Távolságkorreláció

Szakdolgozat

Készítette:

Fegyverneki Tamás

Alkalmazott matematikus MSc

Sztochasztika szakirány

Témavezető:

Székely Gábor

programigazgató, National Science Foundation

tudományos kutató, MTA Rényi Alfréd Kutatóintézet

Eötvös Loránd Tudományegyetem

Természettudományi Kar

2012

Tartalomjegyzék

1. fejezet

Bevezetés

Véletlen mennyiségek egymásra gyakorolt hatása szükségszerűen vizsgálandó dolog a hétköznapi életben, legyen akár szó orvosi diagnosztikai modellezésről, ahol tünetegyüttesek előfordulására következtethetünk a feltárt struktúrákból, vagy gazdasági folyamatok egymásra gyakorolt hatásáról, ahol köznapi szinten érezhető változásokat eredményez egy jobban megismert kapcsolat két vizsgált adatsor között. Ennek a feladatnak megoldására számos jó és kevésbé jó eszköz áll rendelkezésünkre, melyek előnyökkel és hátrányokkal is egyaránt rendelkeznek. Sok klasszikus módszer és vizsgálat alapul a klasszikus Pearson-féle korreláción, ami a leggyakrabban használt összefüggőségi mérőszám, nyilvánvaló hátrányai ellenére, talán épp azért, mert az alkalmas alternatívákkal szemben könnyen alkalmazható és jól számolható. A több dimenzióban használt klasszikusnak tekinthető módszerek esetében is sokszor csak igen speciális esetekben tudunk jó tulajdonságokról beszélni.

A Székely Gábor nevéhez köthető távolságkorreláció áthidal egy igen fontos problémát, a függetlenség jó karakterizálását. Míg például a klasszikus korreláció esetében közismert, hogy 0 értéket vesz fel akár erősen összefüggő változók esetében is, hiszen csak lineáris kapcsolatot mér, a távolságkorreláció mindenféle kapcsolatot leír, és ezáltal alkalmas eszközt ad a függetlenség jó kezelésére és az összefüggőség mérésére. Ezen kívül gyakorlati szempontból szinte mindenféle függetlenséget használó problémára átfogó eszközt ad olyan értelemben, hogy bármilyen dimenzióban definiálható mérőszámról van szó, amely alkalmazhatóságát nem korlátozza adott esetben a megfigyelések száma sem, ellentétben néhány klasszikus eszközzel.

A továbbiakban Székely, Rizzo és Bakirov[23] nyomán először áttekintjük a távolságkorreláció alapötletét, azaz általánosabban a távolságon alapuló összefüggőségi mérőszámokat, melyek lényege, hogy valószínűségi változók együttes karakterisztikus függvényének és marginális karakterisztikus függvényeik szorzatának valamilyen értelemben vett távolságát mérik, így használva ki, hogy ez a távolság pontosan a függetlenség esetén lesz 0. Az általánosabb kép után definiáljuk a távolságkovarianciát és -korrelációt, speciálisan megválasztva a metrikát. Ezek után a gyakorlati alkalmazásokhoz definiáljuk a tapasztalati megfelelőit a konstruált mérőszámoknak. A tapasztalati

és elméleti távolságkorreláció praktikus előnye, hogy struktúrájában teljesen analóg a klasszikus pearsoni korrelációval, így könnyen áttekinthető néhány, szintén az előzővel analóg tulajdonságuk. Ezek után igazoljuk a tapasztalati távolságkorreláció definíciójának jogosságát és néhány tulajdonságát, ami szintén a gyakorlati alkalmazhatóságát igazolja.

A következő fejezetben rátérünk a függetlenség jó karakterizációja által indokolt statisztikai eszköztár felépítésére, áttekintve a tapasztalati távolságkovariancia jó tulajdonságait, például gyenge konvergenciáját egy normálisokból álló kvadratikus alakhoz, majd ezek segítségével próbát konstruálunk a függetlenség tesztelésére. További tulajdonságokat tekintünk át a próba konzisztenciájáról és elérhető szignifikanciaszintjéről. Ezek után a szórásanalízis egy távolságalapon történő alternatíváját vizsgáljuk meg, valamint röviden kitérünk a definiált eszközök torzítására.

Ezt követően további alternatívákat és általánosítási lehetőségeket tekintünk át a távolságkorrelációra Bakirov[1] alapján, megváltoztatva a normát és így a távolságot, amellyel a távolságkorreláció definíciójához jutottunk, majd Székely és Rizzo[20] alapján tekintve egy folyamatalapú megközelítést a függetlenségre, majd látjuk, hogy véges szórású esetben ez a más megközelítés pontosan visszaadja a távolságkorrelációt. Ezek után tovább általánosítjuk a definiáltakat, megadva az összefüggőségi mérőszámok egy paraméteres osztályát, téve ezt úgy, hogy a távolságkoncepció paraméteres alakjából nem csak a távolságkorrelációt eredményező speciális esetet tekintjük, majd kiterjesztjük euklideszi terekről tetszőleges véges dimenziós normált terekre, majd pedig Hilbert-terekre a mérőszámot Kosorok[9] nyomán, tapasztalva, hogy általános esetben Hilbert-tereken sajnos nem tudjuk a függetlenséget úgy karakterizálni, mint eredetileg. Ezek után áttekintünk még röviden pár kiterjesztési vagy javítási lehetőséget, legvégül pedig röviden összehasonlítjuk Rényi Alfréd[16] összefüggőségi mérőszámokra adott követelményeivel a kapottakat, hangsúlyt fektetve a normális eloszlások esetére.

2. fejezet

Távolságkorreláció

2.1. Távolság alapú összefüggőségi mérőszámok

2.1.1. Jelölések és definíciók

Olyan mérőszámot szeretnénk definiálni, amely esetén a klasszikus Pearson-féle korrelációval ellentétben a 0 értékből következtethetünk a függetlenségre, ehhez az első lépés a távolság alapú összefüggőségi mérőszámok definiálása. Ehhez tekintsük először a következőket:

A továbbiakban legyenek p és q pozitív egészek mellett $X \in \mathbb{R}^p$ és $Y \in \mathbb{R}^q$ valószínűségi vektorváltozók, továbbá ezen változók karakterisztikus függvényeit és együttes karakterisztikus függvényüket jelölje rendre f_X, f_Y és $f_{X,Y}$! Az u és v vektorok skaláris szorzatát jelölje $\langle u, v\rangle$! A komplex értékű g függvényre legyen $|g|^2 = g\overline{g}$, ahol $\overline{g}$ a g függvény komplex konjugáltját jelöli! Az $\mathbb{R}^p$ téren az euklideszi normát jelölje most $|x|_p$, ahol $x \in \mathbb{R}^p$!

2.1. definíció. Tetszőleges γ $\mathbb{R}^p \times \mathbb{R}^q$ téren értelmezett komplex függvényre legyen az $\mathbb{R}^{p+q}$ térbeli függvények súlyozott L_2 terén a $\|\cdot\|_w$ norma a következő:

$$\|\gamma(u,v)\|_w^2 = \int_{\mathbb{R}^{p+q}} |\gamma(u,v)|^2 w(u,v)dudv \tag{2.1}$$

ahol $w(u,v)$ egy tetszőleges pozitív súlyfüggvény, amelyre létezik az integrál.

Ennek a normának a segítségével definiálunk most egy összefüggőségi mérőszámot:

2.2. definíció. Legyen olyan w súlyfüggvény mellett, ahol a megfelelő integrál létezik

$$\mathcal{V}^2(X,Y;w) = \|f_{X,Y}(u,v) - f_X(u)f_Y(v)\|_w^2 = \tag{2.2}$$

$$= \int_{\mathbb{R}^{p+q}} |f_{X,Y}(u,v) - f_X(u)f_Y(v)|^2 w(u,v)dudv, \tag{2.3}$$

és teljesen hasonlóan legyen

$$\mathcal{V}^2(X;w) = \mathcal{V}^2(X,X;w) = \int_{\mathbb{R}^{2p}} |f_{X,X}(u,v) - f_X(u)f_X(v)|^2 w(u,v)dudv \tag{2.4}$$

Analóg módon a klasszikus korrelációhoz, a fenti definíció alapján definiálhatunk egy $\mathcal{R}_w$ távolság alapú korrelációt az

$$\mathcal{R}_w(X,Y) = \frac{\mathcal{V}(X,Y;w)}{\sqrt{\mathcal{V}(X;w)\mathcal{V}(Y;w)}} \tag{2.5}$$

alakban. Látható a definícióból, hogy $\mathcal{V}^2(X,Y;w) = 0$ akkor és csak akkor, ha X és Y függetlenek.

2.1.2. A súlyfüggvény megválasztása

A 2.2 definícióval eljutottunk egy olyan mérőszámig, ami láthatóan jól karakterizálja a függetlenséget, de felmerül vele a kérdés, hogyan válasszuk meg a w súlyfüggvényt. Az $\mathcal{R}_w$ mérőszámról szeretnénk, ha skálainvariáns lenne abban az értelemben, hogy tetszőleges pozitív ε szám mellett (X,Y) és $(\varepsilon X, \varepsilon Y)$ valószínűségiváltozó-párokra ne változzon az értéke. Mivel továbbra is az egyik cél a függetlenség jó karakterizálása, ésszerű követelmény, hogy $\mathcal{R}_w$ legyen pozitív összefüggő valószínűségi változók esetén. További szempontokat is áttekintünk a későbbiekben Rényi Alfréd[16] alapján.

Belátható véges szórású X és Y mellett a karakterisztikus függvények Taylor-sorfejtéséből, hogy integrálható w súlyfüggvény mellett

$$\lim_{\varepsilon\to 0} \frac{\mathcal{V}^2(\varepsilon X, \varepsilon Y; w)}{\sqrt{\mathcal{V}^2(\varepsilon X; w)\mathcal{V}^2(\varepsilon Y; w)}} = \rho^2(X,Y), \tag{2.6}$$

ahol ρ a klasszikus Pearson-féle korreláció, tehát integrálható súlyfüggvény és összefüggő X és Y mellett is tetszőlegesen közel lehet nullához $\mathcal{R}_w(X,Y)$ értéke.

Egy Székely, Rizzo és Bakirov[23] által definiált súlyfüggvény segítségével azonban elérhető a fent említett skálainvariancia tetszőleges esetben és a pozitivitás összefüggő esetekben. Ehhez először tekintsük a következő lemmát [21] alapján:

2.3. lemma. $0 < \alpha < 2$ *mellett* $\forall x \in \mathbb{R}^d$

$$\int_{\mathbb{R}^d} \frac{1-\cos\langle u, x\rangle}{|u|_d^{d+\alpha}} du = C(d,\alpha)|x|^\alpha, \tag{2.7}$$

ahol

$$C(d,\alpha) = \frac{2\pi^{\frac{d}{2}}\Gamma(1-\alpha/2)}{\alpha 2^\alpha \Gamma((d+\alpha)/2)}, \tag{2.8}$$

ahol $\Gamma(\cdot)$ *a gamma-függvény.*
Az integrálok a 0*-an és* ∞*-ben a*

$$\lim_{\varepsilon\to 0} \int_{\mathbb{R}^d \setminus \{\varepsilon B + \varepsilon^{-1} B^c\}} \tag{2.9}$$

értelemben tekintendőek, ahol B *az* $\mathbb{R}^d$*-beli origó középpontú egységgömb,* B^c *pedig a komplementere.*

Bizonyítás. Legyen

$$A = \int_{\mathbb{R}^{d-1}} \frac{dz_2 \dots dz_d}{(1 + z_2^2 + z_3^2 + \dots + z_d^2)^{\frac{d+\alpha}{2}}} \tag{2.10}$$

Ekkor [13] alapján

$$\int_{x^2 \le r^2} f(x^2)dx = \frac{2\pi^{n/2}}{\Gamma(\frac{n}{2})} \int_0^r t^{n-1} f(t^2)dt, \tag{2.11}$$

ahol $x^2 \le r^2$ valójában az $x_1^2 + x_2^2 + \dots + x_n^2 \le r^2$ tartományt jelöli,

$$\int_0^\infty \frac{x^\alpha - 1}{(x+z)^r} dx = z^{\alpha-p} B(\alpha, p-\alpha), \tag{2.12}$$

ahol $|argz| < \pi$ és $0 < Re(\alpha) < Re(p)$ és $B(a,b) = \frac{\Gamma(a)+\Gamma(b)}{\Gamma(a+b)}$ a béta-függvény, továbbá

$$\int_0^\infty x^{\alpha-1} \sin^n bx dx = I_{\alpha,n}, \tag{2.13}$$

ahol $I_{\alpha,n}$ definícióját lásd [13] 2.5.3.13., vagy ennek speciális esetével,

$$\int_0^\infty x^{\alpha-1} \sin^2 bx dx = -\frac{\Gamma(\alpha)\cos\frac{\alpha\pi}{2}}{2^{\alpha+1}b^\alpha} \tag{2.14}$$

a megfelelő feltételek mellett, lásd [6] 319.15.
Ezekkel

$$A = \frac{2\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d-1}{2})} \int_0^\infty \frac{x^{d-2}dx}{(1+x^2)^{\frac{d+\alpha}{2}}} = \frac{2\pi^{\frac{d-1}{2}}\Gamma(\frac{\alpha+1}{2})}{\Gamma(\frac{d+\alpha}{2})}, \tag{2.15}$$

továbbá

$$\frac{d}{da}\left(\int_0^\infty \frac{1-\cos ax}{x^{\alpha+1}} dx \right) = a^{\alpha-1} \int_0^\infty \frac{\sin x}{x^\alpha} dx = \tag{2.16}$$

$$= a^{\alpha-1} \frac{\sqrt{\pi}\Gamma(1-\alpha/2)}{2^\alpha \Gamma(\frac{\alpha+1}{2})}, \tag{2.17}$$

ezekkel pedig

$$C(d,\alpha) = A \cdot \int_{-\infty}^\infty \frac{1-\cos z_1}{|z_1|^{\alpha+1}} dz_1 = \tag{2.18}$$

$$\frac{2\pi^{\frac{d-1}{2}}\Gamma(\frac{\alpha+1}{2})}{\Gamma(\frac{d+\alpha}{2})} \cdot \frac{2\sqrt{\pi}\Gamma(1-\alpha/2)}{\alpha 2^\alpha \Gamma(\frac{\alpha+1}{2})} = \frac{2\pi^{\frac{d}{2}}\Gamma(1-\frac{\alpha}{2})}{\alpha 2^\alpha \Gamma(\frac{d+\alpha}{2})} \tag{2.19}$$

□

Vegyük most 2.3 lemma legegyszerűbb esetét, azaz amikor $\alpha = 1$ mellett $\Gamma(1/2) = \sqrt{\pi}$, így

$$c_d = C(d,1) = \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)}. \tag{2.20}$$

Ekkor legyen a súlyfüggvény eszerint

$$w(u,v) = \left(c_p c_q |u|_p^{1+p} |v|_q^{1+q}\right)^{-1}. \tag{2.21}$$

Jelölje az index nélküli $\|\cdot\|$ norma a 2.1 definícióbeli normát a 2.21 súlyfüggvénnyel, és analóg módon $\mathcal{V}^2(X,Y)$ a 2.2 definícióban leírt mérőszámot ugyanezzel a súlyfüggvénnyel!

$\|f_{X,Y}(u,v) - f_X(u)f_Y(v)\|^2$ végességére elégséges feltétel, hogy $E|X|_p$ és $E|Y|_q$ véges legyen:

$$|f_{X,Y}(u,v) - f_x(u)f_Y(v)|^2 = \left[E\left(e^{i\langle u,X\rangle} - f_X(u)\right)\left(e^{i\langle v,Y\rangle} - f_Y(v)\right)\right]^2 \leq \tag{2.22}$$

$$\leq E\left[e^{i\langle u,X\rangle} - f_X(u)\right]^2 E\left[e^{i\langle v,Y\rangle} - f_Y(v)\right]^2 = \tag{2.23}$$

$$= \left(1 - |f_X(u)|^2\right)\left(1 - |f_Y(v)|^2\right), \tag{2.24}$$

Ha $E(|X|_p + |Y|_q) < \infty$, akkor

$$\int_{\mathbb{R}^{p+q}} |f_{X,Y}(u,v) - f_x(u)f_Y(v)|^2 \left(c_p c_q |u|_p^{1+p} |v|_q^{1+q}\right)^{-1} du dv \leq \tag{2.25}$$

$$\leq \int_{\mathbb{R}^p} \frac{1 - |f_X(u)|^2}{c_p |u|_p^{1+p}} du \int_{\mathbb{R}^q} \frac{1 - |f_Y(v)|^2}{c_q |v|_q^{1+q}} dv = \tag{2.26}$$

$$= E\left[\int_{\mathbb{R}^p} \frac{1 - \cos\langle u, X - X'\rangle}{c_p |u|_p^{1+p}} du\right] \cdot E\left[\int_{\mathbb{R}^q} \frac{1 - \cos\langle v, Y - Y'\rangle}{c_q |v|_q^{1+q}} dv\right] = \tag{2.27}$$

$$= E|X - X'|_p E|Y - Y'|_q < \infty, \tag{2.28}$$

ahol X és X' függetlenek és azonos eloszlásúak, és ugyanígy Y és Y' is.

Ez a megadott súlyfüggvény tehát alkalmas arra, hogy összefüggőségi mérőszámot definiáljunk vele.

2.2. Távolságkorreláció

2.2.1. Definíciók

A 2.21 súlyfüggvénnyel számolt távolság alapú mérőszámokat definiálja Székely, Rizzo és Bakirov[23] a következő módon:

2.4. definíció. Az X és Y véges várhatóértékű valószínűségi vektorváltozók távolságkovarianciáján azt a nemnegatív $\mathcal{V}(X,Y)$ számot értjük, amelyre

$$\mathcal{V}^2(X,Y) = \|f_{X,Y}(u,v) - f_X(u)f_Y(v)\|^2 = \tag{2.29}$$

$$= \frac{1}{c_p c_q} \int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(u,v) - f_x(u)f_Y(v)|^2}{|u|_p^{1+p}|v|_q^{1+q}} dudv. \tag{2.30}$$

Teljesen hasonlóan a klasszikus kovarianca esetéhez, itt is definiálható a szórásnégyzet analógiájára

$$\mathcal{V}^2(X) = \mathcal{V}^2(X,X) = \|f_{X,X}(u,v) - f_X(u)f_X(v)\|^2 \tag{2.31}$$

2.5. definíció. Az X és Y véges várhatóértékű valószínűségi vektorváltozók távolságkorrelációján azt a nemnegatív $\mathcal{R}(X,Y)$ számot értjük, amelyre

$$\mathcal{R}^2(X,Y) = \begin{cases} \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}, & \mathcal{V}^2(X)\mathcal{V}^2(Y) > 0 \\ 0, & \mathcal{V}^2(X)\mathcal{V}^2(Y) = 0. \end{cases} \tag{2.32}$$

Látható az analógia a klasszikus Pearson-féle korreláció definíciója, és a Székely-féle távolságkorreláció definíciója között. A 2.4 és 2.5 definíciókbeli mennyiségeket az angol nyelvű terminológia alapján jelöljük röviden rendre a dCov, dVar és dCor jelölésekkel!

2.2.2. Tapasztalati távolságkorreláció

Motiváció és definíció

A távolságkorreláció elméleti definícióján túl gyakorlatban használható függetlenségi tesztekhez szükséges definiálni tapasztalati távolságkovarianciát és -korrelációt. Ezek felépítéséhez kézenfekvő ötlet a tapasztalati karakterisztikus függvényeket felhasználni. A továbbiakban legyen $(\mathbf{X}, \mathbf{Y}) = \{(X_k, Y_k) : k = 1 \dots n)\}$ egy n elemű véletlen minta az $X \in \mathbb{R}^p$ és $Y \in \mathbb{R}^q$ valószínűségi vektorváltozók együttes eloszlásából. Ekkor a tapasztalati karakterisztikus függvény

$$f_{X,Y}^n(u,v) = \frac{1}{n} \sum_{k=1}^{n} \exp\{i\langle u, X_k\rangle + i\langle v, Y_k\rangle\}, \tag{2.33}$$

a marginális tapasztalati karakterisztikus függvények pedig

$$f_X^n(u) = \frac{1}{n} \sum_{k=1}^{n} \exp\{i\langle u, X_k\rangle\}, \tag{2.34}$$

$$f_Y^n(v) = \frac{1}{n} \sum_{k=1}^{n} \exp\{i\langle v, Y_k\rangle\}. \tag{2.35}$$

Ezekkel a mennyiségekkel természetes lenne a tapasztalati távolságkovariancia definiálására a $\|f_{X,Y}^n(u,v) - f_X^n(u)f_Y^n(v)\|$ mennyiséget használni. Ezzel szemben Székely,

Rizzo és Bakirov[23] definíciói a megfelelő mennyiségekre a következők:

Legyen $k, l = 1, \ldots, n$ esetén

$$a_{kl} = |X_k - X_l|_p, \tag{2.36}$$

$$\bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}, \tag{2.37}$$

$$\bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl}, \tag{2.38}$$

$$\bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}, \tag{2.39}$$

$$A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} - \bar{a}_{\cdot\cdot}, \tag{2.40}$$

és teljesen hasonlóan definiáljuk a b_{kl}, $\bar{b}_{k\cdot}$, $\bar{b}_{\cdot l}$, $\bar{b}_{\cdot\cdot}$, B_{kl} mennyiségeket!

2.6. definíció. A tapasztalati távolságkovariancia az a nemnegatív $\mathcal{V}_n(\mathbf{X}, \mathbf{Y})$ szám, amelyre

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}, \tag{2.41}$$

és hasonlóan

$$\mathcal{V}_n^2(\mathbf{X}) = \mathcal{V}_n^2(\mathbf{X}, \mathbf{X}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}^2. \tag{2.42}$$

2.7. definíció. A tapasztalati távolságkorreláció az a nemnegatív $\mathcal{R}_n(\mathbf{X}, \mathbf{Y})$ szám, amelyre

$$\mathcal{R}_n^2(\mathbf{X}, \mathbf{Y}) = \begin{cases} \frac{\mathcal{V}_n^2(\mathbf{X},\mathbf{Y})}{\sqrt{\mathcal{V}_n^2(\mathbf{X})\mathcal{V}_n^2(\mathbf{Y})}}, & \mathcal{V}_n^2(\mathbf{X})\mathcal{V}_n^2(\mathbf{Y}) > 0, \\ 0, & \mathcal{V}_n^2(\mathbf{X})\mathcal{V}_n^2(\mathbf{Y}) = 0, \end{cases} \tag{2.43}$$

Tulajdonságok

A definíció motivációját alapvetően nehéz látni, de az alábbi tétel megerősíti jogosságát:

2.8. tétel. *Ha* $(\mathbf{X}, \mathbf{Y})$ *egy* n*-elemű minta* (X, Y) *együttes eloszlásából, akkor*

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = \|f_{X,Y}^n(u, v) - f_X^n(u)f_Y^n(v)\|^2 \tag{2.44}$$

Bizonyítás. (Vázlat)

A bizonyítás lényege, hogy a $\|f_{X,Y}^n(u, v) - f_X^n(u)f_Y^n(v)\|^2$ integrált a 2.3 lemma segítségével kifejezi. A tapasztalati távolságkorrelációval a cél függetlenség tesztelésére

alkalmas statisztikai próba konstruálása, ehhez a bizonyítás közben kapott lényeges köztes eredmény, hogy a fenti normanégyzet a következő alakban felírható:

$$\|f^n_{X,Y}(u,v) - f^n_X(u)f^n_Y(v)\|^2 = S_1 + S_2 - 2S_3, \tag{2.45}$$

ahol

$$S_1 = \frac{1}{n^2}\sum_{k,l=1}^{n} |X_k - X_l|_p |Y_k - Y_l|_q, \tag{2.46}$$

$$S_2 = \frac{1}{n^2}\sum_{k,l=1}^{n} |X_k - X_l|_p \frac{1}{n^2}\sum_{k,l=1}^{n} |Y_k - Y_l|_q, \tag{2.47}$$

$$S_3 = \frac{1}{n^3}\sum_{k=1}^{n}\sum_{l,m=1}^{n} |X_k - X_l|_p |Y_k - Y_m|_q. \tag{2.48}$$

Ezek után algebrai átalakításokkal belátható, hogy $\mathcal{V}^2_n(\mathbf{X},\mathbf{Y}) = S_1 + S_2 - 2S_3$.
A teljes bizonyítást lásd[23] □

2.9. megjegyzés. A tapasztalati távolságkorreláció és -kovariancia definíciójának gyakorlati előnyei vannak. A természetszerűen adódó ötlettel szemben a számítási idő bizonyos tesztelések esetén akár 98%-os időmegtakarítást is eredményezhet Székely és Rizzo[20] tapasztalatai szerint.

2.10. tétel. *Ha $E|X|_p < \infty$ és $E|Y|_q < \infty$, akkor majdnem mindenütt*

$$\lim_{n\to\infty} \mathcal{V}_n(\mathbf{X},\mathbf{Y}) = \mathcal{V}(X,Y). \tag{2.49}$$

Bizonyítás. Legyen

$$\xi_n(u,v) = \frac{1}{n}\sum_{k=1}^{n} e^{i\langle u,X_k\rangle + i\langle v,Y_k\rangle} - \frac{1}{n}\sum_{k=1}^{n} e^{i\langle u,X_k\rangle}\frac{1}{n}\sum_{k=1}^{n} e^{i\langle v,Y_k\rangle}, \tag{2.50}$$

ezzel $\mathcal{V}^2_n = \|\xi_n(u,v)\|^2$.

$$\xi_n(u,v) = \frac{1}{n}\sum_{k=1}^{n} U_k V_k - \frac{1}{n}\sum_{k=1}^{n} U_k \frac{1}{n}\sum_{k=1}^{n} V_k, \tag{2.51}$$

ahol $U_k = \exp\{i\langle u, X_k\rangle\} - f_X(u)$ és $V_k = \exp\{i\langle v, Y_k\rangle\} - f_Y(v)$.
Tetszőleges $\delta > 0$ szám esetén legyen

$$D(\delta) = \{(u,v) : \delta \le |u|_p \le 1/\delta, \delta \le |v|_q \le 1/\delta\}, \tag{2.52}$$

és

$$\mathcal{V}^2_{n,\delta} = \int_{D(\delta)} |\xi_n(u,v)|^2 d\omega, \tag{2.53}$$

ahol

$$d\omega = \left(c_p c_q |u|_p^{1+p} |v|_q^{1+q}\right)^{-1} du dv. \tag{2.54}$$

Tetszőleges rögzített δ pozitív szám esetén $w(u,v)$ korlátos a $D(\delta)$ tartományon, tehát $\mathcal{V}_{n,\delta}^2$ kombinációja korlátos valószínűségi változók von Mises-féle V-statisztikáinak (lásd [18],[24]). Tetszőleges $\delta > 0$ esetén a von Mises-féle V-statisztikákra vonatkozó nagy számok erős törvénye szerint majdnem mindenütt

$$\lim_{n\to\infty} \mathcal{V}_{n,\delta}^2 = \mathcal{V}_{\cdot,\delta}^2 = \int_{D(\delta)} |f_{X,Y}(u,v) - f_X(u) f_Y(v)|^2 d\omega. \tag{2.55}$$

$\mathcal{V}_{n,\delta}^2$ definíciójából látszik, hogy $\delta \to 0$ esetén $\mathcal{V}^2$-hez tart. Ekkor még azt kell belátnunk, hogy majdnem mindenütt

$$\limsup_{\delta\to 0} \limsup_{n\to\infty} |\mathcal{V}_{n,\delta}^2 - \mathcal{V}_n^2| = 0. \tag{2.56}$$

Tetszőleges pozitív δ esetén

$$|\mathcal{V}_{n,\delta}^2 - \mathcal{V}_n^2| \le \int_{|u|_p<\delta} |\xi_n(u,v)|^2 d\omega + \int_{|u|_p>1/\delta} |\xi_n(u,v)|^2 d\omega + \int_{|v|_q<\delta} |\xi_n(u,v)|^2 d\omega + \int_{|v|_q>1/\delta} |\xi_n(u,v)|^2 d\omega. \tag{2.57}$$

Ekkor

$$\xi_n(u,v) = \frac{1}{n}\sum_{k=1}^n U_k V_k - \frac{1}{n}\sum_{k=1}^n U_k \frac{1}{n}\sum_{k=1}^n V_k \tag{2.58}$$

miatt

$$|\xi_n(u,v)|^2 \le 2\left|\frac{1}{n}\sum_{k=1}^n U_k V_k\right|^2 + 2\left|\frac{1}{n}\sum_{k=1}^n U_k \frac{1}{n}\sum_{k=1}^n V_k\right|^2 \le \tag{2.59}$$

$$\le \frac{4}{n}\sum_{k=1}^n |U_k|^2 \frac{1}{n}\sum_{k=1}^n |V_k|^2. \tag{2.60}$$

Ekkor a 2.57 jobb oldalának első tagjára

$$\int_{|u|_p<\delta} |\xi_n(u,v)|^2 d\omega \le \frac{4}{n}\sum_{k=1}^n \int_{|u|_p<\delta} \frac{|U_k|^2 du}{c_p |u|_p^{1+p}} \frac{1}{n}\sum_{k=1}^n \int_{\mathbb{R}^q} \frac{|V_k|^2 dv}{c_q |v|_q^{1+q}}. \tag{2.61}$$

$|V_k|^2 = 1 + |f_Y(v)|^2 - e^{i\langle v, Y_k\rangle} \bar{f_Y}(v) - e^{-i\langle v, Y_k\rangle} f_Y(v)$, ezt az integrálba helyettesítve

$$\int_{\mathbb{R}^q} \frac{|V_k|^2 dv}{c_q |v|_q^{1+q}} = 2E_Y |Y_k - Y| - E|Y - Y'| \le 2(|Y_k| + E|Y|). \tag{2.62}$$

Itt E_Y az Y eloszlása szerinti várható érték, Y' pedig az Y-tól független, vele azonos eloszlású valószínűségi változó. A $z = (z_1, \ldots, z_p) \in \mathbb{R}^p$ vektorra legyen

$$G(y) = \int_{|z|<y} \frac{1 - \cos(z_1)}{|z|^{1+p}} dz. \tag{2.63}$$

Ezzel a másik integrálra

$$\int_{|u|_p<\delta} \frac{|U_k|^2 du}{c_p |u|_p^{1+p}} = 2E_X|X_k - X|G(|X_k - X|\delta) - E|X - X'|G(|X - X'|\delta) \leq \quad (2.64)$$

$$\leq 2E_X|X_k - X|G(|X_k - X|\delta). \quad (2.65)$$

Ezeket behelyettesítve az integrálra vonatkozó felső becslésbe

$$\int_{|u|_p<\delta} |\xi_n(u,v)|^2 d\omega \leq 2\frac{4}{n}\sum_{k=1}^{n}(|Y_k| + E|Y|) \cdot 2\frac{1}{n}\sum_{k=1}^{n} E_X|X_k - X|G(|X_k - X|\delta) \quad (2.66)$$

Ekkor

$$\limsup_{n\to\infty} \int_{|u|_p<\delta} |\xi_n(u,v)|^2 d\omega \leq 2 \cdot 4 \cdot (E|Y| + E|Y|) \cdot 2 \cdot E(|X_1 - X_2|)G(|X_1 - X_2|\delta) \quad (2.67)$$

a nagy számok erős törvénye szerint majdnem mindenütt, ekkor viszont alkalmazhatjuk rá a Lebesgue-tételt, mivel $G(y)$ értékkészletének c_p felső korlátja, és

$$\lim_{y\to 0} G(y) = 0, \quad (2.68)$$

így majdnem mindenütt

$$\limsup_{\delta\to 0} \limsup_{n\to\infty} \int_{|u|_p<\delta} |\xi_n(u,v)|^2 d\omega = 0. \quad (2.69)$$

Mivel $|U_k|^2 \leq 4$ és $\frac{1}{n}\sum_{k=1}^{n}|U_k|^2 \leq 4$, a 2.57 jobb oldalának második tagjára

$$\int_{|u|_p>1/\delta} |\xi_n(u,v)|^2 d\omega \leq \frac{4}{n}\sum_{k=1}^{n}\int_{|u|_p>1/\delta} \frac{|U_k|^2 du}{c_p|u|_p^{1+p}} \frac{1}{n}\sum_{k=1}^{n}\int_{\mathbb{R}^q} \frac{|V_k|^2 dv}{c_q|v|_q^{1+q}} \leq \quad (2.70)$$

$$\leq 16\int_{|u|_p>1/\delta} \frac{du}{c_p|u|_p^{1+p}} \frac{1}{n}\sum_{k=1}^{n}\int_{\mathbb{R}^q} \frac{|V_k|^2 dv}{c_q|v|_q^{1+q}} \leq \quad (2.71)$$

$$\leq 16\delta\frac{2}{n}\sum_{k=1}^{n}(|Y_k| + E|Y|), \quad (2.72)$$

és az előzőhöz hasonlóan így

$$\limsup_{\delta\to 0} \limsup_{n\to\infty} \int_{|u|_p>1/\delta} |\xi_n(u,v)|^2 d\omega = 0 \quad (2.73)$$

Ekkor teljesen hasonlóan a 2.57 maradék két tagjára végigszámolható ugyanez. Ekkor tehát

$$\lim_{\delta\to 0}\lim_{n\to\infty} \mathcal{V}_{n,\delta}^2 = \mathcal{V}^2 \quad (2.74)$$

és

$$\limsup_{\delta\to 0} \limsup_{n\to\infty} |\mathcal{V}_{n,\delta}^2 - \mathcal{V}_n^2| = 0, \quad (2.75)$$

tehát a tételt beláttuk.

□

A tételből pedig egyszerű meggondolással adódik a következő:

2.11. következmény. *Ha* $E|X|_p < \infty$ *és* $E|Y|_q < \infty$*, akkor majdnem mindenütt*

$$\lim_{n\to\infty} \mathcal{R}_n^2(\mathbf{X}, \mathbf{Y}) = \mathcal{R}^2(X, Y). \tag{2.76}$$

Az alábbi tételben összefoglalunk néhány, a definícióból nem feltétlenül világos tulajdonságot a tapasztalati mennyiségekről:

2.12. tétel. *A tapasztalati mennyiségekre*

1. $\mathcal{V}_n(\mathbf{X}, \mathbf{Y}) \geq 0$
2. $0 \leq \mathcal{R}_n(\mathbf{X}, \mathbf{Y}) \leq 1$
3. $\mathcal{V}_n(\mathbf{X}) = 0$ *akkor és csak akkor, ha minden megfigyelés a mintában egyenlő*
4. *Ha* $\mathcal{R}_n(\mathbf{X}, \mathbf{Y}) = 1$*, akkor létezik olyan a vektor, b nem nulla valós szám és egy C ortogonális mátrix, hogy* $\mathbf{Y} = a + b\mathbf{X}C$*.*

Bizonyítás. 1. Ha $\mathcal{V}_n(\mathbf{X}) = 0$, akkor $A_{kl} = 0$ minden k és l esetén, így

$$A_{kk} = a_{kk} - \bar{a}_{k\cdot} - \bar{a}_{\cdot k} - \bar{a}_{\cdot\cdot} = 0 \tag{2.77}$$

$$\bar{a}_{k\cdot} = \bar{a}_{\cdot k} = \frac{\bar{a}_{\cdot\cdot}}{2}, \tag{2.78}$$

hasonlóan

$$\bar{a}_{l\cdot} = \bar{a}_{\cdot l} = \frac{\bar{a}_{\cdot\cdot}}{2}, \tag{2.79}$$

így

$$0 = A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} - \bar{a}_{\cdot\cdot} = a_{kl} = |X_k - X_l|_p, \tag{2.80}$$

tehát $k, l = 1, \ldots, n$ $|X_k - X_l|_p = 0 \Rightarrow X_1 = \ldots = X_n$.

A másik irány nyilvánvaló.

2. teljesen hasonló a 2.13 tétel $\mathcal{R}$-re vonatkozó hasonló pontjának bizonyításához.

3. A 2.8 tételből nyilvánvaló.

4. Ha $\mathcal{R}_n(\mathbf{X}, \mathbf{Y}) = 1$, akkor a definícióból látható, hogy X és Y majdnem mindenütt hasonlóak abban az értelemben, hogy valamilyen $\varepsilon \neq 0$ számra X és εY izometrikusak. Ekkor az $\mathbf{X}$ és $\mathbf{Y}$ által kifeszített lineáris alterek dimenziói majdnem mindenütt egyenlőek. Feltehetjük, hogy $\mathbf{X}$ és $\mathbf{Y}$ ugyanabban az euklideszi térben vannak, és $\mathbb{R}^p$-t feszítik ki. Ha $\mathcal{R}_n(\mathbf{X}, \mathbf{Y}) = 1$, akkor

$$\mathcal{R}_n(\mathbf{X}, \mathbf{Y}) = \frac{\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})}{\sqrt{\mathcal{V}_n^2(\mathbf{X})\mathcal{V}_n^2(\mathbf{Y})}} = 1 \tag{2.81}$$

miatt

$$\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) = \sqrt{\mathcal{V}_n^2(\mathbf{X})\mathcal{V}_n^2(\mathbf{Y})} \tag{2.82}$$

$$\frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl} = \sqrt{\frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}^2 \frac{1}{n^2}\sum_{k,l=1}^{n} B_{kl}^2} \tag{2.83}$$

Viszont a Cauchy-Schwarz-Bunyakovszkij egyenlőtlenség szerint

$$\left(\sum_{k,l=1}^{n} A_{kl}B_{kl}\right)^2 \leq \sum_{k,l=1}^{n} A_{kl}^2 \sum_{k,l=1}^{n} B_{kl}^2, \tag{2.84}$$

tehát a 2.82 egyenlőség csak akkor teljesülhet, ha $A_{kl} = c \cdot B_{kl}$ valamilyen c számra. Tegyük fel most, hogy $|c| = 1$, ekkor valamilyen d_k, d_l konstansok mellett minden k, l esetén

$$|X_k - X_l|_p = |Y_k - Y_l|_q + d_k + d_l, \tag{2.85}$$

és a $k = l$ esetből kapjuk, hogy $d_k = 0$ minden k-ra.
Mivel a két minta izometrikus, $\mathbf{Y}$ megkapható $\mathbf{X}$-ből eltolás, forgatás és tükrözés után, így $\mathbf{Y} = a + b\mathbf{X}C$ valamilyen a vektor, $b = c$ szám és C ortogonális mátrix mellett. Ha $|c| \neq 1$ és $c \neq 0$,akkor ugyanezt a meggondolást alkalmazva $c\mathbf{X}$-re és $\mathbf{Y}$-ra az állítást beláttuk.

□

2.2.3. Tulajdonságok

Az alábbi tételben áttekintjük a dCov, dCor és dVar mennyiségek néhány tulajdonságát. A definíciókból és eddigi tulajdonságokból jól látható analógia a Pearson-féle korrelációval az alábbi tételből is megfigyelhető.

2.13. tétel. *Az $X \in \mathbb{R}^p$ és $Y \in \mathbb{R}^q$ véges várhatóértékű valószínűségi vektorváltozókra*

1. $0 \leq \mathcal{R}(X,Y) \leq 1$ *és* $\mathcal{R} = 0$ *akkor és csak akkor, ha X és Y függetlenek.*

2. *Minden $a_1 \in \mathbb{R}^p$, $a_2 \in \mathbb{R}^q$ vektorra, b_1, b_2 számra és $C_1 \in \mathbb{R}^p$, $C_2 \in \mathbb{R}^q$ ortonormált mátrixra*

$$\mathcal{V}(a_1 + b_1C_1X, a_2 + b_2C_2Y) = \sqrt{|b_1b_2|}\mathcal{V}(X,Y). \tag{2.86}$$

3. *Ha (X_1, Y_1) vektorváltozó független (X_2, Y_2) vektorváltozótól, akkor*

$$\mathcal{V}(X_1 + X_2, Y_1 + Y_2) \leq \mathcal{V}(X_1, Y_1) + \mathcal{V}(X_2, Y_2). \tag{2.87}$$

Egyenlőség akkor és csak akkor áll fenn, ha X_1 és Y_1 is konstans vagy X_2 és Y_2 is konstans vagy X_1, X_2, Y_1, Y_2 függetlenek.

4. *Ha* $\mathcal{V}(X) = 0$, *akkor* $X = E(X)$ *majdnem mindenütt.*

5. *Minden* $a \in \mathbb{R}^p$ *vektorra,* b *számra és* $C \in \mathbb{R}^p$ *ortonormált mátrixra*

$$\mathcal{V}(a + bCX) = |b|\mathcal{V}(X). \tag{2.88}$$

6. *Ha* X *és* Y *függetlenek, akkor*

$$\mathcal{V}(X + Y) \leq \mathcal{V}(X) + \mathcal{V}(Y). \tag{2.89}$$

Egyenlőség akkor és csak akkor áll fenn, ha X *vagy* Y *konstans.*

Bizonyítás. 1. A függetlenségre vonatkozó állítás a definícióból nyilvánvaló. Legyen $U = e^{i\langle u,X\rangle} - f_X(u)$ és $V = e^{i\langle v,Y\rangle} - f_Y(v)$! Ekkor

$$|f_{X,Y}(u,v) - f_X(u)f_Y(v)|^2 = |E(UV)|^2 \leq (E(|U||V|))^2 \leq \tag{2.90}$$

$$\leq E(|U|^2|V|^2) = (1 - |f_X(u)|^2)(1 - |f_Y(v)|^2). \tag{2.91}$$

Ezt behelyettesítve a dCor definíciójába

$$\int_{\mathbb{R}^{p+q}} |f_{X,Y}(u,v) - f_X(u)f_Y(v)|^2 d\omega \leq \int_{\mathbb{R}^{p+q}} |(1 - |f_X(u)|^2)(1 - |f_Y(v)|^2)|^2 d\omega, \tag{2.92}$$

tehát $0 \leq \mathcal{R}(X,Y) \leq 1$.

2. Az állítás a definícióból nyilvánvaló.

3.

$$\mathcal{V}(X_1 + X_2, Y_1 + Y_2) = \tag{2.93}$$

$$\|f_{X_1+X_2,Y_1+Y_2}(u,v) - f_{X_1+X_2}(u)f_{Y_1+Y_2}(v)\| = \tag{2.94}$$

$$= \|f_{X_1,Y_1}(u,v)f_{X_2,Y_2}(u,v) - f_{X_1}(u)f_{X_2}(u)f_{Y_1}(v)f_{Y_2}(v)\| \leq \tag{2.95}$$

$$\leq \|f_{X_1,Y_1}(u,v)\big(f_{X_2,Y_2}(u,v) - f_{X_2}(u)f_{Y_2}(v)\big)\| + \tag{2.96}$$

$$+\|f_{X_2}(u)f_{Y_2}(v)\big(f_{X_1,Y_1}(u,v) - f_{X_1}(u)f_{Y_1}(v)\big)\| \leq \tag{2.97}$$

$$\leq \|f_{X_2,Y_2}(u,v) - f_{X_2}(u)f_{Y_2}(v)\| + \|f_{X_1,Y_1}(u,v) - f_{X_1}(u)f_{Y_1}(v)\| = \tag{2.98}$$

$$= \mathcal{V}(X_1,Y_1) + \mathcal{V}(X_2,Y_2) \tag{2.99}$$

Ha X_1 és $Y_1$is konstans vagy X_2 és Y_2 is konstans vagy X_1, X_2, Y_1, Y_2 függetlenek, akkor láthatóan egyenlőséget kapunk mindkét egyenlőtlenségnél.
Ha az egyenlőség fennáll, tegyük fel, hogy X_1 és Y_1 sem konstans egyidejűleg és X_2 és Y_2 sem konstans egyidejűleg! Ekkor a második egyenlőtlenségben csak akkor kaphatunk egyenlőséget, ha X_1 és Y_1 függetlenek és X_2 és Y_2 is függetlenek, továbbá a feltételeink szerint (X_1, Y_1) és (X_2, Y_2) függetlenek, ezzel pontosan az állítást kaptuk.

4. Ha $\mathcal{V}(X) = 0$, akkor

$$\mathcal{V}^2(X) = \int_{\mathbb{R}^p \times \mathbb{R}^p} \frac{|f_{X,X}(u,v) - f_X(u)f_X(v)|^2}{c_p^2 |u|_p^{p+1} |s|_p^{p+1}} dudv = 0, \tag{2.100}$$

azaz $\forall u, v \quad : \quad f_{X,X}(u,v) = f_X(u+v) = f_X(u)f_X(v)$, tehát $f_X(u) = e^{i\langle u,c\rangle}$ valamilyen $c \in \mathbb{R}^p$ konstans vektorral, tehát X vektorváltozó konstans majdnem mindenütt.

5. Az állítás a definícióból nyilvánvaló.

6. Alkalmazzuk a dCov-ra vonatkozó hasonló állítást $X = X_1 = Y_1$ és $Y = X_2 = Y_2$ mellett, ekkor pontosan ezt az állítást kapjuk.

□

2.3. Statisztikai módszerek távolságalapon

Az alábbi fejezetben áttekintünk néhány lehetőséget, hogyan lehet a távolságalapon definiált mennyiségeket statisztikai vizsgálatokra használni.

2.3.1. Függetlenségvizsgálat dCov használatával

Ebben a fejezetben függetlenségvizsgálatra a korábban definiált tapasztalati távolságkovariancia tulajdonságait kihasználva szeretnénk jó próbastatisztikát és próbát definiálni. Ehhez a kérdéses mennyiségaszimptotikus tulajdonságairól tekintsük át a lehetőségeket!

Legyen ζ egy nulla várható értékű komplex Gauss-folyamat az

$$R(t,t_0) = \big(f_X(u-u_0) - f_X(u)\overline{f_X(u_0)}\big)\big(f_Y(v-v_0) - f_Y(v)\overline{f_Y(v_0)}\big) \tag{2.101}$$

kovarianciafüggvénnyel, ahol $t = (u,v)$ és $t_0 = (u_0, v_0)$ $\mathbb{R}^p \times \mathbb{R}^q$-beli vektorok. A függetlenségvizsgálat próbastatisztikájaként $\frac{n\mathcal{V}_n^2}{S_2}$-t szeretnénk használni, ahol S_2-t a 2.8 tétel bizonyításvázlatában definiáltuk a 2.47 pontban. Elsőként lássunk egy tételt $n\mathcal{V}_n^2$ gyenge konvergenciájáról!

2.14. tétel. *Ha $X \in \mathbb{R}^p$ és $Y \in \mathbb{R}^q$ vektorváltozók függetlenek, és $E(|X|_p + |Y|_q) < \infty$, akkor*

$$n\mathcal{V}_n^2 \underset{n\to\infty}{\overset{D}{\to}} \|\zeta(u,v)\|^2. \tag{2.102}$$

Bizonyítás. Legyen $\zeta_n(t)$ az empirikus folyamat, azaz

$$\zeta_n(t) = \zeta_n(u,v) = \sqrt{n}\xi_n(u,v) = \sqrt{n}\big(f_{X,Y}^n(u,v) - f_X^n(u)f_Y^n(v)\big) \tag{2.103}$$

A függetlenség miatt $E(\zeta_n(t)) = 0$ és

$$E(\zeta_n(t)\overline{\zeta_n(t_0)}) = \frac{n-1}{n} \cdot R(t, t_0) \tag{2.104}$$

$u = u_0$ mellett pedig

$$E|\zeta_n(t)|^2 = \frac{n-1}{n}(1 - |f_X(u)|^2)(1 - |f_Y(v)|^2). \tag{2.105}$$

Tekintsük most a következő $\{Q_n(\delta)\}$ valószínűségiváltozó-sorozatot a δ nemnegatív szám mellett! Rögzített $\varepsilon > 0$ mellett legyen $\{D_k\}_{k=1}^n$ a 2.10 tétel bizonyításában definiált 2.52-beli $D(\delta)$ tartomány egy N-elemű partíciója legfeljebb ε átmérőjű mérhető halmazokra! Legyen

$$Q_n(\delta) = \sum_{k=1}^{N(\varepsilon)} \int_{D_k} |\zeta_n|^2 d\omega. \tag{2.106}$$

Rögzített $M > 0$ mellett legyen

$$\beta(\varepsilon) = \sup_{\substack{t,\, t_0 \\ \max\{|u|, |u_0|, |v|, |v_0|\} < M \\ |u-u_0|^2 + |v-v_0|^2 < \varepsilon^2}} E\big||\zeta_n(t)|^2 - |\zeta_n(t_0)|^2\big|, \tag{2.107}$$

Ekkor $lim_{\varepsilon \to 0}\beta(\varepsilon) = 0$ minden rögzített pozitív M mellett, és rögzített $\delta > 0$ esetén

$$E\left| \int_{D(\delta)} |\zeta_n(t)|^2 d\omega - Q_n(\delta) \right| \leq \beta(\varepsilon) \int_D |\zeta_n(t)|^2 d\omega \underset{\varepsilon \to 0}{\to} 0 \tag{2.108}$$

Továbbá a 2.10 tétel bizonyításában szereplő felső becslésekhez teljesen hasonló módon

$$\left| \int_D |\zeta_n(t)|^2 d\omega - \int_{\mathbb{R}^{p+q}} |\zeta_n(t)|^2 d\omega \right| \leq \tag{2.109}$$

$$\leq \int_{|u|_p < \delta} |\zeta_n(u,v)|^2 d\omega + \int_{|u|_p > 1/\delta} |\zeta_n(u,v)|^2 d\omega + \tag{2.110}$$

$$+ \int_{|v|_q < \delta} |\zeta_n(u,v)|^2 d\omega + \int_{|v|_q > 1/\delta} |\zeta_n(u,v)|^2 d\omega. \tag{2.111}$$

és

$$E\left[\int_{|u|_p < \delta} |\zeta_n(u,v)|^2 d\omega + \int_{|u|_p > 1/\delta} |\zeta_n(u,v)|^2 d\omega \right] \leq \tag{2.112}$$

$$\leq \frac{n-1}{n}(E|X_1 - X_2|G(|X_1 - X_2|\delta) + w_p\delta)E|Y_1 - Y_2|\delta \to_{\delta \to 0} 0, \tag{2.113}$$

ahol w_p egy csak p-től függő konstans, és ugyanígy

$$E\left[\int_{|v|_q < \delta} |\zeta_n(u,v)|^2 d\omega + \int_{|v|_q > 1/\delta} |\zeta_n(u,v)|^2 d\omega \right] \underset{\delta \to 0}{\to} 0. \tag{2.114}$$

Hasonló egyenlőtlenségek igazak a

$$Q(\delta) = \sum_{k=1}^{N(\varepsilon)} \int_{D_k} |\zeta|^2 d\omega \tag{2.115}$$

valószínűségi változóra is. Ekkor minden $\delta > 0$ esetén a centrális határeloszlás-tétel szerint $Q_n(\delta)$ eloszlásban tart $Q(\delta)$-hoz $n \to \infty$ esetén, $E|Q_n(\delta) - \zeta_n| \leq \delta$ és $E|Q(\delta) - \zeta| \leq \delta$, ezekből pedig

$$n\mathcal{V}_n^2 = \|\zeta_n\|^2 \underset{n\to\infty}{\overset{D}{\to}} \|\zeta(u,v)\|^2. \tag{2.116}$$

□

2.15. következmény. *Ha $X \in \mathbb{R}^p$ és $Y \in \mathbb{R}^q$ vektorváltozók függetlenek, akkor*

$$\frac{n\mathcal{V}_n^2}{S_2} \underset{n\to\infty}{\overset{D}{\to}} Q, \tag{2.117}$$

ahol

$$Q \overset{D}{=} \sum_{j=1}^{\infty} \lambda_j Z_j^2, \tag{2.118}$$

ahol Z_j független standard normálisok, a λ_j konstansokat pedig (X, Y) eloszlása határozza meg, és hogy $E(Q) = 1$.

2.16. következmény. *Ha $X \in \mathbb{R}^p$ és $Y \in \mathbb{R}^q$ vektorváltozók összefüggőek, akkor*

$$\frac{n\mathcal{V}_n^2}{S_2} \underset{n\to\infty}{\overset{P}{\to}} \infty, \tag{2.119}$$

Bizonyítás. A 2.10 tétel szerint

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) \underset{n\to\infty}{\overset{m.m.}{\to}} \mathcal{V}^2(X, Y), \tag{2.120}$$

és az összefüggőség miatt $\mathcal{V}^2 > 0$, így

$$n\mathcal{V}_n^2 \underset{n\to\infty}{\to} \infty \tag{2.121}$$

sztochasztikusan, S_2 pedig egy számhoz tart a nagy számok erős törvénye szerint majdnem mindenütt, így

$$\frac{n\mathcal{V}_n^2}{S_2} \underset{n\to\infty}{\overset{P}{\to}} \infty. \tag{2.122}$$

□

Tehát a próbánk olyan, hogy magas $\frac{n\mathcal{V}_n^2}{S_2}$ értékek esetén elutasítjuk a függetlenség hipotézisét. Ennek a próbának a konkretizálását és az elérhető szignifikanciaszintet írja le a következő tétel:

2.17. tétel. *Legyen $T(X, Y, \alpha, n)$ az a próba, ami elutasítja a függetlenség hipotézisét*

$$\frac{n\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})}{S_2} > \left(\Phi^{-1}(1-\alpha/2)\right)^2 \tag{2.123}$$

esetén, ahol Φ a standard normális eloszlásfüggvény. Jelölje $\alpha(X, Y, n)$ a próba elérhető szignifikanciaszintjét! Ha $E(|X|_p + |Y|_q) < \infty$, akkor $\forall 0 < \alpha < 0.215$ esetén

1. $\lim_{n\to\infty} \alpha(X, Y, n) \leq \alpha$
2. $\sup_{X,Y}\{\lim_{n\to\infty} \alpha(X, Y, n) : \mathcal{V}(X, Y) = 0\} = \alpha$.

A bizonyításhoz először tekintsük át Székely és Bakirov[19] nyomán a következőket! A 2.118-hoz hasonlóan egy $X \sim N(0, R)$ n-dimenziós valószínűségi vektorváltozó és tetszőleges $A \in \mathbb{R}^{n\times n}$ mátrix esetén $Q_n = \langle X, AX\rangle$ kvadratikus alak előállítható

$$Q_n = \sum_{j=1}^{n} \lambda_j Z_j^2 \tag{2.124}$$

alakban, ahol λ_j az $R^{\frac{1}{2}}(A + A^T)R^{\frac{1}{2}}/2$ mátrix sajátértékei. Legyen A pozitív szemidefinit, így $\forall j \geq 1 \lambda_j \geq 0$. Ugyanígy

$$Q = \sum_{j=1}^{\infty} \lambda_j Z_j^2. \tag{2.125}$$

Legyen

$$I_n(x) = \inf_{\{Q_n \geq 0 | E(Q_n)=1\}} (Q_n \leq x) \tag{2.126}$$

és

$$I(x) = \inf_{\{Q \geq 0 | E(Q)=1\}} (Q \leq x), \tag{2.127}$$

Legyen χ_n^2 egy n szabadsági fokú χ^2 eloszlású valószínűségi változó, ekkor jelölje $x(d)$ a $P(d^{-1}\chi_d^2 \leq x)$ és $P((d+1)^{-1}\chi_{d+1}^2 \leq x)$ eloszlásfüggvények görbéinek metszéspontját minden d pozitív egészre, és legyen $x(0) = \infty$! Ekkor $x(d)$ egyértelmű és d-ben monoton csökkenve tart 1-hez. Székely és Bakirov[19] alapján

2.18. tétel.

$$I_n(x) = \begin{cases} P(d^{-1}\chi_d^2 \leq x), & \forall x \in [x(d), x(d-1)); d = 1, \ldots, n-1, \\ P(n^{-1}\chi_n^2 \leq x), & \forall x \in [0, x(n-1)), \end{cases} \tag{2.128}$$

és

$$I(x) = \begin{cases} P(d^{-1}\chi_d^2 \leq x), & \forall x \in [x(d), x(d-1)); d = 1, 2, \ldots, \\ 0, & \forall x < 1, \\ 0.5, & x = 1. \end{cases} \tag{2.129}$$

2.19. megjegyzés. Székely és Bakirov[19] belátták továbbá, hogy ha

$$p_d = P(d^{-1}\chi^2_d > x) = P((d+1)^{-1}\chi^2_{d+1} > x), \tag{2.130}$$

akkor $\forall x(d) \leq x < x(d-1) : P(Q > x) \leq p_d$

Ekkor a 2.17 tétel bizonyítása:

Bizonyítás. 1. A 2.18 tétel speciális eseteként $x(1) = 1.536397\ldots$ érték mellett az előző megjegyzés alapján $p_1 = 0.2151549\ldots$ értékkel $\forall x(1) \leq x < x(0) = \infty :$ $P(Q > x) \leq 0.215$, azaz a Q kvadratikus alakra $E(Q) = 1$ mellett $\forall 0 < \alpha \leq 0.215$

$$P\Big[Q \geq \left(\Phi^{-1}(1-\alpha/2)\right)^2\Big] \leq \alpha \tag{2.131}$$

2. Független X és Y indikátorokra a centrális határeloszlás-tétellel $\sqrt{n}\mathcal{R}_n(\mathbf{X}, \mathbf{Y})$ aszimptotikusan normális, mivel ebben az esetben a dCor megegyezik a klasszikus pearsoni korreláció abszolútértékével, ekkor tehát Q egyetlen tagból áll, $Q = Z^2$, és így elértük az α felső korlátot.

□

Szimulációs eredmények

Székely, Rizzo és Bakirov[23] Monte Carlo szimulációkon keresztül összehasonlították a Wilks-lambda[26] és két Puri és Sen[14] féle likelihood-hányados próbák erejét és az itt leírt távolságkovariancia alapú függetlenségvizsgálat erejét különböző példákon $p = q = 5$ esetekben. A szimulációk szerint X és Y marginális standard normális eloszlása esetén a tavolságkovarianciás teszt ereje közel van az ebben az esetben optimális Wilks-lambda likelihood-hányados próba erejéhez, marginális t-eloszlások esetén is hasonlóan jól teljesít, és lényegesen kisebb elsőfajúhiba-valószínűségekkel mint a Wilks-lambda próba.

X standard többváltozós normális esetben, $Y_{kj} = X_{kj}\varepsilon_{kj}$ mellett, ahol ε_{kj} független standard normálisok, és X-től is függetlenek, a Puri és Sen féle próbák ereje a mintaelemszám növelésével nem növekedik, a távolságkovarianciás próba lényegesen erősebbnek bizonyult, mint a likelihood-hányados próbák. $Y_{kj} = \log(X^2_{kj})$ esetén a helyzet teljesen hasonló, azaz míg a dCov-próba szimulációi elég nagy erőt jeleznek, a likelihood-hányados próbák ereje mégcsak nem is növekszik a mintaelemszám növelése mellett. Levonható tehát a következtetés, hogy a konstruált próba nemlineáris kapcsolatok esetén lehet lényegesen erősebb, mint a klasszikus likelihood-hányados próbák, többváltozós normális esetben pedig elég közel lehet a kérdéses próbákhoz.

2.20. megjegyzés. A fent említett jó tulajdonságokhoz nem feltétlenül szükséges pontosan ezt a próbát használni. Bakirov, Rizzo és Székely[1] 2006-ban konstruáltak egy alapjaiban teljesen megegyező, ám a normabeli integráláshoz más súlyfüggvényt használó mérőszámot, amellyel hasonló vizsgálati sikereket értek el. A megfelelő súlyfüggvényre és normára a 2.4 fejezetben pontosabban kitérünk.

2.3.2. Szórásanalízis kiterjesztése

Távolságalapú mérőszámokat használhatunk a klasszikus szórásanalízis módszereinek kiterjesztésére Rizzo és Székely[17] nyomán. Az itt definiált próba a klasszikus módszerekkel ellentétben nem csak a mintaelemszámnál kisebb dimenziókban alkalmazható, valamint nem függ a minták eloszlásától sem, továbbá a hibák szórásáról való homoszkedaszticitási feltételt sem szüskéges feltennünk. A klasszikus és többváltozós szórásanalízisben adott K darab populációnk, ezekre vizsgáljuk a várható értékeik egyenlőségét, azaz nullhipotézisünk

$$H_0 \quad : \quad \mu_1 = \ldots = \mu_K, \tag{2.132}$$

ellenhipotézisünk

$$H_1 \quad : \quad \exists j \neq k \quad : \quad \mu_j \neq \mu_k, \tag{2.133}$$

ahol $\mu_i, i = 1, \ldots, K$ a populációk várható értékei vagy többváltozós esetben várhatóérték-vektorai, azaz

$$x_{1,1}, \ldots, x_{1,n_1} \sim N(\mu_1, \sigma^2) \tag{2.134}$$

$$x_{2,1}, \ldots, x_{2,n_2} \sim N(\mu_2, \sigma^2) \tag{2.135}$$

$$\vdots \tag{2.136}$$

$$x_{K,1}, \ldots, x_{K,n_K} \sim N(\mu_K, \sigma^2), \tag{2.137}$$

$$\tag{2.138}$$

és

$$\bar{x_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{i,j} \quad i = 1, \ldots, K \tag{2.139}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{K} \sum_{j=1}^{n_i} x_{i,j}. \tag{2.140}$$

Ekkor a klasszikus ANOVA esetén a teljes szóródást felbontjuk a következő módon:

$$\sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x})^2 = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (x_{i,j} - \bar{x_i})^2 + \sum_{i=1}^{K} n_i (\bar{x_i} - \bar{x})^2. \tag{2.141}$$

Itt a jobb oldal első tagját az angol nyelvű terminológiának megfelelően jelöljük SSE-vel, a másodikat SST-vel!

Rizzo és Székely kiterjesztik a klasszikus, akár többváltozós szórásanalízist egy általánosabb hipotézis vizsgálatára a teljes szóródás egy más, távolságon alapuló felbontás bevezetésével. Legyen most a mintánk K darab $\mathbb{R}^p$-beli független minta rendre az $F_1, ldots, F_K$ eloszlásfüggvényekkel megadott eloszlásokból. Legyen most a nullhipotézis

$$H_0 \quad : \quad F_1 = \ldots F_K, \tag{2.142}$$

az ellenhipotézis pedig

$$H_1 \quad : \quad \exists 1 \leq j < k \leq K : F_j \neq F_k. \tag{2.143}$$

A klasszikus ANOVA esetén a minták eloszlásairól feltettük a normalitást, most F_i eloszlásfüggvényekre semmilyen megkötést nem teszünk. A módszer lényege, hogy a teljes szóródást a klasszikus módszerben szereplő négyzetes eltérésekkel analóg távolságkomponensekre (DISCO az angol nyelvű terminológiából) bontjuk, amik alkalmasak lesznek 2.142 nullhipotézis tesztelésére. A távolságok most a mintaelemek közti euklideszi távolságokon alapulnak, pontosabban azok hatványain, ahol az α hatványkitevő a $(0, 2]$ intervallumból való. Az $\alpha = 2$ esetben visszakapjuk a klasszikus ANOVA felbontását, a többi esetben pedig az általánosabb nullhipotézisre konzisztens staisztikai próbát kapunk az ANOVA próbastatisztikájával analóg eszközzel.

2.21. definíció. Legyen $A = \{a_1, \ldots, a_{n_1}\}$ és $B = \{b_1, \ldots, b_{n_2}\}$ mintákra a d_α távolság

$$d_\alpha(A, B) = \frac{n_1 n_2}{n_1 + n_2}[2g_\alpha(A, B) - g_\alpha(A, A) - g_\alpha(B, B)], \tag{2.144}$$

ahol

$$g_\alpha(A, B) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} |a_i - b_j|^\alpha \tag{2.145}$$

Az $\alpha = 2$ esetben egydimenziós esetben a d_2 távolság pontosan a szórást adja vissza, és ha az előző minták átlagait rendre $\bar{a}$-val és $\bar{b}$-vel jelöljük,

$$d_2(A, B) = 2 \cdot SST = 2[n_1(\bar{a} - \bar{c})^2 + n_2(\bar{b} - \bar{c})^2], \tag{2.146}$$

ahol $\bar{c} = \frac{n_1\bar{a} + n_2\bar{b}}{n_1 + n_2}$.

A d_α távolságon alapuló statisztika, ami a szórásanalízisbeli SST-t helyettesíti, súlyozott összege a szórásoknak. Definiáljuk a minták közötti szóródást a következőképpen:

2.22. definíció. Legyenek $A_1, \ldots, A_K$ rendre $n_1, \ldots, n_K$ elemű, p-dimenziós minták, $N = \sum n_i$! Ekkor legyen kiegyensúlyozott, azaz $n_1 = n_2 = \ldots = n_K$ esetben

$$V_\alpha = V_\alpha(A_1, \ldots, A_K) = \frac{1}{K} \sum_{1 \leq j < k \leq K} d_\alpha(A_j, A_k), \tag{2.147}$$

kiegyensúlyozatlan esetben pedig

$$V_\alpha = V_\alpha(A_1, \ldots, A_K) = \sum_{1 \leq j < k \leq K} \Big(\frac{n_j + n_k}{2N}\Big) d_\alpha(A_j, A_k), \tag{2.148}$$

2.23. megjegyzés. $K = 2, p = 1, \alpha = 2$ esetben $V_2 = SST$.

Legyen a teljes szóródás

$$T_\alpha = T_\alpha(A_1, \ldots, A_K) = \frac{N}{2} g_\alpha(A, A), \tag{2.149}$$

ahol A az összevont minta, a mintán belüli szóródás pedig

$$W_\alpha = W_\alpha(A_1, \ldots, A_K) = \sum_{j=1}^{K} \frac{n_j}{2} g_\alpha(A_j, A_j) \tag{2.150}$$

Legyenek X és X' független azonos eloszlásúak és Y és Y' is független azonos eloszlásúak, amelyek X-től is függetlenek.

2.24. definíció. Legyen α olyan szám, amelyre $E|X|^\alpha + E|Y|^\alpha < \infty$, ekkor definiáljuk a az $\varepsilon(X, Y)$ távolságot X és Y között az alábbi módon:

$$\varepsilon(X, Y) = 2E|X - Y|^\alpha - E|X - X'|^\alpha - E|Y - Y'|^\alpha. \tag{2.151}$$

Ekkor Rizzo és Székely[17] nyomán a következő tételt mondhatjuk ki:

2.25. tétel. *Tegyük fel, hogy X és X' $\mathbb{R}^p$-beli független azonos eloszlású valószínűségi vektorváltozók F eloszlásfüggvénnyel, Y és Y' $\mathbb{R}^p$-beli független azonos eloszlású valószínűségi vektorváltozók G eloszlásfüggvénnyel és X és Y függetlenek. Ha $\alpha \in (0, 2]$ olyan, hogy $E|X|^\alpha + E|Y|^\alpha < \infty$, akkor*

1. *$\varepsilon(X, Y) \geq 0$.*
2. *Ha $0 < \alpha < 2$, akkor $\varepsilon(X, Y) = 0$ akkor és csak akkor, ha $X =^D Y$.*
3. *Ha $\alpha = 2$, akkor $\varepsilon(X, Y) = 0$ akkor és csak akkor, ha $E(X) = E(Y)$.*

2.26. következmény. *$K \geq 2$ és $\alpha \in (0, 2]$ esetén a következők igazak:*

1. *$V_\alpha(A_1, \ldots, A_K) \geq 0$.*
2. *Ha $0 < \alpha < 2$, akkor $V_\alpha(A_1, \ldots, A_K) = 0$ akkor és csak akkor, ha $A_1 = \ldots = A_K$.*
3. *$V_2(A_1, \ldots, A_K) = 0$ akkor és csak akkor, ha $A_i, i = 1, \ldots, K$ ugyanolyan várható értékűek.*

Bizonyítás. Legyen $A_j = \{a_1, \ldots, a_{n_j}\}$ és $B_j = \{b_1, \ldots, b_{n_k}\}$! Legyenek X és X' függetlenek és egyenletesek A_j-n, hasonlóan Y és Y' függetlenek és egyenletesek A_k-n! Ekkor

$$E|X - Y|^\alpha = g_\alpha(A_j, A_k), \tag{2.152}$$

$$E|X - X'|^\alpha = g_\alpha(A_j, A_j), \tag{2.153}$$

$$E|Y - Y'|^\alpha = g_\alpha(A_k, A_k), \tag{2.154}$$

$$\tag{2.155}$$

és

$$\frac{n_j n_k}{n_j + n_k}\epsilon_\alpha(X, Y) = d_\alpha(A_j, A_k). \tag{2.156}$$

Ekkor alkalmazva erre az esetre 2.25 tételt, $K = 2$ esetben megkapjuk a következmény első két állítását, innen teljes indukcióval $K > 2$-re is igaz. A 2.25 tétel harmadik pontjából pedig következik a harmadik következmény.

□

A fenti tulajdonságokból itt is látható analógia a klasszikus módszerekkel, így a fenti jó tulajdonságokkal rendelkező tagokra való felbontás jogossága is egyre jobban látszik. A felbontás létezéséről Rizzo és Székely[17] a következő tételt bizonyították, amely létjogosultságot ad a kiterjesztett esetre való módszernek.

2.27. tétel. *$K \geq 2$ esetén*

$$T_\alpha(A_1, \ldots, A_K) = V_\alpha(A_1, \ldots, A_K) + W_\alpha(A_1, \ldots, A_K). \tag{2.157}$$

2.28. megjegyzés. $p = 1$ esetben a korábban látottak és az előző tétel szerint $\alpha = 2$ mellett $T_2 = V_2 + W_2 = SST + SSE$, azaz megegyezik a klasszikus ANOVA módszerbeli felbontással.

Tegyük fel mostantól, hogy $A_1, \ldots, A_K$ rendre $n_1, \ldots, n_K$ elemű független véletlen minták az $X_1, \ldots, X_K$ változók eloszlásaiból! Az ANOVA-beli szórásalapú felbontáshoz hasonlóan V_α és W_α becslései $E|X_j - X_j'|^\alpha$-nak, $g_\alpha(A_j, A_j)$ is becslése ennek, de nem torzítatlan, a torzítatlanságot a klasszikus esethez hasonlóan az $\frac{n_j}{n_j - 1} g_\alpha(A_j, A_j)$ becsléssel érhetjük el. A kiterjesztett 2.142 nullhipotézis mellett

$$E(V_\alpha) = \frac{K-1}{2} E|X_j - X_j'|^\alpha, \tag{2.158}$$

$$E(W_\alpha) = \frac{N-K}{2} E|X_j - X_j'|^\alpha. \tag{2.159}$$

Ahogy a klasszikus ANOVA esetében a próbastatisztikánk

$$F = \frac{SST/(K-1)}{SSE/(N-K)}, \tag{2.160}$$

teljesen analóg módon itt most legyen

$$F_\alpha = \frac{V_\alpha/(K-1)}{W_\alpha/(N-K)}. \tag{2.161}$$

Speciálisan tudjuk, hogy 2.132 nullhipotézis mellett $F \sim F(K-1, N-K)$, de általánosságban F_α nem lesz F-eloszlású. Viszont F_α is nemnegatív, és nagy értékeire elvetjük a 2.142 nullhipotézist. Erre a korábbiakhoz hasonló határeloszlás-tulajdonságokat írhatunk fel. Tetszőleges $\alpha \in (0, 2)$ esetén a 2.142 nullhipotézis mellett $d_\alpha(A_j, A_j)$

tart egy 2.118 alakú kvadratikus alakhoz, ezért a nullhipotézis mellett $V_\alpha/(K-1)$ eloszlásban tart egy ilyen kvadratikus alakhoz, továbbá a nagy számok törvénye szerint $W_\alpha/(N-K)$ sztochasztikusan tart egy konstans számhoz. Ekkor tehát pontosan ugyanúgy, mint korábban $\frac{n\mathcal{V}_n^2}{S_2}$ a 2.14 tétel következményében,

$$F_\alpha = \frac{V_\alpha/(K-1)}{W_\alpha/(N-K)} \xrightarrow{D} Q, \tag{2.162}$$

ahol Q pontosan olyan, mint 2.118 pontban. Ez alapján a DISCO próbára is pontosan ugyanolyan állítást mondhatunk ki Székely és Bakirov[19] nyomán, mint a dCov alapú függetlenségvizsgálat esetén, tehát a DISCO próba elutasítja a 2.142 nullhipotézist, ha F_α az β értékhez tartozó jobb oldali kritikus tartományba esik, amelyet ismét $\forall 0 < \beta \leq 0.215, E(Q) = 1$:

$$P\big(Q \geq (\Phi^{-1}(1-\frac{\beta}{2}))^2\big) \leq \beta \tag{2.163}$$

alapján határozhatunk meg.

A kiterjesztett problémára adott távolságalapú módszer további előnye a konzisztencia, amelyet pontosabban az alábbi tételben lát be Rizzo és Székely[17]:

2.29. tétel. *Ha $0 < \alpha < 2$, a kiterjesztett nullhipotézisre vonatkozó DISCO próba konzisztens a véges szórású próbák körében.*

2.30. megjegyzés. Az $\alpha = 2$ esetben nem kapunk konzisztens próbát.

2.31. megjegyzés. Akkor is használható a fenti elv, ha akár az első momentumok sem léteznek. Tetszőleges eloszlásra, amelynek létezik ε-momentuma valamely $\varepsilon > 0$ számra, a $0 < \alpha < \frac{\varepsilon}{2}$ választás elégséges a konzisztenciához, mivel $E|X-Y|^{2\alpha} < \infty$.

A próbához felmerülő természetes kérdés, hogyan is kellene megválasztani az α paramétert. A legegyszerűbb választás $\alpha = 1$ eset, mivel egyrészt ez az alkalmas intervallum közepe, másrészt egyváltozós esetben $\alpha = 1$ mellett g_α linearizálható, és így az egyébként fennálló $O(N^2)$ számításigényt $O(N(\log N))$-re lehet csökkenteni. Vastag farkú eloszlások esetén az adatokon alapuló kicsi α értéket lehet választani. Például ha X Pareto-eloszlású, azaz sűrűségfüggvénye az x helyen

$$f(x) = \frac{k\sigma^k}{x^{k+1}}, x > \sigma, \tag{2.164}$$

akkor X véges első momentuma csak $k > 1$ esetén létezik, a második momentum pedig csak $k > 2$ mellett. Ebben az esetben a megfelelő Pareto-modell kiválasztása közben a k paraméter maximum likelihood becslésével határozható meg a megfelelő α érték, amelyre X^α véges szórású Pareto.

Vastag farkú stabil eloszlások esetén $\alpha < 1$ választás lehet indokolt.

Szimulációs eredmények

A többszempontú szórásanalízis (MANOVA) esetére a módszer hasonlóan kiterjeszthető (lásd [17]). Rizzo és Székely Monte Carlo szimulációkon keresztül hasonlította össze a távolságalapú próba tapasztalható erejét a Pillai[12] és Wilks[25] féle paraméteres MANOVA próbákkal, $\alpha = 1$ paraméter mellett. A próba nagyobb erőt mutat, és erőben nagyobb növekedést a dimenzió növelésével 4 szabadsági fokú t-eloszlás esetén, mint az említett MANOVA módszerek, továbbá lényeges különbség, hogy míg a szórásanalitikus tesztek csak akkor működnek, ha a dimenzió nem nagyobb a megfigyelések számánál, a DISCO teszt tetszőleges dimenzióban alkalmazható. A konstruált próba marginális gamma eloszlások esetén is jobban teljesít, mint a MANOVA módszerek, és hasonlóan szerepel a dimenzió növelésével is.

2.3.3. Torzítatlanság

A definiált összefüggőségi mérőszám sajnos nem torzítatlan, és a torzítás nőhet is a dimenzió függvényében (lásd [4]). Ennek tükrében definiálandó a távolságkorreláció torzítatlan becslése. Látható, hogy $E(\mathcal{V}_n^2(\mathbf{X},\mathbf{Y})) = \frac{n-1}{n^2}\big((n-2)\mathcal{V}^2(X,Y) + E|X - X'|_p E|Y - Y'|_q\big)$, ez alapján definiálhatjuk a következőket:

2.32. definíció. A korrigált tapasztalati távolságkovariancia legyen az a nemnegatív $U_n(\mathbf{X},\mathbf{Y})$ szám, amelyre

$$U_n^2(\mathbf{X},\mathbf{Y}) = \frac{n^2}{(n-1)(n-2)}\left[\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) - \frac{S_2}{n-1}\right], n \geq 3 \tag{2.165}$$

2.33. definíció. A korrigált tapasztalati távolságkorreláció az a nemnegatív $C_n(\mathbf{X},\mathbf{Y})$ szám, amelyre $n \geq 3$ esetén

$$C_n^2(\mathbf{X},\mathbf{Y}) = \begin{cases} \frac{U_n^2(\mathbf{X},\mathbf{Y})}{\sqrt{U_n^2(\mathbf{X},\mathbf{X})U_n^2(\mathbf{Y},\mathbf{Y})}}, & U_n^2(\mathbf{X},\mathbf{X})U_n^2(\mathbf{Y},\mathbf{Y}) \neq 0 \\ 0, & U_n^2(\mathbf{X},\mathbf{X})U_n^2(\mathbf{Y},\mathbf{Y}) = 0 \end{cases} \tag{2.166}$$

$n = 1, 2$ esetén legyen $C_n = 1$

Ekkor a fentiekre kimondható 2.14 tételhez hasonló határeloszlástétel, pontosabban annak következményeképpen látható, hogy

$$\frac{nU_n^2}{S_2} = \frac{n^2}{(n-1)(n-2)}\left[\frac{n\mathcal{V}_n^2(\mathbf{X},\mathbf{Y})}{S_2} - \frac{n}{n-1}\right] \underset{n\to\infty}{\overset{D}{\to}} \sum_{j=0}^{\infty} \lambda_j (Z_j^2 - 1), \tag{2.167}$$

ahol Z_i pontosan olyan, mint 2.118 kvadratikus alakban. Ezzel pedig már torzítatlan becslésként lehet számolni.

2.4. További lehetőségek és kiterjesztések

A következő fejezetben áttekintjük a távolságkorreláció néhány általánosítási lehetőségét és néhány egyéb módszert akár a távolságkorreláció tulajdonságainak javítására, akár más megközelítésére a korreláció javításának és a függetlenség karakterizálásának.

2.4.1. Más súlyfüggvény választása

A 2.1.2 fejezetben tárgyalt súlyfüggvény választását a 2.3 lemma indokolta, viszont egyáltalán nem tudunk olyat mondani, hogy ez lenne az egyetlen jó vagy esetleg a legjobb választás. Az alábbiakban áttekintünk egy korábbi eredményt egy lehetséges súlyfüggvényre, ami szintén jó tulajdonságokkal rendelkezik.

A 2.1 definícióban megadott normában a súlyfüggvényt nem feltétlenül kell tehát a 2.21 pontban megadottként választanunk, legyen most a norma súlyfüggvénye $q(u,v)$, ahol

$$q(u,v) = (|u|_p^2 + |v|_q^2)^{-\frac{1+p+q}{2}}. \tag{2.168}$$

Ekkor definiáljuk az $\mathcal{I}(X,Y)$ összefüggőségi mérőszámot hasonlóan a távolságkovarianciához, azaz

$$\mathcal{I}(X,Y) = \mathcal{R}_q(X,Y) = \frac{\|f_{X,Y}(u,v) - f_X(u)f_Y(v)\|_q}{\|\sqrt{(1-|f_X(u)|^2)(1-|f_Y(v)|^2)}\|_q} \tag{2.169}$$

Ezen kívül definiáljuk 2.2.2 fejezethez hasonlóan a tapasztalati mennyiségeket is:

Legyen $Z = (X,Y)$, $Z_j = (X_j, Y_j), j = 1, \ldots, n$ egy minta Z eloszlásából, és legyen $Z_{kl} = (X_{kl}, Y_{kl})$. Tegyük fel, hogy X és Y nem azonosan nullák, és egyik sem konstans majdnem mindenütt! Ekkor legyen

$$z_d = \frac{1}{n^2}\sum_{k,l=1}^{n} |Z_{kk} - Z_{ll}|_{p+q}, \tag{2.170}$$

$$z = \frac{1}{n^4}\sum_{k,l=1}^{n}\sum_{i,j=1}^{n} |Z_{kl} - Z_{ij}|_{p+q}, \tag{2.171}$$

$$x = \frac{1}{n^2}\sum_{k,l=1}^{n} |X_k - X_l|_p, \tag{2.172}$$

$$y = \frac{1}{n^2}\sum_{k,l=1}^{n} |Y_k - Y_l|_q, \tag{2.173}$$

$$\bar{z} = \frac{1}{n^3}\sum_{k=1}^{n}\sum_{i,j=1}^{n} |Z_{kk} - Z_{ij}|_{p+q}, \tag{2.174}$$

és ezekkel legyen

$$\mathcal{I}_n = \sqrt{\frac{2\bar{z} - z_d - z}{x + y - z}}. \tag{2.175}$$

Ezekre a következő tulajdonságokat bizonyítja Bakirov, Rizzo és Székely [1], hasonlóan a korábbiakhoz:

2.34. tétel. *1. Ha* $E(|X|_p + |Y|_q) < \infty$*, akkor majdnem mindenütt*

$$\lim_{n\to\infty} \mathcal{I}_n = \mathcal{I}. \tag{2.176}$$

2. $0 \leq \mathcal{I}_n \leq 1$

3. $p = q = 1$ *esetben, ha* X *és* Y *csak két értéket vesznek fel, akkor* $\mathcal{I}$ *megegyezik a klasszikus korreláció abszolútértékével.*

4. $\mathcal{I}$ *akkor és csak akkor, ha* X *és* Y *függetlenek.*

5. $\mathcal{I} = 1$ *akkor és csak akkor, ha létezik egy* A *véletlen halmaz és* a, b, c, d *konstans vektorok, amelyekre* $X = a + b\chi_A$ *és* $Y = c + d\chi_A$*, ahol* χ_A *az* A *halmaz indikátora.*

6. Ha X *és* Y *függetlenek és* $E(|X|_p + |Y|_q) < \infty$*, akkor*

$$\lim_{n\to\infty} n\mathcal{I}_n^2 \overset{D}{=} Q, \tag{2.177}$$

ahol Q *itt is a 2.118 pontbeli alakú, és* $E(Q) = 1$*.*

Néhány tulajdonság nyilvánvaló a definíciókból és a távolságalapú mérőszámokról látott korábbi tulajdonságokból, a többit a tapasztalati távolságkovarianciára vonatkozó tételekhez egészen hasonlóan lehet belátni. Fontos köztes eredmény, hogy $\mathcal{I}_n$ definícióját kapjuk vissza, ha $\mathcal{I}$-ben a karakterisztikus függvényeket lecseréljük a tapasztalati karakterisztikus függvényekre.

2.35. következmény. *A* $\sqrt{n}\mathcal{I}_n$ *próbastatisztika konzisztens próbát határoz meg a függetlenség vizsgálatára.*

A függetlenségvizsgálatra pedig hasonlót állíthatunk,mint a távolságkorreláció esetében, azaz

2.36. tétel. *Minden* $0 < \alpha \leq 0.215$ *paraméter esetén annak a próbának, amelyel utasítja a függetlenség hipotézisét*

$$\sqrt{n}\mathcal{I}_n \geq \Phi^{-1}\Big(1 - \frac{\alpha}{2}\Big) \tag{2.178}$$

esetén, az elérhető szignifikanciaszintje α*.*

A tétel bizonyítása pontosan ugyanúgy elvégezhető, mint 2.17 tétel esetén.

Szimulációs vizsgálatokkal látható(lásd [1]), hogy ez a fajta távolságalapú mérőszám hasonló jó tulajdonságokkal bír,mint a távolságkorreláció, ugyanis többváltozós normális esetben szintén közel lesz a Wilks-lambda teszthez, valamint sok esetben nagyobb

erőt ér el, mint a klasszikus próbák. A távolságkorrelációhoz hasonlóan tetszőleges dimenzióban alkalmazható, amit nem korlátoz a megfigyelések száma, mint más esetben, és teljesen konzisztens próbát ad.

Mint láthattuk, ezzel a súlyfüggvénnyel is elég jó tulajdonságú összefüggőségi mérőszámot definiálhatunk távolságalapon, amely sok gyakorlati esetben jobban teljesít, mint a klasszikus módszerek, bár mint láttuk, nem rendelkezik a távolságkorrelációhoz hasonló, a pearsoni korrelációval teljesen analóg felépítéssel, és lényegesen bonyolultabban kiszámítható próbastatisztikát eredményez függetlenségvizsgálat esetén, de például szintén rendelkezik a 2.1.2 fejezet elején említett skálainvarianciával és egyéb jó tulajdonságokkal.

2.37. megjegyzés. $\mathcal{I}_n$-nek használható olyan változata is, amely invariáns X és Y affin transzformációira, ha $\mathcal{I}_n$ definíciójába X_k és Y_k helyett rendre a $S_X^{-1/2}X_k$ és $S_Y^{-1/2}Y_k$ mennyiségeket helyettesítjük, ahol S_X és S_Y rendre a minták kovarianciamátrixai. Az így kapott mennyiségre a tételekben leírt tulajdonságok analóg módon fennállnak, mivel lényegében ez szintén egy általánosan tekinthető távolságalapú mérőszám a

$$(|S_X^{1/2}u|_p^2 + |S_Y^{1/2}v|_q^2)^{-\frac{1+p+q}{2}} \tag{2.179}$$

súlyfüggvénnyel.

2.4.2. Brown-kovariancia

2.38. definíció. Legyen X egy valós valószínűségi változó F_X eloszlásfüggvénnyel, $\{(U(t)\}_{t\in\mathbb{R}}$ pedig egy tőle független valósértékű sztochasztikus folyamat! Ekkor az X U-ra centrált változata

$$X_U = U(X) - \int_{-\infty}^{\infty} U(t)dF_X(t) = U(X) - E(U(X)|U), \tag{2.180}$$

ahol a feltételes várható érték létezik.

2.39. megjegyzés. Ha Id az identitás, azaz $t \in \mathbb{R} : Id(t) = t$, akkor $X_{Id} = X - EX$.

Legyen W egydimenziós Wiener-folyamat 0 várható értékkel és

$$|s| + |t| - |s - t| = 2\min(s, t), \quad t, s \geq 0 \tag{2.181}$$

kovarianciafüggvénnyel. Ez a függvény a standard Wiener kovarianciafüggvényének kétszerese, ami a későbbi számításokat könnyíti meg.

A klasszikus Pearson-féle kovariancia négyzete

$$E^2[(X - EX)(Y - EY)] = E[(X - EX)(X' - EX')(Y - EY)(Y' - EY')]. \tag{2.182}$$

Ennek a kifejezésnek az analógiájára bevezetünk egy általánosítást a kovarianciára Székely és Rizzo[20] nyomán. Először tekintsünk egy másik speciális esetet!

2.40. definíció. Legyenek X és Y valós valószínűségi változók! Ekkor a $\mathcal{W}(X,Y)$ Brown-kovariancián vagy Wiener-kovariancián azt a nemnegatív számot értjük, amelyre

$$\mathcal{W}^2(X,Y) = Cov_W^2(X,Y) = E[X_W X'_W Y_{W'} Y'_{W'}], \tag{2.183}$$

ahol W és W' 0 várható értékű és 2.181 kovarianciafüggvényű Wiener-folyamatok, és (W,W') független (X,X',Y,Y')-től.

2.41. megjegyzés. Amennyiben W Wiener-folyamat helyett a definícióban az Id identitás folyamatot használjuk, akkor

$$Cov_{Id}^2(X,Y) = E[X_{Id} X'_{Id} Y_{Id'} Y'_{Id'}] = Cov^2(X,Y), \tag{2.184}$$

azaz visszakapjuk a klasszikus kovariancia négyzetét.

Az itt definiált kovarianciáról látható, hogy két valós valószínűségi változó között minden lehetséges kapcsolat mérésére alkalmas, ellentétben a kizárólag a lineáris kapcsolatokat jól karakterizáló klasszikus módszerrel. A fenti definíciót tetszőleges dimenziójú sztochasztikus mezőkre általánosíthatjuk a következő módon:

2.42. definíció. Legyenek $X \in \mathbb{R}^p$ és $Y \in \mathbb{R}^q$ valószínűségi vektorváltozók, U egy sztochasztikus mező $\mathbb{R}^p$-n, amely független X-től, V egy sztochasztikus mező $\mathbb{R}^q$-n, amely független Y-tól! Ekkor X és Y (U,V)-kovarianciája az a nemnegatív szám, amelyre

$$Cov_{U,V}^2(X,Y) = E[X_U X'_U Y_V Y'_V], \tag{2.185}$$

ahol a várható érték létezik és véges.

Ezzel tehát speciálisan visszakapjuk a klasszikus definíciót is, és a Brown-kovarianciát is a megfelelő tereken értelmezett független folyamatok segítségével.

Teljesen hasonlóan értelmezhetünk itt is szórásnégyzet-analóg mennyiséget

$$\mathcal{W}(X) = \mathcal{W}(X,X) = Cov_W(X,X) = Var_W(X) \tag{2.186}$$

módon. Ezzel pedig definiálhatunk a korrelációnak megfelelő mérőszámot:

2.43. definíció. Legyen az X és Y változók Brown-korrelációja az a nemnegatív szám, amelyre

$$Cor_W(X,Y) = \begin{cases} \frac{\mathcal{W}(X,Y)}{\sqrt{\mathcal{W}(X)\mathcal{W}(Y)}}, & \mathcal{W}(X)\mathcal{W}(Y) \neq 0, \\ 0, & \mathcal{W}(X)\mathcal{W}(Y) = 0. \end{cases} \tag{2.187}$$

A definiált mennyiségekről a következőket mondhatjuk továbbra is Székely és Rizzo [20] alapján:

2.44. tétel. *Ha $X \in \mathbb{R}^p$ és $Y \in \mathbb{R}^q$ valószínűségi vektorváltozókra $E(|X|_p^2+|Y|_q^2) < \infty$, W és W' független, rendre $\mathbb{R}^p$ és $\mathbb{R}^q$-beli, a megfelelő dimenzióbeli euklideszi normákban megadott 2.181 pontbeli kovarianciafüggvényű és 0 várhatóértékű Wiener-folyamatok, akkor $E[X_W X'_W Y_{W'} Y'_{W'}]$ nemnegatív, véges, és*

$$\mathcal{W}^2(X,Y) = E[X_W X'_W Y_{W'} Y'_{W'}] = \tag{2.188}$$

$$E|X-X'|_p|Y-Y'|_q + E|X-X'|_p|Y-Y'|_q - \tag{2.189}$$

$$-E|X-X'|_p|Y-Y''|_q - E|X-X''|_p|Y-Y'|_q, \tag{2.190}$$

ahol $(X,Y),(X',Y')$ és (X'',Y'') függetlenek.

Ezzel a tétellel pedig beláthatjuk a következő lényeges tételt:

2.45. tétel. *Ha $X \in \mathbb{R}^p$ és $Y \in \mathbb{R}^q$ vektorváltozókra $E(|X|_p^2 + |Y|_q^2) < \infty$, akkor*

$$\mathcal{W}(X,Y) = \mathcal{V}(X,Y). \tag{2.191}$$

Bizonyítás. A Brown-kovariancia és a dCov is nemnegatív, így elegendő a négyzeteik egyenlőségét megmutatni. $\mathcal{V}^2$ kifejezéséhez használjuk a 2.3 lemmát! Ekkor az integrálalak számlálójában

$$E[\cos\langle u, X-X'\rangle \cos\langle v, Y-Y'\rangle] \tag{2.192}$$

alakú tagokat kapunk. Ekkor

$$\cos u \cos v = 1 - (1-\cos u) - (1-\cos v) + (1-\cos u)(1-\cos v) \tag{2.193}$$

azonosság szerint átírva a várhatóértékbeli koszinuszok szorzatát, a kapott tényezők alakja

$$E\int_{R^{p+q}} \frac{[1-\cos\langle u, X-X'\rangle][1\cos\langle v, Y-Y'\rangle]}{|u|_p^{p+1}|v|_q^{q+1}} du dv = \tag{2.194}$$

$$= E\left[\int_{R^p} \frac{[1-\cos\langle u, X-X'\rangle]}{|u|_p^{p+1}} du \cdot \int_{\mathbb{R}^q} \frac{[1\cos\langle v, Y-Y'\rangle]}{|v|_q^{q+1}}\right] = \tag{2.195}$$

$$= c_p c_q E|X-X'|_p E|Y-Y'|_q \tag{2.196}$$

Tehát a megfelelő tagokat véve egyszerűsítés után

$$\mathcal{V}^{\in}(X,Y) = E|X-X'|_p|Y-Y'|_q + E|X-X'|_p|Y-Y'|_q - \tag{2.197}$$

$$-E|X-X'|_p|Y-Y''|_q - E|X-X''|_p|Y-Y'|_q, \tag{2.198}$$

ami pontosan a 2.44 tételbeli alakja $\mathcal{W}^2(X,Y)$-nak. □

Tehát véges szórás esetén a Brown-kovariancia megegyezik a távolságkovarianciával, így a definiált (U,V)-kovariancia általánosítása a távolságkovarianciának is ilyen értelemben. Speciális esetként kaphatjuk tehát ebből Wiener-folyamatok választásával, valamint $p = q = 1$ mellett és identitásfolyamatok mellett megkaphatjuk a klasszikus Pearson-féle kovariancia abszolútértékét.

2.4.3. Más paraméter választása

A távolságkovariancia definiálásánál a 2.3 lemma alapján választottunk súlyfüggvényt, azt is abban a speciális esetben használva, amikor $\alpha = 1$. A kapott eredményeket általánosíthatjuk ezáltal oly módon, hogy nem kötjük meg ennek az α paraméternek az értékét, ezzel meghatározva a távolságalapú összefüggőségi mérőszámok egy paraméteres családját. Tekintsük tehát a következő definíciókat:

2.46. definíció. *$E(|X|_p^\alpha + |Y|_q^\alpha) < \infty$ esetén az X és Y valószínűségi vektorváltozók α-távolságkovarianciáján azt a nemnegatív $\mathcal{V}(X,Y)$ számot értjük, amelyre*

$$\mathcal{V}^{2(\alpha)}(X,Y) = \|f_{X,Y}(u,v) - f_X(u)f_Y(v)\|_\alpha^2 = \tag{2.199}$$

$$= \frac{1}{C(p,\alpha)C(q,\alpha)} \int_{\mathbb{R}^{p+q}} \frac{|f_{X,Y}(u,v) - f_x(u)f_Y(v)|^2}{|u|_p^{\alpha+p}|v|_q^{\alpha+q}} dudv. \tag{2.200}$$

Teljesen hasonlóan a klasszikus kovarianca esetéhez, itt is definiálható a szórásnégyzet analógiájára

$$\mathcal{V}^{2(\alpha)}(X) = \mathcal{V}^{2(\alpha)}(X,X) = \|f_{X,X}(u,v) - f_X(u)f_X(v)\|_\alpha^2 \tag{2.201}$$

2.47. definíció. *$E(|X|_p^\alpha + |Y|_q^\alpha) < \infty$ esetén az X és Y valószínűségi vektorváltozók α-távolságkorrelációján azt a nemnegatív $\mathcal{R}(X,Y)$ számot értjük, amelyre*

$$\mathcal{R}^{2(\alpha)}(X,Y) = \begin{cases} \frac{\mathcal{V}^{2(\alpha)}(X,Y)}{\sqrt{\mathcal{V}^{2(\alpha)}(X)\mathcal{V}^{2(\alpha)}(Y)}}, & \mathcal{V}^{2(\alpha)}(X)\mathcal{V}^{2(\alpha)}(Y) > 0 \\ 0, & \mathcal{V}^{2(\alpha)}(X)\mathcal{V}^{2(\alpha)}(Y) = 0. \end{cases} \tag{2.202}$$

Ezekre a mennyiségekre analóg módon definiálhatóak a tapasztalati mennyiségek is, egyszerűen az euklideszi norma helyett mindenütt annak α kitevőjű hatványát véve, azaz $a_{kl} = |X_k - X_l|_p^\alpha$ és $b_{kl} = |Y_k - Y_l|_q^\alpha$ mellett teljesen hasonlóan 2.6 felépítéséhez. Erre az általánosított mérőszámra a tulajdonságok jelentős részét is hasonlóan tudjuk általánosítani a $\|\cdot\|$ norma $\|\cdot\|_\alpha$ normára cserélésével, $\alpha \in (0,2)$ mellett. Tehát 2.10 és 2.14 tételekhez hasonlóan, az előző jelölésekkel analóg módon

2.48. tétel. *Ha $E|X|_p^\alpha < \infty$ és $E|Y|_q^\alpha < \infty$, akkor majdnem mindenütt*

$$\lim_{n\to\infty} \mathcal{V}_n^{(\alpha)}(\mathbf{X},\mathbf{Y}) = \mathcal{V}^{(\alpha)}(X,Y). \tag{2.203}$$

2.49. tétel. *Ha $X \in \mathbb{R}^p$ és $Y \in \mathbb{R}^q$ vektorváltozók függetlenek, és $0 < \alpha < 2$ mellett $E(|X|_p^\alpha + |Y|_q^\alpha) < \infty$, akkor*

$$n\mathcal{V}_n^{2(\alpha)} \underset{n\to\infty}{\overset{D}{\to}} \|\zeta(u,v)\|_\alpha^2. \tag{2.204}$$

Ezeken kívül pedig $0 < \alpha < 2$, $E(|X|_p^\alpha + |Y|_q^\alpha) < \infty$ esetben a konzisztencia is megmarad. Az $\alpha = 2$ esetben, ha $p = q = 1$, akkor $\mathcal{R}^{(2)} = |\rho|$ és $\mathcal{V}_n^{(2)} = 2|\hat{\sigma}_{xy}|$, ahol $\hat{\sigma}_{xy}$ a $Cov(X, Y)$ maximum-likelihood becslése.

Ugyanezt az általánosítást megkaphatjuk a 2.4.2 fejezetben leírt Brown-kovariancia kiterjesztéseként is, az úgynevezett Hurst-indexszel való paraméterezéssel. Ha veszünk egy frakcionált Wiener-folyamatot, azaz egy 0 várhatóértékű Gauss-folyamatot a

$$E(W_H(u)W_H(v)) = \frac{1}{2}\big(|u|^{2H} + |v|^{2H} - |u - v|^{2H}\big) \tag{2.205}$$

kovarianciafüggvénnyel, akkor a $H \in (0, 1)$ indexet nevezzük Hurst-indexnek. Erre a folyamatra $H = \frac{1}{2}$ esetén visszakapjuk a standard Wiener-folyamatot, $H > \frac{1}{2}$ esetén olyan folyamatot kapunk, amelyre a növekmények pozitívan korreláltak, $H < \frac{1}{2}$ esetén pedig a növekmények negatívan korreláltak. Ezt használva most a Brown-kovarianciát definiáló d-dimenziós Wiener-folyamat, azaz sztochasztikus mező esetén legyen a kovarianciafüggvény $u, v \in \mathbb{R}^d$ mellett

$$E(W_H^d(u)W_H^d(v)) = |u|^{2H} + |v|^{2H} - |u - v|^{2H}. \tag{2.206}$$

Ilyen módon ha (W, W') és (X, X', Y, Y') függetlenek, akkor $0 < H, H^* < 1$ és $E(|X|_p^{4H} + |Y|_q^{4H^*}) < \infty$ mellett a korrábiakhoz hasonlóan

$$Cov^2_{W_H^p, W_H'^q}(X, Y) = \tag{2.207}$$

$$E|X - X'|_p^{2H}|Y - Y'|_q^{2H^*} + E|X - X'|_p^{2H}|Y - Y'|_q^{2H^*} - \tag{2.208}$$

$$-E|X - X'|_p^{2H}|Y - Y''|_q^{2H^*} - E|X - X''|_p^{2H}|Y - Y'|_q^{2H^*}. \tag{2.209}$$

Ebből látható, hogy $2H$ és $2H^*$ teljesen analóg az α szerepével.

2.4.4. Kiterjesztés tetszőleges normált terekre és Hilbert-terekre

A távolságkorreláció definíciójában most nem a súlyfüggvényt vagy annak paraméterezését változtatjuk meg, hanem az euklideszi normák helyett általánosabb esetet véve kiterjesztjük a definíciót Hilbert-terekre a véges dimenziós euklideszi terekről Kosorok[9] nyomán. Legyen teljesen analóg módon a 2.8 tétel bizonyításvázlatában szereplő S_1, S_2, S_3 definícióival X és Y valváltozókra, amelyek most tetszőleges normált terekből valók, rendre a $\|\cdot\|_X$ és $\|\cdot\|_Y$ normákkal,

$$T_1 = \frac{1}{n^2}\sum_{k,l=1^n} \|X_k - X_l\|_X \|Y_k - Y_l\|_Y, \tag{2.210}$$

$$T_2 = \frac{1}{n^2}\sum_{k,l=1^n} \|X_k - X_l\|_X \cdot \frac{1}{n^2}\sum_{k,l=1^n} \|Y_k - Y_l\|_Y, \tag{2.211}$$

$$T_3 = \frac{1}{n^3}\sum_{k=1}^{n}\sum_{l,m=1}^{n} \|X_k - X_l\|_X \|Y_k - Y_m\|_Y, \tag{2.212}$$

és ezekkel legyen $V_n(X,Y) = T_1 + T_2 - 2T_3$, valamint legyen

$$T_{10} = E\big(\|X_1 - X_2\|_X \|Y_1 - Y_2\|_Y\big), \tag{2.213}$$

$$T_{20} = E\big(\|X_1 - X_2\|_X\big) E\big(\|Y_1 - Y_2\|_Y\big), \tag{2.214}$$

$$T_{30} = E\big(\|X_1 - X_2\|_X \|Y_1 - Y_3\|_Y\big), \tag{2.215}$$

és $V_0(X,Y) = T_{10} + T_{20} - 2T_{30}$, valamint $V_n(X) = V_n(X,X)$ és $V_0(X) = V_0(X,X)$! Ezekkel definiáljuk R_n és R_0 mennyiségeket analóg módon a korrelációs definíciókhoz. Ekkor természetesen adódik a következő:

2.50. lemma. *Ha* $E(\|X\|_X^4 + \|Y\|_Y^4) < \infty$*, akkor*

$$V_n(X,Y) \xrightarrow{P} V_0(X,Y), \tag{2.216}$$

$$V_n(X) \xrightarrow{P} V_0(X). \tag{2.217}$$

$$\tag{2.218}$$

A távolságkorreláció tulajdonságait is szeretnénk megőrizni értelemszerűen a kiterjesztéssel, viszont azokat véges dimenzióban láttuk, euklideszi normákkal. Most cseréljük ki az euklideszi normákat $\|x\|_{A,p} = \sqrt{x^T A x}$ és $\|y\|_{B,q} = \sqrt{y^T B y}$ normákra, ahol $A \in \mathbb{R}^{p\times p}$ és $B \in \mathbb{R}^{q\times q}$ szimmetrikus pozitív definit mátrixok. Az ezekkel a normákkal kapott mennyiségeket jelöljük ˜ jellel, azaz $\tilde{T}_1, \tilde{T}_2, \ldots$! Ekkor ezekre a következőket mondhatjuk:

2.51. lemma. *Legyen A és B szimmetrikus és pozitív definit! Ekkor*

1. $\tilde{V}_n(X,Y)$ *és* $\tilde{V}_0(X,Y)$ *nemnegatívak*
2. $\tilde{V}_n(X,Y) \le \sqrt{\tilde{V}_n(X)\tilde{V}_n(Y)}$ *és* $\tilde{V}_0(X,Y) \le \sqrt{\tilde{V}_0(X)\tilde{V}_0(Y)}$
3. $\tilde{R}_n = 0$ *akkor és csak akkor, ha az X-ből vagy Y-ból származó minta konstans*
4. $\tilde{R}_0(X,Y) = 0$ *akkor és csak akkor, ha X és Y függetlenek*
5. $\tilde{V}_0(X,Y) = 0$ *akkor és csak akkor, ha* $V_0(X,Y) = 0$.

Bizonyítás. Legyen $U = A^{\frac{1}{2}}X$ és $V = B^{\frac{1}{2}}Y$, ekkor $|U|_p = \|X\|_{A,p}$ és $|V|_q = \|Y\|_{B,q}$. Ekkor a tapasztalati távolságkovariancia tulajdonságai alapján, ha U-ra és V-re alkalmazzuk a korábbiakat, az első négy pont teljesül. Az utolsó állításban pedig mivel $A^{\frac{1}{2}}$ és $B^{\frac{1}{2}}$ pozitív definitek, U és V pontsan akkor független, ha X és Y független, az állítás igaz. □

A tévolságkorrelációt és tulajdonságait úgy szeretnénk kiterjeszteni, hogy véges dimenziós euklideszi terek sorozataival approximálható Hilbert-tereket veszünk. Tegyük fel, hogy X egy valószínűségi változó a H_X Hilbert-térben, amelyen a skaláris szorzatot jelölje $\langle\cdot,\cdot\rangle_X$, a megfelelő normát pedig $\|\cdot\|_X$!

2.52. definíció. Az X valószínűségi változó végesen approximálható, ha létezik

$$\{X_m\}_{m\geq 1} \in H_X \tag{2.219}$$

sorozat, hogy létezik $M_m : H_X \mapsto \mathbb{R}^{p_m}$ lineáris leképezés, amelyre p_m monton növő és $M_m^* M_m$ szimmetrikus és pozitív definit $\mathbb{R}^{p_m}$-en, amelyre

$$X_m = M_m(U_m) \tag{2.220}$$

valamilyen euklideszi altérből származó U_m valószínűségiváltozó-sorozatra úgy, hogy

$$E\|X_m - X\|_X^2 \underset{m\to\infty}{\to} 0. \tag{2.221}$$

Feltehetjük az általánosság megszorítása nélkül, hogy $M_m^* M_m$ az identitás. Ekkor a következőket mondhatjuk:

2.53. lemma. *Legyen X és Y végesen approximálható valószínűségi változó valamilye Hilbert-térben! Ekkor*

1. *$V_n(X,Y)$ és $V_0(X,Y)$ nemnegatívak*
2. *$V_n(X,Y) \leq \sqrt{V_n(X)V_n(Y)}$ és $V_0(X,Y) \leq \sqrt{V_0(X)V_0(Y)}$*

Bizonyítás. Legyen X_m és Y_m a két végesen approximáló sorozat, azaz $E\|X_m - X\|_X^2 \to_{m\to\infty} 0$ és $E\|Y_m - Y\|_Y^2 \to_{m\to\infty} 0$. Ekkor $V_0(X_m, Y_m) \to V_0(X,Y)$, tehát mivel a közelítő sorozatokra igazak az állítások a korábbiak szerint, $V_0(X,Y) \geq 0$. Teljesen ugyanígy látható a V_0-ra vonatkozó többi állítás is.

Legyen egy (X_i, Y_i) n-elemű mintánk! Ekkor létezik $\{(X_{im}, Y_{im})\}_{i=1,\dots,n}$ sorozat, hogy

$$\sum_{i=1}^{n} (E\|X_i - X_{im}\|_X^2 + E\|Y_i - Y_{im}\|_Y^2) \to 0 \tag{2.222}$$

a végesen approximálhatóság miatt. Tekintsük $V_n(X,Y)$ helyett $V_N^{(m)}(X,Y)$-t, ahol a megfigyeléseket az m-edik közelítéssel helyettesítjük! Ekkor mivel a várhatóértékbeli konvergenciából következik a sztochasztikus,

$$V_n^{(m)}(X,Y) \underset{m\to\infty}{\overset{P}{\to}} V_n(X,Y), \tag{2.223}$$

és a közelítő sorozattal igazak az állítások, így $V_n(X,Y)$-ra is igaz a lemma. □

A kiterjesztés akkor lenne igazán jó, ha a függetlenség jó karakterizációja is megmaradna a bővebb terekre. Hilbert terekre sajnos csak speciális esetben tudjuk ezt mondani, de az alábbi lemma általánosan is igaz.

2.54. lemma. *Legyen X és Y valószínűségi változók végesen approximálható Hilbert-terekben! Ekkor ha $R_0(X,Y) > 0$, akkor X és Y összefüggőek.*

Bizonyítás. Tegyük fel indirekt, hogy a feltételek teljesülnek, de X és Y függetlenek! Ekkor a végesen approximálhatóság miatt létezik (X_m, Y_m) sorozat úgy, hogy X_m és Y_m függetlenek, $E\|X_m - X\|_X^2 \to_{m\to\infty} 0$ és $E\|Y_m - Y\|_Y^2 \to_{m\to\infty} 0$. Ekkor $R_0(X_m, Y_m) = 0$ és $R_0(X_m, Y_m) \to R_0(X_m, Y_m)$ miatt ellentmondásra jutottunk. □

Ezen túl viszont egyelőre nem látjuk, hogy a kiterjesztett távolságkovariancia akkor és csak akkor lehet 0, ha X és Y függetlenek. Speciális esetben viszont mondhatunk valamit erről az esetről Kosorok[9] példája nyomán.

2.55. definíció. Legyenek X és Y végesen approximálható Hilbert-terekből vett valószínűségi változók. Legyen $M : H_X \mapsto H_X$ és $N : H_Y \mapsto H_Y$ lineáris leképezések, amelyekre M^*M és N^*N is a megfelelő identitás, és $MX = X_1 + X_2$, $NY = Y_1 + Y_2$, ahol $i = 1, 2$ $X_i \in H_X^{(i)}, Y_i \in H_Y^{(i)}$, és $H_X^{(i)}$ és $H_Y^{(i)}$ rendre véges dimenziós alterei H_X-nek és H_Y-nak, valamint X_2 és Y_2 függetlenek és (X_1, Y_1)-től is függetlenek! Ekkor azt mondjuk, hogy X és Y legfeljebb végesen összefüggőek.

2.56. lemma. *Legyenek X és Y legfeljebb végesen összefüggő valószínűségi változók! Ekkor $R_0(X, Y) \geq 0$ és az egyenlőtlenség akkor és csak akkor szigorú, ha X és Y összefüggőek.*

Az általános esetre való megfontolásokhoz Lyons[10] kiterjesztésén keresztül juthatunk el, először tetszőleges metrikus terekre belátva az általánosítás jó tulajdonságait. Legyen ehhez először $(\mathcal{X}, d)$ metrikus tér, amelyen jelölje $M(\mathcal{X})$ a véges Borel-mértékeket, és $M_1(\mathcal{X})$ ebben a valószínűségi mértékek halmazát! Ekkor $\mu \in M(\mathcal{X})$ első momentuma véges, ha valamely $y \in \mathcal{X}$ mellett $\int_{\mathcal{X}} d(o, x) d|\mu|(x) < \infty$ teljesül. Ha μ és μ' első momentuma véges, akkor

$$\int d(x, x') d(|\mu| \times |\mu'|)(x, x') < \infty, \tag{2.224}$$

$$\int d(x, x') d|\mu|(x) d|\mu'|(x') < \infty. \tag{2.225}$$

Ekkor legyen

$$a_\mu(x) = \int d(x, x') d\mu(x), \tag{2.226}$$

$$D(\mu) = \int d(x, x') d\mu^2(x, x'). \tag{2.227}$$

Ekkor ezek léteznek és végesek, ha μ első momentuma véges, és így legyen ekkor

$$d_\mu(x, x') = d(x, x') - a_\mu(x) - a_\mu(x') + D(\mu). \tag{2.228}$$

Ekkor ha tetszőleges metrikus térben μ valószínűségi mértéknek véges az első momentuma, azaz $d(x, x') \in L^1(\mu \times \mu)$, akkor $d_\mu(x, x') \in L^2(\mu \times \mu)$.

Legyen most $(\mathcal{Y}, d)$ egy másik metrikus tér, és legyen $\theta \in M_1(\mathcal{X} \times \mathcal{Y})$ olyan valószínűségi mérték, amely $\mathcal{X}$-beli μ és $\mathcal{Y}$-beli ν marginálisai véges első momentumúak! Ekkor az általánosítást a fenti mennyiségekkel elvégezhetjük lényegében úgy, mint a tapasztalati mennyiségeknél definiáltak folytonos és általánosított esetei, így építve fel az analógiát a távolságkorrelációval.

2.57. definíció.

$$dcov(\theta) = \int \delta_\theta\big((x,y),(x',y')\big)d\theta^2\big((x,y),(x',y')\big), \tag{2.229}$$

ahol

$$\delta_\theta\big((x,y),(x',y')\big) = d_\mu(x,x')d_\nu(y,y'). \tag{2.230}$$

2.58. megjegyzés. Látható a definícióból, hogy az L^2-beliség miatt $dcov(\theta)$ létezik, továbbá az is, hogy ha θ valószínűségi mérték a marginálisok szorzata, akkor $dcov(\theta) = 0$, a megfordítás viszont általában nem igaz.

2.59. megjegyzés. A fenti definíció euklideszi terekkel és a definiált távolságfogalmakkal pontosan a 2.4 definícióbeli dCov négyzetét adja.

2.60. megjegyzés. Ha az X és Y valószínűségi változók együttes eloszlása θ, akkor legyen jelölésben $dcov(X,Y) = dcov(\theta)$, és ekkor látható, hogy $dcov(\theta)$ felírható a

$$E((d(X,X') - a_\mu(X) - a_\mu(X') + D(\mu))(d(Y,Y') - a_\nu(Y) - a_\nu(Y') + D(\nu))) \tag{2.231}$$

alakban.

Ekkor $|dcov(X,Y)| \leq \sqrt{dcov(X,X)dcov(Y,Y)} \leq D(\mu)D(\nu)$, és $dcov(X,X) = D(\mu)$ akkor és csak akkor, ha μ legfeljebb két pontra koncentrált mérték, így hasonlóan a véges dimenziós euklideszi terekhez, itt is tudunk korrelációhoz hasonló mérőszámot deifniálni. Ehhez hasonlóan más, korábbi tulajdonságok is kiterjednek az így általánosított mennyiségre Lyons[10] nyomán.

2.61. tétel. *1. Ha $dcov(X,X) = 0$, akkor X eloszlása egyetlen pontra koncentrált mérték*

2. Ha μ és ν nem csak egy pontra koncentráltak és

$$|dcov(X,Y)| = \sqrt{dcov(X,X)dcov(Y,Y)}, \tag{2.232}$$

akkor a $c > 0$ számra létezik $f : (X) \to (Y)$ folytonos leképezés, hogy minden x, x'-ra a μ tartójából

$$d(x,x') = cd(f(x),f(x')), \tag{2.233}$$

$$y = f(x) \tag{2.234}$$

θ-majdnem minden (x,y)-ra.

Ezen kívül szeretnénk az általánosított esetben is függetlenségvizsgálati próbát konstruálni, ehhez Lyons a határeloszlásokra vonatkozó tételeket is kiterjeszti metrikus terekre, például 2.10 tétel analógiájára. A továbbiakban feltételezzük, hogy μ és ν nem egy pontra koncentrált mértékek.

2.62. tétel. *Legyenek $\mathcal{X}$, $\mathcal{Y}$, θ mint eddig, és legyen θ_n a tapasztalati eloszlása az első n mintának egy független azonos θ-eloszlású sorozatból! Ekkor $dcov(\theta_n) \to dcov(\theta)$ majdnem mindenütt.*

Hasonlóan kiterjeszthető a 2.14 tétel is.

2.63. tétel. *Legyenek a λ_i számok a következő leképezés multiplicitással vett sajátértékei:*

$$L^2(\theta) \ni F \mapsto \Big((x,y) \mapsto \int \delta_\theta\big((x,y)(x',y')\big)F(x',y')d\theta(x',y')\Big). \tag{2.235}$$

Ekkor a függetlenség hipotézise mellett

$$n \cdot dcov(\theta_n) \xrightarrow{D} \sum_i \lambda_i Z_i^2, \tag{2.236}$$

ahol az euklideszi térhez hasonlóan Z_i független azonos standard normálisok, és $\sum_i \lambda_i = D(\mu)D(\nu)$.

2.64. következmény. *Legyenek μ_n és ν_n a θ_n marginálisai! Ekkor a függetlenség hipotézise mellett*

$$\frac{n \cdot dcov(\theta_n)}{D(\mu_n)D(\nu_n)} \xrightarrow{D} \frac{\sum_i \lambda_i Z_i^2}{D(\mu)D(\nu)}, \tag{2.237}$$

ahol a jobb oldal várható értéke 1. Ha $dcov(\theta) \neq 0$, akkor a bal oldal $\pm\infty$-hez tart majdnem mindenütt.

Továbbra is az a kérdés maradt nyitott, mikor lesz igaz, hogy $dcov(\theta) = 0$ esetből következik $\theta = \mu \times \nu$. Ezt az esetet az előző tételek sem fedik le. Lyons megmutatta, hogy úgynevezett erősen negatív típusú metrikus terek esetében sosem áll elő az az eset, hogy $dcov(\theta) = 0$ és $\theta \neq \mu \times \nu$, más esetekben viszont igen.

2.65. definíció. Legyen $n \geq 1$ mellett $x_1, \ldots, x_n \in \mathcal{X}$ és $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, ahol $\sum \alpha_i = 0$. Ekkor azt mondjuk, hogy $(\mathcal{X}, d)$ metrikus tér negatív típusú, ha

$$\sum_{i,j \leq n} \alpha_i \alpha_j d(x_i, x_j) \leq 0. \tag{2.238}$$

2.66. megjegyzés. Legyen K az az $n \times n$-es távolságmátrix, amelyre $K_{i,j} = d(x_i, x_j)$! Ekkor a negatív típusra vonatkozó egyenlőtlenség pontosan azt jelenti, hogy K feltételesen negatív szemidefinit. Ekkor ha P az $\mathbb{R}^n$-ről való merőleges vetítés a konstans

vektorok ortogonális kiegészítő alterére, akkor $\bar{K} = PKP$ az a mátrix, amelyre $\bar{K}_{i,j} = d_{\mu_n}(x_i, x_j)$, ahol μ_n az $x_1, \ldots, x_n$ tapasztalati eloszlása. Ekkor $\bar{K}$ is negatív szemidefinit.

Ha $\mathcal{X}$ és $\mathcal{Y}$ negatív típusú metrikus terek d-vel, és a vonatkozó távolságmátrixaik rendre K és L, akkor

$$0 \leq tr(\bar{K}\bar{L}) = tr\big(\sqrt{-\bar{K}}\sqrt{-\bar{L}}\sqrt{-\bar{L}}\sqrt{-\bar{K}}\big) = n^2 \cdot dcov(\theta_n) \tag{2.239}$$

Legyen most $\mu_1, \mu_2 \in M_1(\mathcal{X})$ két véges első momentumú mérték! Ekkor ezek véges tartójú valószínűségi mértékekkel való közelítésekor belátható, hogy ha $\mathcal{X}$ negatív típusú, akkor

$$D(\mu_1 - \mu_2) \leq 0. \tag{2.240}$$

2.67. definíció. Az $(\mathcal{X}, d)$ metrikus tér erősen negatív típusú, ha negatív típusú, és a 2.240 egyenlőtlenségben csak akkor áll fenn egyenlőség, ha $\mu_1 = \mu_2$.

A tetszőleges metrikus térre való kiterjesztésen túl szeretnénk a definíciókat és tulajdonságaikat Hilbert-terekre is kiterjeszteni, ezt a meglévő terek és metrikák beágyazásával érjük el. Az általánosság megszorítása nélkül tekinthetünk most csak valós Hilbert-tereket. Belátható, hogy $\mathcal{X}$ akkor és csak akkor negatív típusú, ha létezik H Hilbert-tér és $\phi : \mathcal{X} \to H$ leképezés, hogy $\forall x, x' \in \mathcal{X}\ \ d(x, x') = \|\phi(x) - \phi(x')\|^2$, továbbá tudjuk, hogy H szeparábilis, ha $\mathcal{X}$ az. Lyons példaként említi az ilyen beágyazásokra a Riesz-, Fourier-, Brown- és Crofton-beágyazást, ahol az első kettővel például belátható $\mathbb{R}^n$ negatív típusúsága, és a Fourier-beágyazás a 2.3 lemma szerinti távolságot adja majd euklideszi terek esetében, a Brown-beágyazás pedig hasonlóan az euklideszi terekben a Brown-kovariancia esetét. Két tetszőleges $\phi_1, \phi_2 : (\mathcal{X}, d^{\frac{1}{2}}) \to H$ izometrikus beágyazást ekvivalensnek tekinthetünk, ha létezik $g : H_1 \to H_2$ izometria, amellyel $\phi_2 = g \circ \phi_1$, ahol H_i a ϕ_i leképezés képterének lezártja.

Tekintsünk most egy ilyen ϕ beágyazást! Ekkor legyen

$$\beta_\phi : \mu \mapsto \int \phi(x) d\mu(x), \tag{2.241}$$

ahol $\mu \in M(\mathcal{X})$ mértéknek véges az első momentuma! Ekkor ha $\mathcal{X}$ és $\mathcal{Y}$ negatív típusú metrikus terek a ϕ és ψ beágyazásokkal, akkor jelölésben legyen ezek tenzorszorzata $(x, y) \mapsto \phi(x) \otimes \psi(y)$, $\mathcal{X} \times \mathcal{Y} \to H \otimes H$. Ekkor ezekkel a következők láthatók be, továbbra is Lyons[10] alapján:

2.68. lemma. *Legyen $(\mathcal{X}, d)$ negatív típusú a ϕ beágyazással! Ekkor $(\mathcal{X}, d)$ akkor és csak akkor erősen negatív típusú, ha β_ϕ injektív $M_1(\mathcal{X})$ véges első momentumú részhalmazán.*

2.69. lemma. *Legyenek $(\mathcal{X}, d)$, $(\mathcal{Y}, d)$ negatív típusú metrikus terek a ϕ és ψ beágyazásokkal, és legyenek $\theta \in M_1(\mathcal{X} \times \mathcal{Y})$ marginálisai a véges első momentumú $\mu \in M_1(\mathcal{X})$ és*

$\nu \in M_1(\mathcal{Y})$ mértékek! Ekkor $\theta \circ (\phi \otimes \psi)^{-1}$ első momentuma véges, így $\beta_{\phi\otimes\psi}(\theta)$ értelmes, és

$$dcov(\theta) = 4\|\beta_{\phi\otimes\psi}(\theta - \mu \times \nu)\|^2 \tag{2.242}$$

2.70. megjegyzés. Ha $\mathcal{X}$ és $\mathcal{Y}$ euklideszi terek, akkor Fourier-beágyazásokkal pontosan a dCov négyzetét kapjuk, Brown-beágyazásokkal pedig a Brown-kovarianciáét, tehát belátható a 2.69 lemmával, hogy a távolságkovariancia és a Brown-kovariancia megegyezik. Ez annyival több a korábbiaknál, hogy míg az eddigi vonatkozó 2.45 tételben a szórásnégyzet végességét is feltettük, az itt látottak szerint a várható értékek végessége is elégséges.

2.71. lemma. *Legyenek $(\mathcal{X}, d)$, $(\mathcal{Y}, d)$ negatív típusú metrikus terek a ϕ és ψ beágyazásokkal, és β_ϕ és β_ψ legyenek injektívek rendre az $\mathcal{X}$ és $\mathcal{Y}$ véges első momentumú mértékein! Ekkor $\beta_{\phi\otimes\psi}$ injektív lesz $\mathcal{X} \times \mathcal{Y}$ véges első momentumú mértékeinek halmazán.*

2.72. lemma. *Legyen $(\mathcal{X}, d)$ erősen negatív típusú metrikus tér. Ekkor létezik ϕ beágyazás úgy, hogy β_ϕ injektív az $M(\mathcal{X})$ véges első momentumú részhalmazán.*

2.73. megjegyzés. A 2.71 és 2.72 lemmákban nem tesszük fel, hogy a valószínűségi mértékeken legyenek injektívek a leképezések, csak a véges első momentumot.

Ekkor a 2.69, 2.71 és 2.72 lemmák következményeként beláthatóak az általánosításra vonatkozó fontos tételek a függetlenség jó karakterizálására.

2.74. tétel. *Legyenek $(\mathcal{X}, d)$, $(\mathcal{Y}, d)$ erősen negatív típusú metrikus terek, továbbá a $\theta \in M_1(\mathcal{X} \times \mathcal{Y})$ marginálisai legyenek véges első momentumúak! Ekkor ha $dcov(\theta) = 0$, akkor θ a marginálisok szorzata.*

Ezzel a függetlenséget jól karakterizáltuk erősen negatív típusú metrikus terekre, és a kiterjesztett határeloszlás-tételekkel az euklideszi esetekhez analóg módon konzisztens próbát is adhatunk, teljesen megegyezően dCov esetével. Nem erősen negatív típusú terekben viszont nem tudunk ilyet mondani, sőt, az alábbi tétel ennél erősebbet is állít:

2.75. tétel.

1. *Ha $(\mathcal{X}, d)$ nem negatív típusú, akkor tetszőleges legalább két pontot tartalmazó $\mathcal{Y}$ esetén létezik $\theta \in M_1(\mathcal{X}, \mathcal{Y})$, amelynek marginálisai véges első momentumúak, hogy $dcov(\theta) < 0$.*

2. *Ha $(\mathcal{X}, d)$ nem erősen negatív típusú, akkor tetszőleges legalább két pontot tartalmazó $\mathcal{Y}$ esetén létezik $\theta \in M_1(\mathcal{X}, \mathcal{Y})$, amelynek marginálisai véges első momentumúak, hogy $dcov(\theta) = 0$ és θ nem a szorzatmérték.*

Ezen túlmenően Hilbert-terek esetében is szeretnénk valamit mondani. Látható, hogy euklideszi terek esetében teljesülnek az erős negatív típikusság feltételei, ha csak a véges tartójú mértékeken tekintünk mindent, és ugyanígy igaz ez Hilbert-terekre is. Teljesen általános esetben Lyons[10] a következő lényeges tételt bizonyítja:

2.76. tétel. *Minden szeparábilis Hilbert-tér erősen negatív típusú.*

2.77. megjegyzés. Nem szeparábilis Hilbert-terek akkor és csak akkor lesznek erősen negatív típusúak, ha dimenziójuk nullmértékű abban az értelemben, hogy az egyetlen olyan Borel-mérték, amely egy legfeljebb ilyen dimenziójú tér minden pontjához a 0 értéket rendeli, az az azonosan nulla mérték (lásd [11]).

Ezzel tehát kiterjesztettük a távolságkovarianciát Hilbert-terek esetére, és közben enyhítettünk a Brow-kovarianciával való egyenlőség feltételein is. Fontos megjegyezni még, hogy a tulajdonságok csak erősen negatív típusú metrikus tereken működnek jól, de ha ez nem áll fenn, a metrika módosításával elérhető, pontosabban belátható az, hogy ha $(\mathcal{X}, d)$ negatív típusú metrikus tér, de nem erősen, akkor $r \in (0, 1)$ mellett $(\mathcal{X}, d^r)$ erősen negatív típusú.

2.4.5. Egyéb lehetőségek

Függetlenítés a peremeloszlásoktól

Rémillard[15] szerint a távolságkovarianciának két lényeges gyengesége van, a véges várható érték feltétele és az, hogy az összefüggőség függ a peremeloszlásoktól. Ennek áthidalására javasolja folytonos peremeloszlások esetén, hogy az egyes változókat eloszlásfüggvényeikbe helyettesítve kapott egyenletes eloszlású változókat vizsgáljuk, azaz ha

$$U^{(i)} = F_{X^{(i)}}(X^{(i)}), i = 1, \ldots, p, \tag{2.243}$$

$$V^{(i)} = F_{Y^{(i)}}(Y^{(i)}), i = 1, \ldots, q, \tag{2.244}$$

$$U = (U^{(1)}, \ldots, U^{(p)}), V = (V^{(1)}, \ldots, V^{(q)}), \tag{2.245}$$

akkor X és Y akkor és csak akkor függetlenek, ha U és V azok, valamint $dCov(U, V)$ csak az U és V kapcsolatát jellemző kopulától függ. Az így átalakított módszerre a tapasztalati mennyiségek definiálhatóak úgy, hogy a mintaelemeket a normált rangjukkal helyettesítjük, és ekkor belátható rájuk a 2.14 tételhez hasonló állítás.

Idősorelemzés

A távolságkovarianciát idősorelemzésben is egyszerűen alkalmazhatjuk, definiálva egy Z_i stacionárius idősor esetén az autokovariancia analógiájára $dCov(Z_i, Z_{i+k})$ mennyiséget, és itt is kimondható gyenge konvergencia, ha végesek a várható értékek, és a hibák eloszlása fehér zajt ad.

Affin transzformációk

A Bakirov[1] által definiált statisztikára vonatkozó 2.37 megjegyzéshez hasonlóan a távolságkorrelációnak is definiálhatjuk az affin transzformációkra invariáns változatát. Helyettesítsük $\mathcal{R}_n$ definíciójába X_k és Y_k helyett rendre a $S_X^{-1/2}X_k$ és $S_Y^{-1/2}Y_k$ mennyiségeket helyettesítjük, ahol S_X és S_Y az $\mathbb{X}$ és $\mathbb{Y}$ minták kovarianciamátrixai. Ekkor ugyanúgy itt is igazak a korábbi tételek és állítások, egyszerűen ez egy új távolságalapú összefüggőségi mérőszám a

$$(c_p c_q |S_X^{1/2}u|_p^{1+p} + |S_Y^{1/2}v|_q^{1+q})^{-1} \tag{2.246}$$

súlyfüggvénnyel.

2.5. A távolságkorreláció és Rényi követelményei

A megfelelő súlyfüggvény kiválasztásánál felmerült, hogy olyan mérőszámot szeretnénk, amely skálainvariáns és összefüggő esetben pozitív. Egy összefüggőségi mérőszám tulajdonságaira további kézenfekvő szempontok is adódhatnak, amit szeretnénk megkövetelni. Természetesen más felhasználási szempontok alapján nem minden esetben ugyanazok az elvárásaink adódnak, és így általánosan sem nagyon lehet meghatározni, mik azok a tulajdonságok, amiket mindenképp szeretnénk egy ilyen mérőszámtól.

Rényi[16] meghatározott hét követelményt, amik természetesen nehezen tekinthetők kizáró oknak egy mérőszám esetében, de valójában mind eléggé ésszerű és természetes elvárás egy jól használható eszköztől.

Legyen ξ és η két valószínűségi változó, amelyek közül egyik sem konstans majdnem mindenütt, és jelölje összefüggőségük mérőszámát $\delta(\xi, \eta)$! Ekkor Rényi követelményei a következők:

A) $\delta(\xi, \eta)$ legyen definiálva minden olyan ξ, η valószínűségiváltozó-párra, amelyek nem 1 valószínűséggel konstansok
B) $\delta(\xi, \eta) = \delta(\eta, \xi)$, azaz legyen szimmetrikus
C) $0 \leq \delta(\xi, \eta) \leq 1$
D) $\delta(\xi, \eta) = 0$ akkor és csak akkor, ha ξ és η függetlenek
E) $\delta(\xi, \eta) = 1$, ha ξ és η között erős összefüggés van, például ha $\xi = g(\eta)$ vagy $\eta = f(\xi)$ teljesül, ahol $f(x)$ és $g(x)$ Borel-mérhető függvények
F) Ha az $f(x)$ és $g(x)$ Borel-mérhető $\mathbb{R} \to \mathbb{R}$ bijekciók, akkor $\delta(f(\xi), g(\eta)) = \delta(\xi, \eta)$
G) Ha ξ és η együttes eloszlása normális, akkor legyen $\delta(\xi, \eta) = |\rho(\xi, \eta)|$.

2.78. megjegyzés. Az E) jelű feltétel esetében felmerülhet, hogy miért nem követeljük meg, hogy legyen $\delta(\xi, \eta) = 1$ akkor és csak akkor, ha fennáll az erős összefüggés. Ezt a lehetőséget Rényi elvetette, mivel túlságosan korlátozó erejűnek találta.

Látható, hogy a távolságkorreláció nem felel meg minden követelménynek, hiszen például nincs minden valószínűségiváltozó-párra értelmezve, mivel csak véges várható értékű esetben definiált. Ennek ellenére mégis alkalmazhatóbb, mint egyes alternatívák, amelyek esetleg megfelelnek. Bickel és Xu[2] is felvetették az összehasonlítását a Gebelein[7] által bevezetett és általánosan Rényi[16] által karakterizált maximálkorrelációval, amely teljesíti mind a hét követelményt, és Breiman és Friedman[3] váltakozó feltételes várható érték algoritmusának segítségével szimulációkban össze is hasonlították. Igaz, hogy bár a maximálkorreláció is elég jól teljesít szimulációkban, és megfelel a követelményeknek, de sok esetben sokkal bonyolultabban vagy éppen sehogy sem számolható. A távolságkorreláció ezzel szemben egy praktikus szempontokból is könnyen alkalmazható eszközt ad, ráadásul sok további jó tulajdonsággal is rendelkezik, mint láttuk.

A tulajdonságok közül a G) jelű, utolsó pont esetében is látható, hogy a távolságkorreláció nem teljesíti, viszont tudunk Székely és Rizzo[23] nyomán elég jó kapcsolatot teremteni a korrelációval.

2.79. tétel. *Legyen X és Y két standard normális eloszlású valószínűségi változó, és legyen $\rho(X,Y)=\rho$! Ekkor*

1. $\mathcal{R}(X,Y)\leq|\rho|$
2. $\mathcal{R}^2(X,Y)=\frac{\rho\arcsin\rho+\sqrt{1-\rho^2}-\rho\arcsin\frac{\rho}{2}-\sqrt{4-\rho^2}+1}{1+\frac{\pi}{3}-\sqrt{3}}$
3. $\inf_{\rho\neq0}\frac{\mathcal{R}(X,Y)}{|\rho|}=\lim_{\rho\to0}\frac{\mathcal{R}(X,Y)}{|\rho|}=\frac{1}{2(1+\frac{\pi}{3}-\sqrt{3})^{\frac{1}{2}}}\approx0.89066$

Bizonyítás. Legyen

$$F(\rho)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{|f_{X,Y}(u,v)-f_X(u)f_Y(v)|^2}{u^2v^2}dudv, \tag{2.247}$$

ekkor ezzel, mivel most $p=q=1$,

$$\mathcal{V}^2(X,Y)=\frac{F(\rho)}{c_1^2}=\frac{F(\rho)}{\pi^2}, \tag{2.248}$$

$$\mathcal{R}^2(X,Y)=\frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}=\frac{F(\rho)}{F(1)}. \tag{2.249}$$

Ekkor ezzel

$$F(\rho) = \int_{\mathbb{R}^2} \frac{|e^{-\frac{u^2+v^2}{2}-\rho uv} - e^{-\frac{u^2}{2}} e^{-\frac{v^2}{2}}|^2}{u^2 v^2} dudv = \tag{2.250}$$

$$= \int_{\mathbb{R}^2} e^{-u^2-v^2} \big(1 - 2e^{-\rho uv} + e^{-2\rho uv}\big) \frac{1}{u^2 v^2} dudv = \tag{2.251}$$

$$= \int_{\mathbb{R}^2} e^{-u^2-v^2} \sum_{n=2}^{\infty} \frac{2^n - 2}{n!} (-\rho uv)^n \frac{1}{u^2 v^2} dudv = \tag{2.252}$$

$$= \int_{\mathbb{R}^2} e^{-u^2-v^2} \sum_{k=1}^{\infty} \frac{2^{2k} - 2}{(2k)!} (-\rho uv)^2 k \frac{1}{u^2 v^2} dudv = \tag{2.253}$$

$$= \rho \left[\sum_{k=1}^{\infty} \frac{2^{2k} - 2}{(2k)!} \rho^{2(k-1)} \int_{\mathbb{R}^2} e^{-u^2-v^2} (uv)^{2(k-1)} dudv \right]. \tag{2.254}$$

Ezzel tehát $F(\rho) = \rho^2 G(\rho)$, ahol G ρ-ban monoton növő, mivel nemnegatív tagok összege, és így $G(\rho) \leq G(1)$, tehát

$$\mathcal{R}^2(X, Y) = \rho^2 \frac{G(\rho)}{G(1)} \leq \rho^2, \tag{2.255}$$

ezzel az első állítást beláttuk.

$F(0) = F'(0) = 0$ miatt $F(\rho) = \int_0^\rho \int_0^x F''(z) dzdx$ A másdik derivált pedig

$$F''(z) = \frac{d^2}{dz^2} \int_{\mathbb{R}^2} e^{-u^2-v^2} \big(1 - 2e^{-zuv} + e^{-2zuv}\big) \frac{1}{u^2 v^2} dudv = 4V(z) - 2V\Big(\frac{z}{2}\Big), \tag{2.256}$$

$$V(z) = \int_{\mathbb{R}^2} e^{-u^2-v^2-2zuv} dudv = \frac{\pi}{\sqrt{1-z^2}}. \tag{2.257}$$

Ezzel pedig

$$F(\rho) = \int_0^\rho \int_0^x \left(\frac{4\pi}{\sqrt{1-z^2}} - \frac{2\pi}{\sqrt{1-\frac{z^2}{4}}} \right) dzdx = \tag{2.258}$$

$$= 4\pi \int_0^\rho \arcsin(x) - \arcsin\Big(\frac{x}{2}\Big) dx = \tag{2.259}$$

$$= 4\pi(\rho \arcsin \rho + \sqrt{1-\rho^2} - \rho \arcsin \frac{\rho}{2} - \sqrt{4-\rho^2} + 1) \tag{2.260}$$

Ekkor behelyettesítve $\frac{F(\rho)}{F(1)}$-be megkapjuk a második állítást, és ebből következik, hogy

$$\lim_{|\rho| \to 0} G(\rho) = \frac{1}{2(1 + \frac{\pi}{3} - \sqrt{3})^{\frac{1}{2}}} \tag{2.261}$$

□

Láthatjuk tehát, hogy bár normális együttes eloszlás mellett nem egyezik meg a távolságkorreláció és a klasszikus korreláció értéke, $\mathcal{R}(X, Y)$ a ρ függvényeként felírható.

3. fejezet

Összefoglalás

A fentiekben bevezetett mérőszámról láttuk, hogy jól általánosítja a korrelációs együtthatót, és tetszőleges dimenzióban értelmezhető, valamint mindenféle kapcsolatot mér, és így jól karakterizálja valószínűségi vektorváltozók függetlenségét, miközben megtartja a klasszikus korrelációnál megismert felépítést annak minden jó tulajdonságával. A definiált tapasztalati változatok megegyeznek a tapasztalati karakterisztikus függvényekkel felírt természetes lehetőséggel. A definiált $\frac{n\mathcal{V}_n^2}{S_2}$ statisztika normális kvadratikus alakhoz való konvergenciája és a tapasztalati mennyiségeknek az elviekhez való konvergenciája alkalmassá teszi a definiáltakat egy jó függetlenségvizsgálati próba konstruálására, sőt valójában egy teljes paraméteres mérőszámosztályra láttuk ezt. A kapott próba ráadásul konzisztens, valamint elérhető szignifikanciaszintjéről is pontos tételt lehet kimondani. Áttekintettünk néhány javítási lehetőséget, és ezáltal láttuk, hogy lehet skálainvariáns, affin transzformációkra invariáns mérőszámot meghatározni, ami tetszőleges dimenzióban értelmezhető, és erre meg tudtuk határozni az akár torzítatlan becslést is adó statisztikáinkat, amelyekkel jó teszteket végezhetünk. A távolságkorreláció talán nem felel meg minden mérőszámokkal szemben támasztott követelménynek hiánytalanul, de nagy részének igen, és ezeken kívül további hasznos tulajdonsággal bír, ráadásul a gyakorlatba is egyszerű eszközökkel átültethető, gyorsan kiszámítható eszköz az összefüggőségi struktúrák vizsgálatában.

Köszönetnyilvánítás

Köszönettel tartozom témavezetőmnek, Székely Gábornak türelméért és irányadó tanácsaiért, valamint a legfrissebb eredmények rendelkezésre bocsájtásáért, további köszönettel tartozom Móri Tamásnak a téma javaslásáért és a munkában való elindításért, valamint a technikai segítségekért.

Irodalomjegyzék

[1] Bakirov, N. K.; Rizzo, M. L.; Székely, G. J. (2006)A multivariate nonparametric test of independence. *Journal of Multivariate Analysis* **97** 1742-1756

[2] Bickel, P. J.; Xu,Y. (2009) Discussion of : Brownian distance covariance. *Annals of Applied Statistics* **3** 1266-1269

[3] Breiman, L.; Friedman J. (1985) Estimating optimal transformations for multiple regression and correlation. *Journal of American Statistical Association* **80** 580-598

[4] Cope, L. (2009) Discussion of : Brownian distance covariance. *Annals of Applied Statistics* **3** 1279-1281

[5] Csörgő, S. (1985) Testing independence by the empirical characteristic function. *Journal of Multivariate Analysis* **16** 290-299

[6] Erdélyi, A. et al. (1954) Tables of Integral Transforms vol I. *McGraw Hill, New York*

[7] Gebelein, H. (1947) Das statistische Problem der Korrelation als Variation- und Eigenwertproblem und sein Zusammenhang mit Ausgleichsrechnung. *Zeitschrift für angewandte Mathematik und Mechanik* **21** (1947) 364-379

[8] Mardia, K.V.; Kent, J. T.; Bibby, J.M. (1997) Multivariate Analysis *Academic Press, San Diego*

[9] Kosorok, M. R. (2009) Discussion of : Brownian distance covariance. *Annals of Applied Statistics* **3** 1270-1278

[10] Lyons, R. (2012) Distance Covariance is Metric Spaces. *Kézirat*

[11] Marczewski, E.; Sikorski, R. (1948) Measures in non-separable metric spaces *Colloq. Math.* **1** 133-139

[12] Pillai, K. C. S. (1955) Some new test criteria in multivariate analysis. *Annals of Mathematical statistics* **26** 117-121

[13] Prudnikov, A. P.; Brychkov, A.; Marichev, O. I. (1986) Integrals and Series : Volume 1; Elementary Functions *Gordon and Breach Science Publishers*

[14] Puri, M. L.; Sen, P. K. (1971) Nonparamethric Methods in Mutivariate Analysis *Wiley, New York*

[15] Rémillard, B (2009) Discussion of : Brownian distance covariance. *Annals of Applied Statistics* **3** 1295-1298

[16] Rényi, A. (1959) On measures of dependence. *Acta Math. Acad. Sci. Hung.* **10** 441-451.

[17] Rizzo, M. L.; Székely, G. J. (2010) DISCO Analysis : A nonparametric extension of analysis of variance. *Annals of Applied Statistics* **4** 1034-1055

[18] Rosenblatt, M. (1952) Limit theorems associated with variants of the von Mises statistic. *Annals of Mathematical Statistics* **23** 617-623

[19] Székely, G. J.; Bakirov, N. K. (2003) Extremal probabilities for Gaussian quadratic forms. *Probability Theory Related Fields* **126** 184-202

[20] Székely, G. J.; Rizzo, M. L. (2009) Brownian distance covariance. *Annals of Applied Statistics* **3** 1236-1265

[21] Székely, G. J.; Rizzo, M. L. (2005) Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *Journal of Classification* **22** 151-183

[22] Székely, G. J.; Rizzo, M. L. (2009) Rejoinder : Brownian distance covariance. *Annals of Applied Statistics* **3** 1303-1308

[23] Székely, G. J.; Rizzo, M. L.; Bakirov, N. K. (2007) Measuring and testing dependence by correlation of distances. *Annals of Statistics* **35** 2769-2794

[24] von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Annals of Mathematical Statistics* **18** 309-348

[25] Wilks, S. S. (1932) Certain generalizations in the analysis of variance. *Biometrika* **24** 471-494

[26] Wilks, S. S. (1935) On the independence of k sets of normally distributed statistical variables. *Econometrica* **3** 309-326