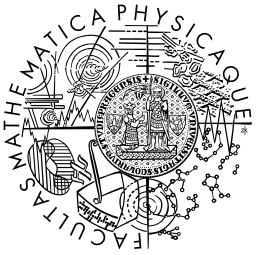


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Monika Matoušková

Míry závislosti

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Zbyněk Pawlas, Ph.D.
Studijní program: Matematika
Studijní obor: Obecná matematika

Praha 2020

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Ráda bych poděkovala vedoucímu práce doc. RNDr. Zbyňku Pawlasovi, Ph.D. za ochotu, trpělivost, cenné rady, připomínky, návrhy na zlepšení a především čas strávený nad opakovaným čtením práce.

Název práce: Míry závislosti

Autor: Monika Matoušková

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Zbyněk Pawlas, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Nejčastěji se setkáváme se základní mírou závislosti, s korelačním koeficientem. Ten ovšem může být roven nule pro dvě závislé náhodné veličiny. V práci se zaměřujeme na dvě míry závislosti, které se rovnají nule právě tehdy, když jsou veličiny nezávislé. Porovnáváme je s Pearsonovým korelačním koeficientem. Jako první zavádíme maximální korelaci, která jde většinou obtížně vypočítat, proto definujeme maximální polynomiální korelaci, jejíž výpočet je snadnější a je neklesající ve stupni polynomu. Druhá zavedená míra je vzdálenostní korelace, u níž uvádíme různé způsoby vyjádření, které se hodí k výpočtu. U obou měr diskutujeme, co se děje v případě sdruženého normálního rozdělení a na závěr ukazujeme na několika příkladech výpočet zavedených měr závislosti.

Klíčová slova: závislost, korelace, korelační koeficient, maximální korelace, vzdálenostní korelace

Title: Measures of dependence

Author: Monika Matoušková

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Zbyněk Pawlas, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The most common measure of dependence is the correlation coefficient. Its problem is that it can be zero for two dependent random variables. We will discuss two measures of dependence, which are equal to zero if and only if the two random variables are independent. We will compare them with Pearson's correlation coefficient. The first one will be the maximal correlation, which is often difficult to calculate. That is why we define the maximal polynomial correlation, which is easier to calculate and is non-decreasing in a degree of a polynomial. We also define the distance correlation and we discuss other ways of the expression of distance correlation, which can be used in the calculation. We deal with the case of normal distribution and we show some calculations of these measures of dependence.

Keywords: dependence, correlation, correlation coefficient, maximal correlation, distance correlation

Obsah

Úvod

Téma závislosti se objevuje v běžném životě, ať už jsou to meteorologické jevy, medicína, politika a samozřejmě také ekonomika. V minulosti bylo toto téma opomíjeno. Až v druhé polovině dvacátého století se mu vědci začali více věnovat. Do té doby jako jediná míra závislosti sloužil korelační koeficient, který však mnohdy vedl k chybným statistickým závěrům. Také proto se začaly hledat jiné způsoby, jak měřit sílu závislosti, které našly své uplatnění i ve statistice.

Základní mírou závislosti zůstává Pearsonův korelační koeficient. Jeho hlavní nedostatek je, že může být roven nule pro dvě závislé náhodné veličiny. Existují další míry závislosti, které se rovnají nule právě tehdy, když jsou veličiny nezávislé. Na dvě takové míry se v práci zaměřujeme a porovnáváme je s Pearsonovým korelačním koeficientem.

První z nich se nazývá maximální korelace. Na rozdíl od Pearsonova korelačního koeficientu je definována pro libovolné dvě nedegenerované náhodné veličiny. Existuje však jen málo případů, kdy ji lze přímo spočítat. Zavedeme proto pojem maximální polynomiální korelace, která se spočítá snadněji, a při zvyšování stupně polynomů se blížíme maximální korelaci. Dále se podíváme na případ, kdy náhodné veličiny mají sdružené normální rozdělení.

Druhá míra závislosti, kterou se v práci zabýváme, je vzdálenostní korelace. Její druhou mocninu definujeme pro dvě reálné náhodné veličiny s konečným prvním momentem analogicky jako Pearsonův korelační koeficient. Dokážeme některé její vlastnosti včetně případu se sdruženým normálním rozdělením a také diskrétního rozdělení, kdy obě náhodné veličiny nabývají pouze dvou hodnot.

V první kapitole uvedeme čtenáře do problematiky měření závislosti a zadefinujeme Pearsonův korelační koeficient. V druhé kapitole už se zaměříme na pojem maximální korelace, její vlastnosti a možnosti výpočtu. Třetí kapitola nám přiblíží korelaci vzdálenostní, nejprve zavedeme pojmy jako norma ve váženém L_2 prostoru a charakteristická funkce, které nás dovedou k definici dané míry závislosti. Budeme se zabývat alternativními způsoby vyjádření, jenž se dají využít k jejímu výpočtu. V závěrečné kapitole pak uvedeme několik příkladů, kde ilustrujeme techniky počítání zavedených měr závislosti.

1. Korelační koeficient

Na začátek uvedeme definici nezávislosti, neboť s ní nadále budeme pracovat.

Definice 1. *Nechť X, Y jsou reálné náhodné veličiny. Řekneme, že jsou nezávislé, jestliže pro všechna $A, B \in \mathcal{B}$ platí*

$$\mathbb{P}([X \in A] \cap [Y \in B]) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

Vyslovíme lemma, které dále využijeme. Můžeme ho najít i s důkazem v Andělově knize (Anděl, 2005, Věta 2.4).

Lemma 1. *Nechť $(X, Y)^T$ je reálný náhodný vektor. Náhodné veličiny X, Y jsou nezávislé právě tehdy, když $F_{X,Y}(x, y) = F_X(x)F_Y(y)$ pro všechna $x, y \in \mathbb{R}$, kde $F_{X,Y}$ je sdružená distribuční funkce náhodného vektoru $(X, Y)^T$ a F_X, F_Y marginální distribuční funkce X a Y.*

Jako první míru závislosti zmíníme tu nejčastěji používanou, a to Pearsonův korelační koeficient.

Definice 2. *Nechť X a Y jsou reálné náhodné veličiny takové, že* $\operatorname{var} X \in (0, \infty)$, $\operatorname{var} Y \in (0, \infty)$. *Pak definujeme Pearsonův korelační koeficient X a Y jako*

$$\rho(X, Y) = \frac{\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y}{\sqrt{\operatorname{var} X}\sqrt{\operatorname{var} Y}}.$$

Vlastnosti této míry závislosti budeme dále také využívat, zde jsou ty nejdůležitější: korelační koeficient nabývá hodnot z intervalu $[-1, 1]$ a pokud jsou náhodné veličiny X a Y nezávislé, pak $\rho(X, Y) = 0$.

Příklad. Ukážeme, že ze vztahu $\rho(X, Y) = 0$ neplyne nezávislost X a Y. Mějme náhodnou veličinu X, která má rovnoměrné rozdělení na intervalu $(-1, 1)$, a položme $Y = X^2$. Veličiny jsou zřejmě závislé, ale

$$\rho(X, Y) = \frac{\mathbb{E}X^3 - \mathbb{E}X\mathbb{E}X^2}{\sqrt{\operatorname{var} X}\sqrt{\operatorname{var} X^2}} = \frac{0 - 0 \cdot \frac{1}{3}}{\sqrt{\frac{1}{3}}\sqrt{\frac{4}{45}}} = 0.$$

Poznámka. V případě, že náhodné veličiny mají sdružené normální rozdělení, plyne z nulovosti korelačního koeficientu jejich nezávislost.

2. Maximální korelace

Druhá míra závislosti, kterou se budeme v práci zabývat, se nazývá maximální korelace, poprvé ji ve svých článcích zavádějí Hirschfeld (1935), Gebelein (1941) a Rényi (1959). Její definice je převzata z publikace autorů Papadatose a Xifarové (Papadatos a Xifara, 2013).

Definice 3. *Nechť X a Y jsou náhodné veličiny. Pak definujeme maximální korelaci X a Y jako*

$$R(X,Y) = \sup \rho(f(X), g(Y)),$$

kde supremum bereme přes všechny borelovsky měřitelné funkce f, $g : \mathbb{R} \to \mathbb{R}$ takové, že $0 < \operatorname{var} f(X) < \infty$, $0 < \operatorname{var} g(Y) < \infty$.

2.1 Vlastnosti maximální korelace

Následující poznámky a lemma 2 byly sepsány na základě článku Rényiho (Rényi, 1959), důkaz lemmatu je v této publikaci pouze naznačen, zde je proveden podrobně.

Poznámka. Maximální korelace je definována pro libovolné X, Y nedegenerované a nabývá hodnot z intervalu $[0, 1]$.

Poznámka. Z definice 3 triviálně plyne, že je-li $f(X) = g(Y)$ pro nějaké f, g borelovsky měřitelné funkce, pak $R(X, Y) = 1$.

V příkladu z kapitoly 1 máme $\rho(X, Y) = 0$ a $R(X, Y) = 1$.

Poznámka. Z definice maximální korelace také dostáváme, že jsou-li f, g bijekce, pak $R(f(X), g(Y)) = R(X, Y)$.

Lemma 2. *Nechť X a Y jsou náhodné veličiny. Potom $R(X, Y) = 0$ právě tehdy, když X a Y jsou nezávislé.*

Důkaz. Nejprve ukážeme, že platí implikace zprava doleva. Jsou-li X, Y nezávislé, pak pro libovolné dvě borelovsky měřitelné funkce f, g jsou také $f(X)$ a $g(Y)$ nezávislé, a tedy $\rho(f(X), g(Y)) = 0$, pokud je výraz definován.

Nyní ověříme platnost druhé implikace. Zvolme pro libovolné a, b reálné

$$f(x) = \begin{cases} 1, & x \le a \\ 0, & x > a \end{cases}, \quad g(y) = \begin{cases} 1, & y \le b \\ 0, & y > b \end{cases}.$$

Pak dostáváme

$$f(X) = \mathbb{1}\{X \le a\},\ g(Y) = \mathbb{1}\{Y \le b\},$$

z čehož plyne

$$\mathbb{E}f(X)g(Y) - \mathbb{E}f(X)\mathbb{E}g(Y) = \mathbb{P}(X \le a, Y \le b) - \mathbb{P}(X \le a)\mathbb{P}(Y \le b) = 0.$$

Tedy $\mathbb{P}(X \le a, Y \le b) = \mathbb{P}(X \le a)\mathbb{P}(Y \le b)$ pro libovolné $a, b \in \mathbb{R}$, a z lemmatu 1 plyne nezávislost X a Y. □

Následující v této sekci je výsledkem práce autorky.

Lemma 3. *Nechť X, Y jsou náhodné veličiny, f, g borelovsky měřitelné funkce takové, že* $\operatorname{var} f(X) \in (0,\infty)$, $\operatorname{var} g(Y) \in (0,\infty)$. *Položme*

$$\widetilde{f}(X) = \frac{f(X) - \mathbb{E}f(X)}{\sqrt{\operatorname{var} f(X)}}, \quad \widetilde{g}(Y) = \frac{g(Y) - \mathbb{E}g(Y)}{\sqrt{\operatorname{var} g(Y)}}.$$

Pak $\rho(f(X), g(Y)) = \rho(\widetilde{f}(X), \widetilde{g}(Y))$.

Důkaz. Platí

$$\begin{aligned}
&\rho\left(\widetilde{f}(X), \widetilde{g}(Y)\right) = \mathbb{E}\widetilde{f}(X)\widetilde{g}(Y) \\
&= \mathbb{E}\left[\frac{f(X)g(Y) - f(X)\mathbb{E}g(Y) - g(Y)\mathbb{E}f(X) + \mathbb{E}f(X)\mathbb{E}g(Y)}{\sqrt{\operatorname{var} f(X)}\sqrt{\operatorname{var} g(Y)}}\right] \\
&= \frac{\mathbb{E}f(X)g(Y) - \mathbb{E}f(X)\mathbb{E}g(Y)}{\sqrt{\operatorname{var} f(X)}\sqrt{\operatorname{var} g(Y)}} = \rho(f(X), g(Y)).
\end{aligned}$$

□

Poznámka. Lemma 3 nám říká, že v definici 3 se stačí omezit na funkce f, g, pro něž $f(X)$ a $g(Y)$ mají střední hodnotu 0 a rozptyl 1.

Lemma 4. *Pro všechny X, Y náhodné veličiny s konečnými a kladnými rozptyly platí $R(X,Y) \geq |\rho(X,Y)|$.*

Důkaz. Pokud $f = g = id$, pak zřejmě $\rho(f(X), g(Y)) = \rho(X,Y)$. Nyní jestliže $\rho(X,Y) < 0$, položme $f(X) = -X$ a $g(Y) = Y$, potom

$$\rho(f(X), g(Y)) = -\rho(X,Y) > 0.$$

Díky tomu, že maximální korelace je supremum přes všechny měřitelné funkce, pro které výraz existuje, nerovnost platí. □

Lemma 5. *Nechť X, Y jsou náhodné veličiny s konečnými a kladnými rozptyly, f, g lineární funkce, pak platí*

$$\rho(f(X), g(Y)) \in \{-\rho(X,Y), \rho(X,Y)\}.$$

Důkaz. Nechť $b, d \in \mathbb{R}$, $a, c \neq 0$, $f(x) = ax + b$ a $g(y) = cy + d$. Počítejme

$$\begin{aligned}
&\rho(f(X), g(Y)) \\
&= \frac{\mathbb{E}[acXY + adX + bcY + bd] - (ac\mathbb{E}X\mathbb{E}Y + ad\mathbb{E}X + bc\mathbb{E}Y + bd)}{\sqrt{a^2 \operatorname{var} X}\sqrt{c^2 \operatorname{var} Y}} \\
&= \frac{ac(\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y)}{|ac|\sqrt{\operatorname{var} X}\sqrt{\operatorname{var} Y}} = \operatorname{sgn}(ac)\rho(X,Y).
\end{aligned}$$

□

Poznámka. Z lemmatu 5 plyne, že pro náhodné veličiny X, Y s $\operatorname{var} X \in (0,\infty)$, $\operatorname{var} Y \in (0,\infty)$ platí $\sup \rho(f(X), g(Y)) = |\rho(X,Y)|$, pokud supremum bereme přes všechny lineární funkce.

Lemma 6. *Pokud náhodné veličiny X a Y nabývají pouze dvou hodnot s.j., pak $R(X,Y) = |\rho(X,Y)|$.*

Důkaz. Nechť X nabývá hodnot x_1, x_2 s.j., $x_1 < x_2$ a Y nabývá hodnot y_1, y_2 s.j., $y_1 < y_2$. Nechť f, g jsou libovolné borelovsky měřitelné funkce. Náhodná veličina $f(X)$ závisí s.j. jen na hodnotách $f(x_1)$, $f(x_2)$. Položme

$$\widetilde{f}(x) = \frac{(f(x_2) - f(x_1))\,x + f(x_1)x_2 - f(x_2)x_1}{x_2 - x_1}, \tag{2.1}$$

analogicky $\widetilde{g}(y)$. Pak $\widetilde{f}(x)$ a $\widetilde{g}(y)$ jsou lineární funkce, platí $\rho(\widetilde{f}(X), \widetilde{g}(Y)) = \rho(f(X), g(Y))$, neboť výrazy závisí pouze na hodnotách funkcí v bodech x_1, x_2 a y_1, y_2, ve kterých se rovnají. Z lemmatu 5 plyne požadovaný vztah. □

Pokud X nabývá pouze dvou hodnot s.j. a Y tří hodnot s.j., v podkapitole 4.2 se nachází příklad, kdy $0 < |\rho(X,Y)| < R(X,Y) < 1$.

2.2 Maximální polynomiální korelace

Zavedeme nyní pojem maximální polynomiální korelace stupně n. Výsledky v této sekci jsou prací autorky.

Definice 4. *Nechť X a Y jsou náhodné veličiny. Pak definujeme maximální polynomiální korelaci X a Y stupně n, $n \in \mathbb{N}$, jako*

$$R_n(X,Y) = \sup \rho(f(X), g(Y)),$$

kde supremum bereme přes všechny polynomy f, g stupně nejvýše n takové, že $0 < \operatorname{var} f(X) < \infty$, $0 < \operatorname{var} g(Y) < \infty$.

Poznámka. Z definice 4 plyne, že pokud $f(X) = g(Y)$ pro nějaké f, g polynomy stupně nejvýše n, pak $R_n(X,Y) = R(X,Y) = 1$.

Lemma 7. *Jsou-li X, Y náhodné veličiny s konečnými nenulovými rozptyly, pak $R_1(X,Y) = |\rho(X,Y)|$ a $R_1(X,Y) \le R_2(X,Y) \le \cdots \le R(X,Y)$.*

Důkaz. První rovnost dostáváme z lemmatu 5 a řetězec nerovností je zřejmý z definice maximální polynomiální korelace. □

Následující lemma je zobecněním lemmatu 6.

Lemma 8. *Pro nedegenerované náhodné veličiny X, Y, které nabývají nejvýše n hodnot, platí $R(X,Y) = R_{n-1}(X,Y)$.*

Důkaz. Podobně jako v lemmatu 6 platí, že pokud náhodná veličina nabývá nejvýše n hodnot, funkce této náhodné veličiny závisí pouze na n hodnotách (funkčních hodnotách v těchto bodech) a těmito hodnotami můžeme proložit polynom stupně $n-1$. Máme-li tedy X, Y náhodné veličiny nabývající nejvýše n hodnot, můžeme libovolnou funkci $f(X)$ splňující $\operatorname{var} f(X) \in (0, \infty)$ nahradit polynomem $\widetilde{f}(X)$ stupně $n-1$ s popsanou vlastností, podobně $g(Y)$ polynomem $\widetilde{g}(Y)$. A opět platí $\rho(f(X), g(Y)) = \rho(\widetilde{f}(X), \widetilde{g}(Y))$. □

Podkapitola 4.1 uvádí návod, jak spočítat R_2 ve speciálním případě, kdy jeden polynom je stupně 2 a druhý stupně 1.

Nyní uvedeme obecný návod, jak počítat $R_2(X, Y)$.

Věta 9. *Nechť X a Y jsou náhodné veličiny s konečnými čtvrtými momenty. Pak*

$$R_2(X,Y) = \sup\left(a_2b_2\operatorname{cov}(X^2,Y^2) + a_2b_1\operatorname{cov}(X^2,Y) + a_1b_2\operatorname{cov}(X,Y^2) + a_1b_1\operatorname{cov}(X,Y)\right),$$

kde supremum bereme přes koeficienty a_2, a_1, b_2 a b_1 splňující

$$a_2^2\operatorname{var}X^2 + 2a_1a_2\operatorname{cov}(X^2,X) + a_1^2\operatorname{var}X = 1 \tag{2.2}$$

a

$$b_2^2\operatorname{var}Y^2 + 2b_1b_2\operatorname{cov}(Y^2,Y) + b_1^2\operatorname{var}Y = 1. \tag{2.3}$$

Důkaz. Z lemmatu 3 víme, že se při výpočtu $R_2(X,Y)$ stačí omezit na funkce se střední hodnotou 0 a rozptylem 1. Předpokládejme, že polynomy mají tvar $f(x) = a_2x^2+a_1x+a_0$ a $g(y) = b_2y^2+b_1y+b_0$. Podmínka $\mathbb{E}f(X) = 0$ bude splněna, když $a_0 = -a_2\mathbb{E}X^2 - a_1\mathbb{E}X$. Obdobně $b_0 = -b_2\mathbb{E}Y^2 - b_1\mathbb{E}Y$ zaručí, že $\mathbb{E}g(Y) = 0$. Požadavek jednotkových rozptylů dá následující omezení pro koeficienty a_2, a_1, b_2 a b_1:

$$\begin{aligned}\mathbb{E}f(X)^2 &= \mathbb{E}\left[(a_2(X^2-\mathbb{E}X^2) + a_1(X-\mathbb{E}X))^2\right]\\ &= a_2^2\operatorname{var}X^2 + 2a_1a_2\operatorname{cov}(X^2,X) + a_1^2\operatorname{var}X = 1\end{aligned}$$

a

$$\mathbb{E}g(Y)^2 = b_2^2\operatorname{var}Y^2 + 2b_1b_2\operatorname{cov}(Y^2,Y) + b_1^2\operatorname{var}Y = 1.$$

Dostali jsme omezení (2.2) a (2.3). Dále

$$\begin{aligned}&\rho(f(X),g(Y)) = \mathbb{E}f(X)g(Y)\\ &= \mathbb{E}\left[(a_2(X^2-\mathbb{E}X^2) + a_1(X-\mathbb{E}X))(b_2(Y^2-\mathbb{E}Y^2) + b_1(Y-\mathbb{E}Y))\right]\\ &= a_2b_2\operatorname{cov}(X^2,Y^2) + a_2b_1\operatorname{cov}(X^2,Y) + a_1b_2\operatorname{cov}(X,Y^2) + a_1b_1\operatorname{cov}(X,Y).\end{aligned}$$

Potřebujeme tedy najít supremum funkce

$$a_2b_2\operatorname{cov}(X^2,Y^2) + a_2b_1\operatorname{cov}(X^2,Y) + a_1b_2\operatorname{cov}(X,Y^2) + a_1b_1\operatorname{cov}(X,Y)$$

přes koeficienty a_2, a_1, b_2 a b_1 splňující (2.2) a (2.3). □

Jako příklad využití věty 9 uvažujme následující větu.

Věta 10. *Nechť X, Y jsou náhodné veličiny, které mají sdružené normální rozdělení s nulovou střední hodnotou a varianční maticí $\left(\begin{smallmatrix}1 & \rho\\ \rho & 1\end{smallmatrix}\right)$. Potom platí $R_2(X,Y) = |\rho|$.*

Důkaz. Případ, kdy $\rho \in \{-1, 1\}$ je zřejmý, uvažujme tedy pouze $\rho \in (-1, 1)$. Pro $\rho = 0$ jsou X a Y nezávislé podle poznámky v kapitole 1, a $R_2(X,Y) = 0$, nulu tedy také uvažovat nebudeme.

K výpočtu využijeme větu 9. Máme $\mathbb{E}X = \mathbb{E}Y = 0$ a $\mathbb{E}X^2 = \mathbb{E}Y^2 = 1$. Napočítejme další momenty, které budeme potřebovat:

$$\mathbb{E}X^3 = \int_{-\infty}^{\infty} \frac{x^3}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2} \,\mathrm{d}x = 0,$$

neboť integrujeme lichou funkci přes celé $\mathbb{R}$. Dále spočítáme čtvrtý moment X:

$$\begin{aligned}\mathbb{E}X^4 &= \int_{-\infty}^{\infty} \frac{x^4}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2} \,\mathrm{d}x = 2\int_{0}^{\infty} \frac{x^4}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2} \,\mathrm{d}x \\ &= \frac{4}{\sqrt{\pi}} \int_0^{\infty} y^{3/2} \mathrm{e}^{-y} \,\mathrm{d}y = \frac{4}{\sqrt{\pi}} \Gamma\left(\frac{5}{2}\right) = 3,\end{aligned}$$

kde jsme zavedli substituci $y = \frac{x^2}{2}$. Nyní vyjádříme $\mathbb{E}X^2Y^2$:

$$\begin{aligned}\mathbb{E}X^2Y^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2y^2 \frac{1}{2\pi\sqrt{1-\rho^2}} \mathrm{e}^{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}} \,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} \frac{y^2}{\sqrt{2\pi}} \mathrm{e}^{-\frac{y^2}{2}} \int_{-\infty}^{\infty} \frac{x^2}{\sqrt{2\pi}\sqrt{1-\rho^2}} \mathrm{e}^{-\frac{(x-\rho y)^2}{2(1-\rho^2)}} \,\mathrm{d}x\,\mathrm{d}y \\ &= (1-\rho^2)\int_{-\infty}^{\infty} \frac{y^2}{\sqrt{2\pi}} \mathrm{e}^{-\frac{y^2}{2}} \,\mathrm{d}y + \rho^2 \int_{-\infty}^{\infty} \frac{y^4}{\sqrt{2\pi}} \mathrm{e}^{-\frac{y^2}{2}} \,\mathrm{d}y \\ &= (1-\rho^2) + 3\rho^2 = 1 + 2\rho^2.\end{aligned}$$

Na druhém řádku jsme si všimli, že jsme v integrálu podle x dostali druhý moment rozdělení $\mathrm{N}(\rho y, 1-\rho^2)$, tedy $1-\rho^2+\rho^2y^2$. Podobně v dalším výpočtu dostaneme první moment rozdělení $\mathrm{N}(\rho y, 1-\rho^2)$:

$$\begin{aligned}\mathbb{E}XY^2 &= \int_{-\infty}^{\infty} \frac{y^2}{\sqrt{2\pi}} \mathrm{e}^{-\frac{y^2}{2}} \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}\sqrt{1-\rho^2}} \mathrm{e}^{-\frac{(x-\rho y)^2}{2(1-\rho^2)}} \,\mathrm{d}x\,\mathrm{d}y \\ &= \rho \int_{-\infty}^{\infty} \frac{y^3}{\sqrt{2\pi}} \mathrm{e}^{-\frac{y^2}{2}} \,\mathrm{d}y = 0.\end{aligned}$$

Z předchozích výpočtů dostaneme: $\operatorname{var} Y^2 = \operatorname{var} X^2 = \mathbb{E}X^4 - (\mathbb{E}X^2)^2 = 2$, $\operatorname{cov}(Y^2, Y) = \operatorname{cov}(X^2, X) = \mathbb{E}X^3 - \mathbb{E}X\mathbb{E}X^2 = 0$, $\operatorname{cov}(X^2, Y^2) = \mathbb{E}X^2Y^2 - \mathbb{E}X^2\mathbb{E}Y^2 = 2\rho^2$ a $\operatorname{cov}(X^2, Y) = \operatorname{cov}(X, Y^2) = \mathbb{E}XY^2 - \mathbb{E}X\mathbb{E}Y^2 = 0$.

Podle věty 9 chceme maximalizovat $2\rho^2a_2b_2 + \rho a_1b_1$ přes koeficienty splňující $2a_2^2 + a_1^2 = 1$ a $2b_2^2 + b_1^2 = 1$, což odpovídá podmínkám (2.2) a (2.3). Stačí se omezit na případy $a_2 \geq 0$, $b_2 \geq 0$, $a_1 \geq 0$ a $\rho b_1 \geq 0$. Pak $a_1 = \sqrt{1-2a_2^2}$ a $b_1 = \operatorname{sgn}(\rho)\sqrt{1-2b_2^2}$.

Úlohu jsme tak převedli na maximalizaci funkce dvou proměnných

$$h(a_2, b_2) = 2\rho^2a_2b_2 + |\rho|\sqrt{1-2a_2^2}\sqrt{1-2b_2^2}$$

přes $0 \leq a_2 \leq \frac{\sqrt{2}}{2}$ a $0 \leq b_2 \leq \frac{\sqrt{2}}{2}$. Standardním postupem najdeme maximum

(funkce je symetrická, uvedeme jen parciální derivace podle a_2):

$$\frac{\partial h}{\partial a_2}(a_2, b_2) = 2\rho^2 b_2 - 2|\rho| a_2 \frac{\sqrt{1-2b_2^2}}{\sqrt{1-2a_2^2}} = 0,$$

$$\frac{\partial^2 h}{\partial a_2^2}(a_2, b_2) = -2|\rho| \frac{\sqrt{1-2b_2^2}}{\sqrt{1-2a_2^2}} - 4a_2^2|\rho| \frac{\sqrt{1-2b_2^2}}{(1-2a_2^2)^{3/2}} = -2|\rho| \frac{\sqrt{1-2b_2^2}}{(1-2a_2^2)^{3/2}},$$

$$\frac{\partial^2 h}{\partial a_2 \partial b_2}(a_2, b_2) = \frac{\partial^2 h}{\partial b_2 \partial a_2}(a_2, b_2) = 2\rho^2 + \frac{4|\rho| a_2 b_2}{\sqrt{1-2a_2^2}\sqrt{1-2b_2^2}}.$$

Z první rovnice a $\frac{\partial h}{\partial b_2}(a_2, b_2) = 0$ dostaneme $b_2^2 = a_2^2/(2a_2^2 + \rho^2(1-2a_2^2))$ a $\rho^2 a_2^2 = a_2^2/\rho^2$. Protože nepředpokládáme, že $\rho \in \{-1, 1\}$, musí být $a_2 = b_2 = 0$. V bodě $(0,0)$ je matice druhých parciálních derivací $\left(\begin{smallmatrix} -2|\rho| & 2\rho^2 \\ 2\rho^2 & -2|\rho| \end{smallmatrix}\right)$, po symetrické úpravě dostaneme matici $\left(\begin{smallmatrix} -2|\rho| & 0 \\ 0 & -2|\rho|+2|\rho|^3 \end{smallmatrix}\right)$, tedy matice je negativně definitní, neboť $|\rho| \in (0,1)$, a pro $a_2 = b_2 = 0$ se nabývá maximum, máme $h(0,0) = |\rho|$. □

Poznámka. Věta 10 platí pro obecné náhodné veličiny se sdruženým normálním rozdělením, předpoklad o nulovosti střední hodnoty a tvaru varianční matice jsou uvedeny pro zjednodušení výpočtu.

Věta 10 platí i v obecnější variantě, vztah lze totiž dokázat také pro maximální korelaci, pokud máme náhodné veličiny se sdruženým normálním rozdělením. Uvedeme zde však jen tvrzení bez důkazu, neboť důkaz je komplikovaný, lze jej najít např. v publikaci Gebeleina (Gebelein, 1941) nebo v novějším článku Papadatose a Xifarové (Papadatos a Xifara, 2013).

Věta 11. *Nechť X, Y jsou náhodné veličiny, které mají sdružené normální rozdělení, pak $R(X,Y) = |\rho(X,Y)|$.*

3. Vzdálenostní korelace

Další míra závislosti, kterou v této práci diskutujeme, je vzdálenostní korelace. Jako první ji ve svém článku prezentují Székely, Rizzo a Bakirov (2007).

3.1 Definice vzdálenostní korelace

Definice vzdálenostní korelace je založena na normě ve váženém L_2 prostoru, která je převzata z publikace Székelyho a kol. (Székely a kol., 2007).

Definice 5. *Pro komplexní funkci γ definovanou na $\mathbb{R}^2$ definujeme $\|\cdot\|_\omega$-normu ve váženém L_2 prostoru funkcí na $\mathbb{R}^2$ jako*

$$\|\gamma(t,s)\|_\omega^2 = \int_{\mathbb{R}^2} |\gamma(t,s)|^2 \omega(t,s)\,\mathrm{d}t\,\mathrm{d}s,$$

kde $\omega(t,s)$ je libovolná kladná váhová funkce, pro kterou tento integrál existuje.

Uvedeme zde také definici charakteristické funkce, protože ji dále budeme využívat.

Definice 6. *Nechť X je k-rozměrný reálný náhodný vektor, $k \in \mathbb{N}$, definujeme charakteristickou funkci náhodného vektoru X jako $f_X(t) = \mathbb{E}\,\mathrm{e}^{\mathrm{i}\langle t, X\rangle}$ pro $t \in \mathbb{R}^k$.*

Následující poznámky a definice 7 a 8 jsou zpracovány podle publikace Székelyho a kol. (Székely a kol., 2007).

Pomocí $\|\cdot\|_\omega$-normy použitím vhodné váhové funkce ω definujeme míru závislosti $\mathcal{V}^2(X,Y;\omega) = \|f_{X,Y}(t,s) - f_X(t)f_Y(s)\|_\omega^2$. Pro tuto míru závislosti platí $\mathcal{V}^2(X,Y;\omega) = 0$ právě tehdy, když X, Y jsou nezávislé. Pokud $\mathcal{V}^2(X,Y;\omega)$ vydělíme $\sqrt{\mathcal{V}^2(X;\omega)}\sqrt{\mathcal{V}^2(Y;\omega)}$, kde $\mathcal{V}^2(X;\omega) = \|f_{X,X}(t,s) - f_X(t)f_X(s)\|_\omega^2$, dostaneme analogii korelačního koeficientu, která nabývá pouze nezáporných hodnot (v případě, že $\mathcal{V}^2(X;\omega)\mathcal{V}^2(Y;\omega) = 0$, dodefinujeme nulou).

Poznámka. Tato míra závislosti je definována dokonce pro dva reálné náhodné vektory libovolné dimenze, tato práce ale obsahuje pouze případ pro dvě náhodné veličiny.

Vlastnosti této míry závislosti pak závisí na váhové funkci, kterou použijeme. Jednou z možných voleb váhové funkce je $\omega(t,s) = \frac{1}{\pi^2 t^2 s^2}$. Normu využívající zvolené váhové funkce budeme značit $\|\cdot\|$. Příslušnou míru závislosti označíme $D(X,Y)$ a říkáme jí vzdálenostní korelace.

Definice 7. *Vzdálenostní kovariance náhodných veličin X, Y s konečným prvním momentem je nezáporné číslo $\mathcal{V}(X,Y)$ definované výrazem*

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \|f_{X,Y}(t,s) - f_X(t)f_Y(s)\|^2 \\ &= \frac{1}{\pi^2}\int_{\mathbb{R}^2} \frac{|f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2}{t^2 s^2}\,\mathrm{d}t\,\mathrm{d}s.\end{aligned}$$

Dále definujeme vzdálenostní rozptyl jako odmocninu z výrazu

$$\mathcal{V}^2(X) = \|f_{X,X}(t,s) - f_X(t)f_X(s)\|^2.$$

Definice 8. *Vzdálenostní korelace náhodných veličin X, Y s konečným prvním momentem je nezáporná hodnota $D(X,Y)$ definovaná výrazem*

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

3.2 Vlastnosti vzdálenostní kovariance a vzdálenostního rozptylu

Nyní uvedeme dva způsoby vyjádření druhých mocnin vzdálenostní kovariance a vzdálenostního rozptylu. Druhý způsob nám dá, že vzdálenostní korelace nabývá hodnot z intervalu $[0,1]$.

Následující věta je výsledkem zpracování třetí části článku autorů Székelyho a Rizzové (Székely a Rizzo, 2009), důkaz je zde proveden podrobněji.

Věta 12. *Mějme náhodné veličiny X a Y s charakteristickými funkcemi $f_X(t) = \mathbb{E}\,\mathrm{e}^{\mathrm{i}\,tX}$ a $f_Y(s) = \mathbb{E}\,\mathrm{e}^{\mathrm{i}\,sY}$. Nechť $(X,Y)^T$, $(X',Y')^T$ a $(X'',Y'')^T$ jsou nezávislé stejně rozdělené náhodné vektory. Pak platí*

$$\mathcal{V}^2(X,Y) = \mathbb{E}|X-X'||Y-Y'| + \mathbb{E}|X-X'|\mathbb{E}|Y-Y'| - 2\mathbb{E}|X-X'||Y-Y''|$$

a

$$\mathcal{V}^2(X) = \mathbb{E}|X-X'|^2 + (\mathbb{E}|X-X'|)^2 - 2\mathbb{E}|X-X'||X-X''|.$$

Důkaz. Nechť $(X,Y)^T$, $(X',Y')^T$ a $(X'',Y'')^T$ jsou ze znění věty, můžeme psát

$$\begin{aligned} |f_X(t)|^2 &= f_X(t)\overline{f_X(t)} = f_X(t)f_X(-t) = \mathbb{E}\,\mathrm{e}^{\mathrm{i}\,tX}\,\mathbb{E}\,\mathrm{e}^{-\mathrm{i}\,tX} = \mathbb{E}\,\mathrm{e}^{\mathrm{i}\,t(X-X')} = f_{X-X'}(t) \\ &= \mathbb{E}\cos t(X-X'), \end{aligned}$$

podobně $|f_Y(s)|^2 = \mathbb{E}\cos s(Y-Y')$ a

$$\begin{aligned} |f_{X,Y}(t,s)|^2 &= f_{X-X',Y-Y'}(t,s) = \mathbb{E}\cos(t(X-X') + s(Y-Y')) \\ &= \mathbb{E}\cos t(X-X')\cos s(Y-Y') - \mathbb{E}\sin t(X-X')\sin s(Y-Y'). \end{aligned}$$

Dále

$$f_{X-X',Y-Y''}(t,s) = \mathbb{E}\,\mathrm{e}^{\mathrm{i}(tX+sY)}\,\mathbb{E}\,\mathrm{e}^{-\mathrm{i}\,tX'}\,\mathbb{E}\,\mathrm{e}^{-\mathrm{i}\,sY''} = f_{X,Y}(t,s)\overline{f_X(t)f_Y(s)}$$

a reálná část je

$$\begin{aligned} &\operatorname{Re} f_{X-X',Y-Y''}(t,s) = \mathbb{E}\cos(t(X-X') + s(Y-Y'')) \\ &= \mathbb{E}\cos t(X-X')\cos s(Y-Y'') - \mathbb{E}\sin t(X-X')\sin s(Y-Y''). \end{aligned}$$

Pro komplexní čísla a, b máme $|a-b|^2 = |a|^2 + |b|^2 - 2\operatorname{Re} a\bar{b}$, dále víme, že $\cos u\cos v = 1 - (1-\cos u) - (1-\cos v) + (1-\cos u)(1-\cos v)$, tedy můžeme psát

$$\begin{aligned} |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 &= \mathbb{E}\cos t(X-X')\cos s(Y-Y') \\ &\quad - \mathbb{E}\sin t(X-X')\sin s(Y-Y') + \mathbb{E}\cos t(X-X')\mathbb{E}\cos s(Y-Y') \\ &\quad - 2\mathbb{E}\cos t(X-X')\cos s(Y-Y'') + 2\mathbb{E}\sin t(X-X')\sin s(Y-Y'') \\ &= \mathbb{E}\left[(1-\cos t(X-X'))(1-\cos s(Y-Y'))\right] \\ &\quad + \mathbb{E}\left[1-\cos t(X-X')\right]\mathbb{E}\left[1-\cos s(Y-Y')\right] \\ &\quad - 2\mathbb{E}\left[(1-\cos t(X-X'))(1-\cos s(Y-Y''))\right] \\ &\quad - \mathbb{E}\sin t(X-X')\sin s(Y-Y') + 2\mathbb{E}\sin t(X-X')\sin s(Y-Y''). \end{aligned}$$

Využijeme-li toho, že

$$\frac{1}{\pi}\int_{\mathbb{R}}\frac{1-\cos tx}{t^2}\,\mathrm{d}t = |x|,$$

dostaneme

$$\frac{1}{\pi^2}\int_{\mathbb{R}^2}\frac{\mathbb{E}\left[(1-\cos t(X-X'))(1-\cos s(Y-Y'))\right]}{t^2s^2}\,\mathrm{d}t\,\mathrm{d}s = \mathbb{E}|X-X'||Y-Y'|$$

a

$$\frac{1}{\pi^2}\int_{\mathbb{R}^2}\frac{\mathbb{E}\left[1-\cos t(X-X')\right]\mathbb{E}\left[1-\cos s(Y-Y')\right]}{t^2s^2}\,\mathrm{d}t\,\mathrm{d}s = \mathbb{E}|X-X'|\mathbb{E}|Y-Y'|.$$

Protože integrál členů obsahujících funkce sinus je nulový, máme dohromady

$$\mathcal{V}^2(X,Y) = \mathbb{E}|X-X'||Y-Y'| + \mathbb{E}|X-X'|\mathbb{E}|Y-Y'| - 2\mathbb{E}|X-X'||Y-Y''|.$$

Podobně platí

$$\mathcal{V}^2(X) = \mathbb{E}|X-X'|^2 + \left(\mathbb{E}|X-X'|\right)^2 - 2\mathbb{E}|X-X'||X-X''|.$$

□

Znění další věty je převzato z publikace Lyonse (Lyons, 2013), důkaz je vlastní.

Věta 13. *Máme-li* $(X,Y)^T$, $(X',Y')^T$ *nezávislé stejně rozdělené náhodné vektory, pak lze* $\mathcal{V}^2(X,Y)$ *vyjádřit jako*

$$\mathcal{V}^2(X,Y) = \mathbb{E}g(X,X')g(Y,Y'),$$

kde

$$g(X,X') = |X-X'| - \mathbb{E}\left[|X-X'|\,|\,X'\right] - \mathbb{E}\left[|X-X'|\,|\,X\right] + \mathbb{E}|X-X'|.$$

Podobně platí $\mathcal{V}^2(X) = \mathbb{E}g(X,X')^2$ *a* $\mathcal{V}^2(Y) = \mathbb{E}g(Y,Y')^2$.

Důkaz. Nechť $(X,Y)^T$, $(X',Y')^T$ a $(X'',Y'')^T$ jsou nezávislé stejně rozdělené náhodné vektory. Rozepišme výraz $\mathbb{E}g(X,X')g(Y,Y')$:

$$\begin{aligned}
\mathbb{E}g(X,X')g(Y,Y') &= \mathbb{E}|X-X'||Y-Y'| - \mathbb{E}\left[|X-X'|\mathbb{E}\left[|Y-Y'|\,|\,Y'\right]\right] \\
&\quad - \mathbb{E}\left[|X-X'|\mathbb{E}\left[|Y-Y'|\,|\,Y\right]\right] + \mathbb{E}|X-X'|\mathbb{E}|Y-Y'| \\
&\quad - \mathbb{E}\left[|Y-Y'|\mathbb{E}\left[|X-X'|\,|\,X'\right]\right] + \mathbb{E}\left[\mathbb{E}\left[|X-X'|\,|\,X'\right]\mathbb{E}\left[|Y-Y'|\,|\,Y'\right]\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[|X-X'|\,|\,X'\right]\mathbb{E}\left[|Y-Y'|\,|\,Y\right]\right] - \mathbb{E}|X-X'|\mathbb{E}|Y-Y'| \\
&\quad - \mathbb{E}\left[|Y-Y'|\mathbb{E}\left[|X-X'|\,|\,X\right]\right] + \mathbb{E}\left[\mathbb{E}\left[|X-X'|\,|\,X\right]\mathbb{E}\left[|Y-Y'|\,|\,Y'\right]\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[|X-X'|\,|\,X\right]\mathbb{E}\left[|Y-Y'|\,|\,Y\right]\right] - \mathbb{E}|X-X'|\mathbb{E}|Y-Y'| \\
&\quad + \mathbb{E}|X-X'|\left(\mathbb{E}|Y-Y'| - \mathbb{E}|Y-Y'| - \mathbb{E}|Y-Y'| + \mathbb{E}|Y-Y'|\right) \\
&= \mathbb{E}|X-X'||Y-Y'| + \mathbb{E}|X-X'|\mathbb{E}|Y-Y'| - 2\mathbb{E}\left[|Y-Y'|\mathbb{E}\left[|X-X'|\,|\,X\right]\right] \\
&\quad - 2\mathbb{E}\left[|X-X'|\mathbb{E}\left[|Y-Y'|\,|\,Y\right]\right] + 2\mathbb{E}\left[\mathbb{E}\left[|X-X'|\,|\,X\right]\mathbb{E}\left[|Y-Y'|\,|\,Y\right]\right],
\end{aligned}$$

kde jsme si všimli, že $\mathbb{E}\left[|X-X'|\,|\,X\right]$ je funkce X, $\mathbb{E}\left[|Y-Y'|\,|\,Y'\right]$ je funkce Y', tedy tyto dvě podmíněné střední hodnoty jsou nezávislé náhodné veličiny, proto

$$\mathbb{E}\left[\mathbb{E}\left[|X-X'|\,|\,X\right]\mathbb{E}\left[|Y-Y'|\,|\,Y'\right]\right] = \mathbb{E}|X-X'|\mathbb{E}|Y-Y'|,$$

podobně

$$\mathbb{E}\left[\mathbb{E}\left[|X-X'|\,|\,X'\right]\mathbb{E}\left[|Y-Y'|\,|\,Y\right]\right]=\mathbb{E}|X-X'|\mathbb{E}|Y-Y'|.$$

Dále z toho, že $(X,Y)^T$ a $(X',Y')^T$ jsou stejně rozdělené, plyne

$$\begin{aligned}\mathbb{E}\left[|Y-Y'|\mathbb{E}\left[|X-X'|\,|\,X'\right]\right]&=\mathbb{E}\left[|Y-Y'|\mathbb{E}\left[|X-X'|\,|\,X\right]\right],\\ \mathbb{E}\left[|X-X'|\mathbb{E}\left[|Y-Y'|\,|\,Y'\right]\right]&=\mathbb{E}\left[|X-X'|\mathbb{E}\left[|Y-Y'|\,|\,Y\right]\right]\end{aligned}$$

a

$$\mathbb{E}\left[\mathbb{E}\left[|X-X'|\,|\,X'\right]\mathbb{E}\left[|Y-Y'|\,|\,Y'\right]\right]=\mathbb{E}\left[\mathbb{E}\left[|X-X'|\,|\,X\right]\mathbb{E}\left[|Y-Y'|\,|\,Y\right]\right].$$

Nyní označme $m_1(x)=\mathbb{E}|x-X'|$ a $m_2(y)=\mathbb{E}|y-Y'|$. Pak

$$\begin{aligned}&\mathbb{E}\left[|X-X'|\mathbb{E}\left[|Y-Y'|\,|\,Y\right]\right]=\mathbb{E}[m_2(Y)|X-X'|]\\ &=\mathbb{E}\left[\mathbb{E}\left[m_2(Y)|X-X'|\,|\,(X,Y)^T\right]\right]=\mathbb{E}\left[m_2(Y)\mathbb{E}[|X-X'|\,|\,X]\right]\\ &=\mathbb{E}m_1(X)m_2(Y),\end{aligned}$$

podobně

$$\mathbb{E}\left[|Y-Y'|\mathbb{E}\left[|X-X'|\,|\,X\right]\right]=\mathbb{E}m_1(X)m_2(Y)$$

a dále taky

$$\mathbb{E}|X-X'||Y-Y''|=\mathbb{E}\left[\mathbb{E}\left[|X-X'||Y-Y''|\,|\,(X,Y)^T\right]\right]=\mathbb{E}m_1(X)m_2(Y).$$

Z toho plyne

$$\mathbb{E}g(X,X')g(Y,Y')=\mathbb{E}|X-X'||Y-Y'|+\mathbb{E}|X-X'|\mathbb{E}|Y-Y'|-2\mathbb{E}|X-X'||Y-Y''|.$$

Dospěli jsme k vyjádření druhé mocniny vzdálenostní kovariance z věty 12, tedy tvrzení o vzdálenostní kovarianci je dokázáno. Tvrzení o vzdálenostním rozptylu se dokáže obdobně. □

3.3 Vlastnosti vzdálenostní korelace

Následující lemma je přímým důsledkem věty 13, můžeme ho najít i s důkazem v publikaci Lyonse (Lyons, 2013).

Lemma 14. *Nechť X, Y jsou náhodné veličiny, pak $D^2(X,Y)\in[0,1]$.*

Důkaz. Tvrzení plyne z vyjádření z věty 13 a Cauchyho–Schwarzovy nerovnosti:

$$\mathcal{V}^2(X,Y)=\mathbb{E}g(X,X')g(Y,Y')\leq\sqrt{\mathbb{E}g(X,X')^2}\sqrt{\mathbb{E}g(Y,Y')^2}=\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}.$$

□

Příští větu a její důkaz autorka zpracovala na základě článku Székelyho a kol. (Székely a kol., 2007).

Věta 15. *Pokud X, Y mají sdružené normální rozdělení s nulovou střední hodnotou a varianční maticí $\left(\begin{smallmatrix}1&\rho\\ \rho&1\end{smallmatrix}\right)$, pak $D(X,Y)\leq|\rho|$.*

Důkaz. Nechť X, Y jsou jako ve znění věty. Charakteristická funkce náhodného vektoru $(X,Y)^T$ je $f_{X,Y}(t,s) = \mathrm{e}^{-(t^2+s^2)/2-\rho ts}$ a náhodné veličiny X je $f_X(t) = \mathrm{e}^{-t^2/2}$, analogicky pro Y. Zavedeme funkci

$$F(\rho) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(\mathrm{e}^{-(t^2+s^2)/2-\rho ts} - \mathrm{e}^{-t^2/2}\,\mathrm{e}^{-s^2/2}\right)^2 \frac{\mathrm{d}t}{t^2}\,\frac{\mathrm{d}s}{s^2}.$$

Je to druhá mocnina vzdálenostní kovariance vynásobená π^2. Tento výraz upravíme:

$$\begin{aligned} F(\rho) &= \int_{\mathbb{R}^2} \left(\mathrm{e}^{-(t^2+s^2)-2\rho ts} - 2\,\mathrm{e}^{-(t^2+s^2)-\rho ts} + \mathrm{e}^{-(t^2+s^2)}\right) \frac{\mathrm{d}t}{t^2}\,\frac{\mathrm{d}s}{s^2} \\ &= \int_{\mathbb{R}^2} \mathrm{e}^{-(t^2+s^2)} \left(\mathrm{e}^{-2\rho ts} - 2\,\mathrm{e}^{-\rho ts} + 1\right) \frac{\mathrm{d}t}{t^2}\,\frac{\mathrm{d}s}{s^2}. \end{aligned}$$

Nyní exponenciály uvnitř závorky rozvineme do řad:

$$\mathrm{e}^{-2\rho ts} - 2\,\mathrm{e}^{-\rho ts} + 1 = \sum_{k=0}^{\infty} \frac{(-2\rho ts)^k}{k!} - 2\sum_{k=0}^{\infty} \frac{(-\rho ts)^k}{k!} + 1 = \sum_{k=2}^{\infty} \frac{(-\rho ts)^k}{k!}(2^k - 2),$$

neboť nultý člen sumy je -1 a první je 0. Pokračujme v úpravě $F(\rho)$:

$$\begin{aligned} F(\rho) &= \int_{\mathbb{R}^2} \mathrm{e}^{-(t^2+s^2)} \sum_{k=2}^{\infty} \frac{(-\rho ts)^k}{k!}(2^k - 2)\,\frac{\mathrm{d}t}{t^2}\,\frac{\mathrm{d}s}{s^2} \\ &= \sum_{k=2}^{\infty} \int_{\mathbb{R}^2} \mathrm{e}^{-(t^2+s^2)}\,\frac{(-\rho ts)^k}{k!}(2^k - 2)\,\frac{\mathrm{d}t}{t^2}\,\frac{\mathrm{d}s}{s^2} \\ &= \sum_{n=1}^{\infty} \int_{\mathbb{R}^2} \mathrm{e}^{-(t^2+s^2)}\,\frac{(-\rho ts)^{2n}}{(2n)!}(2^{2n} - 2)\,\frac{\mathrm{d}t}{t^2}\,\frac{\mathrm{d}s}{s^2} \\ &= \rho^2 \left(\sum_{n=1}^{\infty} \frac{2^{2n}-2}{(2n)!}\rho^{2(n-1)} \int_{\mathbb{R}^2} \mathrm{e}^{-(t^2+s^2)}(ts)^{2(n-1)}\,\mathrm{d}t\,\mathrm{d}s\right). \end{aligned}$$

Zůstanou pouze sudé členy, protože pro k liché dostaneme integrál liché funkce přes celé $\mathbb{R}$, tedy nulu. Máme $F(\rho) = \rho^2 G(\rho)$, kde $G(\rho)$ je suma, jejíž všechny členy jsou nezáporné a měřitelné funkce, z čehož plyne korektnost prohození sumy a integrálu. Dále $G(\rho)$ je neklesající v ρ, tedy $G(\rho) \le G(1)$. Proto

$$D^2(X,Y) = \frac{F(\rho)}{F(1)} = \rho^2 \frac{G(\rho)}{G(1)} \le \rho^2.$$

Neboli $D(X,Y) \le |\rho|$. □

Pokud X, Y mají sdružené normální rozdělení jako ve větě 15, lze vzdálenostní korelaci X a Y analyticky vyjádřit jako funkci ρ. Vzorec lze najít v Székely a kol. (2007).

Následující lemma je vlastní prací autorky.

Lemma 16. *Pokud X a Y jsou náhodné veličiny, které nabývají dvou hodnot s.j., pak $D(X,Y) = |\rho(X,Y)|$.*

Důkaz. Mějme dvě náhodné veličiny X, Y, jejichž sdružené rozdělení je definované následující tabulkou pravděpodobností:

$Y\backslash X$	x_1	x_2	
y_1	a	b	$a+b$
y_2	c	$1-(a+b+c)$	$1-(a+b)$
	$a+c$	$1-(a+c)$	1

Tabulka 3.1: Tabulka pravděpodobností.

Bez újmy na obecnosti předpokládejme, že $x_1 < x_2$ a $y_1 < y_2$. Spočítejme kovarianci a rozptyl X (rozptyl Y vyjde stejně, jen místo c bude b a místo x_1, x_2 bude y_1, y_2) a poté korelační koeficient:

$$\begin{aligned}
&\operatorname{cov}(X,Y) = x_1y_1(a-(a+c)(a+b)) + x_1y_2(c-(a+c)(1-a-b))\\
&\quad + x_2y_1(b-(a+b)(1-a-c)) + x_2y_2(1-a-b-c-(1-a-b)(1-a-c))\\
&= (a-(a+b)(a+c))(x_1-x_2)(y_1-y_2),\\
&\operatorname{var} X = x_1^2(a+c-(a+c)^2) - 2x_1x_2(a+c)(1-a-c)\\
&\quad + x_2^2(1-a-c-(1-a-c)^2) = (a+c-(a+c)^2)(x_1-x_2)^2,\\
&\rho(X,Y) = \frac{a-(a+b)(a+c)}{\sqrt{(a+b)-(a+b)^2}\sqrt{(a+c)-(a+c)^2}}.
\end{aligned}$$

Vyjádřeme $|f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2$:

$$\begin{aligned}
&\mathbb{E}\,\mathrm{e}^{\mathrm{i}tX+\mathrm{i}sY} = a\,\mathrm{e}^{\mathrm{i}x_1t+\mathrm{i}y_1s} + b\,\mathrm{e}^{\mathrm{i}x_2t+\mathrm{i}y_1s} + c\,\mathrm{e}^{\mathrm{i}x_1t+\mathrm{i}y_2s} + (1-a-b-c)\,\mathrm{e}^{\mathrm{i}x_2t+\mathrm{i}y_2s},\\
&\mathbb{E}\,\mathrm{e}^{\mathrm{i}tX} = (a+c)\,\mathrm{e}^{\mathrm{i}x_1t} + (1-a-c)\,\mathrm{e}^{\mathrm{i}x_2t},\\
&\mathbb{E}\,\mathrm{e}^{\mathrm{i}sY} = (a+b)\,\mathrm{e}^{\mathrm{i}y_1s} + (1-a-b)\,\mathrm{e}^{\mathrm{i}y_2s},\\
&\left|\mathbb{E}\,\mathrm{e}^{\mathrm{i}tX+\mathrm{i}sY} - \mathbb{E}\,\mathrm{e}^{\mathrm{i}tX}\,\mathbb{E}\,\mathrm{e}^{\mathrm{i}sY}\right|^2 = |(a-(a+b)(a+c))\,\mathrm{e}^{\mathrm{i}x_1t+\mathrm{i}y_1s}\\
&\quad + (b-(a+b)(1-a-c))\,\mathrm{e}^{\mathrm{i}x_2t+\mathrm{i}y_1s} + (c-(a+c)(1-a-b))\,\mathrm{e}^{\mathrm{i}x_1t+\mathrm{i}y_2s}\\
&\quad + (1-a-b-c-(1-a-b)(1-a-c))\,\mathrm{e}^{\mathrm{i}x_2t+\mathrm{i}y_2s}\,|^2\\
&= (a-(a+b)(a+c))^2|\,\mathrm{e}^{\mathrm{i}x_1t} - \mathrm{e}^{\mathrm{i}x_2t}\,|^2|\,\mathrm{e}^{\mathrm{i}y_1s} - \mathrm{e}^{\mathrm{i}y_2s}\,|^2.
\end{aligned}$$

Podobně vypočítáme, že

$$\left|\mathbb{E}\,\mathrm{e}^{\mathrm{i}tX+\mathrm{i}sX} - \mathbb{E}\,\mathrm{e}^{\mathrm{i}tX}\,\mathbb{E}\,\mathrm{e}^{\mathrm{i}sX}\right|^2 = (a+c-(a+c)^2)^2|\,\mathrm{e}^{\mathrm{i}x_1t} - \mathrm{e}^{\mathrm{i}x_2t}\,|^2|\,\mathrm{e}^{\mathrm{i}x_1s} - \mathrm{e}^{\mathrm{i}x_2s}\,|^2,$$

$$\left|\mathbb{E}\,\mathrm{e}^{\mathrm{i}tY+\mathrm{i}sY} - \mathbb{E}\,\mathrm{e}^{\mathrm{i}tY}\,\mathbb{E}\,\mathrm{e}^{\mathrm{i}sY}\right|^2 = (a+b-(a+b)^2)^2|\,\mathrm{e}^{\mathrm{i}y_1t} - \mathrm{e}^{\mathrm{i}y_2t}\,|^2|\,\mathrm{e}^{\mathrm{i}y_1s} - \mathrm{e}^{\mathrm{i}y_2s}\,|^2.$$

Teď už můžeme vyjádřit druhé mocniny vzdálenostní kovariance a rozptylu (vzdálenostní rozptyl opět stačí pro X, pro Y analogicky):

$$\begin{aligned}
&\mathcal{V}^2(X,Y) = \frac{1}{\pi^2}\int_{\mathbb{R}^2} \frac{(a-(a+b)(a+c))^2|\,\mathrm{e}^{\mathrm{i}x_1t} - \mathrm{e}^{\mathrm{i}x_2t}\,|^2|\,\mathrm{e}^{\mathrm{i}y_1s} - \mathrm{e}^{\mathrm{i}y_2s}\,|^2}{t^2s^2}\,\mathrm{d}t\,\mathrm{d}s\\
&= \frac{1}{\pi^2}(a-(a+b)(a+c))^2\int_{-\infty}^{\infty}\frac{|\,\mathrm{e}^{\mathrm{i}x_1t} - \mathrm{e}^{\mathrm{i}x_2t}\,|^2}{t^2}\,\mathrm{d}t\int_{-\infty}^{\infty}\frac{|\,\mathrm{e}^{\mathrm{i}y_1s} - \mathrm{e}^{\mathrm{i}y_2s}\,|^2}{s^2}\,\mathrm{d}s,\\
&\mathcal{V}^2(X) = \frac{1}{\pi^2}\int_{\mathbb{R}^2} \frac{(a+c-(a+c)^2)^2|\,\mathrm{e}^{\mathrm{i}x_1t} - \mathrm{e}^{\mathrm{i}x_2t}\,|^2|\,\mathrm{e}^{\mathrm{i}x_1s} - \mathrm{e}^{\mathrm{i}x_2s}\,|^2}{t^2s^2}\,\mathrm{d}t\,\mathrm{d}s\\
&= \frac{1}{\pi^2}(a+c-(a+c)^2)^2\left(\int_{-\infty}^{\infty}\frac{|\,\mathrm{e}^{\mathrm{i}x_1t} - \mathrm{e}^{\mathrm{i}x_2t}\,|^2}{t^2}\,\mathrm{d}t\right)^2.
\end{aligned}$$

Nyní už je zřejmé, že $D^2(X,Y) = \rho^2(X,Y)$, tedy $D(X,Y) = |\rho(X,Y)|$. □

4. Příklady

4.1 Výpočet korelace polynomů stupňů 1 a 2

Nechť X nabývá pouze dvou hodnot s.j., označme je x_1 a x_2, pak libovolná borelovsky měřitelná funkce $f(X)$ závisí pouze na bodech $f(x_1)$ a $f(x_2)$. Těmito body proložíme lineární funkci $ax + b$, $a, b \in \mathbb{R}$, musí tedy platit $f(x_1) = ax_1 + b$ a $f(x_2) = ax_2 + b$. Z tohoto je zřejmé, že $f(X)$ se dá nahradit příslušnou lineární funkcí $aX + b$. Její tvar je dán vzorcem (2.1). Podobně pokud Y nabývá pouze tří hodnot s.j., pak také $g(Y)$ nabývá tří hodnot s.j., tyto body lze proložit kvadratickou funkcí a $g(Y)$ nahradit $cY^2 + dY + e$ pro nějaké $c, d, e \in \mathbb{R}$.

Platí tedy $\rho(f(X), g(Y)) = \rho(aX + b, cY^2 + dY + e)$. Nyní odvodíme vzorec pro výpočet tohoto korelačního koeficientu:

$$\begin{aligned}
&\operatorname{cov}(aX + b, cY^2 + dY + e) = ac \operatorname{cov}(X, Y^2) + ad \operatorname{cov}(X, Y) \\
&= ac(\mathbb{E}XY^2 - \mathbb{E}X\mathbb{E}Y^2) + ad(\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y), \\
&\operatorname{var}(aX + b) = a^2 \operatorname{var} X, \\
&\operatorname{var}(cY^2 + dY + e) = c^2 \operatorname{var} Y^2 + 2cd \operatorname{cov}(Y^2, Y) + d^2 \operatorname{var} Y \\
&= c^2(\mathbb{E}Y^4 - (\mathbb{E}Y^2)^2) + 2cd(\mathbb{E}Y^3 - \mathbb{E}Y^2\mathbb{E}Y) + d^2(\mathbb{E}Y^2 - (\mathbb{E}Y)^2).
\end{aligned}$$

Dohromady

$$\begin{aligned}
&\rho(aX + b, cY^2 + dY + e) \\
&= \frac{ac(\mathbb{E}XY^2 - \mathbb{E}X\mathbb{E}Y^2) + ad(\mathbb{E}XY - \mathbb{E}X\mathbb{E}Y)}{\sqrt{a^2 \operatorname{var} X \, (c^2 \operatorname{var} Y^2 + 2cd(\mathbb{E}Y^3 - \mathbb{E}Y^2\mathbb{E}Y) + d^2 \operatorname{var} Y)}}.
\end{aligned}$$

Tento výraz nezávisí na b ani na e. Pro výpočet maximální korelace X, Y bude stačit omezit se na případ, kdy $a \neq 0$, neboť pro $a = 0$ je $\rho(aX+b, cY^2+dY+e) = 0$. Z výrazu lze dále vytknout $\frac{a}{|a|} = \operatorname{sgn}(a)$, na tom maximální korelace nezávisí, neboť pokud $\operatorname{sgn}(a) = -1$, volbou $\widetilde{f} = -f$ dostáváme $\widetilde{f}(X) = \widetilde{a}X + \widetilde{b}$, kde $\widetilde{a} = -a$ a $\widetilde{b} = -b$, tedy $\operatorname{sgn}(\widetilde{a}) = 1$. Vzorec jsme tak ještě o něco zjednodušili. Z předchozího je zřejmé, že $R(X, Y) = \sup \rho(aX + b, cY^2 + dY + e)$, kde supremum bereme přes všechna $c, d \in \mathbb{R}$. Stačí se omezit na $c \neq 0$, neboť pro $c = 0$ dostáváme $\rho(aX + b, dY + e) = \operatorname{sgn}(ad)\rho(X, Y)$ podle lemmatu 5.

4.2 Diskrétní příklad na maximální korelaci

Z lemmatu 4 víme, že platí $0 \leq |\rho(X, Y)| \leq R(X, Y) \leq 1$. V první kapitole jsme uvedli příklad, kdy v první a poslední nerovnosti nastává rovnost. Nyní uvedeme příklad, kdy jsou všechny nerovnosti ostré.

Definujme sdružené rozdělení X a Y následující tabulkou pravděpodobností:

$Y\backslash X$	0	1	
0	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$
1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$
2	$\frac{1}{6}$	0	$\frac{1}{6}$
	$\frac{1}{2}$	$\frac{1}{2}$	1

Tabulka 4.1: Tabulka pravděpodobností.

Napočítejme nyní všechny momenty, které budeme k výpočtu potřebovat: $\mathbb{E}X = \frac{1}{2}$, $\mathbb{E}X^2 = \frac{1}{2}$, $\mathbb{E}Y = \frac{2}{3}$, $\mathbb{E}Y^2 = 1$, $\mathbb{E}Y^3 = \frac{5}{3}$, $\mathbb{E}Y^4 = 3$, $\mathbb{E}XY = \frac{1}{6}$, $\mathbb{E}XY^2 = \frac{1}{6}$.

Můžeme spočítat korelační koeficient X a Y:

$$\rho(X,Y) = \frac{-\frac{1}{6}}{\sqrt{\frac{1}{4}}\sqrt{\frac{5}{9}}} = -\frac{1}{\sqrt{5}}.$$

K výpočtu maximální korelace použijeme vzorec, který jsme odvodili v podkapitole 4.1, budeme uvažovat $b, d, e \in \mathbb{R}$ a $a, c \neq 0$:

$$\frac{\rho(aX+b, cY^2+dY+e)}{\operatorname{sgn}(a)} = \frac{-\frac{1}{3}c - \frac{1}{6}d}{\frac{1}{2}\sqrt{2c^2+2cd+\frac{5}{9}d^2}} = \frac{-2c-d}{\sqrt{18c^2+18cd+5d^2}} = x.$$

Poslední rovnost vynásobíme jmenovatelem na levé straně a umocníme na druhou:

$$\begin{aligned} 4c^2+4cd+d^2 &= x^2(18c^2+18cd+5d^2), \\ 0 &= c^2(18x^2-4) + cd(18x^2-4) + d^2(5x^2-1). \end{aligned}$$

Nyní vydělíme c^2, uvažujeme totiž $c \neq 0$. Obdržíme

$$0 = (18x^2-4) + \frac{d}{c}(18x^2-4) + \left(\frac{d}{c}\right)^2 (5x^2-1).$$

Spočítejme diskriminant kvadratické rovnice pro neznámou $\frac{d}{c}$:

$$D = 324x^4 - 144x^2 + 16 - 4(90x^4 - 38x^2 + 4) = -36x^4 + 8x^2.$$

Položme $D = 0$, získáme kořeny $0, \frac{\sqrt{2}}{3}, -\frac{\sqrt{2}}{3}$. Tedy $D \geq 0$ pro $|x| \leq \frac{\sqrt{2}}{3}$, pro $|x| > \frac{\sqrt{2}}{3}$ je $D < 0$ a neexistuje reálné řešení rovnice. Chceme získat maximální $|x| \leq 1$, takové, že rovnice má řešení, neboli $D \geq 0$. Takovým řešením je $|x| = \frac{\sqrt{2}}{3}$. Vidíme, že hledané maximum $\rho(aX+b, cY^2+dY+e)$ přes $b, d, e \in \mathbb{R}$, $a, c \neq 0$ je $\operatorname{sgn}(a)x = \frac{\sqrt{2}}{3}$. Vypočítali jsme, že pro výše definované X a Y platí $R(X,Y) = \frac{\sqrt{2}}{3}$.

Tedy je splněno $0 < |\rho(X,Y)| < R(X,Y) < 1$.

4.3 Diskrétní příklad na vzdálenostní korelaci

Nyní ukážeme, že neplatí $D(X,Y) \leq |\rho(X,Y)|$ nebo naopak $D(X,Y) \geq |\rho(X,Y)|$, neboť mohou nastat obě situace. Mějme sdružené rozdělení X, Y definované následující tabulkou:

$Y\backslash X$	0	1	
0	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{3}{5}$
1	$\frac{1}{10}$	0	$\frac{1}{10}$
2	$\frac{1}{5}$	$\frac{1}{10}$	$\frac{3}{10}$
	$\frac{7}{10}$	$\frac{3}{10}$	1

Tabulka 4.2: Tabulka pravděpodobností.

Vyčíslíme korelační koeficient

$$\rho(X,Y) = \frac{\frac{2}{10} - \frac{3}{10}\cdot\frac{7}{10}}{\sqrt{\frac{3}{10} - \frac{9}{100}}\sqrt{\frac{13}{10} - \frac{49}{100}}} = \frac{-\frac{1}{100}}{\frac{1}{100}\sqrt{21}\sqrt{81}} = -\frac{1}{9\sqrt{21}}.$$

Spočítáme vzdálenostní korelaci, k jejímu výpočtu využijeme větu 12. Mějme $(X',Y')^T$ a $(X'',Y'')^T$ nezávislé kopie náhodného vektoru $(X,Y)^T$. Počítejme

$$\begin{aligned}\mathbb{E}|X-X'| &= \mathbb{P}(X=0,X'=1) + \mathbb{P}(X=1,X'=0) = 2\mathbb{P}(X=0)\mathbb{P}(X=1)\\ &= 2\left(\frac{7}{10}\cdot\frac{3}{10}\right) = \frac{21}{50}.\end{aligned}$$

Zřejmě $\mathbb{E}|X-X'|^2 = \mathbb{E}|X-X'|$. Dále

$$\begin{aligned}\mathbb{E}|Y-Y'| &= \mathbb{P}(Y=0,Y'=1) + 2\mathbb{P}(Y=0,Y'=2) + \mathbb{P}(Y=1,Y'=0)\\ &\quad + \mathbb{P}(Y=1,Y'=2) + 2\mathbb{P}(Y=2,Y'=0) + \mathbb{P}(Y=2,Y'=1)\\ &= 2\mathbb{P}(Y=0)\mathbb{P}(Y=1) + 4\mathbb{P}(Y=0)\mathbb{P}(Y=2) + 2\mathbb{P}(Y=1)\mathbb{P}(Y=2)\\ &= 2\left(\frac{3}{5}\cdot\frac{1}{10}\right) + 4\left(\frac{3}{5}\cdot\frac{3}{10}\right) + 2\left(\frac{1}{10}\cdot\frac{3}{10}\right) = \frac{9}{10}.\end{aligned}$$

Podobně dostaneme, že $\mathbb{E}|Y-Y'|^2 = \frac{81}{50}$. Nyní určíme $\mathbb{E}|X-X'||Y-Y'|$. Existuje 36 kombinací, jakých hodnot můžou nabývat vektory $(X,Y)^T$ a $(X',Y')^T$, jen 12 z nich však vede na nenulovou hodnotu $|X-X'||Y-Y'|$, snadno spočítáme, že $\mathbb{E}|X-X'||Y-Y'| = \frac{19}{50}$, a podobně pak $\mathbb{E}|X-X'||Y-Y''| = \frac{189}{500}$. Pro hodnoty X, X' a X'' existuje 8 kombinací, z nichž 2 dávají nenulovou hodnotu $|X-X'||X-X''|$. Z toho vidíme, že $\mathbb{E}|X-X'||X-X''| = \frac{7}{10}\left(\frac{3}{10}\right)^2 + \left(\frac{7}{10}\right)^2\frac{3}{10} = \frac{21}{100}$. Podobně pro hodnoty Y, Y' a Y'' existuje 27 kombinací, pro 12 z nich je hodnota $|Y-Y'||Y-Y''|$ nenulová. Dopočítáme, že $\mathbb{E}|Y-Y'||Y-Y''| = \frac{441}{500}$.

Teď můžeme vyjádřit druhou mocninu vzdálenostní kovariance:

$$\begin{aligned}\mathcal{V}^2(X,Y) &= \mathbb{E}|X-X'||Y-Y'| + \mathbb{E}|X-X'|\mathbb{E}|Y-Y'| - 2\mathbb{E}|X-X'||Y-Y''|\\ &= \frac{19}{50} + \frac{21}{50}\cdot\frac{9}{10} - 2\left(\frac{189}{500}\right) = \frac{1}{500}.\end{aligned}$$

Dále vyčíslíme druhé mocniny vzdálenostních rozptylů:

$$\begin{aligned}\mathcal{V}^2(X) &= \mathbb{E}|X-X'|^2 + (\mathbb{E}|X-X'|)^2 - 2\mathbb{E}|X-X'||X-X''| \\ &= \frac{21}{50} + \left(\frac{21}{50}\right)^2 - 2\left(\frac{21}{100}\right) = \frac{441}{2500}, \\ \mathcal{V}^2(Y) &= \frac{81}{50} + \left(\frac{9}{10}\right)^2 - 2\left(\frac{441}{500}\right) = \frac{333}{500}.\end{aligned}$$

Dohromady

$$D^2(X,Y) = \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}} = \frac{\frac{1}{500}}{\sqrt{\frac{441}{2500}}\sqrt{\frac{333}{500}}} = \frac{\sqrt{5}}{21\sqrt{333}}.$$

Získali jsme $0 < |\rho(X,Y)| < D(X,Y) < 1$. Pokud bychom vypočítali maximální korelaci, dostaneme $R(X,Y) = \frac{1}{\sqrt{21}}$, dohromady tedy je $0 < |\rho(X,Y)| < D(X,Y) < R(X,Y) < 1$.

Vypočítáme-li stejným způsobem druhou mocninu vzdálenostní korelace v příkladu 4.2, zjistíme, že $D^2(X,Y) = \frac{1}{\sqrt{31}}$ a platí tedy $0 < D(X,Y) < |\rho(X,Y)| < R(X,Y) < 1$.

4.4 Spojitý příklad na vzdálenostní korelaci

Uvedeme nyní příklad, který je zobecněním příkladu v kapitole 1, kde jsme ukázali, že pro závislé veličiny může být korelační koeficient nulový. Uvažujme náhodnou veličinu X s rovnoměrným rozdělením na $(-1,1)$. Položme $Y_n = X^n$ pro $n \in \mathbb{N}$.

Ihned vidíme, že z poznámky v podkapitole 2.2 plyne vztah $R_n(X,Y_n) = R(X,Y_n) = 1$.

Zřejmě platí

$$\mathbb{E}X^k = \frac{1}{2}\int_{-1}^{1} u^k \,\mathrm{d}u = \begin{cases} 0 & \text{pro } k \text{ liché}, \\ \frac{1}{k+1} & \text{pro } k \text{ sudé}. \end{cases}$$

Proto

$$\rho(X,Y_n) = \begin{cases} \frac{\sqrt{3(2n+1)}}{n+2} & \text{pro } n \text{ liché}, \\ 0 & \text{pro } n \text{ sudé}. \end{cases}$$

Tedy pro $n = 2$ je korelační koeficient nula, což odpovídá příkladu v kapitole 1.

K výpočtu vzdálenostní korelace pomocí vyjádření z věty 12 se hodí spočítat

$$\mathbb{E}|X^k - (X')^k| = \int_{-1}^{1}\int_{-1}^{1} \frac{1}{4}|u^k - v^k| \,\mathrm{d}v \,\mathrm{d}u = \frac{1}{2}\int_{0}^{1}\int_{-1}^{1} |u^k - v^k| \,\mathrm{d}v \,\mathrm{d}u.$$

Pro k liché vyjde

$$\begin{aligned}\mathbb{E}|X^k - (X')^k| &= \frac{1}{2}\int_0^1 \left(\int_{-1}^{u}(u^k - v^k)\,\mathrm{d}v + \int_u^1 (v^k - u^k)\,\mathrm{d}v\right)\mathrm{d}u \\ &= \frac{1}{k+1}\int_0^1 (ku^{k+1} + 1)\,\mathrm{d}u = \frac{2}{k+2}.\end{aligned}$$

Zatímco pro k sudé

$$\begin{aligned}\mathbb{E}|X^k-(X')^k| &= \int_0^1\int_0^1 |u^k-v^k|\,\mathrm{d}v\,\mathrm{d}u\\ &= \int_0^1\left(\int_0^u (u^k-v^k)\,\mathrm{d}v+\int_u^1 (v^k-u^k)\,\mathrm{d}v\right)\mathrm{d}u\\ &= \int_0^1\left(\frac{2k}{k+1}u^{k+1}-u^k+\frac{1}{k+1}\right)\mathrm{d}u=\frac{2k}{(k+1)(k+2)}.\end{aligned}$$

Dále potřebujeme

$$\begin{aligned}\mathbb{E}|X^k-(X')^k|^2 &= \int_{-1}^1\int_{-1}^1 \frac{1}{4}(u^k-v^k)^2\,\mathrm{d}v\,\mathrm{d}u\\ &= \frac{1}{2}\int_{-1}^1 u^{2k}\,\mathrm{d}u-\frac{1}{2}\left(\int_{-1}^1 u^k\,\mathrm{d}u\right)^2+\frac{1}{2}\int_{-1}^1 v^{2k}\,\mathrm{d}v\\ &= \frac{2}{2k+1}-\frac{2}{(k+1)^2}\mathbb{1}\{k \text{ sudé}\}.\end{aligned}$$

Podobně se dá dopočítat $\mathbb{E}|X-X'||X^k-(X')^k|$, $\mathbb{E}|X-X'||X^k-(X'')^k|$ a $\mathbb{E}|X^k-(X')^k||X^k-(X'')^k|$. Uvedeme pouze výsledky:

$$\mathbb{E}|X-X'||X^k-(X')^k| = \begin{cases}\frac{2}{k+2} & \text{pro } k \text{ liché},\\ \frac{2k}{(k+1)(k+3)} & \text{pro } k \text{ sudé},\end{cases}$$

$$\mathbb{E}|X-X'||X^k-(X'')^k| = \begin{cases}\frac{5k+16}{3(k+2)(k+4)} & \text{pro } k \text{ liché},\\ \frac{k(5k^2+30k+46)}{3(k+1)(k+2)(k+3)(k+4)} & \text{pro } k \text{ sudé},\end{cases}$$

$$\mathbb{E}|X^k-(X')^k||X^k-(X'')^k| = \begin{cases}\frac{k+6}{(k+2)(2k+3)} & \text{pro } k \text{ liché},\\ \frac{k^2(2k^2+11k+8)}{(k+1)^2(k+2)(2k+1)(2k+3)} & \text{pro } k \text{ sudé}.\end{cases}$$

Můžeme vyčíslit druhou mocninu vzdálenostního rozptylu X jako

$$\mathcal{V}^2(X)=\frac{2}{3}+\frac{4}{9}-2\cdot\frac{7}{15}=\frac{8}{45}.$$

Dále vyjádřeme druhou mocninu vzdálenostní kovariance X a Y_n a rozptylu Y_n nejprve pro n liché

$$\begin{aligned}\mathcal{V}^2(X,Y_n) &= \frac{2}{n+2}+\frac{2}{3}\cdot\frac{2}{n+2}-2\cdot\frac{5n+16}{3(n+2)(n+4)}=\frac{8}{3(n+2)(n+4)},\\ \mathcal{V}^2(Y_n) &= \frac{2}{2n+1}+\left(\frac{2}{n+2}\right)^2-2\cdot\frac{n+6}{(n+2)(2n+3)}\\ &= \frac{4(n^2+2n+3)}{(n+2)^2(2n+1)(2n+3)}.\end{aligned}$$

Pro n sudé

$$\begin{aligned}\mathcal{V}^2(X,Y_n) &= \frac{2n}{(n+1)(n+3)} + \frac{2}{3}\cdot\frac{2n}{(n+1)(n+2)} \\ &\quad - 2\cdot\frac{n(5n^2+30n+46)}{3(n+1)(n+2)(n+3)(n+4)} \\ &= \frac{4n}{3(n+2)(n+3)(n+4)}, \\ \mathcal{V}^2(Y_n) &= \frac{2}{2n+1} - \frac{2}{(n+1)^2} + \left(\frac{2n}{(n+1)(n+2)}\right)^2 \\ &\quad - 2\cdot\frac{n^2(2n^2+11n+8)}{(n+1)^2(n+2)(2n+1)(2n+3)} \\ &= \frac{4n^2}{(n+1)(n+2)^2(2n+3)}.\end{aligned}$$

Nyní není problém vyjádřit $D^2(X,Y)$, pro n liché dostaneme

$$D^2(X,Y_n) = \frac{\sqrt{10}}{n+4}\sqrt{\frac{(2n+1)(2n+3)}{n^2+2n+3}}.$$

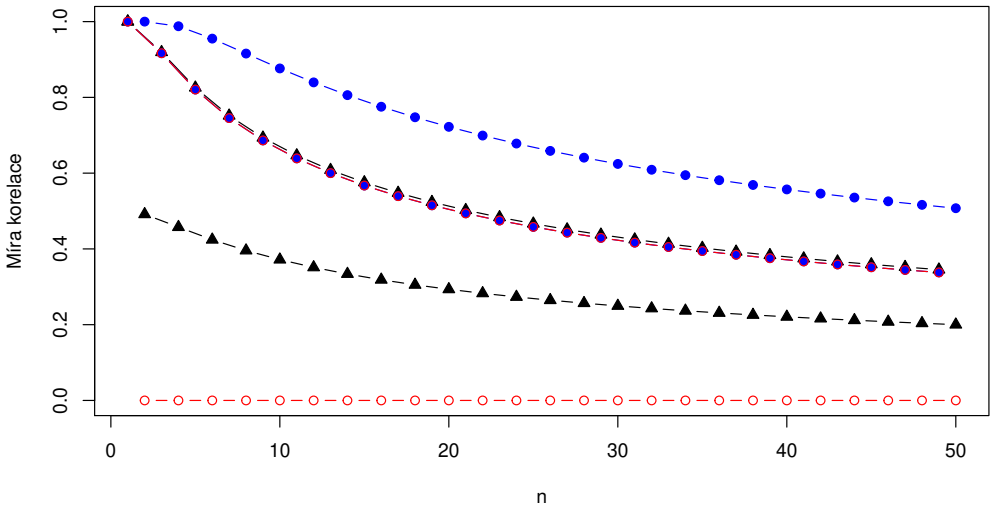
Pro n sudé

$$D^2(X,Y_n) = \frac{\sqrt{10}}{2(n+3)(n+4)}\sqrt{(n+1)(2n+3)}.$$

Vypočítáme-li $R_2(X,Y_n)$ pomocí věty 9, dostaneme úlohu na maximalizaci podobně jako ve větě 10. Po vyřešení získáme pro n liché $R_2(X,Y_n) = \rho(X,Y_n)$ a pro n sudé

$$R_2(X,Y_n) = \sqrt{5(3n+1)}\frac{\sqrt{4n^2+3n+13}}{(n+3)(2n+3)}.$$

Na závěr vykreslíme hodnoty $D(X,Y_n)$ a porovnáme je s hodnotami $\rho(X,Y_n)$ a $R_2(X,Y_n)$. Povšimněme si rozdílnosti chování koeficientů pro n sudé a liché. Pro liché n hodnoty koeficientů téměř splývají, $R_2(X,Y_n)$ vychází stejně jako $\rho(X,Y_n)$, $D(X,Y_n)$ je rovno $\rho(X,Y_n)$ pro $n=1$, pro vyšší lichá n je o něco větší než $\rho(X,Y_n)$. V případě sudých n je $\rho(X,Y_n)=0$, $D(X,Y_n)$ je mezi hodnotami $\rho(X,Y_n)$ pro n liché a nulou a $R_2(X,Y_n)$ je větší než hodnoty $\rho(X,Y_n)$ pro n liché. Zajímavé také je, že koeficienty s rostoucím n klesají.



Obrázek 4.1: Graf porovnání $D(X, X^n)$ (černě), $\rho(X, X^n)$ (červeně) a $R_2(X, X^n)$ (modře), kde X má rovnoměrné rozdělení na intervalu $(-1, 1)$, pro liché n modrá a červená čára splývají.

Závěr

V této práci jsme se zabývali mírami závislosti mezi dvěma náhodnými veličinami. Nejprve jsme zadefinovali pojem nezávislosti náhodných veličin a uvedli jsme definici základní míry závislosti, Pearsonova korelačního koeficientu, s nímž další zavedené srovnáváme. Uvedli jsme typický příklad, který dokazuje neplatnost implikace, že nulovost korelačního koeficientu dává nezávislost náhodných veličin, v obecném případě.

Dále jsme zavedli maximální korelaci a po shrnutí vlastností, které triviálně plynou z definice, jsme ukázali, že maximální korelace je rovna nule právě tehdy, když náhodné veličiny jsou nezávislé. Navázali jsme lemmatem, které ukazuje, že se při jejím výpočtu stačí omezit na funkce s nulovou střední hodnotou a rozptylem jedna. Jako další je dokázáno, že maximální korelace je větší nebo rovná absolutní hodnotě z korelačního koeficientu a rovnost platí například, pokud náhodné veličiny nabývají pouze dvou hodnot.

Následně jsme uvedli definici maximální polynomiální korelace a návod jak ji spočítat pro stupeň 2, který jsme využili pro výpočet toho, že v případě sdruženého normálního rozdělení vyjde absolutní hodnota z korelačního koeficientu.

Po zavedení pojmů normy ve váženém L_2 prostoru a charakteristické funkce jsme definovali vzdálenostní kovarianci, rozptyl a korelaci. Pro potřeby výpočtu a odhadu vzdálenostní korelace shora jsme ukázali dva možné způsoby vyjádření a díky tomu pak dokázali, že vzdálenostní korelace nabývá hodnot mezi nulou a jedničkou. Na základě vyjádření z definice jsme ukázali, že v případě sdruženého normálního rozdělení je vzdálenostní korelace menší nebo rovná absolutní hodnotě z korelačního koeficientu a pokud náhodné veličiny nabývají dvou hodnot, pak platí rovnost.

Na závěr jsme uvedli několik příkladů, nejdříve na diskrétní rozdělení, kde jsme ukázali, že pokud jedna náhodná veličina nabývá dvou hodnot a druhá tří, může platit, že maximální korelace je ostře větší než absolutní hodnota z korelačního koeficientu a obě míry jsou ostře mezi nulou a jedničkou. Dále, že pro vzdálenostní korelaci neplatí nerovnost s absolutní hodnotou z korelačního koeficientu ani na jednu stranu. Jako poslední uvádíme příklad na spojité rozdělení, je to zobecnění příkladu, který je uveden už v první kapitole, kde ukazujeme, že pro závislé náhodné veličiny může korelační koeficient vyjít nula. Napočítali jsme na něm vzdálenostní korelaci a porovnali v obrázku s Pearsonovým korelačním koeficientem a s maximální polynomiální korelací stupně 2.

Seznam použité literatury

Anděl, J. (2005). *Základy matematické statistiky*. Vydání první. Matfyzpress, Praha. ISBN 80-86732-40-1.

Gebelein, H. (1941). Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, **21**(6), 364–379. doi: 10.1002/zamm.19410210604. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/zamm.19410210604.

Hirschfeld, H. O. (1935). A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society*, **31**(4), 520–524. doi: 10.1017/S0305004100013517. URL https://doi.org/10.1017/S0305004100013517.

Lyons, R. (2013). Distance covariance in metric spaces. *Annals of Probability*, **41** (5), 3284–3305. doi: 10.1214/12-AOP803. URL https://doi.org/10.1214/12-AOP803.

Papadatos, N. a Xifara, T. (2013). A simple method for obtaining the maximal correlation coefficient and related characterizations. *Journal of Multivariate Analysis*, **118**, 102–114. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2013.03.017. URL http://sciencedirect.com/science/article/pii/S0047259X13000444.

Rényi, A. (1959). On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, **10**(3-4), 441–451. doi: 10.1007/BF02024507. URL https://doi.org/10.1007/BF02024507.

Székely, G. J., Rizzo, M. L. a Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Annals of Statistics*, **35**(6), 2769–2794. doi: 10.1214/009053607000000505. URL https://doi.org/10.1214/009053607000000505.

Székely, G. J. a Rizzo, M. L. (2009). Brownian distance covariance. *Annals of Applied Statistics*, **3**(4), 1236–1265. doi: 10.1214/09-AOAS312. URL https://doi.org/10.1214/09-AOAS312.