dvar}(X) := \operatorname{dcov}(X, X)$. Se, ademais, se supón que $(\mathscr{X}, d_{\mathscr{X}})$ e $(\mathscr{Y}, d_{\mathscr{Y}})$ son de tipo negativo:

$$\operatorname{dcov}(X, Y) = 4\,\|\beta_{\varphi\times\psi}(\theta - \mu \times \nu)\|^2_{\mathcal{H}_1\otimes\mathcal{H}_2} \geq 0.$$

Neste contexto, e supoñendo que o denominador non se anule, pódese definir a *correlación de distancias en espazos métricos*:

$$\operatorname{dcor}(X, Y) := \frac{\operatorname{dcov}(X, Y)}{\sqrt{\operatorname{dvar}(X)\operatorname{dvar}(Y)}} \in [0, 1].$$

Para ver cando está *dcor* ben definida, hai que estudar cando $\operatorname{dvar}(X)$ acada os valores extremos do seu rango $[0, D(\mu)^2]$. Pódese comprobar que:

$$\operatorname{dvar}(X) = 0 \Leftrightarrow \exists\, x \in \mathscr{X},\ \mu = \delta_x \text{ “}\mu\text{−case seguro”};$$

$$\operatorname{dvar}(X) = D(\mu)^2 \Leftrightarrow \exists\, x, y \in \mathscr{X},\ \mu = \frac{\delta_x + \delta_y}{2} \text{ “}\mu\text{−case seguro”.}$$

Polo tanto, en casos non dexenerados pódese falar de correlación de distancias tranquilamente.

2.5.2. *dcor* en espazos euclidianos

Segundo o visto no apartado anterior *dcor* ten como rango $[0, 1]$ e anúlase se e só se hai independencia, o que lembra a propiedade vista para espazos euclidianos (§ 2.1). De feito, é posible demostrar (usando as representacións en espazos de Hilbert vistas en 2.4.2) que, cando $(\mathscr{X}, d_{\mathscr{X}})$ e $(\mathscr{Y}, d_{\mathscr{Y}})$ son espazos euclidianos (de dimensión finita), a noción de correlación de distancias vista en 2.1 (Székely *et al.*, 2007) estende o cadrado da de 2.5.1 (Lyons, 2013):

$$\operatorname{dcov}(X, Y) = \operatorname{dCov}(X, Y)^2 \Rightarrow \operatorname{dcor}(X, Y) = \operatorname{dCor}(X, Y)^2.$$

Ademais, sen máis que desenvolver o produto de esperanzas a que se reduce $\operatorname{dcov}(X, Y)$ cando $\theta \in M_1^{2,2}(\mathscr{X} \times \mathscr{Y})$, tense a extensión da covarianza de distancias browniana (Székely e Rizzo, 2009, teoremas 7–8) a espazos métricos xerais:

$$\operatorname{dcov}(X, Y) = \mathrm{E}[d_{\mathscr{X}}(X, X')d_{\mathscr{Y}}(Y, Y')] + \mathrm{E}[d_{\mathscr{X}}(X, X')]\,\mathrm{E}[d_{\mathscr{Y}}(Y, Y')] - 2\,\mathrm{E}[d_{\mathscr{X}}(X, X')d_{\mathscr{Y}}(Y, Y'')].$$

En conclusión, *dcov* xeneraliza de maneira satisfactoria o cadrado de *dCov*.

2.6. Contraste de independencia non paramétrico en espazos métricos

2.6.1. Núcleo asociado a *dcov*

O seguinte «núcleo» (aplicación) vai ser transcendental no desenvolvemento posterior:

$$\begin{aligned} h : (\mathscr{X} \times \mathscr{Y})^6 &\longrightarrow \mathbb{R} \\ \big((x_i, y_i)\big)_{i=1}^6 &\mapsto f_{\mathscr{X}}(x_1, x_2, x_3, x_4)\, f_{\mathscr{Y}}(y_1, y_2, y_5, y_6); \end{aligned}$$

onde, para $\mathscr{Z} \in \{\mathscr{X}, \mathscr{Y}\}$,

$$f_{\mathscr{Z}}(z) := d_{\mathscr{Z}}(z_1, z_2) + d_{\mathscr{Z}}(z_3, z_4) - d_{\mathscr{Z}}(z_1, z_3) - d_{\mathscr{Z}}(z_2, z_4),\ z \in \mathscr{Z}^4.$$

As funcións $f_{\mathscr{Z}}$ e h son claramente medibles e, para probar a súa integrabilidade, cómpre acudir a certas desigualdades combinatorias baseadas na desigualdade triangular, cuxo manexo é simple mais sumamente laborioso (véxanse as páxs. 148–150 de Jakobsen [2017] para unha versión corrixida das formuladas por Lyons [2013]). Unha vez analizado que ten sentido facelo, calcular as integrais das devanditas funcións resulta inmediato. En primeiro lugar:

$$\int_{(\mathscr{X}\times\mathscr{Y})^2} f_{\mathscr{X}}(x_1, x_2, x_3, x_4)\, \mathrm{d}\theta^2((x_3, y_3), (x_4, y_4)) \overset{\text{ACOV}}{=}$$

$$= d_{\mathscr{X}}(x_1, x_2) - a_\mu(x_1) - a_\mu(x_2) + D(\mu) \equiv d_\mu(x_1, x_2),\ (x_1, x_2) \in \mathscr{X}^2$$

para calquera $\theta \in M_1^{1,1}(\mathscr{X} \times \mathscr{Y})$ con marxinais (μ, ν). Tendo en conta que, *mutatis mutandi*, o mesmo é certo para $f_{\mathscr{Y}}$, dedúcese que:

$$\mathrm{dcov}(\theta) = \int_{(\mathscr{X}\times\mathscr{Y})^2} d_\mu(x_1, x_2) d_\nu(y_1, y_2)\, \mathrm{d}\theta^2\, ((x_1, y_1), (x_2, y_2)) = \int_{(\mathscr{X}\times\mathscr{Y})^6} h\, \mathrm{d}\theta^6.$$

Isto significa que, se Z é un vector que contén 6 elementos aleatorios independentes e identicamente distribuídos a $(X, Y) \sim \theta$,

$$\mathrm{dcov}(\theta) = \mathrm{E}[h(Z)] \equiv \mathrm{E}\left[h\left((X_i, Y_i)_{i=1}^6\right)\right]$$

e, polo tanto, a súa versión empírica vai ser un $V-$estatístico, como os estudados (erroneamente) por Lyons (2013), tal e como se verá en 2.6.2.

2.6.2. Distancia de covarianzas empírica e $U-$estatístico asociado

Antes de nada, lémbrese que, para $n \in \mathbb{Z}^+$, a *medida empírica* asociada ás observacións $\{(X_i, Y_i)\}_{i=1}^n$ i.i.d. $(X, Y) \sim \theta$, é:

$$\theta_n := \frac{1}{n}\sum_{i=1}^n \delta_{(X_i, Y_i)} : \Omega \longrightarrow M_1^{1,1}(\mathscr{X} \times \mathscr{Y}).$$

Facendo algunhas operacións elementais, chégase a que o estimador natural

$$\widehat{\mathrm{dcov}}(\theta) := \mathrm{dcov}(\theta_n)$$

é o $V-$estatístico con núcleo (non simétrico) h:

$$\operatorname{dcov}(\theta_n) = \frac{1}{n^6}\sum_{i_1=1}^{n}\cdots\sum_{i_6=1}^{n} h\left((X_{i_\lambda}, Y_{i_\lambda})_{\lambda=1}^{6}\right) \equiv V_n^6(h).$$

Por outra banda, é considerar o $U-$estatístico correspondente como estimador alternativo, que ademais vai presentar un comportamento satisfactorio baixo condicións moito menos restritivas que $\operatorname{dcov}(\theta_n)$. Así pois, se $n \geq 7$, ten sentido definir:

$$\tilde{U}_n^6(R) := \frac{1}{6!\binom{n}{6}} \sum_{\{i_\lambda\}_\lambda \subset [1,n]\cap\mathbb{Z} \text{ distintos}} h\left((X_{i_\lambda}, Y_{i_\lambda})_{\lambda=1}^{6}\right);$$

onde a vírgula indica que, en lugar dos habituais núcleos simétricos, estase a traballar cun que non o é. Porén, a partir de h pódese obter $\bar{h}$ simétrico do seguinte xeito:

$$\bar{h}(z) := \frac{1}{6!}\sum_{\sigma\in S_6} h\left(z_{\sigma(j)}\right)_{j=1}^{6} \equiv \frac{1}{6!}\sum_{\sigma\in S_6} h(z_\sigma),\ z \in (\mathscr{X}\times\mathscr{Y})^6;$$

sendo $S_6 := \{\sigma : [1,6]\cap\mathbb{Z} \longrightarrow [1,6]\cap\mathbb{Z} :\ \sigma \text{ bixectiva}\}$ o grupo simétrico de orde 6. E a definición de $\bar{h}$ vén ao caso porque, ademais, claramente verifica:

$$\tilde{U}_n^6(h) = \frac{1}{\binom{n}{6}} \sum_{i_1<\ldots<i_6} \bar{h}\left((X_{i_\lambda}, Y_{i_\lambda})_{\lambda=1}^{6}\right).$$

Tamén se ten o resultado análogo para o $V-$estatístico:

$$\forall\,\sigma\in S_6,\ \operatorname{dcov}(\theta_n) \equiv V_n^6(h) = \int_{(\mathscr{X}\times\mathscr{Y})^6} h(z)\,\mathrm{d}\theta_n^6(z) \overset{\text{Fubini}}{=}$$

$$= \int_{(\mathscr{X}\times\mathscr{Y})^6} h(z)\,\mathrm{d}\theta_n^6(z_{\sigma^{-1}}) \overset{\text{ACOV}}{=} \int_{(\mathscr{X}\times\mathscr{Y})^6} h(z_\sigma)\,\mathrm{d}\theta_n^6(z) = V_n^6(\bar{h})$$

e, ademais, o razoamento anterior permite demostrar que $\operatorname{dcov}(\theta) = \int \bar{h}\,\mathrm{d}\theta^6$.

Agora que xa se pode traballar cos habituais núcleos simétricos, é posible aplicar a *lei forte dos grandes números* (LFGN) para $U-$estatísticos (Hoeffding, 1961) e deducir directamente que, se $\theta \in M_1^{1,1}(\mathscr{X}\times\mathscr{Y})$,

$$\tilde{U}_n^6(h) \xrightarrow[n\to\infty]{c.s.} \operatorname{dcov}(\theta).$$

Lyons (2013) confundiu as hipóteses do devandito teorema de Hoeffding coas da LFGN para $V-$estatísticos (Giné e Zinn, 1992, páx. 274), segundo se indica tacitamente en Lyons (2018). As condicións máis débiles baixo as cales se ten aplicado a LFGN para $V-$estatísticos neste contexto veñen dadas por: $\theta \in M_1^{5/3,5/3}(\mathscr{X}\times\mathscr{Y})$. (Jakobsen, 2017, teorema 5.5). Noutras palabras, a finitude dos momentos de orde $\frac{5}{3}$ é suficiente para que:

$$V_n^6(h) \xrightarrow[n\to\infty]{c.s.} \operatorname{dcov}(\theta).$$

2.6.3. Distribución nula do estatístico de contrate

Se $\theta \in M_1^{1,1}(\mathscr{X} \times \mathscr{Y})$ é igual ao produto das súas marxinais e estas son non dexeneradas, a distribución nula asintótica dos estimadores introducidos en 2.6.2 é:

$$nV_n^6(h) \xrightarrow[n\to\infty]{\mathcal{D}} \sum_{i=1}^{\infty} \lambda_i (Z_i^2 - 1) + D(\mu)D(\nu);$$
$$n\tilde{U}_n^6(h) \xrightarrow[n\to\infty]{\mathcal{D}} \sum_{i=1}^{\infty} \lambda_i (Z_i^2 - 1);$$

onde $\{Z_i\}_{i\in\mathbb{N}^*}$ i.i.d. $\mathbb{N}(0,1)$ e $\{\lambda_i\}_{i\in\mathbb{N}^*}$ son os autovalores (con multiplicidade) do operador linear $S : \mathcal{L}^2(\theta) \longrightarrow \mathcal{L}^2(\theta)$ que leva f en $S(f) : \mathscr{X} \times \mathscr{Y} \longrightarrow \mathbb{R}$, definida como:

$$S(f)(x,y) := \int_{\mathscr{X}\times\mathscr{Y}} d_\mu(x,x') d_\nu(y,y') f(x',y') \,\mathrm{d}\theta(x',y'),\ (x,y) \in \mathscr{X} \times \mathscr{Y}.$$

A demostración orixinal do resultado para o $V-$estatístico (Lyons, 2013) incluía un razoamento erróneo para concluír que $\sum_{i=1}^{\infty} \lambda_i = D(\mu)D(\nu)$. Lyons (2018) afirma que isto si que é certo cando os espazos marxinais son de tipo negativo, aínda que cunha xustificación un tanto abstrusa. De ser isto certo, coincidiría exactamente coa distribución asintótica de Székely *et al.* (2007).

En calquera caso, isto carece de utilidade práctica inmediata unha vez máis (cf.§ 2.1), ao depender $\{\lambda_i\}_i$ de θ (descoñecida) e non ser facilmente estimables. A solución máis lóxica é, coma en 2.1, unha estratexia de remostraxe. Se se desexa que esta teña uns sólidos fundamentos teóricos e que estea garantida a súa consistencia, unha opción é usar *bootstrap* naíf para aproximar os niveis críticos para o rexeitamento do test e xustificalo en base aos resultados de Arcones e Giné (1992), xa que $\bar{h}$ verifica a condición de integrabilidade requirida por estes autores.

Capítulo 3

Proposta dun test baseado na correlación de distancias

Ao longo do capítulo 2 sentáronse as bases teóricas para o uso da correlación de distancias en certos espazos métricos. Unha vez se dispón deste aval, ten sentido propoñer extensións e aplicacións das propostas de Lyons (2018). Posiblemente por este motivo, a comunidade científica non comezou a producir avances neste eido ata moi recentemente, sendo as dúas contribucións máis destacables as relativas ás series de tempo (Davis *et al.*, 2018) e aos procesos estocásticos discretizados (Dehling *et al.*, 2018).

O que se vai propoñer no presente capítulo é a particularización do marco teórico a espazos de cardinal 3 (§ 3.1), para logo deseñar un procedemento adaptado á detección da epistase (§ 3.2) que ademais solucione as deficiencias dos tests de Cai e Liu (2016) vistas en 1.3.5. Ademais, en 3.3 discutiranse algúns detalles técnicos.

Finalmente, en 3.4 explícanse as principais dificultades informáticas que se atoparon, salientando que na bibliografía especializada non hai constancia de que ninguén teña realizado unha proposta como a deste capítulo con anterioridade.

3.1. Correlación de distancias en espazos de cardinal 3

Evidentemente, nun espazo finito, a finitude dos momentos (de calquera orde) e a separabilidade non son un problema. Ademais, acudindo ás definicións (Klebanov, 2005; Lyons, 2013), é posible atopar unha demostración directa e orixinal (resolvendo varios sistemas de inecuacións moi laboriosos, que se omiten por rutineiros) de que, se $(\mathscr{X}, d_{\mathscr{X}})$ só ten tres puntos, é necesariamente de tipo negativo forte. Esta proba, en principio, non sería necesaria porque bastaría observar que claramente $\left(\mathscr{X}, \sqrt{d_{\mathscr{X}}}\right)$ admite unha inmersión nun espazo de Hilbert (en base ao teorema de Schoenberg antes mencionado), o que claramente ocorre (vértices dun triángulo en $\mathbb{R}^2$, tamén tras aplicar a transformación raíz cadrada). Con todo, é bo comprobar que, nun caso en que a estrutura métrica é moi sinxela, os abstractos argumentos de teoremas coma este tamén se simplifican.

Sexa $\mathscr{X} := \{0, 1, 2\}$ o conxunto dos tres posibles xenotipos para cada SNP. O coñecemento biolóxico indica que non hai ningunha base para considerar que $2 \in \mathscr{X}$ copias do alelo menor afectan o dobre que unha (Bush e Moore, 2012), ben sexa aumentando ou diminuíndo a susceptibilidade a padecer algún trastorno psiquiátrico. De feito, hai casos nos que a susceptibilidade a un trastorno psiquiátrico pode ser máxima en heterocigose (Costas *et al.*, 2011), situación que se reflicte mediante $1 \in \mathscr{X}$.

En situacións coma estas non hai ningunha razón para considerar necesariamente a distancia euclidiana:

$$d_E(0, 2) = 2d_E(0, 1) = 2d_E(1, 2),$$

en lugar de espazos métricos máis xenerais que un «linear». É aquí onde resulta útil a correlación de distancias para estender as ideas de Cai e Liu (2016). Como xa se comentou con anterioridade, o carácter marcadamente discreto dos datos de SNP é un aliciente máis para abandonar a idea da correlación linear.

A priori, non se está buscando ningún tipo de interacción específico: o obxectivo é simplemente detectar a epistase. Por este motivo, e tamén para reducir os inxentes custos informáticos, a distancia utilizada en todas as simulacións será a «equilátera»:

$$d(0, 1) = d(1, 2) = d(0, 2) = 1.$$

Por outra parte, vanse definir tres métricas «dexeneradas» (dous dos vértices do triángulo coinciden), que serán especialmente ilustrativas para os datos reais, permitindo interpretar que clase de modelo alélico se está estudando:

- Recesivo (distancia "0=1"): $d(0,1) = 0;\ \ d(0,2) = d(1,2) = 1$.

- Heterocigoto (distancia "0=2"): $d(0,2) = 0;\ \ d(0,1) = d(2,1) = 1$.

- Dominante (distancia "1=2"): $d(1,2) = 0;\ \ d(1,0) = d(2,0) = 1$.

3.2. Contraste de hipóteses proposto

Para simplificar a notación, sexan X e Y variables aleatorias con soporte $\{0,1,2\}$, correspondentes a dous SNP distintos, para as que se dispón dunha m.a.s. conxunta de tamaño $n \in \mathbb{Z}^+$:

$$(X_1, Y_1), \dots, (X_n, Y_n) \text{ i.i.d. } (X, Y).$$

Nesta situación, hai varias formas de calibrar a distribución nula do estatístico de contraste $\widehat{\text{dcov}}(X,Y)$ para o test

$$H_{0XY} \ : \ X, Y \text{ independientes.}$$

Loxicamente, debe incorporarse ao plan de remostraxe a información de que se dispón: neste caso, a independencia de X e Y baixo H_{0XY}. Desta forma, o aconsellable non é «remostrear» en $\{(X_i, Y_i)\}_i$, senón en $\mathcal{X} := \{X_i\}_i$ e $\mathcal{Y} := \{Y_i\}_i$ de maneira separada (*permutation tests*). Así, abondará calcular $B \in \mathbb{Z}^+$ estatísticos da forma $\widehat{\text{dcov}}(\mathcal{X}^*, \mathcal{Y}^*)$ para aproximar a distribución na mostraxe baixo H_0 da covarianza de distancias (empírica) por Monte–Carlo.

Desta forma, dados p SNP, cómpre realizar $p^2 - p$ contrastes de independencia, a metade no grupo de casos e a outra metade nos controis. Nunha segunda etapa, rexeitarase a ausencia de epistase entre un par de SNP se se rexeitou a independencia para sans e non para enfermos, ou viceversa.

Ao realizar o proceso en dúas etapas, o nivel de significación utilizado na primeira non pode ser o valor nominal. Para efectuar tal control da FWER, utilizarase a corrección de Bonferroni. Inicialmente, tamén se considerou a de Šidák (1967), pero na práctica comprobouse que conduce a resultados moi similares.

3.3. Discusión metodolóxica do contraste proposto

Á parte do bo comportamento da técnica introducida en 3.2 que se vai expoñer nos capítulos 4 e 5, o máis salientable é que a aplicación desta metodoloxía estatística a datos discretos (xenómicos ou non) non conta con precedentes na bibliografía, afirmación que se realiza tras ter revisado todos os artigos que citan a Lyons (2013) (máis de cen) e lanzado unha busca entre os que contan con Székely *et al.* (2007) entre as súas referencias (case un milleiro).

Unha pregunta importante é por que se optou por deseñar un procedemento en dúas etapas, en lugar de testar a igualdade de correlacións de distancias directamente. A xustificación deste feito vén dada da mesma forma que en Cai e Liu (2016), é dicir, seguindo os argumentos de De la Fuente (2010), entre outros (como xa se comentou na introdución de 1.3). No entanto, non todas as diferencias significativas entre correlacións son indicativas de epistase, senón unicamente aquelas en que un dos valores é próximo a cero e o outro non, en contra do suposto por Cai e Liu. Ademais, é especialmente importante realizar esta distinción no caso da correlación de distancias porque, se ben a súa nulidade caracteriza a independencia, a interpretación do «grande» ou «pequena» que sexa non permite analizar a intensidade da dependencia de maneira tan sinxela coma no caso linear, senón que se producen algúns fenómenos «contraintuitivos» (Székely e Rizzo, 2013).

Por outra parte, no que atangue á técnica de remostraxe escollida, comprobouse experimentalmente que a dos *tests de permutacións* a máis axeitada, comparando o contraste resultante en termos de potencia co *bootstrap* naíf proposto por Jakobsen (2017, páx. 99). Isto non é estraño se se ten en conta o ben que funciona esta estratexia en espazos euclidianos (Székely *et al.*, 2007; Székely e Rizzo, 2017), pero ten o inconveniente de que non se atopou na bibliografía ningunha xustificación teórica sólida (como a de Arcones e Giné [1992] para o *bootstrap*), se ben é certo que son poucos os casos en que a remostraxe é inconsistente.

Tamén debe aclararse que en artigos como Cai e Liu (2016) e Székely *et al.* (2007) defenden que o número de remostras B é pouco importante para os seus métodos, sempre e cando non sexa esaxeradamente pequeno. Por isto, e tendo en conta que o tempo de execución é $O(B)$, optouse por incorporar ao contraste que se presenta neste traballo a «receita» para B en función do tamaño mostral n de Székely *et al.* (2007):

$$B(n) = 200 + \lfloor 5000/n \rfloor$$

Con posterioridade a esta decisión, realizáronse numerosas comparativas e compro-

bouse que, efectivamente, é irrelevante para a potencia e a calibración de niveis de significación (sempre e cando o nivel nominal non sexa extremadamente pequeno).

Porén, se o obxectivo fose o control da FDR, a cousa cambiaría: tanto Benjamini e Hochberg (1995) como Benjamini e Yekutieli (2001) baseánse na ordenación de p-valores e, en consecuencia, deben eludirse valores de B demasiado pequenos en relación ao número de contrastes efectuados, evitando así unha acumulación excesiva de niveis críticos nulos. Desafortunadamente, á hora de analizar datos reais (coma os do capítulo 5), isto resulta inviable (véxanse as cuestións empíricas en 3.4). Ante esta situación, optouse por unha filosofía de FWER (en lugar de FDR), a cal ten como única pequena desvantaxe que hai que obviar as febles estruturas de dependencias entre os $\binom{L}{2}$ tests efectuados (tanto por coincidir un dos SNP como polo desequilibrio de ligamento). Malia isto, esta parece a elección máis sensata.

3.4. Dificultades informáticas

Neste apartado explícase o importante desafío desde o punto de vista informático que supuxo a implementación do test presentado en 3.2 e a maneira en que se superou.

Loxicamente, os inconvenientes inherentes aos GWAS (alta dimensión e baixo tamaño mostral) son comúns a toda técnica que intente realizar inferencia estatística a partir deles. A modo ilustrativo, na Táboa 3.1 compáranse os tempos de execución do código orixinal en *R* (R Core Team, 2018) e as versións optimizadas, todos eles no superordenador Finisterrae II do CESGA. Á vista destas cantidades, queda plenamente xustificada a necesidade do uso de computación de altas prestacións (HPC), especialmente se un ten en conta que un GWAS coma os do capítulo 5 podería contar ata con $L = 10\,000$ e que o tempo de execución é unha función linear de $\binom{L}{2}$ e, polo tanto, $O(L^2)$.

	Simulación, $R = 10^3$	GWAS, $L = 1000$	GWAS, $L = 4117$ *
R secuencial	12 h 19 min	42 días 1 h	2 anos
R e C secuencial	3 h 59 min	2 días 1 h	29 días 1 h
R e C paralelo	50 min	2 h 41 min	9 h 48 min

Táboa 3.1: *Comparativa de tempos de execución.*
Os tempos para o GWAS grande son estimacións.

Na versión paralela, usouse a mínima cantidade de recursos de *hardware* que permitía

un tempo de execución razoable en cada caso: 12 *cores* para os modelos simulados, 48 procesadores para $L = 1000$ e 240 (planificados) para $L = 4\,117$. Ademais, o código optimizouse de dúas maneiras distintas. Inicialmente afrontáronse os modelos simulados, para os que bastou o uso dun esquema de memoria compartida mediante a biblioteca OpenMP, onde se reparte o cómputo entre os núcleos e fíos existentes nun mesmo procesador.

Pola súa parte, os datos reais supuxeron un desafío moito maior. En consecuencia, acudiuse a un esquema de paralelización «amo-escravo» (memoria distribuída segundo o paradigma *master/slave*), a través de MPI (Gropp *et al.*, 2014). Consiste en:

1. Un procesador (*amo*) chama as rutinas de *R* para cargar os datos matriciais, dividilos e distribuílos entre os diferentes procesadores (*escravos*).
2. Cada procesador traballa co seu fragmento de matriz, executando as iteracións (contrastes de independencia para unha fracción do total de pares de SNP) que ten asignadas.
3. Unha vez que cada escravo concluíu a súa parte, envíalla ao amo.
4. Finalmente, o amo reconstrúe a matriz de p-valores final, que permitirá procesar os resultados en *R*.

Os programas optimizados combinan unha interface na linguaxe de programación *R* cun «núcleo» en *C*, o cal realiza as operacións máis custosas a baixo nivel. Todas elas involucran vectores e matrices de gran tamaño, polo que para acelerar os cálculos tamén se empregaron as bibliotecas MKL e as instrucións SSE de Intel®.

Non foi posible o uso de *software* preexistente, xa que os algoritmos relativos á correlación de distancias máis eficientes (coma Huo e Székely, 2015) só están deseñados para o caso euclidiano e, máis aínda, os poucos programas que involucran outras métricas non son públicos nin adaptables á estrutura dos espazos de tres puntos que se manexan neste capítulo.

Tras unha pescuda bibliográfica intensiva entre os artigos que citan a Székely *et al.* (2007) e Lyons (2013), parece xustificado afirmar que non hai constancia do uso da correlación de distancias en espazos discretos nin, en particular, da súa aplicación á busca de interaccións entre SNP. Así mesmo, tampouco se atopou ningún traballo previo a este TFM en que se realicen *tests múltiples a grande escala* empregando ningunha das técnicas derivadas da *enerxía dos datos* (Székely e Rizzo, 2017).

Capítulo 4

Estudo de simulación

Neste capítulo preséntase un breve resumo dos estudos de simulación realizados, incluíndo unicamente unha pequena selección das táboas e gráficas máis representativas. A § 4.1 describe os modelos estatísticos que se crearon para este propósito. Isto foi necesario porque, como se comentou en 1.2, non existe consenso na forma de avaliar o rendemento de detectores de epistase a nivel teórico e, ademais, son moi escasos os artigos cuxos modelos son reproducibles. Por último, en 4.2 expóñense algúns dos resultados máis salientables que se obtiveron, acompañados dunha breve discusión.

4.1. Deseño de modelos poboacionais para a validación do método

Vanse definir varios modelos teóricos para a interacción entre (Z_i, Z_j), onde Z designa X ou Y, segundo o caso. Ao longo desta sección expoñeranse algúns dos máis representativos de entre aqueles que se crearon para os estudos de simulación.

Á hora de elixir as frecuencias marxinais, optouse por respectar o equilibrio de Hardy–Weinberg (Hardy, 1908; Weinberg, 1908), xa que os SNP da base de datos de esquizofrenia o cumpren, ao ser este un dos filtros impostos aos polimorfismos seleccionados (§ 5.1). Ademais, cada vez que se xera o modelo sortéase a frecuencia do alelo menor de maneira uniforme no intervalo [0'05,0'2], cuxo límite inferior responde

ao observado na base de datos e o superior permite que os modelos non sexan excesivamente «favorables», dificultando a detección das interaccións e outorgando maior solidez ás curvas de potencia que se obteñan.

A necesidade de crear tales modelos *ad hoc* responde ao feito de que nos artigos que estudan a epistase entre SNP é moi infrecuente ver modelos simulados e, cando os hai (como en Marchini *et al.* [2005], por exemplo), son demasiado simplistas e en ningún caso permiten regular a intensidade da interacción para estudar a robustez fronte a diversas alternativas.

Modelo `indep`

O modelo máis evidente é aquel en que a probabilidade de cada xenotipo é o produto das marxinais (hai independencia), tal e como ilustra a correspondente táboa de continxencia 3×3 (Táboa 4.1).

$Z_i \setminus Z_j$	0	1	2	
0	pr	ps	$p(1-r-s)$	p
1	qr	qs	$q(1-r-s)$	q
2	$(1-p-q)r$	$(1-p-q)s$	$(1-p-q)(1-r-s)$	$1-p-q$
	r	s	$1-r-s$	1

Táboa 4.1: *Táboa de continxencia para o modelo* `indep`.

Modelos `qexp` e `rexp`

Aquí, a magnitude da dependencia vén dada polo afastado de 1 que estea o parámetro $e \in [1, +\infty[$, tal e como o reflicten as Táboas 4.2–4.3. Ademais, para un mesmo valor de e, a intensidade da interacción é maior no modelo `qexp` que en `rexp`, a causa do equilibrio de Hardy–Weinberg.

$Z_i \setminus Z_j$	0	1	2	
0	$pr + q^e s - qs$	$ps - q^e s + qs$	$p(1-r-s)$	p
1	$qr - q^e s + qs$	$q^e s$	$q(1-r-s)$	q
2	$(1-p-q)r$	$(1-p-q)s$	$(1-p-q)(1-r-s)$	$1-p-q$
	r	s	$1-r-s$	1

Táboa 4.2: *Táboa de continxencia para* `qexp`*.*

$Z_i \setminus Z_j$	0	1	2	
0	pr	ps	$p(1-r-s)$	p
1	qr	$qs - [(1-p-q) - (1-p-q)^e](1-r-s)$	$q(1-r-s) + [(1-p-q) - (1-p-q)^e](1-r-s)$	q
2	$(1-p-q)r$	$(1-p-q)s + [(1-p-q) - (1-p-q)^e](1-r-s)$	$(1-p-q)^e(1-r-s)$	$1-p-q$
	r	s	$1-r-s$	1

Táboa 4.3: *Táboa de continxencia para* `rexp`*.*

Modelos `qmult` e `rmult`

Ao igual que nos modelos anteriores, a asociación é menos manifesta canto máis próximo a 1 sexa o parámetro, que neste caso é $g \in [0, 1]$ (Táboas 4.4–4.5). Analogamente ao que ocorría antes, a igual g, `qmult` presenta unha interacción máis forte que `rmult`.

$Z_i \setminus Z_j$	0	1	2	
0	$pr - (1-g)qs$	$ps + (1-g)qs$	$p(1-r-s)$	p
1	$qr + (1-g)qs$	gqs	$q(1-r-s)$	q
2	$(1-p-q)r$	$(1-p-q)s$	$(1-p-q)(1-r-s)$	$1-p-q$
	r	s	$1-r-s$	1

Táboa 4.4: *Táboa de continxencia para* `qmult`*.*

$Z_i \setminus Z_j$	0	1	2	
0	pr	ps	$p(1-r-s)$	p
1	qr	$qs-(1-g)(1-p-q)(1-r-s)$	$q(1-r-s)+(1-g)(1-p-q)(1-r-s)$	q
2	$(1-p-q)r$	$(1-p-q)s+(1-g)(1-p-q)(1-r-s)$	$g(1-p-q)(1-r-s)$	$1-p-q$
	r	s	$1-r-s$	1

Táboa 4.5: *Táboa de continxencia para* `rmult`.

4.2. Resultados do estudo de simulación

Dado que os datos que se desexan analizar corresponden a uns 10 000 SNP e 500 individuos de cada grupo (cf. 5.1), é evidente que os custos informáticos son tan altos que superan posibilidades de calquera ordenador doméstico e mesmo dalgunhas arquitecturas convencionais de HPC, que non dan ofrecido resultados nun tempo razoable (Yang *et al.*, 2015; Kam-Thong *et al.*, 2012). Ofrécense máis detalles sobre estas cuestión na § 3.4.

Tendo en conta este feito, e para facilitar a interpretación dos resultados, só se estudou o comportamento do método para un par de SNP simulado, impoñendo diversos modelos paramétricos baixo as hipóteses nula e alternativa, de entre os introducidos con anterioridade (por brevidade, só se inclúe unha fracción dos resultados obtidos nos ensaios realizados). Os resultados que a continuación se expoñen permitirán estudar a efectividade real do método porque se trata dun problema de contrastes múltiples e non dun só test en alta dimensión (véxase Cai [2017] para unha discusión sobre as diferencias metodolóxicas e conceptuais), no cal si que se faría uso de resultados asintóticos cando $p \to \infty$.

Nas Táboas 4.6a–4.6b, recóllese a calibración do nivel de significación para valores nominais usuais.

$\boldsymbol{\alpha}$	0'01	0'02	0'05	0'10
$\hat{\boldsymbol{\alpha}}$	0'01	0'02	0'04	0'07

a) `indep`.

$\boldsymbol{\alpha}$	0'01	0'02	0'05	0'10
$\hat{\boldsymbol{\alpha}}$	0'01	0'02	0'04	0'08

b) `rexp` *con* $e = 10$.

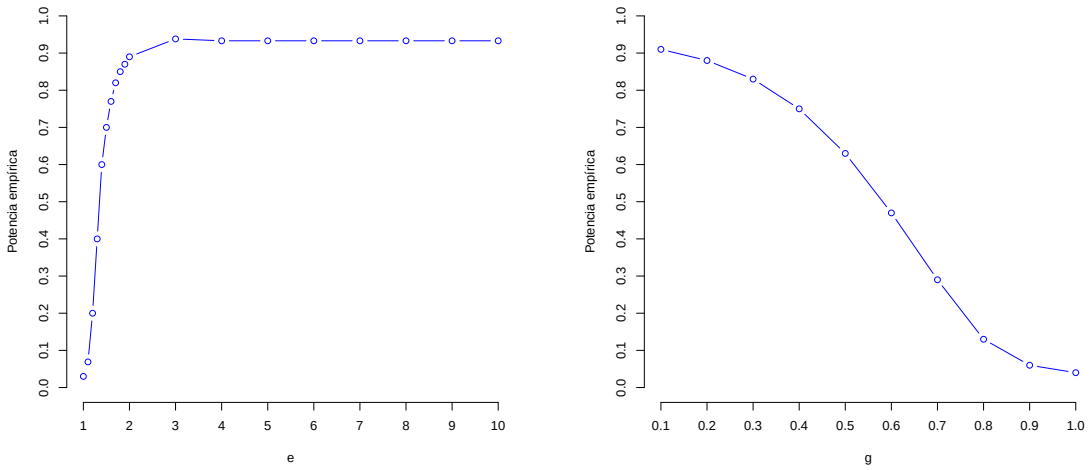
Táboa 4.6: *Nivel significación nominal (α) fronte a potencia empírica baixo a hipótese nula ($\hat{\alpha}$), para dous modelos distintos.*

Por outra parte, nas Figuras 4.1a–4.1b represéntase a potencia empírica do método proposto. En todos os casos realizáronse $R = 1000$ réplicas.

As táboas anteriores (xunto con outras similares, que se omitiron) permiten concluír

que a calibración do nivel de significación é aceptable ou incluso boa. Pola súa parte, as gráficas da Figura 4.1 mostran que a potencia é moi satisfactoria e que presenta o aumento desexable a medida que un se afasta da hipótese nula. Isto último non é sorprendente, xa que a única crítica da potencia do contraste de independencia mediante correlación de distancias da que hai constancia na bibliografía (Ramdas *et al.*, 2015) refírese unicamente aos tests simples en alta dimensión (e non aos múltiples a grande escala), por non mencionar que os achados deste artigo contrastan cos de moitos outros (véxase: Székely e Rizzo, 2017) e, polo tanto, deben tomarse con cautela.

En conclusión, tense un procedemento que, polo menos no plano teórico, non sofre as deficiencias de Cai e Liu (2016) expostas na § 1.3.5.



a) *Potencia empírica cando os modelos son* `qexp` *con parámetro* $e \in \mathbb{Z}^+$ *e* `indep`*.*

b) *Potencia empírica cando os modelos son* `qmult` *con parámetro* $g \in [0, 1]$ *e* `indep`*.*

Figura 4.1: *Potencia empírica cando os modelos son* `indep` *e outro, en función do parámetro que caracteriza este.*

Capítulo 5

Aplicación a un estudo caso-control de esquizofrenia

Comézase (§ 5.1) describindo as bases de datos xenómicos que motivaron o desenvolvemento das técnicas estatísticas presentadas en 3.2. Aínda que a xenotipación non formou parte deste traballo nin foi realizada polo seu autor, optouse por explicar os detalles experimentais para permitir a reproducibilidade dos resultados deste capítulo, os cales se presentan en 5.2 e se discuten en 5.3.

5.1. Bases de datos xenómicas

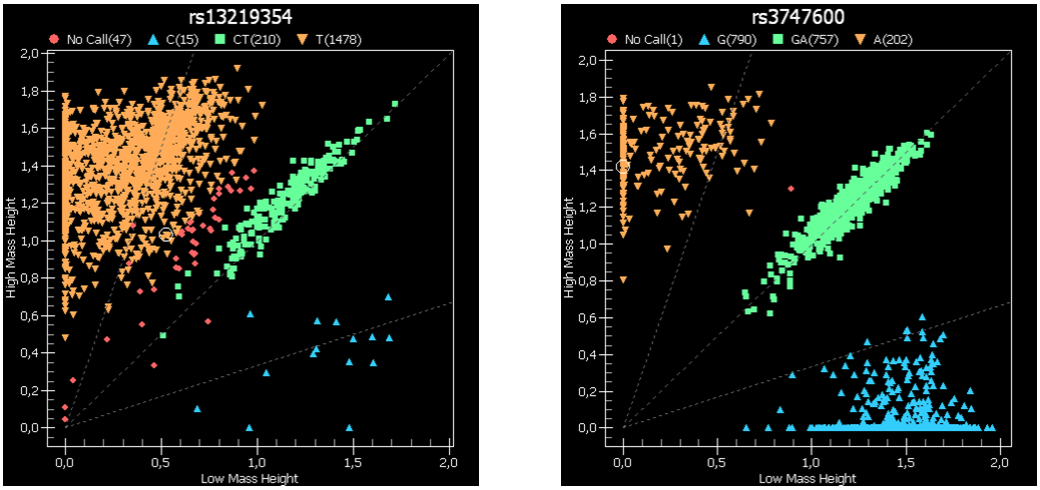
Disponse de dúas bases de datos de SNP, correspondentes a estudos caso-control de esquizofrenia e alcoholismo. En ambos os dous casos, xenotipáronse inicialmente os 588 628 SNP seleccionados polo Consorcio de Xenómica Psiquiátrica (PGC) no deseño do *microarray* denominado *PsychArray-24 v1.1 BeadChip*. Destes SNP, uns 50 000 foron elixidos pola súa presunta relación con trastornos psiquiátricos comúns (a miúdo con máis dun), baseándose en indicios bastante febles. Do resto, aproximadamente a metade son os «*tag* SNP» presentes no *HumanCore-24 BeadChip* e os demais son polimorfismos das rexións exónicas, tomados do *Exome-24 BeadChip*.

Tras a xenotipación inicial, realizáronse varios controis de calidade (QC) convencionais. En concreto, para evitar problemas experimentais, eliminároнse aqueles SNP

que verificasen algunha das seguintes condicións:

- A *frecuencia do alelo menor* (MAF) é inferior ao 1 %.
- As proporcións de xenotipos dos *controis* difiren significativamente das esperables baixo o *equilibrio de Hardy–Weinberg* (nos *casos* non sería inesperado).
- A *call rate* (proporción de datos non faltantes) está por debaixo do 95 %, ou ben é significativamente distinta entre casos e controis.

De non ter imposto as restricións anteriores, estaríanse mantendo polimorfismos «indesexables», nos que non é posible separar con claridade os centroides correspondentes aos dous alelos en cuestión (Figura 5.1a), fronte a unha gran maioría na que a secuenciación permite discriminar xenotipos con rotundidade (Figura 5.1b).



a) Call rate *baixa (confusión entre os xenotipos CT e TT).*

b) Call rate *próxima ao 100 % (boa discriminación entre GG, GA e AA).*

Figura 5.1: *Exemplos de centroides correspondentes a un SNP que non pasaría o QC (esquerda) e a outro que si (dereita).*

Finalmente, tamén se impón un filtro de calidade aos individuos eliminando aqueles para os que a secuenciación fallara en máis dun 5 % dos casos.

Dos SNP autosómicos remanentes tras o QC, escolléronse unicamente aqueles exónicos *missense*, é dicir, aqueles que fan que se pase de codificar un aminoácido a outro distinto (Figura 5.2), porque son os que con maior probabilidade poden presentar interacción entre si. Polo mesmo motivo, tamén se impuxo ás MAF un límite

inferior máis alto do estritamente necesario, xa que se cre que as enfermidades complexas (non mendelianas) están producidas pola combinación de múltiples variantes comúns de escaso efecto marxinal (Bush e Moore, 2012, hipótese CV/CD). Outro argumento para evitar MAF moi pequenas é que ocasionarían que, ante os tamaños mostrais relativamente pequenos que se manexan, apenas se observaría o alelo menor en homocigose (xurdindo así un claro problema de potencia).

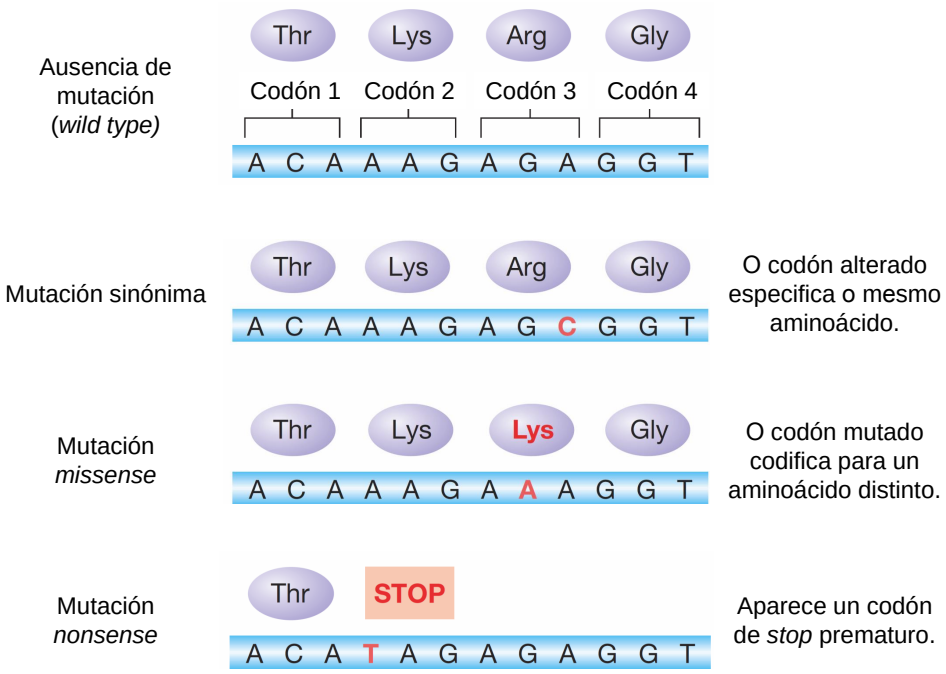


Figura 5.2: *Tipos de SNP codificantes Griffiths (figura adaptada de 2015, páx. 583).*

A primeira base de datos coa que se vai a traballar é a de esquizofrenia, que conta con $n_1 = 585$ casos e $n_2 = 573$ controis. O número de SNP considerados, $L = 4\,117$, é moito menor que o habitual en GWAS destas características porque a todos os polimorfismos considerados se lles esixe:

- MAF $> 10\,\%$. Isto permite reducir a dimensión o suficiente como para que o problema sexa informaticamente abordable nun tempo razoable e, ao mesmo tempo, non tanto como para que sexa imposible descubrir interaccións, dada a notable *sparsity* das matrices de adxacencia das redes epistáticas (Cai, 2017).
- Pertenza a algún dos xenes contidos nos 35 módulos de coexpresión espazo-temporal en cerebro que se coñecen (Fromer *et al.*, 2016). Isto permitirá avaliar

os resultados: nunha situación favorable, espérase que os pares de variantes entre os que se detecten interaccións significativas coincidan no mesmo módulo nunha proporción maior que a que se tería seleccionándoos ao azar. En tal caso, estaríanse «validando bioloxicamente» os achados e o deseño da metodoloxía.

Ademais, nunha segunda etapa, intentarase corroborar os descubrimentos máis destacados mediante buscas bibliográficas.

Por outra parte, a base de datos de alcoholismo conta cun tamaño e características moi similar, pero presenta a desvantaxe de referirse a unha enfermidade psiquiátrica moito menos estudada desde o punto de vista xenómico (Sullivan *et al.*, 2017), polo que dar sentido biolóxico aos resultados sería unha tarefa moito máis complicada. En consecuencia, optouse por estudar primeiro os datos reducidos de esquizofrenia e deixar o resto para o futuro (§ 6.2).

É importante salientar que, tal e como se require para o funcionamento de métodos como o de Cai e Liu (2016) ou o proposto en 3.2, a mostraxe de controis e casos é independente: os primeiros proceden de mostras tomadas hai anos no Centro de Transfusións de Galicia, mentres que os segundos proceden dun estudo moito máis dilatado no tempo levado a cabo no Hospital Clínico Universitario de Santiago (CHUS). Noutro caso (por exemplo, en estudos de familias), habería que adaptar a metodoloxía estatística á estrutura de dependencias correspondente.

5.2. Resultados da aplicación a bases de datos de SNP

Para a análise das bases xenómicas, os custos informáticos son moi elevados (cf. Táboa 3.1), de maneira que para estudar todos os posibles pares requírese un número de procesadores demasiado alto para as posibilidades do autor de este traballo. Ademais, como o obxectivo é confirmar a bondade do método proposto a través de criterios biolóxicos, optouse por reducir a base de datos de esquizofrenia de $L = 4117$ a $L = 1000$ SNP, o que supón unha gran diferencia porque o tempo de execución é $O(L^2)$. Para non introducir ningún nesgo, estes SNP seleccionáronse ao azar (de maneira equiprobable).

5.2.1. Estudo por módulos de coexpresión

Usáronse as catro distancias anteriormente definidas para examinar a base de datos de esquizofrenia, contabilizando a proporción das presuntas interaccións SNP-SNP detectadas nas que ambos polimorfismos coinciden nun dos 35 módulos de coexpresión espazo-temporal en cerebro descritos por Fromer *et al.* (2016).

Só cunha das distancias, a "0=1", a proporción obtida foi significativamente maior que a esperable se se elixiran ao azar (de maneira equiprobable): 0′085 fronte a 0′071, cun nivel crítico (p-valor) menor que calquera dos niveis de significación nominais usuais. Este feito é a xustificación de que no apartado seguinte só se use a métrica recesiva.

5.2.2. Aplicación dos resultados ao xene *SLC39A8*

Finalmente, a análise efectuada vaise particularizar para un SNP en concreto, podendo así obter resultados facilmente interpretables e que poidan ser contrastados acudindo á bibliografía especializada. O polimorfismo elixido é o de posición cromosómica *4:103188709*. A asociación deste SNP foi descuberta por Carrera *et al.* (2012) e confirmada por Ripke *et al.* (2014), Máis aínda, trátase do *missense* máis claramente asociado á esquizofrenia (Costas, 2018).

Empregando a distancia "0=1", 3 interaccións destacan por encima das demais (Táboa 5.1):

SNP_1 (módulo)	SNP_2 (módulo)	p-valor (controis)	p-valor (casos)
2:103149100 (4)	4:103188709 (21)	0'000	0'591
14:88407888 (21)	4:103188709 (21)	0'000	0'409
7:100466441 (16)	4:103188709 (21)	0'000	0'409

Táboa 5.1: *As tres interaccións máis significativas de 4:103188709 con outros SNP.*

Estes resultados interpretaranse de seguido, en 5.3.

5.3. Discusión da análise de SNP

Non é inesperado que a distancia "0=1" sexa a que exhiba un mellor comportamento na práctica: ao fin e ao cabo, é a que mellor permite detectar o efecto de aqueles SNP que seguen un modelo recesivo, é dicir, aqueles para os que o relevante a efectos fenotípicos é a presenza ou ausencia de 2 copias do alelo menor (sendo indiferente se hai unha o ningunha). Tal e como se indicou con anterioridade, é habitual asumir este modelo.

Por outra banda, o estudo particular dun SNP deu lugar á detección de 3 pares de polimorfismos destacables, un dos cales está constituído por *4:103188709* e *7:100466441*. Esta presunta interacción é moi plausible bioloxicamente. Por un lado, *7:100466441* está no xene *TRIP6*, que actúa (Willier *et al.*, 2011) como activador da ruta de sinalización (*signaling pathway*) de NF-κB, implicada en inmunidade. Pola súa parte, *4:103188709* encóntrase no xene *SLC39A8*, que é un inhibidor desa mesma ruta (Liu *et al.*, 2013). É moi pouco probable que esta coincidencia sexa casual, dado o elevado número de rutas existentes. Ademais, esta idea ven reforzada pola existencia de bastantes indicios de que unha resposta inmune anormal podería estar implicada en esquizofrenia (Costas, 2018).

Discutiuse, pois, como de verosímil é un dos descubrimentos. Porén, non pode dicirse o mesmo dos outros dous. É máis, a partir das pescudas bibliográficas realizadas, non se achou ningún indicio de que as presuntas relacións de *4:103188709* con *14:88407888* e *2:103149100* teñan algunha explicación molecular. En consecuencia, o máis razoable é consideralas falsos positivos.

Por outra parte, é lóxico preguntarse cal era a probabilidade *a priori* de que, escollendo un xene ao azar (de maneira equiprobable), sexa posible atopar argumentos para apoiar a súa presunta relación co *SLC39A8*. O pouco que se sabe deste xene indica que actúa na ruta NF-κB e na homeostase de ións metálicos. Nunha revisión exhaustiva de 2011 (White *et al.*, 2011), documentáronse 235 xenes asociados a NF-κB, polo que se pode aproximar burdamente que na actualidade coñeceranse uns 300. No que atinxe aos ións, o número de xenes probadamente relacionados é moito menor: uns 50. Tendo en conta o anterior, a probabilidade de que, dados 3 dos 1000 SNP considerados, algún teña unha interacción plausible con *4:103188709* é de aproximadamente o 4 %.

En consecuencia, aínda que inicialmente poida parecer un pouco decepcionante que só 1 de 3 decubrimentos sexa plausible, á vista desta probabilidade, en absoluto resulta insatisfactorio. Ademais, hai que ter en conta que apenas hai interaccións de orde 2 documentadas en trastornos complexos (aínda menos con datos de SNP) e, en particular, en esquizofrenia non se coñece ningunha a día de hoxe; o que, de

novo, dá idea da dificultade do problema de partida.

Por último, isto permite afirmar que se están solucionando (polo menos parcialmente) as deficiencias dos LCT de Cai e Liu (2016) expostas na Figura 1.1b. Iso si: esta conclusión debe tomarse con certa cautela, posto que sería preciso facer un estudo máis exhaustivo para ter maior seguridade.

Capítulo 6

Balance e conclusións

6.1. Resposta ás cuestións da introdución

Finalmente, estase en condicións de dar resposta ás preguntas que motivaron o traballo (§ 1.4):

1. Estudouse rigorosamente a correlación de distancias en espazos métricos e particularizouse esta análise teórica para espazos marxinais de cardinal 3.
2. O contraste de independencia proposto xeneraliza os LCT como ferramenta de detección de pares de variables aleatorias nos que a natureza da dependencia (non só linear) é distinta en casos e controis.
3. Ao ser válida calquera distancia, o procedemento proposto é extremadamente flexible: permite decidir, a priori, que tipo de interacción buscar. Por exemplo, de entre as métricas estudadas, só unha demostrou un comportamento claramente satisfactorio na práctica.
4. En base aos estudos de simulación realizados, pódese concluír que a calibración do nivel de significación é adecuada e a potencia, considerablemente alta ante diversas alternativas; mellorando os resultados que se obterían cos métodos de Cai e Liu (2016).
5. Estudáronse as interaccións máis significativas para os datos de esquizofrenia e obtívose unha asociación bioloxicamente plausible, o cal sería altamente improbable se o método non funcionase adecuadamente.

6.2. Conclusións e futuras liñas de traballo

O estudo realizado permitiu deseñar un contraste de hipóteses baseado na caracterización xeral da independencia que ofrece a correlación de distancias, estendendo así as ideas dos LCT (Cai e Liu, 2016) a datos ternarios. Este método presenta un comportamento satisfactorio na práctica; tal e como o ilustran, entre outros, os modelos simulados que aquí se presentaron. Ademais, unha das métricas demostrou a súa adecuación aos datos xenómicos considerados, a través da detección dunha proporción significativa de pares de SNP no mesmo módulo de coexpresión, así como do descubrimento dunha interacción dentro da mesma ruta de sinalización.

Con todo, quedan diversos problemas abertos, que foron xurdindo durante o desenvolvemento do traballo. Algunhas liñas de traballo actuais e futuras son:

- Ampliar a validación biolóxica dos resultados experimentais (§ 5.3) á base completa de 4117 SNP, para dar maior solidez ás conclusións obtidas.
- Analizar tamén os datos de alcoholismo do IDIS.
- Estudar bioloxicamente os resultados que do anterior se obteñan.
- Buscar melloras informáticas para reducir aínda máis os tempos de execución.
- A aceleración dos programas tamén permitirían ensaiar un maior número de distancias, e incluso «aprender» da mostra para achar a métrica óptima nalgún sentido.
- Publicar un paquete de *R* que recompile todo o *software* desenvolvido.
- Deseñar novos modelos poboacionais de cara a un estudo de simulación moito máis exhaustivo (e, polo tanto, concluínte) que o que aquí se expuxo.
- Adaptar o contido deste traballo a busca de interaccións entre xenoma mitocondrial (haploide) e nuclear (diploide), aplicándoo aos datos de autismo ofrecidos polo Grupo de Bases Neurobiolóxicas dos Trastornos Mentais do Instituto de Investigación Sanitaria Gregorio Marañón.
- Deducir a expresión simplificada da distribución asintótica da covarianza de distancias para espazos de tres puntos e tratar de achar estimadores razoables dos autovalores que nela aparezan.

Pódese atopar unha versión actualizada deste traballo en: `http://bit.ly/TFM_FCP`.

Bibliografía consultada

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. 3.ª edición. John Wiley & Sons. ISBN 0471360919. Citado na páxina ↑5.

Arcones, M. Á.; Giné, E. (1992). «On the bootstrap of U and V-statistics». *Annals of Statistics* 2: 655–674. DOI: 10.1214/aos/1176348650. Citado nas páxinas ↑28 e ↑32.

Bakirov, N. K.; Rizzo, M. L.; Székely, G. J. (2006). «A multivariate nonparametric test of independence». *Journal of Multivariate Analysis* 97: 1742 –1756. DOI: 10.1016/j.jmva.2005.10.005. Citado na páxina ↑9.

Benjamini , Y.; Hochberg, Y. (1995). «Controlling the false discovery rate: a practical and powerful approach to multiple testing». *Journal of the Royal Statistical Society (Series B)* 57, 289–300. JSTOR: 2346101. Citado nas páxinas ↑5 e ↑33.

Benjamini , Y.; Yekutieli, D. (2001). «The control of the false discovery rate in multiple testing under dependency». *Annals of Statistics* 29, 1165–1188. DOI: 10.1214/aos/1013699998. Citado nas páxinas ↑5 e ↑33.

Berg, C.; Christensen, J. P. R.; Ressel, P. (1984). *Harmonic analysis on semigroups*. 1.ª edición. Springer. ISBN 0387909257. Citado na páxina ↑21.

Biernacka, J. M.; Geske, J.; Jenkins, G. D.; Colby, C.; Rider, D. N.; Karpyak, V. M.; Choi, D.-S.; Fridley, B. L. (2013). «Genome-wide gene-set analysis for identification of pathways associated with alcohol dependence». *International Journal of Neuropsychopharmacology* 16: 271–278. DOI: 10.1017/S1461145712000375. Citado na páxina ↑1.

Billingsley, P. (1995). *Probability and measure*. 3.ª edición. John Wiley & Sons. ISBN 0471007102. Citado na páxina ↑13.

Bishop, Y. M. M.; Fienberg, S. E.; Holland, P. W. (1975). *Discrete multivariate analysis: theory and practice*. MIT Press. ISBN 0262520400 . Citado na páxina ↑20.

Bogachev, V. I. (2007). *Measure theory* (volumes 1–2). 1.ª edición. Springer. ISBN 3540345138. Citado nas páxinas ↑12 e ↑14.

Bush , W. S.; Moore, J. H. (2012). «Genome-wide association studies». *PLoS Computational Biology* 8: *e1002822*. DOI: 10.1371/journal.pcbi.1002822. Citado nas páxinas ↑30 e ↑43.

Cai, T, T. (2017). «Global testing and large-scale multiple testing for high-dimensional covariance structures». *Annual Review of Statistics and Its Application* 4: 1–24. DOI: 10.1146/annurev-statistics-060116-053754. Citado nas páxinas ↑4, ↑38 e ↑43.

Cai, T. T.; Liu, W. (2016). «Large-scale multiple testing of correlations». *Journal of the American Statistical Association* 111: 229–240. DOI: 10.1080/01621459.2014.999157. Citado nas páxinas ↑4, ↑6, ↑7, ↑8, ↑29, ↑30, ↑32, ↑39, ↑44, ↑47, ↑49 e ↑50.

Camacho, D.; De la Fuente, A.; Mendes, P.; (2005). «The origin of correlations in metabolomics data». *Metabolomics* 1: 53–63. DOI: 10.1007/s11306-005-1107-3. Citado na páxina ↑4.

Carrera, N. *et al.* (2012). «Association study of nonsynonymous single nucleotide polymorphisms in schizophrenia». *Biological Psychiatry* 71: 169–177. DOI: 10.1016/j.biopsych.2011.09.032. Citado na páxina ↑45.

Costas, J. (2018). «The highly pleiotropic gene *SLC39A8* as an opportunity to gain insight into the molecular pathogenesis of schizophrenia». *American Journal of Medical Genetics (Part B)* 177: 274–283. DOI: 10.1002/ajmg.b.32545. Citado nas páxinas ↑45 e ↑46.

Costas, J. *et al.* (2011). «Heterozygosity at catechol-O-methyltransferase Val158Met and schizophrenia: New data and meta-analysis». *Journal of Psychiatric Research* 45: 7–14. DOI: 10.1016/j.jpsychires.2010.04.021. Citado na páxina ↑30.

Davis, R. A.; Matsui, M.; Mikosch, T.; Wan, P. (2018). «Applications of distance correlation to time series». *Bernoulli* 24: 3087–3116. DOI: 10.3150/17-BEJ955. Citado na páxina ↑29.

De la Fuente, A. (2010). «From differential expression to differential networking identification of dysfunctional regulatory networks in diseases». *Trends in Genetics* 26: 326–333. DOI: 10.1016/j.tig.2010.05.001. Citado nas páxinas ↑4 e ↑32.

Dehling, H.; Matsui, M.; Mikosch, T.; Samorodnitsky, G.; Tafakori, L. (2018). «Distance covariance for discretized stochastic processes». *Universidade de Copenhague*. ArXiv: 1806.09369. Citado na páxina ↑29.

D'Haeseleer, P.; Liang, S.; Somogyi, R.; (2000). «Genetic network inference: from co-expression clustering to reverse engineering». *Bioinformatics* 16: 707–726. DOI: 10.1093/bioinformatics/16.8.707. Citado na páxina ↑4.

Devor, E. J.; Cloninger, C. R. (1989). «Genetics of alcoholism». *Annual Review of Genetics* 23, 19–36. DOI: 10.1146/annurev.ge.23.120189.000315. Citado na páxina ↑2.

Deza, M. M.; Laurent, M. (1997). *Geometry of cuts and metrics*. 1.ª edición. Springer. ISBN 3642042942. Citado na páxina ↑20.

Emily, M. (2012). «IndOR: a new statistical procedure to test for SNP-SNP epistasis in genome-wide association studies». *Statistics in Medicine* 31: 2359-2373. DOI: 10.1002/sim.5364. Citado na páxina ↑4.

Fromer, M. *et al.* (2016). «Gene expression elucidates functional impact of polygenic risk for schizophrenia». *Nature Neuroscience* 19: 1442–1453. DOI: 10.1038/nn.4399. Citado nas páxinas ↑43 e ↑45.

Giné, E.; Zinn, J. (1992). «Marcinkiewicz type laws of large numbers an dconvergence of moments for $U-$statistics», capítulo de *Probability in Banach Spaces, 8: Proceedings of the Eighth International Conference* (páxs. 273–291). Springer. ISBN 1461267287. Citado na páxina ↑27.

Goldman, D.; Groszi, G.; Ducci, F.; (2005). «The genetics of addictions: uncovering the genes». *Nature Reviews Genetics* 6: 521–530. DOI: 10.1038/nrg1635. Citado na páxina ↑1.

Goudey, B. *et al.* (2013). «GWIS —Model-free, fast and exhaustive search for epistatic interactions in case-control GWAS». *BMC Genomics* 14: S10. DOI: 10.1186/1471-2164-14-S3-S10. Citado na páxina ↑4.

Griffiths, A. J. F. (2015). *Introduction to genetic analysis*. 11.ª edición. Freeman & Co. ISBN 1464109486. Citado na páxina ↑43.

Gropp, W.; Lusk, E.; Skjellum, A. (2014). *Using MPI: Portable parallel programming with the message-passing interface*. 3.ª edición. MIT Press. ISBN 0262527392. Citado na páxina ↑34.

Gusareva, E. S.; Van Steen, K. (2014). «Practical aspects of genome-wide association interaction analysis». *Human Genetics* 133, 1343–1358. DOI: 10.1007/s00439-014-1480-y. Citado nas páxinas ↑2 e ↑3.

Gyenesei, A.; Moody, J.; Semple, C. A. M.; Haley, C. S.; Wei, W.-H. (2012). «High-throughput analysis of epistasis in genome-wide association studies with BiForce». *Bioinformatics* 28: 1957-1964. DOI: 10.1093/bioinformatics/bts304. Citado na páxina ↑4.

Hardy, G. H. (1908). «Mendelian proportions in a mixed population». *Science* 28, 49–50. DOI: 10.1126/science.28.706.49. Citado na páxina ↑35.

Hemani, G.; Theocharidis, A.; Wei, W.-H.; Haley, C. S. (2011). «EpiGPU: Exhaustive pairwise epistasis scans parallelized on consumer level graphics cards». *Bioinformatics* 27: 1462-1465. DOI: 10.1093/bioinformatics/btr172. Citado na páxina ↑4.

Henry, V. J.; Bandrowski, A. E.; Pepin, A.-S.; González, B. J.; Desfeux, A.; (2014). «OMICtools: an informative directory for multi-omic data». *Database* 2014: artigo *bau069*. DOI: 10.1093/database/bau069. Citado na páxina ↑3.

Hoeffding, W. (1961). «The strong law of large numbers for $u-$statistics». *Institute of Statistics Mimeo Series* 302. URL: https://repository.lib.ncsu.edu/handle/1840.4/2128. Citado na páxina ↑27.

Huo, X.; Székely, G. J. (2015). «Fast computing for distance covariance». *Technometrics* 58: 435-447. DOI: 10.1080/00401706.2015.1054435. Citado na páxina ↑34.

Jakobsen, M. E. (2017). *Distance covariance in metric spaces: Non-parametric independence testing in metric spaces*. Universidade de Copenhague. ArXiv: 1706.03490. Citado nas páxinas ↑9, ↑12, ↑15, ↑18, ↑21, ↑22, ↑23, ↑24, ↑26, ↑27 e ↑32.

Kam-Thong, T.; Czamara, D.; Tsuda, K.; Borgwardt, K.; Lewis, C. M.; Erhardt-Lehmann, A.; Hemmer, B.; Rieckmann, P.; Daake, M.; Weber, F.; Wolf, C.; Ziegler, A.; Pütz, B.; Holsboer, F.; Schölkopf, B.; Müller-Myhsok, B. (2011). «EPIBLASTER —fast exhaustive two-locus epistasis detection strategy using graphical processing units». *European Journal of Human Genetics* 19 : 465–471. DOI: 10.1038/ejhg.2010.196. Citado nas páxinas ↑4 e ↑5.

Kam-Thong, T.; Azencott, C.-A.; Cayton, L.; Pütz, B.; Altmann, A.; Karbalai, N.; Sämann, P. G.; Schölkopf, B.; Müller-Myhsok, B.; Borgwardt, K. M. (2012). «GLIDE: GPU-Based linear regression for detection of epistasis». *Human Heredity* 73: 220–236. DOI: 10.1159/000341885. Citado nas páxinas ↑4 e ↑38.

Klebanov, L. B. (2005). *$\mathfrak{N}$-distances and their applications*. The Karolinum Press. ISBN 802461152X. Citado nas páxinas ↑21 e ↑30.

Liu, M.-J.; Bao, S.; Gálvez-Peralta, M.; Pyle, C. J.; Rudawsky, A. C.; Pavlovicz, R. E.; Killilea, D. W.; Li, C.; Nebert, D. W.; Wewers, M. D.; Knell, D. L. (2013). «ZIP8 regulates host defense through zinc-mediated inhibition of NF-κB». *Cell Reports* 3 : 386–400. DOI: 10.1016/j.celrep.2013.01.009. Citado na páxina ↑46.

Lyons, R. (2013). «Distance covariance in metric spaces». *Annals of Probability* 41: 3284–3305. DOI: 10.1214/12-AOP803. Citado nas páxinas ↑9, ↑13, ↑14, ↑15, ↑17, ↑21, ↑22, ↑25, ↑26, ↑27, ↑28, ↑30, ↑32 e ↑34.

Lyons, R. (2018). «Errata to "Distance covariance in metric spaces"». *Annals of Probability* 46: 2400–2405. DOI: 10.1214/17-AOP1233. Citado nas páxinas ↑9, ↑10, ↑13, ↑14, ↑27, ↑28 e ↑29.

Manolio, T. A. *et al.* (2009). «Finding the missing heritability of complex diseases». *Nature* 461: 747–753. DOI: 10.1038/nature08494. Citado na páxina ↑2.

Marchini, J.; Donnelly, P.; Cardon, L. R. (2005). «Genome-wide strategies for detecting multiple loci that influence complex diseases». *Nature Genetics* 37: 413–417. DOI: 10.1038/ng1537. Citado na páxina ↑36.

Moore, J. H.; Hill, D. P. (2015). «Epistasis analysis using artificial intelligence». *Methods in Molecular Biology* 1253: 327–346. DOI: 10.1007/978-1-4939-2155-3_18. Citado na páxina ↑4.

Naor, A. (2010). «L_1 embeddings of the Heisenberg group and fast estimation of graph isoperimetry». *Proceedings of the International Congress of Mathematicians* 3: 1549–1575 ArXiv: 1003.4261. Citado na páxina ↑20.

Niel, C.; Sinoquet, C.; Dina, C.; Rocheleau, G. (2015). «A survey about methods dedicated to epistasis detection». *Frontiers in Genetics* 6: artigo 285. DOI: 10.3389/fgene.2015.00285. Citado na páxina ↑3.

Pettis, B. J. (1938). «On integration in vector spaces». *Transactions of the American Mathematical Society* 44: 277–304. DOI: 10.1090/S0002-9947-1938-1501970-8. Citado na páxina ↑21.

Purcell, S. M. *et al.* (2009). «Common polygenic variation contributes to risk of schizophrenia and bipolar disorder». *Nature* 460: 748–752. DOI: 10.1038/nature08185. Citado na páxina ↑2.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Versión 3.5.1. R Foundation for Statistical Computing, 2018. ISBN 3900051070. Citado na páxina ↑33.

Ramdas, A.; Reddi, S. J.; Póczos, B.; Singh, A.; Wasserman, L. (2015). «On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions». *Proceedings of the 29th AAAI Conference on Artificial Intelligence*: 3571–3577. ArXiv: 1406.2083. Citado na páxina ↑39.

Ripke, S. *et al.* (2014). «Biological insights from 108 schizophrenia-associated genetic loci». *Nature* 511: 421–437. DOI: 10.1038/nature13595. Citado na páxina ↑45.

Ritchie, M. D.; Hahn, L. W.; Roodi, N.; Bailey, L. R.; Dupont, W. D.; Parl, F. F.; Moore, J. H. (2001). «Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer». *American Journal of Human Genetics* 69: 138-147. DOI: 10.1086/321276. Citado na páxina ↑4.

Rudin, W. (1921). *Real and complex analysis*. 3.ª edición. McGraw-Hill. ISBN 0071002766. Citado na páxina ↑13.

Schechter, E. (1996). *Handbook of analysis and its foundations*. 1.ª edición. Academic Press. ISBN 0126227608. Citado na páxina ↑12.

Schoenberg, I. J. (1938). «Metric spaces and positive definite functions». *Transactions of the American Mathematical Society* 44, 522–536. DOI: 10.1090/S0002-9947-1938-1501980-0. Citado na páxina ↑22.

Schoenberg, I. J. (1937). «On certain metric spaces arising from euclidean spaces by a change of metric and their imbedding in Hilbert space». *Annals of Mathematics (Second Series)* 38: 787–793. DOI: 10.2307/1968835. Citado na páxina ↑22.

Schwarz, D. F; König, I. R.; Ziegler, A. (2010). «On safari to random jungle: A fast implementation of random forests for high-dimensional data». *Bioinformatics* 26: 1752–1758. DOI: 10.1093/bioinformatics/btq257. Citado na páxina ↑4.

Šidák , Z. (1967). «Rectangular confidence regions for the means of multivariate normal distributions». *Journal of the American Statistical Association* 62, 626–633. DOI: 10.1080/01621459.1967.10482935. Citado na páxina ↑31.

Singh, D.; Febbo, P. G.; Ross, K.; Jackson, D. G.; Manola, J.; Ladd, C.; Tamayo, P.; Renshaw, A. A.; D'Amico, A. V.; Richie, J. P.; Lander, E. S.; Loda, M.; Kantoff, P. W.; Golub, T. R.; Sellers, W. R. (2002). «Gene expression correlates of clinical prostate cancer behavior». *Cancer Cell* 1: 203–209. DOI: 10.1016/S1535-6108(02)00030-2. Citado na páxina ↑7.

Sullivan, P. F. *et al.* (2017). «Psychiatric genomics: An update and an agenda». *American Journal of Psychiatry* 175: 15–27. DOI: 10.1176/appi.ajp.2017.17030283. Citado nas páxinas ↑1, ↑2 e ↑44.

Sun, Y.; Shang, J.; Liu, J.-X.; Li, S.; Zheng, C.-H. (2017). «EpiACO – a method for identifying epistasis based on ant colony optimization algorithm». *BioData Mining* 10. DOI: 10.1186/s13040-017-0143-7. Citado na páxina ↑4.

Székely, G. J.; Rizzo, M. L. (2009). «Brownian distance covariance». *Annals of Applied Statistics* 4: 1236–1265. DOI: 10.1214/09-AOAS312. Citado nas páxinas ↑9 e ↑25.

Székely, G. J.; Rizzo, M. L. (2010). «DISCO analysis: a nonparametric extensión of analysis of variance». *Annals of Applied Statistics* 2: 1034–1055. DOI: 10.1080/01621459.2014.999157. Citado na páxina ↑9.

Székely, G. J.; Rizzo, M. L. (2012). «On the uniqueness of distance covariance». *Statistics and Probability Letters* 82: 2278–2282. DOI: 10.1016/j.spl.2012.08.007. Citado na páxina ↑10.

Székely , G. J.; Rizzo, M. L. (2013). «The distance correlation t-test of independence in high dimension». *Journal of Multivariate Analysis* 117, 193–213. DOI: 10.1016/j.jmva.2013.02.012. Citado nas páxinas ↑9 e ↑32.

Székely, G. J.; Rizzo, M. L. (2017). «The energy of data». *Annual Review of Statistics and Its Application* 4: 447–479. DOI: 10.1146/annurev-statistics-060116-054026 . Citado nas páxinas ↑9, ↑20, ↑21, ↑32, ↑34 e ↑39.

Székely, G.; Rizzo, M.; Bakirov, N. (2007). «Measuring and testing dependence by correlation of distances». *Annals of Statistics* 35: 2769–2794. DOI: 10.1214/009053607000000505. Citado nas páxinas ↑9, ↑10, ↑25, ↑28, ↑32 e ↑34.

Wan, X.; Yang, C.; Yang, Q.; Xue, H.; Fan, X.; Tang, N. L. S.; Yu, W. (2010). «BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies». *American Journal of Human Genetics* 87: 325-340. DOI: 10.1016/j.ajhg.2010.07.021. Citado na páxina ↑4.

Weinberg, W. (1908). «Über den Nachweis der Vererbung beim Menschen». *Jahreshefte des Vereins für vaterländische Naturkunde in Württemberg* 64, 368–382. Citado na páxina ↑35.

White, K. L.; Rider, D, N.; Kalli, K. R.; Knutson, K. L.; Jarvik, G. P.; Goode, E. L. (2011). «Genomics of the NF-κB signaling pathway: hypothesized role in ovarian cancer». *Cancer Causes and Control* 22: 785–801. DOI: 10.1007/s10552-011-9745-4. Citado na páxina ↑46.

Willier, S.; Butt, E.; Richter, G. H. S.; Burdach, S.; Grunewald, T. G. P. (2011). «Defining the role of TRIP6 in cell physiology and cancer». *Biology of the Cell* 102: 573–591. DOI: 10.1042/BC20110077. Citado na páxina ↑46.

Wilson , W. A. (1935). «On certain types of continuous transformations of metric spaces». *American Journal of Mathematics* 57: 62–68. DOI: 10.2307/2372019. Citado na páxina ↑20.

Yang, G.; Jiang, W.; Yang, Q.; Yu, W. (2015). «PBOOST: A GPU-based tool for parallel permutation tests in genome-wide association studies». *Bioinformatics* 31: 1460–1462. DOI: 10.1093/bioinformatics/btu840. Citado na páxina ↑38.

Zhang, Y.; Liu, J. S. (2007). «Bayesian inference of epistatic interactions in case-control studies». *Nature Genetics* 39: 1167–1173. DOI: 10.1038/ng2110. Citado na páxina ↑4.