Srovnání adekvátnosti metod pro zkoumání závislostí na základě korelací v datech

Martin Rosecký

Leden 2020

1 Úvod

V základních kurzech statistiky a pravděpodobnosti se studenti typicky setkávají s pojmem korelace. Tím je, v drtivé většině případů, míněn tzv. Pearsonův korelační koeficient. Autor bývá někdy vynecháván, pravděpodobně proto, aby posluchač nebyl zbytečně zmaten. Výběr tohoto konkrétního přístupu ke kvantifikaci závislostí se zdá být opodstatněný (přinejmenším s ohledem na obsah těchto kurzů). Jeho praktické využití má ale pravděpodobně více úskalí, než se může na první pohled zdát (zvlášť pro člověka bez hlubších znalostí a zkušeností). Tento text se tedy pokusí upozornit na takovéto problémy, nabídnout možnosti jejich řešení a v neposlední řadě je ověřit.

2 Pearsonův korelační koeficient

Pro to, aby bylo možné diskutovat vlastnosti a charakteristické chování Pearsonova korelačního koeficientu, bude vhodné jej nejprve zavést.
Nechť $X_1, ..., X_n$ je *náhodným výběrem* z rozdělení Q. Pak

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

nazveme *výběrovým průměrem* a

$$M^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

výběrovým rozptylem. Vzhledem k tomu, že M^2 není nestranným odhadem rozptylu, bude dále používán jeho nestranný odhad:

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

Buď $(X_1, Y_1), ..., (X_n, Y_n)$ náhodným výběrem z dvojrozměrného normálního rozdělení. Nechť $\overline{X}$ a S_X^2 resp. $\overline{Y}$ a S_Y^2 jsou charakteristikami výběrů $X_1, ..., X_n$ resp. $Y_1, ..., Y_n$. Dále mějme *výběrovou kovarianci*

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}).$$

Za předpokladu $S_X^2 > 0$ a $S_Y^2 > 0$ je pak *empirický Pearsonův korelační koeficient* dán vzorcem

$$R_P = \frac{S_{XY}}{\sqrt{S_X^2 S_Y^2}}.$$

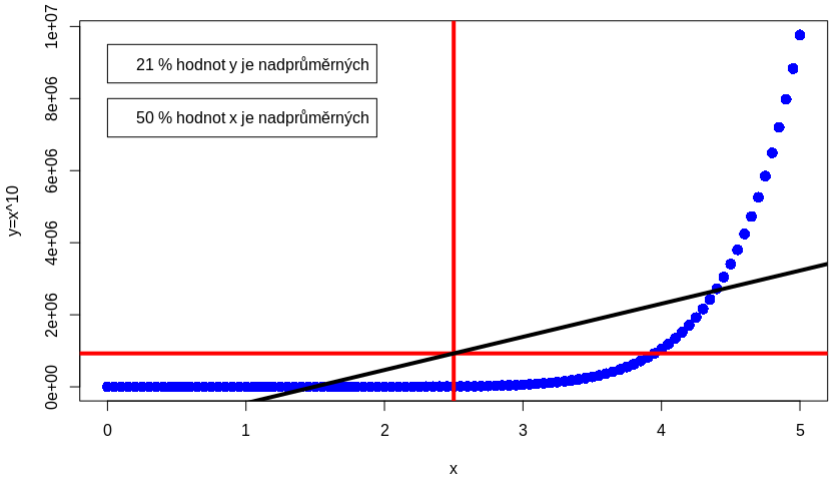
Mezi důležité vlastnosti Pearsonova korelačního koeficientu patří jeho nezávislost na posunutí a škále (důkaz viz [1]). Ze Schwarzovy nerovnosti plyne $-1 \leq R_p \leq 1$. Použití Pearsonova korelačního koeficientu předpokládá lineární závislost mezi zkoumanými proměnnými. Porušení předpokladů může vést k (přinejmenším) zavádějícím výsledkům a tedy i závěrům.

3 Důsledky nekorektního použití Pearsonova korelačního koeficientu

V technické praxi je častým problémem hledání závislosti v datech, příp. posuzování nezávislosti proměnných kvůli předpokladům statistického modelu. V důsledku používání Pearsonova korelačního koeficientu nekorektním způsobem může při hledání závislostí docházet k výrazným chybám při kvantifikaci míry zkoumané závislosti, což může vést k nesprávným závěrům. Stejně tak při použití nekorelovanosti (v Pearsonově smyslu) jako dostatečného argumentu pro posouzení nezávislosti.

Korelace by se v obecném smyslu dala popsat jako měřítko, které dokáže vystihnout toto chování: když jedna proměnná roste, roste i druhá (pozitivní korelace) příp. jedna proměnná roste a druhá klesá (negativní korelace). Pearsonův korelační koeficient ale spíše popisuje jak dobře data tvoří přímku (do jaké míry je závislost v datech lineární). Při pohledu na vzorec pro výpočet Pearsonova korelačního koeficientu je zřejmé, že na jeho chování má vliv pouze kovariance (rozptyly resp. směrodatné odchylky slouží pouze pro škálování na interval $[-1, 1]$). Z pohledu na vzorec pro kovarianci lze popsat její chování následovně: kovariance roste, pokud $X_i > \overline{X}$ a zároveň $Y_i > \overline{Y}$ (příp. pro $X_i < \overline{X}$ a zároveň $Y_i < \overline{Y}$), jinak klesá. Z této vlastnosti je dobře patrná důležitost předpokladu (dvourozměrného) normálního rozdělení. Pokud je totiž porušen (např. tím, že jedna z proměnných má asymetrické rozdělení), bude tím silně ovlivněno chování kovariance jako celku viz obr. 1. Poznamenejme, že pro zobrazenou závislost vychází $Rp = 0.66$.

Zmíněné problémy nejsou ničím novým pod Sluncem. Poměrně známým je příklad Anscombe (1973, [2]), který uvádí čtveřici statistických souborů (viz



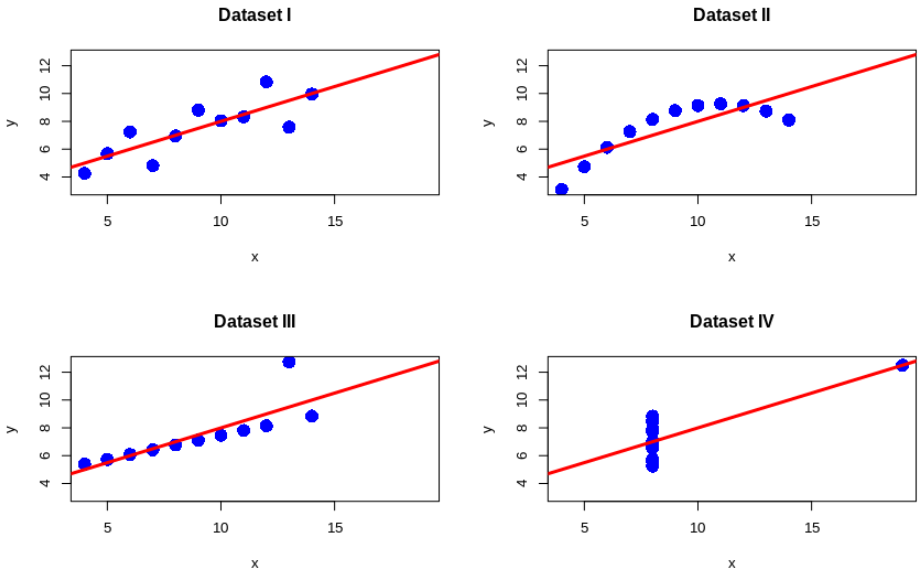
Obrázek 1: Ukázka chování Pearsonova korelačního koeficientu pro asymetrické rozdělení a nelineární závislost. Červeně jsou značeny průměry x resp. y, černě regresní přímka.

obr. 2). Tyto soubory mají stejné, nebo alespoň velmi podobné hodnoty základních sumarizačních charakteristik (výběrové průměry, rozptyly, korelaci) a také (lineární) regresní model (včetně koeficientu determinace) viz tab. 1. Tento příklad, podobně jako projekt Datasaurus ([4], viz obr. 3), primárně nekritizuje používání (Pearsonova) korelačního koeficientu. Snaží se spíše upozornit na nutnost vizualizace dat oproti jejich prostému shrnutí prostřednictvím základních měřítek (viz obě zmíněné datové sady).

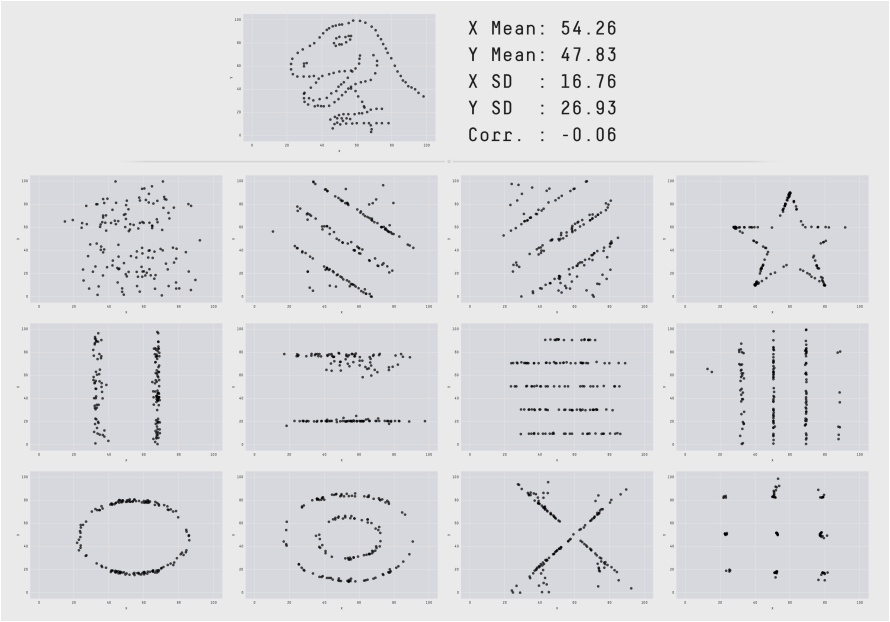
V roce 2019 se Pearsonův korelační koeficient ocitl pod silnou kritikou N. N. Taleba. Přesněji řečeno, šlo spíše o kritiku jeho nesprávného používání (při

Charakteristika	Hodnota	Přesnost
Průměr x	9	přesné
Rozptyl x	11	přesné
Průměr y	7.50	2 desetinná místa
Rozptyl y	4.125	± 0.003
Korelace mezi x a y	0.816	3 desetinná místa
Lineární regresní model	y = 3.00 + 0.500x	2 resp. 3 desetinná místa
R^2	0.67	2 desetinná místa

Tabulka 1: Charakteristiky dat Anscombova kvartetu



Obrázek 2: Anscombův kvartet



Obrázek 3: Datasaurus - bodové grafy a sumarizační charakteristiky

nesplnění předpokladů), zejména v sociálních vědách[1]. Zde byly prezentovány a demonstrovány mj. následující problémy:

- nesprávná interpretace informační hodnoty poskytované korelačním koeficientem
- nesprávné použití korelace pro nelineární závislosti.

V dalším tedy budou zkoumány zmíněné problémy v souvislosti s vybranými měřítky závislosti (s případnou zmínkou zjištěných úskalí jejich použití).

Obecným problémem s korelacemi (zejména při menším množství dat) jsou tzv. *spurious correlations*[2]. Ty odkazují na situace, kdy jsou sice nějaké proměnné silně korelovány, ale neexistuje mezi nimi žádný kauzální vztah (např. počet absolventů PhD studia a množství uranu uloženého v Amerických jaderných elektrárnách). Dalším případem falešných korelací je situace, kdy jsou 2 proměnné (např. počet požárů a množství prodané zmrzliny) „svázány“ prostřednictvím třetí proměnné (zde např. průměrná denní teplota), která buď není nebo nemůže být pozorována.

4 Alternativní přístupy

V této části budou zavedena další vybraná měřítka pro měření závislosti v datech. *Spearmanův korelační koeficient* se používá jako nejčastější alternativa k Pearsonovu, zejména pokud dojde k porušení jeho předpokladů. *Vzdálenostní korelace* je novým a nepříliš známým přístupem k tomuto problému. *Vzájemná informace* je pak navrhována N. N. Talebem (viz odkaz výše).

4.1 Spearmanův korelační koeficient

Buď $(X_1, Y_1), ..., (X_n, Y_n)$ náhodným výběrem ze spojitého dvojrozměrného rozdělení. Dále nechť $R_1, ..., R_n$ jsou pořadí veličin $X_1, ..., X_n$ a $Q_1, ..., Q_n$ odpovídají pořadí veličin $Y_1, ..., Y_n$. Pak je *Spearmanův korelační koeficient* (počítaný z dvojic $(R_1, Q_1), ..., (R_n, Q_n)$) dán vztahem:

$$R_S = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (R_i - Q_i)^2.$$

Pozn. uvedený vzorec vznikne dosazením $(R_1, Q_1), ..., (R_n, Q_n)$ do vzorce pro Pearsonův korelační koeficient. Z výše popsaného je zřejmá hlavní výhoda Spearmanova korelačního koeficientu - jedná se o neparametrický přístup, nevyžaduje tedy splnění speciálních předpokladů. Další velkou výhodou je, že Spearmanův

[1] https://www.academia.edu/39797871/Common_Misapplications_and_Misinterpretations_of_Correlation_in_Social_Science_ (Koncept)

[2] https://www.tylervigen.com/spurious-correlations

korelační koeficient není citivý na odlehlé hodnoty. Největším (a prakticky jediným zásadním) omezením tohoto přístupu je požadavek na monotonii zkoumané závislosti.

4.2 Vzdálenostní korelace (Distance correlation)

Vzdálenostní korelace (R_{DC}) je novým (2007) nástrojem pro měření závislosti mezi náhodnými vektory [5]. R_{DC} a *vzdálenostní kovariance* jsou analogem ke korelaci a kovarianci založené na použití momentových charakteristik. Narozdíl od těchto klasických definic je však R_{DC} resp. vzdálenostní kovariance nulová právě když jsou náhodné vektory nezávislé (nikoliv pouze lineárně nezávislé). Empirická vzdálenostní měřítka závislostí využívají (Euklidovské) vzdálenosti mezi prvky výběru namísto výběrových momentů.

Pro libovolná rozdělení s konečnými prvními momenty R_{DC} zobecňuje pojem korelace ve dvou zásadních ohledech:

- $R_{DC}(X, Y)$ je definovaná pro X a Y v libovolných dimenzích,
- $R_{DC}(X, Y) = 0$ značí nezávislost X a Y.

Vzdálenostní korelace také splňuje podmínku $0 \leq R_{DC}(X, Y) \leq 1$. Pro dvourozměrné normální rozdělení je $R_{DC}(X, Y)$ funkcí R_P a platí $R_{DC}(X, Y) \leq |R_P(X, Y)|$, přičemž rovnost nastává pro $R_P = \pm 1$.
Buďte $X \in \mathbb{R}^p$ resp. $Y \in \mathbb{R}^q$ náhodné vektory a $p, q \in \mathbb{N}^+$. Označme charakteristické funkce X resp. Y jako f_X resp. f_Y a sdruženou charakteristickou funkci X a Y jako $f_{X,Y}$. Vzdálenostní kovarianci ν lze použít k měření vzdálenosti $||f_{X,Y}(t, s) - f_X(t)f_Y(s)|\,|$ a tedy i k testování hypotézy o nezávislosti

$$H_0 : f_{X,Y} = f_X f_Y \qquad \text{vs.} \qquad H_1 : f_{X,Y} \neq f_X f_Y.$$

Empirická vzdálenostní kovariance $\nu_n(\mathbf{X}, \mathbf{Y})$ mezi náhodnými vektory $\mathbf{X}$ a $\mathbf{Y}$ je nezáporné číslo definované vztahem

$$\nu_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl} B_{kl},$$

kde

$$a_{kl} = |X_k - X_l|_p\,, \qquad \overline{a}_{k\cdot} = \frac{1}{n} \sum_{l=1}^{n} a_{kl}, \qquad \overline{a}_{\cdot l} = \frac{1}{n} \sum_{k=1}^{n} a_{kl},$$

$$\overline{a}_{\cdot\cdot} = \frac{1}{n^2} \sum_{k,l=1}^{n} a_{kl}, \qquad A_{kl} = a_{kl} - \overline{a}_{k\cdot} - \overline{a}_{\cdot l} + \overline{a}_{\cdot\cdot},$$

$k, l = 1, ..., n$. Analogicky definujeme $b_{kl} = |Y_k - Y_l|_q$ a $B_{kl} = b_{kl} - -\overline{b}_{k\cdot} - \overline{b}_{\cdot l} + \overline{b}_{\cdot\cdot}$ pro $k, l = 1, ..., n$.
Podobně $\nu_n^2(\mathbf{X}, \mathbf{X})$ je nezáporné číslo definované jako

$$\nu_n^2(\mathbf{X}) = \nu_n^2(\mathbf{X}, \mathbf{X}) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl}^2.$$

Empirická vzdálenostní korelace $R_{DC}(X, Y)$ je odmocninou z výrazu

$$R_{DC}^2(X, Y) = \begin{cases} \frac{\nu_n(\mathbf{X},\mathbf{Y})}{\sqrt{\nu_n(\mathbf{X})\nu_n(\mathbf{Y})}}, & \nu_n(\mathbf{X})\nu_n(\mathbf{Y}) > 0 \\ 0, & \nu_n(\mathbf{X})\nu_n(\mathbf{Y}) = 0. \end{cases} \tag{1}$$

Narozdíl od Pearsonova a Spearmanova korelačního koeficientu pro hodnotu vzdálenostní korelace platí $0 \leq R_{DC}(X, Y) \leq 1$. Empirické výsledky [5] naznačují, že vzdálenostní korelace je citlivá na všechny případy nelineárních či nemonotonních závislostí.

4.3 Vzájemná informace (Mutual Information)

Při zkoumání dat je cílem zisk nějaké informace. Její koncept je však příliš široký na to, aby byl vystihnutelný jednoduchou definicí. S jejím intuitivním uchopením však pomůže pojem *entropie*, který je též užíván v různých odvětvích (objevuje se jak v teorii informací a pravděpodobnosti, tak např. v termomechanice). Entropie je měřítko, které má mnoho vlastností, jež jsou v souladu s intuitivním vnímáním míry informace a lze jej definovat pro libovolné rozdělení pravděpodobnosti. Je ovšem potěšující a do jisté míry fascinující, že věcný význam tohoto pojmu je mezi různými odvětvími vědy totožný, ačkoliv vznikal (alespoň do určité míry) nezávisle [3]. Tento koncept je možno dále rozšířit a definovat tzv. *vzájemnou informaci*, která jest měřítkem množství informací nesené jednou náhodnou proměnnou o jiné náhodné proměnné. Vzájemná informace je pak speciálním případem *relativní entropie*, která umožňuje kvantifikovat vzdálenost mezi dvěma rozděleními pravděpodobnosti. V [3] se uvádí, že tyto pojmy vzešly přirozeným způsobem z problémů a otázek zejména z oblastí komunikace, statistiky, komplexnosti a hazardu.

Pozn. bývá zvykem uvedené pojmy definovat pro diskrétní náhodné veličiny a poté se zabývat spojitým případem. Ačkoliv jsou základní rysy a myšlenky podobné, rozdíl mezi oběma případy není pouze formální. Z důvodu stručnosti a přehlednosti je zde uveden pouze diskrétní případ, přestože korelace se většinou posuzují u proměnných, které považujeme (alespoň svojí podstatou) za spojité. Prakticky se však i u spojitých dat používá diskrétní verze výpočtu (data je nutné před samotným výpočtem diskretizovat). Spojitý případ a další informace je možné dohledat v [3].

Buď X diskrétní náhodná veličina s pravděpodobnostní funkcí

$$p(x) = P(X = x) > 0.$$

Pak *entropie H(X)* diskrétní náhodné veličiny X je definovaná jako

$$H(X) = -\sum_{x \in X} p(x) \log p(x).$$

V případě, že se ve výše uvedené definici užívá logaritmus o základu 2, je entropie vyjádřena v bitech. Snadno se nahlédne, že například entropie při hodu (dokonalou) mincí je rovna 1 bitu. V případě přirozeného logaritmu jsou jednotkou entropie tzv. *naty*. V dalším bude uvažován logaritmus o základu 2 a všechny entropie tedy budou měřeny v bitech. Poznamenejme, že entropie nezávisí na skutečné hodnotě X, ale pouze na pravděpodobnostech.

Mějme diskrétní náhodné veličiny X a Y se sdruženou pravděpodobnostní funkcí $p(x, y)$. Pak jejich *sdruženou entropii H(X,Y)* definujeme vztahem

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y).$$

Pro $(X, Y) \sim p(x, y)$ je *podmíněná entropie H(X,Y)* definována následovně

$$\begin{aligned} H(X|Y) &= \sum_{x \in X} p(x, y) H(Y, X = x) \\ &= -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log p(y|x) \\ &= -\sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y|x) \\ &= -E \log p(Y|X). \end{aligned}$$

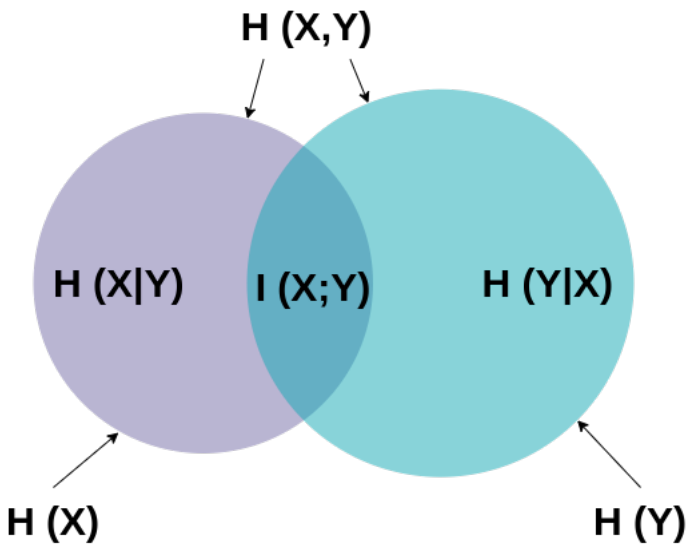
Přirozenost definice sdružené a podmíněné entropie ukazuje fakt, že entropie dvojice náhodných veličin je součtem entropie jedné veličiny a podmíněné entropie druhé veličiny (grafická ilustrace viz obr. 4), tedy

$$H(X, Y) = H(X) + H(Y|X).$$

Relativní entropie (Kullback-Leiblerova vzdálenost) $D(p||q)$ mezi dvěma pravděpodobnostními funkcemi je definovaná vztahem

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(x)}{q(x)}.$$

Buďte X a Y náhodné veličiny se sdruženou pravděpodobnostní funkcí *p(x,y)* a marginálními pravděpodobnostními funkcemi *p(x)* a *p(y)*. Pak *Vzájemná informace I(X;Y)* je relativní entropií mezi sdruženým rozdělením a součinem marginálních rozdělení, tj.



Obrázek 4: Vzájemná informace a entropie - Vennův diagram

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \\ &= D(p(x,y)||p(x)p(y)) \\ &= \mathrm{E}_{p(x,y)}\log\frac{p(X,Y)}{p(X)p(Y)}. \end{aligned}$$

Vzájemnou informaci lze vyjádřit také následovně

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X|Y).$$

Podle první z rovností tedy vzájemná informace *I(X;Y)* vyjadřuje snížení neurčitosti X v důsledku znalosti Y. Z druhé rovnosti je patrné, že X poskytuje o Y právě tolik informací jako Y o X. Vzájemná informace je tedy symetrická relace.

Důležité vlastnosti:

- $I(X;Y) \geq 0$, rovnost nastává právě tehdy když jsou X a Y nezávislé,
- $h(X|Y) \leq h(X)$, rovnost nastává právě tehdy když jsou X a Y nezávislé.

Z hlediska praktické využitelnosti je u vzájemné informace nepříjemné, že narozdíl od korelačních koeficientů není škálovaná. Její využití pro kvantifikaci síly závislosti je tedy poměrně problematické.

5 Praktická demonstrace zkoumaných problémů

V této části budou zkoumány zmíněné problémy z praktického pohledu s případným návrhem řešení.

5.1 Univerzální řešení problému s vizuální kontrolou?

Nejprve se vraťme k problému nutnosti vizualizace dat, konkrétně k Anscombově kvartetu (viz obr. 2). Ve třetí kapitole byla zmíněna neschopnost Pearsonova korelačního koeficientu v rozlišení typově výrazně odlišných závislostí. Hlavní otázka tohoto problému je nasnadě: dokáže některé ze zkoumaných měřítek odhalit, že se jedná o výrazně odlišné závislosti? Jako vodítko nám poslouží výsledky shrnuté v tabulce 2. U Spearmanova korelačního koeficientu můžeme pozorovat demonstraci jeho zmíněných vlastností (problém s posuzováním nemonotonních závislostí - Dataset II, a naopak velmi dobrá práce s odlehlými hodnotami - Dataset III). Vzdálenostní korelace se zdá být vhodnou alternativou pro nemonotonní závislosti, ale silně nadhodnocuje „závislost“ pro Dataset IV. Vzájemná informace se typově chová nejlépe z uvedených měřítek. U prvních tří datových sad vykazuje vyšší hodnoty (jejich správné uspořádání už je spíše filozofickou otázkou) a u poslední naopak udává hodnotu blízkou nule.

	Dataset I	Dataset II	Dataset III	Dataset IV
Pearson	0,82	0.82	0,82	0,82
Spearman	0,82	0.69	0,99	0,50
Vzdálenostní	0,82	0,87	0,91	0,81
Vzájemná infromace	0,71	0,84	1,34	0,10

Tabulka 2: Srovnání hodnot vybraných měřítek pro Anscombův kvartet

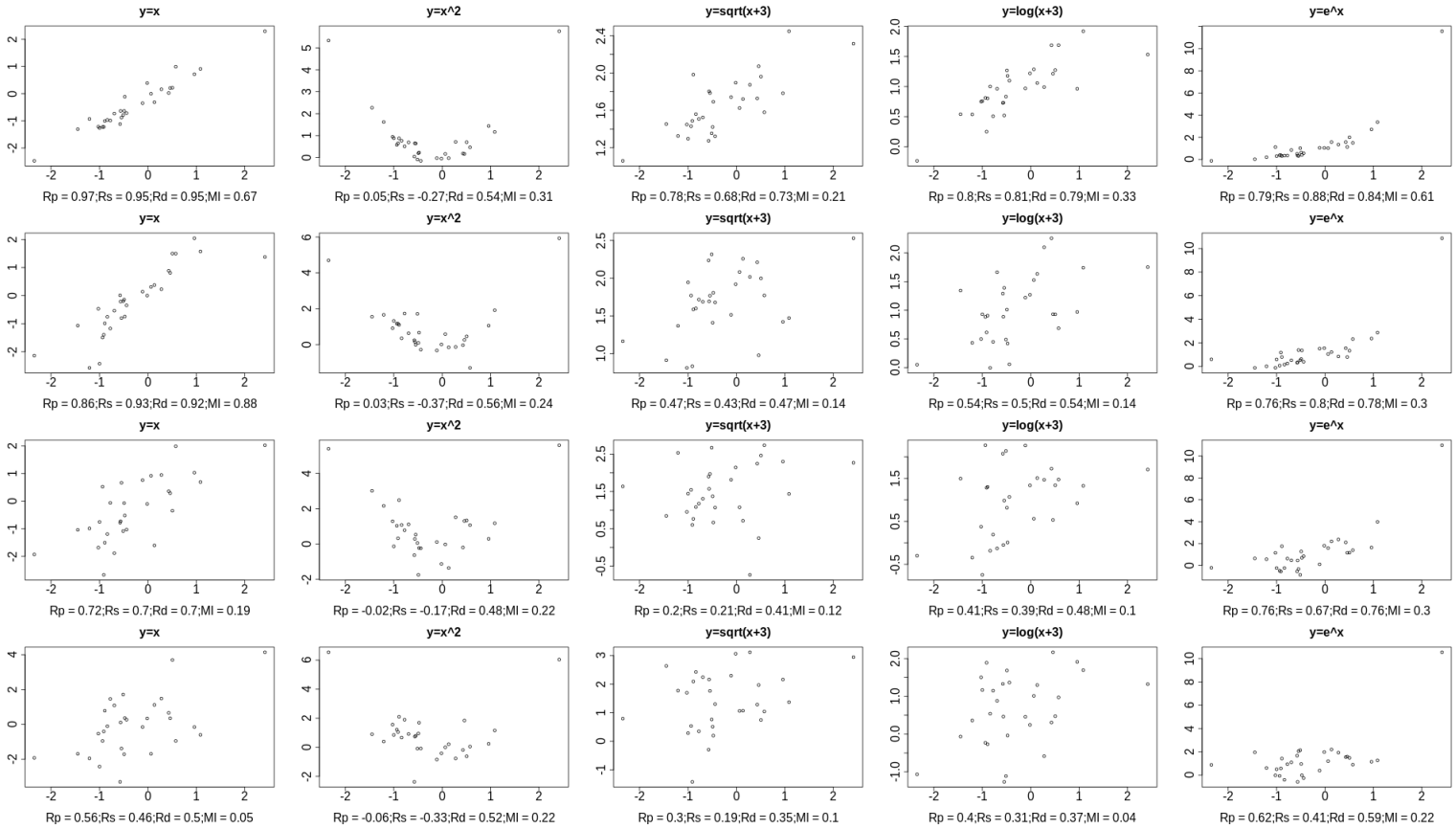
5.2 Citlivost na množství šumu a velikost vzorku u vybraných závislostí

Zde dojde ke zkoumání vhodnosti vybraných měřítek v závislosti na typu závislosti, množství šumu v datech a velikosti datového souboru. Pro demonstraci byly vybrány lineární, kvadratická a exponenciální funkce, odmocnina a logaritmus. Vzorky vznikly náhodným výběrem $X \sim N(0, 1)$, který byl následně použit jako vstup do dané funkční závislosti[3]. Následně byl k funkční hodnotě v daném bodě přidán aditivní šum $e \sim N(0, 1)$ ve zvoleném poměru. V grafech na obr. 5-9 je podíl šumu zvyšován směrem shora dolů a to postupně na 20 %, 33 %, 43 % a 50 %.

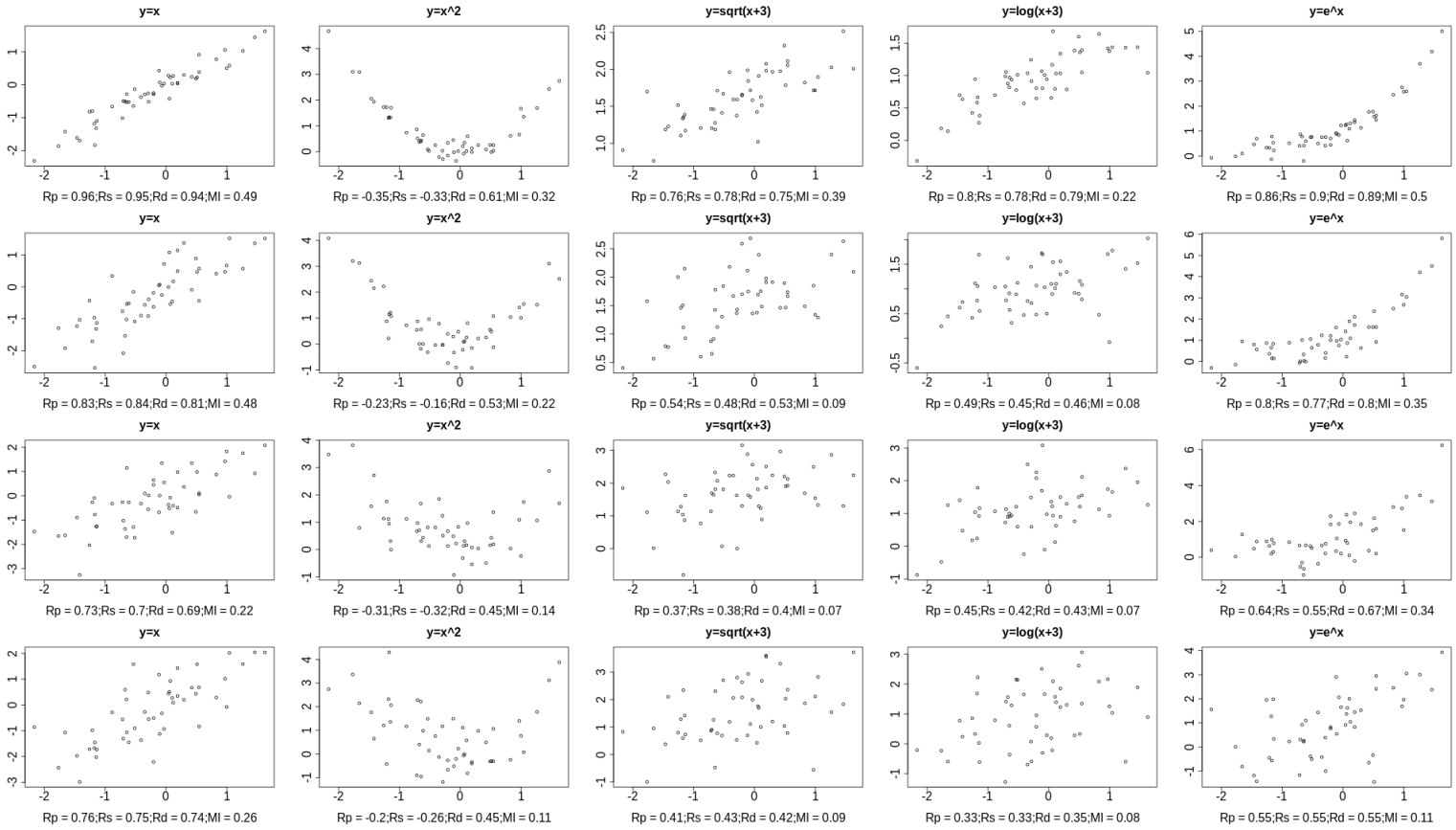
Z pohledu na následující grafy lze vyvodit několik závěrů. Prvním je, že pro menší vzorky (zde zejména soubory o velikosti 30 a 50) je chování všech vy-

[3]Odmocnina a logaritmus musely být (na rozdíl od ostatních funkcí) posunuty na ose x, tak, aby nevznikaly problémy spojené s omezením jejich definičního oboru.

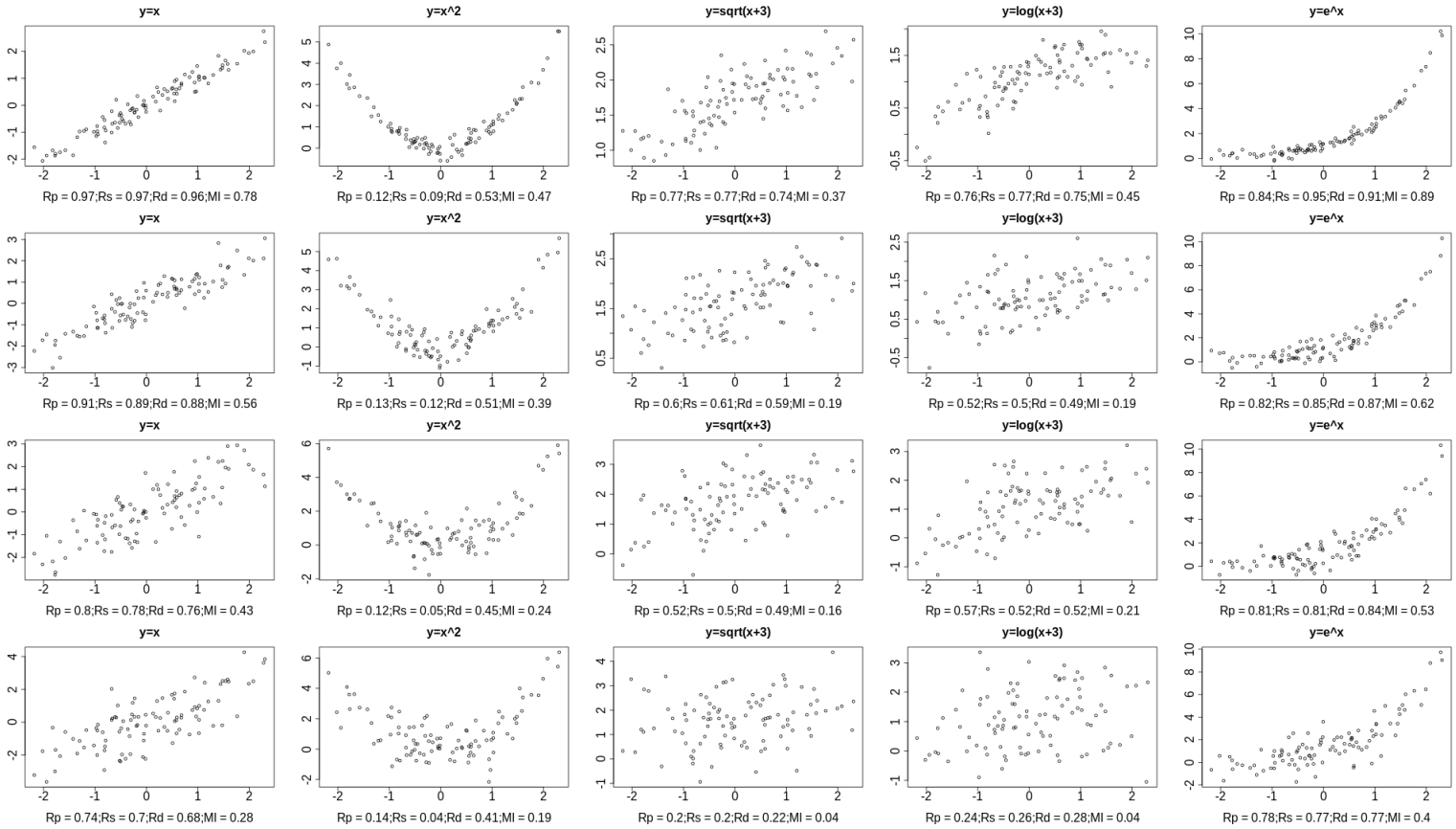
braných měřítek alespoň v nějakém případě nekonzistentní (měřítko vykazuje vyšší hodnoty pro případ, kdy je v datech více šumu). To ukazuje na značný vliv náhody pro malé vzorky. Druhá zajímavá informace se týká vzdálenostní korelace. Ta sice dokáže lépe (než Pearsonova a Spearmanova) zachytit kvadratickou závislost, ale nedokáže ji kvantifikovat dostatečně dobře (jeho hodnota je výrazně vyšší pro silně zašuměnou lineární závislost než pro slabě zašuměnou kvadratickou závislost - viz poslední graf v prvním sloupci a první graf ve druhém). Tato nepříjemnost se objevuje zejména u výběru o větším rozsahu. Dalším pozorováním je, že korelační koeficienty mají tendenci nadhodnocovat korelaci pro lineární a exponenciální závislosti. Zejména u větších datových souborů (se stabilnějším chováním) je vidět, že i když jsou data z 50 % tvořena šumem, hodnoty korelačních koeficientů se pohybují v rozmezí 0,65-0,75. Chování vzájemné informace je v těchto případech přirozenější, ale tomuto problému bude věnována následující sekce.



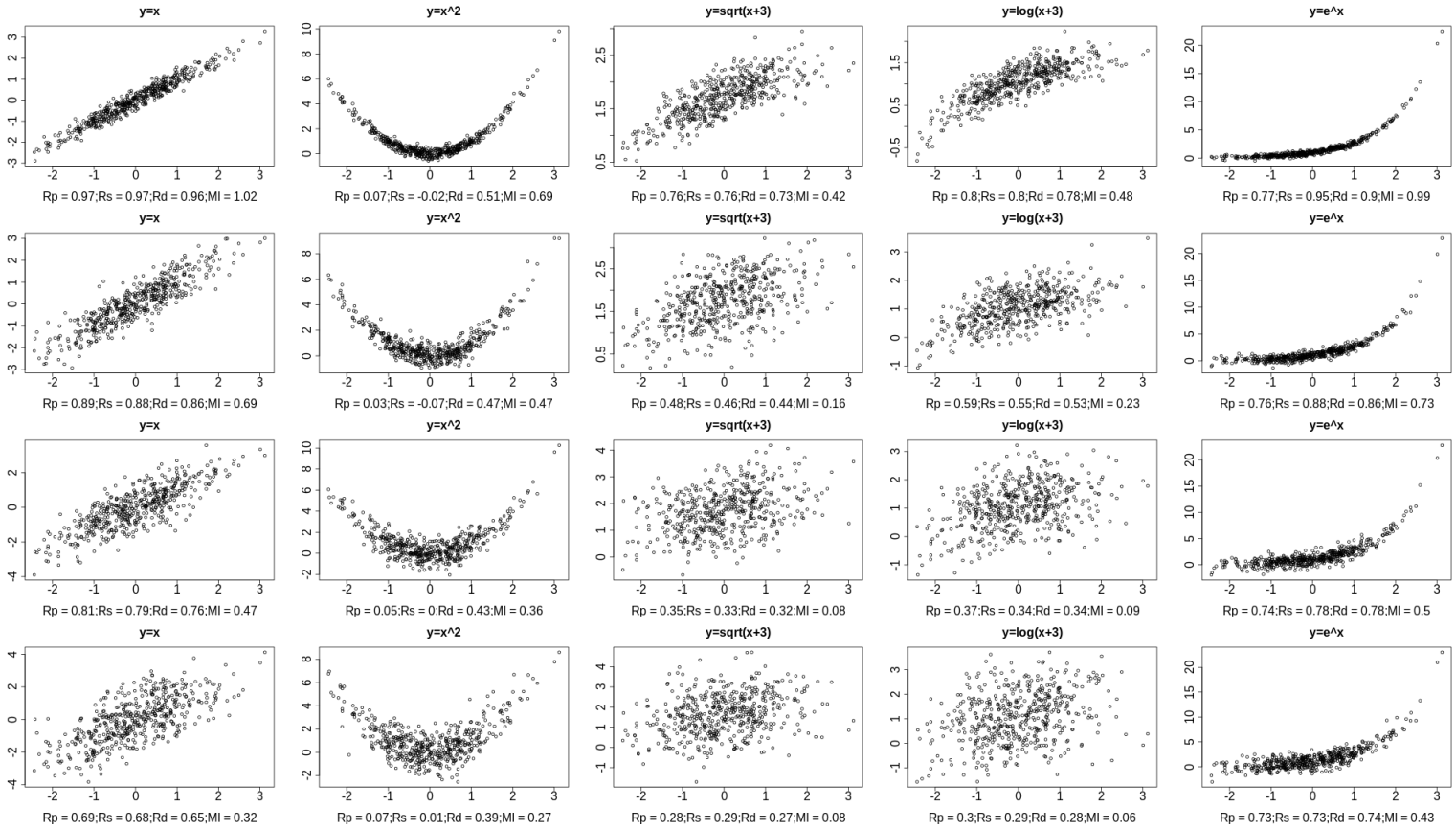
Obrázek 5: Ilustrace hodnot vzájemné informace, Pearsonova, Spearmanova a vzdálenostního korelačního koeficientu pro výběr o velikosti n=30 v závislosti na podílu šumu (shora dolů postupně 20 %, 33 %, 43 % a 50 %).



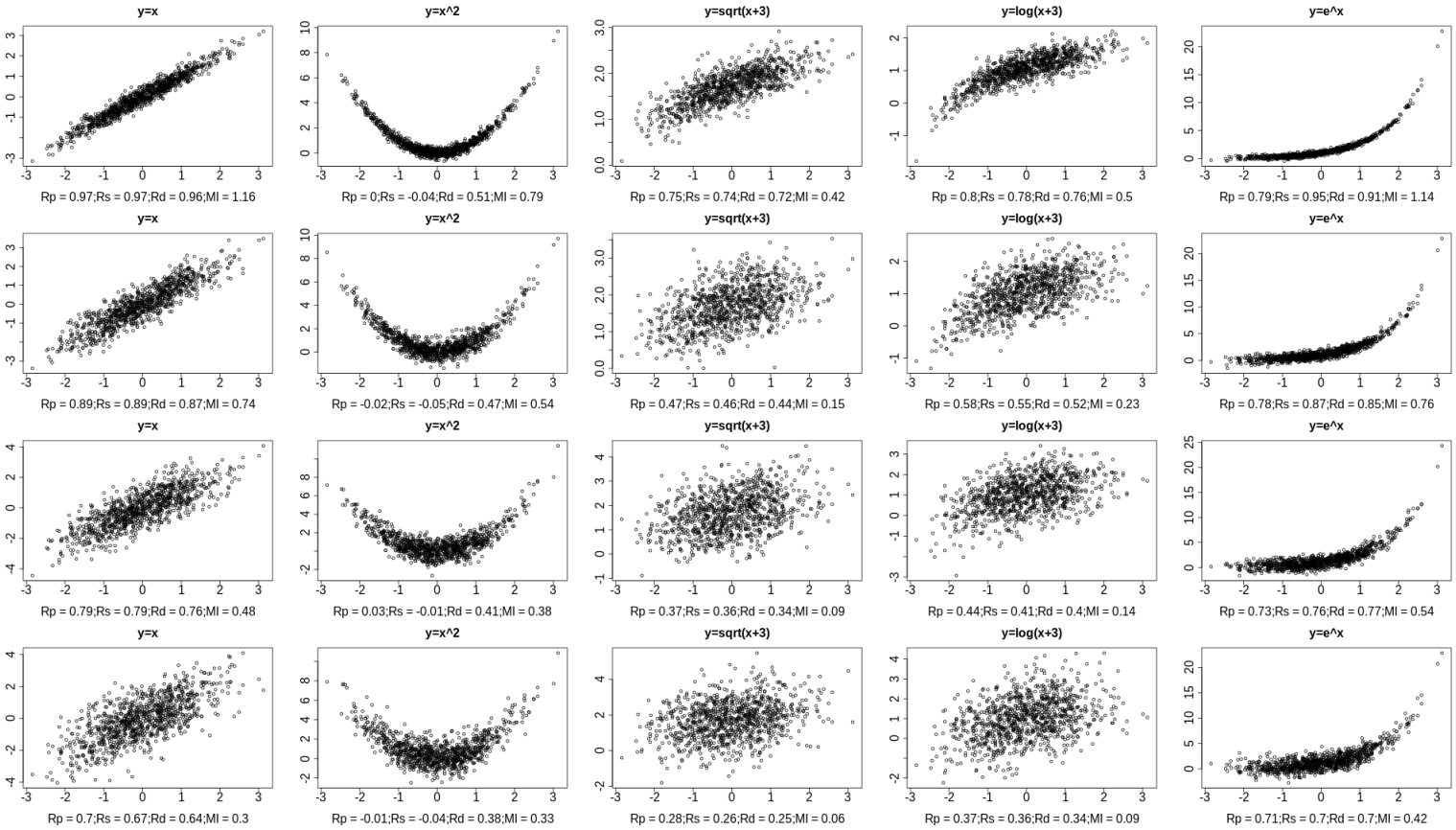
Obrázek 6: Ilustrace hodnot vzájemné informace, Pearsonova, Spearmanova a vzdálenostního korelačního koeficientu pro výběr o velikosti n=50 v závislosti na podílu šumu (shora dolů postupně 20 %, 33 %, 43 % a 50 %).



Obrázek 7: Ilustrace hodnot vzájemné informace, Pearsonova, Spearmanova a vzdálenostního korelačního koeficientu pro výběr o velikosti n=100 v závislosti na podílu šumu (shora dolů postupně 20 %, 33 %, 43 % a 50 %).



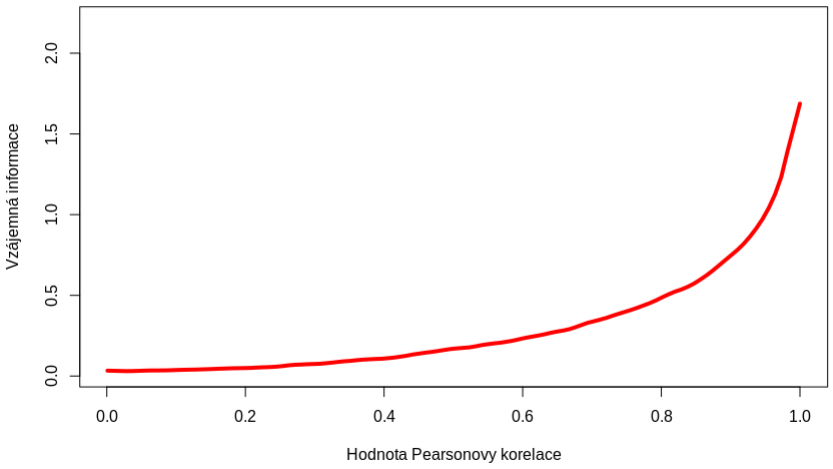
Obrázek 8: Ilustrace hodnot vzájemné informace, Pearsonova, Spearmanova a vzdálenostního korelačního koeficientu pro výběr o velikosti n=500 v závislosti na podílu šumu (shora dolů postupně 20 %, 33 %, 43 % a 50 %).



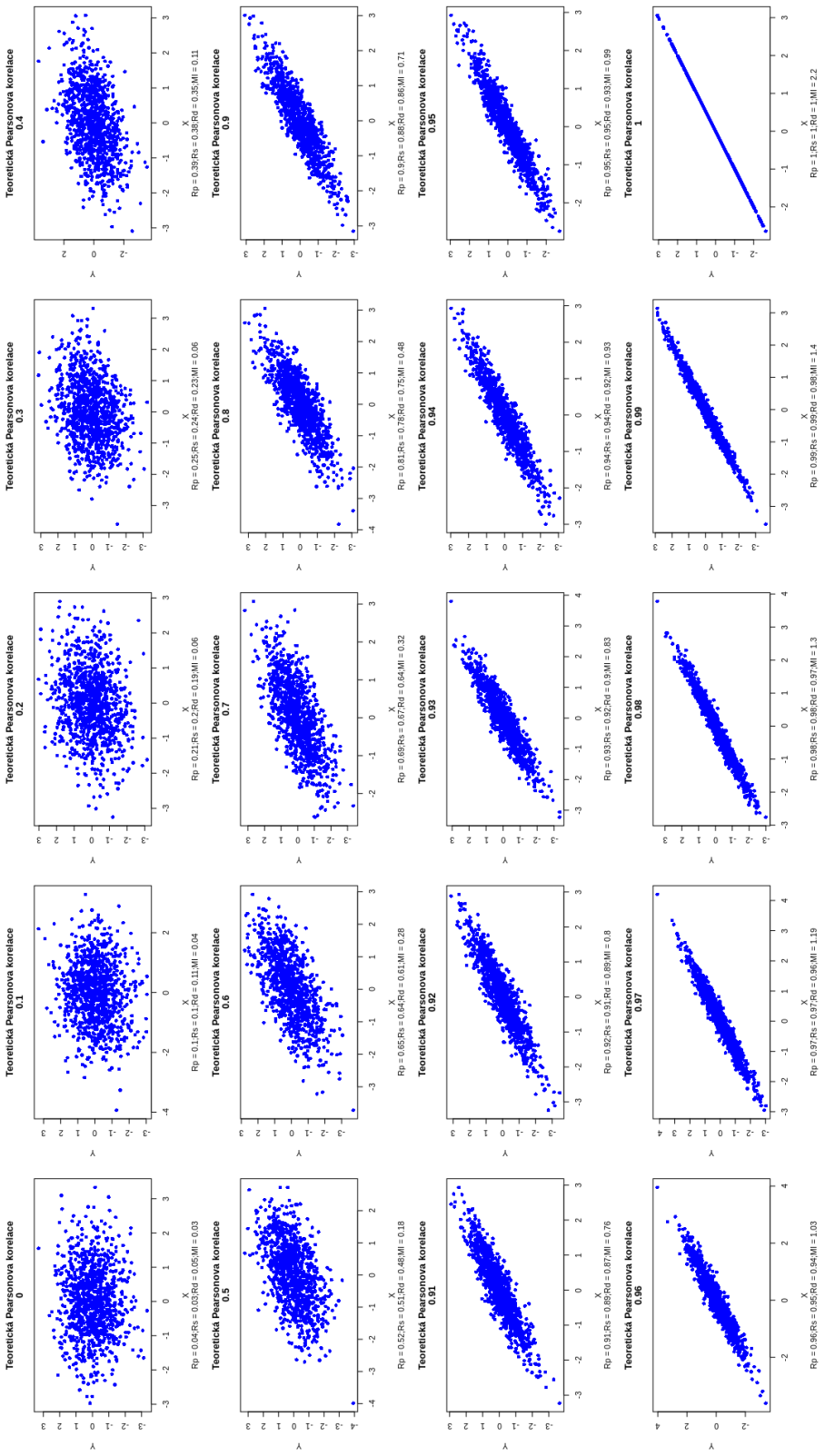
Obrázek 9: Ilustrace hodnot vzájemné informace, Pearsonova, Spearmanova a vzdálenostního korelačního koeficientu pro výběr o velikosti n=1000 v závislosti na podílu šumu (shora dolů postupně 20 %, 33 %, 43 % a 50 %).

5.3 Nelinearita informačního přínosu

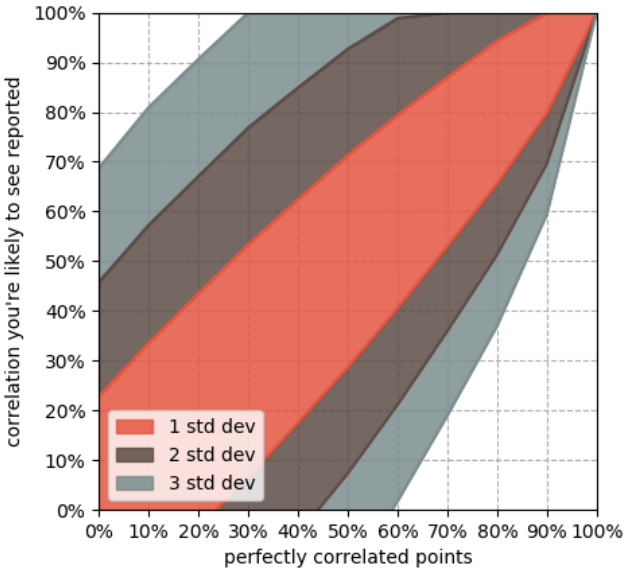
Na konci předchozí části byl zmíněn problém nelinearity informačního přínosu korelačních koeficientů. Grafickou demonstraci tohoto faktu zprostředkovávají obr. 10 a obr. 11. První z nich ukazuje vzájemnou informaci jako funkci (Pearsonova) korelačního koeficientu. Druhý ze zmíněných obrázků poskytuje názornou demonstraci. Z ní je patrné, že např. (Pearsonova) korelace o síle 0,7 má fakticky blíže k nulové korelaci než k dokonalé. Je dobré si povšimnout, že i korelace o síle 0,9 má ještě poměrně daleko do dokonalé korelace. Z uvedeného je mj. možné pozorovat, že i korelace o síle 0,5 (na nemalém množství dat) jsou poměrně diskutabilním zdrojem informací. Toto chování je dáno tím, že korelační koeficienty jsou škálovány na interval $[-1, 1]$ (resp. $[0, 1]$), přičemž kovariance jsou funkce neomezené. Situaci dokresluje obrázek 12, který demonstruje vliv nejistoty na empirickou hodnotu (Pearsonova) korelačního koeficientu. Je z něj patrné, že i při (fakticky) nulové korelaci lze vlivem náhody dosáhnout empirické hodnoty $R_p = 0,7$. Důsledek tohoto chování lze vystihnout i citací N. N. Taleba na toto téma: „Korelace neimplikuje korelaci“.



Obrázek 10: Nelineární informační přínos (Pearsonova) korelačního koeficientu



Obrázek 11: Demonstrace nelinearity informačního přínosu zkoumaných korelačních koeficientů



Obrázek 12: Grafické znázornění vlivu nejistoty na empirickou hodnotu korelačního koeficientu. Převzato z: https://shayallenhill.com/correlation-pitfalls/

6 Závěry a doporučení

Hlavní závěry tohoto textu by se daly shrnout následovně: 1. při splnění předpokladu normality a linearity je Pearsonův korelační koeficient dobrou volbou pro kvantifikaci této závislosti, 2. při porušení předpokladů je vhodné využít Spearmanovu verzi (pokud očekáváme monotonní závislost), 3. vzdálenostní korelace by měla fungovat poměrně spolehlivě i pro nemonotonní závislosti (i když kvantifikace míry závislosti není v některých případech příliš přesná). Ve všech případech je ale nutné brát na vědomí silnou nelinearitu informačního přínosu korelačních koeficientů. Jako možné východisko se jeví využití přístupů založených na vzájemné informaci.

Mezi nejdůležitější poselství tohoto textu patří upozornění na existenci vhodných měřítek pro zkoumání (obecné) nezávislosti dat (vzdálenostní korelace, vzájemná informace). Ačkoliv může tento text na čtenáře působit jako snaha o kompromitaci (zejména Pearsonova) korelačního koeficientu, opak je pravdou. Cílem nebylo ukázat, že některý ze zmíněných nástrojů by neměl být používán, ale spíše jako ukázka problémů spojených (zejména, ale ne výlučně) s porušením jejich předpokladů. Podle autorova názoru je právě taková demonstrace velmi důležitá.

7 Poděkování

Hlavní poděkování patří N. N. Talebovi, který na problém upozornil širokou veřejnost a podložil své myšlenky matematickým aparátem, čímž autora tohoto textu inspiroval k jeho vytvoření. Dále je vhodné poděkovat Shayi Allenu Hillovi a Mattimu Heinovi, jejichž blogy sloužily jako velmi cenný zdroj inspirace. V neposlední řade náleží poděkování i milé kolegyni Kristíně Šramkové za věcné poznámky a odhalení chyb v původní verzi textu.

Reference

[1] Jiří ANDĚL. *Základy matematické statistiky.* Matfyzpress, 2007.

[2] F. J. Anscombe. Graphs in statistical analysis. *The American Statistician*, 27(1):17–21, 1973.

[3] T. M. Cover and Joy A. Thomas. *Elements of information theory.* Wiley-Interscience, Hoboken, N.J., 2nd ed edition, c2006.

[4] Justin Matejka and George Fitzmaurice. Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, page 1290–1294, New York, NY, USA, 2017. Association for Computing Machinery.

[5] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *Ann. Statist.*, 35(6):2769–2794, 12 2007.