VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV TELEKOMUNIKACÍ
DEPARTMENT OF TELECOMMUNICATIONS

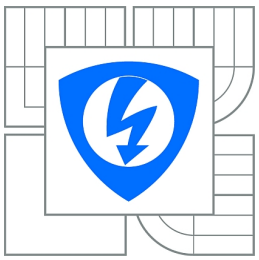
VIZUALIZACE BIOMEDICINSKÝCH DAT V PROSTŘEDÍ MATLAB
BIOMEDICAL DATA VISUALIZATION USING MATLAB

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR
Bc. Vojtěch Zvončák

VEDOUCÍ PRÁCE
SUPERVISOR
Ing. Zoltán Galáž

BRNO 2016



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav telekomunikací

Diplomová práce

magisterský navazující studijní obor
Telekomunikační a informační technika

Student:	Bc. Vojtěch Zvončák	***ID:***	125347
Ročník:	2	***Akademický rok:***	2015/2016

NÁZEV TÉMATU:

Vizualizace biomedicinských dat v prostředí Matlab

POKYNY PRO VYPRACOVÁNÍ:

Zpracování biomedicínských dat je často doprovázeno jistou formou vizualizace. V rámci této diplomové práce budou naprogramovány světově uznávané a často používané metody vizualizace 1D a vícerozměrných dat jako jsou například kvantilové grafy, krabicové grafy, grady trendu, korelační grafy, jádrové odhady hustoty pravděpodobnosti atd. Tyto vizualizace budou naprogramovány v jazyce MATLAB a budou otestovány na testovacích nahrávkách řeči poškozené přítomností Parkinsonovy nemoci. Dále bude vytvořeno uživatelské rozhraní, z kterého bude možné tyto data načítat a přehledně vizualizovat pomocí metod naprogramovaných v této práci.

DOPORUČENÁ LITERATURA:

[1] Arnold, Steven F. 1993. “Gibbs sampling,” Handbook of Statistics, roč. 9, Computational Statistics, C. R. Rao, ed., The Netherlands: Elsevier Science Publishers, s. 599-625.
[2] Bickel, Peter J. and Kjell A. Doksum. 2001. Mathematical Statistics: Basic Ideas and Selected Topics, roč. 1, Second Edition, New York: Prentice Hall.
[3] Diggle, Peter J. 1981. “Some graphical methods in the analysis of spatial point patterns,” Interpreting Multivariate Data, V. Barnett, ed., New York: John Wiley & Sons, s. 55-73.

Termín zadání:	1.2.2016	***Termín odevzdání:***	25.5.2016

Vedoucí práce: Ing. Zoltán Galáž
Konzultanti diplomové práce:

doc. Ing. Jiří Mišurec, CSc.
Předseda oborové rady

UPOZORNĚNÍ:

ABSTRAKT

Diplomová práce se zabývá vizualizací biomedicínských dat v prostředí MATLAB. Práce obsahuje popis programové realizace těchto funkcí: P-P plot, Q-Q plot, histogram, box plot, jádrové odhady hustoty pravděpodobnosti, korelační diagram a popisné statistiky. Jednotlivé funkce jsou buď řešeny od základních fcí prostředí MATLAB, nebo pomocí externích fcí, které jsou vhodně upraveny pro tuto práci. Teoretické základy a praktické realizace uvedených funkcí jsou popsány v konkrétních kapitolách. Předmětem práce je také návrh a realizace uživatelského grafického rozhraní - GUI. GUI zajišťje jednoduché a rychlé volání vizualizačních funkcí a zpracování dat. V práci jsou ukázány výsledné grafy vizualizačních fcí, které byly exportovány z GUI. Nakonec je zde uvedeno další možné rozšíření programu.

KLÍČOVÁ SLOVA

Box plot, graf, grafické rozhraní, GUI, histogram, jádrové odhady, korelační diagram, MATLAB, p-p plot, popisné statistiky, q-q plot, vizualizace.

ABSTRACT

The thesis deals with the visualization of biomedical data in MATLAB environment. The thesis contains following statistical methods and their descriptions: P-P plot, Q-Q plot, histogram, box plot, kernel denstity estimation, scatter plot and several time series metrics. Some functions are programmed from buil-in functions of MATLAB and others using external functions, which are changed to fit to this thesis's purpose. First part of the thesis conserns theoretical background, whereas the second part conserns practical programmed realizations of mentioned functions. The program contains a graphical user interface - GUI, which the thesis describes in detail. The purpose of the GUI is to ensure ease of use and also data processing. The output graphs of GUI are shown in chapter 5. The last part deals with the possible extensions of the program.

KEYWORDS

Box plot, graph, graphical user interface, GUI, histogram, kernel density estimation, MATLAB, p-p plot, q-q plot, scatter plot, time series, visualization.

ZVONČÁK, Vojtěch *Vizualizace biomedicínských dat v prostředí MATLAB*: diplomová práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, 2016. 51 s. Vedoucí práce byl Ing. Zoltán Galáž.

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma „Vizualizace biomedicínských dat v prostředí MATLAB“ jsem vypracoval(a) samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor(ka) uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil(a) autorská práva třetích osob, zejména jsem nezasáhl(a) nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom(a) následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

podpis autora(-ky)

PODĚKOVÁNÍ

Rád bych poděkoval vedoucímu mé diplomové práce, panu Ing. Zoltánu Galážovi, za odborné vedení, všechny konzultace, trpělivost, kvalitní podklady ke studiu a spoustu podnětných návrhů k práci.

Brno ....

podpis autora(-ky)

Faculty of Electrical Engineering
and Communication
Brno University of Technology
Purkynova 118, CZ-61200 Brno
Czech Republic

http://www.six.feec.vutbr.cz

PODĚKOVÁNÍ

Výzkum popsaný v této diplomové práci byl realizován v laboratořích podpořených z projektu SIX; registrační číslo CZ.1.05/2.1.00/03.0072, operační program Výzkum a vývoj pro inovace.

Brno . . . . . .

. . . . . .

podpis autora(-ky)

EVROPSKÁ UNIE
EVROPSKÝ FOND PRO REGIONÁLNÍ ROZVOJ
INVESTICE DO VAŠÍ BUDOUCNOSTI

OBSAH

SEZNAM OBRÁZKŮ

ÚVOD

Práce s biomedicínskými daty je často spojená s jejich vizualizací. Cílem práce je realizovat vizualizaci několika populárních statistických metod. Dalším obsahem práce je vytvořit grafické rozhraní, které umožní jednoduché a rychlé volání těchto funkcí.

Na začátku práce jsou uvedeny kapitoly, které se zabývají teoretickými podklady pro každou vizualizační funkci. Jsou uvedeny základní informace a případně matematické rovnice týkající se dané problematiky.

První kapitola 1 pojednává o několika popisných statistikách. Popisné statistiky nemají grafickou vizualizaci, ale pouze výpis výsledných hodnot4.2.1.

Další teoretická kapitola 2 rozebírá problematiku Gaussova-normálního rozdělení pravděpodobnosti. Tyto poznatky jsou pak uplatněni v programové realizaci pomocí jádrových odhadů hustoty pravděpodobnosti 4.2.6. V podkapitole o testování dat na normalitu 2.2 jsou teoreticky rozebrány zadané vizualizační fce. Jejich programová řešení jsou pak komentována v podkapitole Vizualizace dat4.2.

Poslední teoretická kapitola 3 se zabývá korelační analýzou. Popis její programové fce je uveden v kapitole 4.2.7.

V kapitole Realizace programu 4 je mimo jiné detailně rozebráno zpracování vstupních dat, před jejich samotnou vizualizací. Nakonec je v kapitole 4.3 popsáno samotné GUI.

Výsledné grafické realizace jsou ukázány v kapitole 5. Ke každé vizualizační fci je vykresleno několik demonstračních grafů.

V závěru jsou shrnuty dosažené výsledky práce. V přílohách je navrhuto další možné rozšíření programu. Dále obsahují popis obsahu DVD s krátkým návodem, jak spustit GUI.

1 POPISNÉ STATISTIKY

1.1 Střední hodnota

Při zkoumání určité množiny dat je důležitá informace o typické hodnotě jednoho vzorku. Matematicky je střední hodnota definována jako suma všech prvků v množině dělená celkovým počtem prvků:

$$M = \frac{1}{n}\sum_{i=0}^{n} x_i. \tag{1.1}$$

Kde x_i je hodnota i vzorku a n je celkový počet vzorků.

Tato hodnota je základním parametrem využívaným k výpočtu dalších veličin, které budou uvedeny dále. Pomocí střední hodnoty lze rychle kategorizovat danou skupinu vzorků. Zároveň však dochází ke ztrátě jiných důležitých informací o dané množině [14].

1.2 Medián

Pokud je u obecné množiny čísel provedeno vzestupné seřazení, tak hodnota prostředního čísla se nazývá medián dané množiny. V situaci, že seřazená množina má sudý počet čísel, tak jako medián se určuje hodnota průměru dvou čísel, které jsou nejvíce uprostřed.

Medián tedy bere v úvahu počet všech prvků v množině. Například v situaci, kdy je cílem získat informaci o platu v nějakém subjektu jsou průměr i medián adekvátními metodami. Použitím metody medián je získána informace o částce průměrné osoby. Kdežto použitím metody průměr je výslednou informací průměrná částka všech osob[14].

1.3 Rozsah

Matematicky lze rozsah R množiny X popsat jako

$$R = X_{max} - X_{min}, \tag{1.2}$$

kde X_{max} a X_{min} je maximální a minimální hodnota z množiny X. Tedy jednoduše informuje jaká je vzdálenost mezi prvkem s nejvyšší a nejnižší hodnotou[14].

1.4 Rozptyl

Pro výpočet rozptylu je nejdříve nutné definovat odchylku o_i jako

$$o_i = x_i - M, \tag{1.3}$$

kde x_i je i prvek z množiny X a M je střední hodnota. Rozptyl této množiny je pak definován jako

$$s^2 = \frac{1}{n}\sum_{i=0}^{n} {o_i}^2. \tag{1.4}$$

Kde n představuje celkový počet prvků v množině X, a o_i je odchylka.

Obecně lze pak rozptyl množiny X definovat jako střední hodnotu kvadrátů všech odchylek od střední hodnoty právě této množiny.

Pokud bude množina prvků představovat údaje v jednotce metr, pak rozptyl bude v metru čtverečním. Mocnina dvou byla volena z toho důvodu, že jednotlivé odchylky můžou mít záporné znaménko. V důsledku toho by došlo ke ztrátě informace, jelikož by se jednotlivé odchylky od sebe odečetly. A jelikož cílem je zjistit součet všech odchylek, je každá jednotlivá odchylka umocněna na druhou. Proč je voleno násobení $\frac{1}{n}$ a ne $\frac{1}{n-1}$ lze najít v literatuře [14].

1.5 Směrodatná odchylka

Pro jednodušší interpretaci rozptylu se dá použít směrodatná odchylka, která jíž má stejnou jednotku jako jednotlivé prvky množiny. Je definována jako

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n}\sum_{i=0}^{n} {o_i}^2}. \tag{1.5}$$

Lze říci, že rozptyl a směrodatná odchylka sdělují, jak kvalitní informaci nese konkrétní střední hodnota[14][15].

1.6 Kvantily

Základním operací pro získání kvantilů je seřazení dat. Poté je tato množina dat imaginárně rozdělena po určitých kvantech do kvantilů. Pořadové číslo kvantilu informuje o dalším rozložení hodnot množiny v daném místě kvantilu. Mezi nejčastěji používané kvantily patří kvartily a percentily. Percentil popisuje celou množinu pomocí procent a kvartil pomocí násobku 25 %. Obecný vzorec pro výpočet pozice konkrétního percentilu je

$$P_k = \frac{nk}{100}, \tag{1.6}$$

kde n představuje celkový počet prvků v množině, k je hledaný percentil, číslo 100 rozděluje množinu po procentech[9].

Při použití tohoto vztahu je nutné brát v úvahu, že výsledná pozice může být i desetinné číslo. To znamená, že daný percentil leží mimo pozice jednotlivých hodnot. Pak je nutné hodnotu daného percentilu vhodně dopočítat.

Kvartily na druhou stranu rozdělují množinu do čtyř částí. Při použití notace percentilů, můžeme říct že $Q_1 = P_{25}$, $Q_2 = P_{50}$, $Q_3 = P_{75}$. Slovně tedy kvartil Q_3 říká, že 25 % všech hodnot je nad jeho hodnotou a 75 % je pod jeho hodnotou.

Při práci s kvartily se používá popis rozsahu hodnot ve formě rozdílu prvního kvartilu Q_1 a třetího Q_3. Říká tedy jaký je střední rozsah hodnot 50% prvků z celé množiny. V anglicky mluvící literatuře se značí jako Interquartile range - IQR[9].

2 GAUSSOVO-NORMÁLNÍ ROZLOŽENÍ PRAVDĚPODOBNOSTI

Velká většina testů uvažuje, že medicínská data se kterými pracují, jsou normálně rozložená. To znamená, že se většina hodnot se soustřeďuje s největší pravděpodobností kolem jejich průměru. Proto je potřeba medicínská data testovat na jejich normalitu. K tomuto účelu se využívá funkce hustoty pravděpodobnosti a distribuční funkce.

2.1 Funkce hustoty pravděpodobnosti a distribuční funkce

Funkce hustoty pravděpodobnosti popisuje pravděpodobnost výskytu náhodné veličiny. Umožňuje zjistit pravděpodobnost výskytu čísel z daného intervalu. Platí, že čím širší je interval, tím vyšší pravděpodobnost výskytu veličin v intervalu. Na vertikální ose je vynesena hustota pravděpodobnosti a na horizontální ose náhodná veličina[9][12].

Pro náhodné veličiny s Gaussovým-normálním rozložením má funkce hustoty pravděpodobnosti tvar

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{\frac{(x-\mu)^2}{2\sigma^2}}, \tag{2.1}$$

kde $x \in (-\infty, +\infty)$; $\mu \in (-\infty, +\infty)$ a $\sigma^2 > 0$. Notace μ zde představuje střední hodnotu a σ^2 rozptyl pro náhodnou proměnnou x.

Základní vlastnosti této křivky jsou[11][8]:

- Její hodnota se blíží nule spolu s tím, jak se hodnota x blíží ±nekonečnu.
- Plocha, kterou vykresluje je rovna 1 (představuje součet všech pravděpodobností).
- Je centrována na μ a maximální hodnota této funkce je také na μ.
- Je symetrická kolem hodnoty μ.

Distribuční funkce vychází z funkce hustoty pravděpodobnosti. Její funkční hodnota je dána rovnicí

$$F(x) = P(X \leq x) = \int_{-\infty}^{\infty} f(t)dt, \tag{2.2}$$

kde výraz $P(X \leq x)$ představuje pravděpodobnost, že náhodná proměnná X nabude hodnoty rovné, nebo nižší než x. Pro zjištění tvaru distribuční funkce Gaussova-normálního rozložení lze uvažovat speciální případ funkce hustoty pravděpodobnosti,

kdy průměr $\mu = 0$, rozptyl $\sigma^2 = 1$. Po dosazení tohoto speciálního tvaru fce do rovnice 2.2 vypadá rovnice pro distribuční funkci jako

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{y^2}{2}} dy, \tag{2.3}$$

kde z = X. Distribuční funkce může popsat, jaké je rozložení všech hodnot ve vztahu s veličinou x[11][9].

2.2 Testování normality dat

Test na normalitu se skládá ze tří částí. Nejdříve je provedena vizualizace zvolené funkce, která má jako vstupní množinu hodnot testovaná data. Poté je vykreslena ta samá funkce, která má jako vstupní množinu hodnot data s Gaussovým-normálním rozložením (viz 2.1). A nakonec jsou výsledky porovnány. Níže jsou popsány metody, které jsou vyhodnocovány subjektivně. Proto pro detailnější popis zkoumané distribuce jsou pak v kapitole o korelační analýze popsány další přesnější metody.

2.2.1 P-P plot

Obecný P-P plot graficky zobrazuje závislost dvou distribučních funkcí. Jedna vychází z testovaných dat a druhá z dat s normální distribucí. Pokud budou obě dvě distribuční funkce identické, tak po proložení vykreslených bodů přímkou, bude mít přímka sklon 45° a bude procházet středem souřadnicového systému. Čím více se budou vykreslené body nacházet mimo přímku $y = x$, tím méně se testovaná distribuce dat bude podobat normální[4][16].

2.2.2 Q-Q plot

Q-Q plot zobrazuje závislost kvantilů dvou distribučních funkcí. Z nichž jedna má opět normální rozdělení dat. Platí zde stejné vlastnosti jako u P-P plotu. Čím více jsou jednotlivé body vykresleny blíže přímce $y = x$, tím více jsou si dané množiny dat podobny. Přímku k můžeme popsat pomocí rovnice

$$F^{-1}(p) = \mu + \sigma G^{-1}(p), \tag{2.4}$$

kde F(p),G(p) jsou distribuční funkce dvou množin hodnot a p jsou jednotlivé kvantily.

Při použití obou metod k popisu jedné testované množiny bude P-P plot více citlivý na rozdíly v prostřední části dvou distribucí. A Q-Q plot bude zase více citlivý na rozdíly na koncích dvou distribucí[11][14][16][4].

2.2.3 Histogram

Histogram umožňuje pochopit obecné vlastnosti zkoumané distribuce, jakými jsou například tvar, celkový rozsah hodnot, umístění anebo neobvyklé chování. Rozděluje zkoumanou množinu na několik definovaných intervalů (dále sloupců B). Každý sloupec vlastní informaci o počtu vzorků. Můžeme rozlišit tři základní tipy histogramů:

- Frekvenční,
- s relativní frekvencí,
- se zohledněním funkce hustoty pravděpodobnosti.

U frekvenčního histogramu má každý sloupec stejnou šířku a jeho výšku určuje počet vzorků, které v něm leží. Pro histogram s relativní frekvencí platí, že počet vzorků v každém sloupci je podělen celkovým počtem vzorků.

Pro snadné porovnání histogramu a funkce hustoty pravděpodobnosti je vhodné, aby integrál přes celou plochu histogramu byl roven jedné. A to je vlastností histogramu se zohledněním funkce hustoty pravděpodobnosti, který je definován jako

$$f(x) = \frac{V_k}{nh}, \tag{2.5}$$

kde $x \in B_k$, k je číslo sloupce, V_k je počet vzorků v sloupci B_k, n je počet všech vzorků a h je šířka sloupce[11].

2.2.4 Funkce hustoty pravděpodobnosti

Pro získání funkce hustoty pravděpodobnosti se používá metoda jádrových odhadů. Obecná funkce má tvar

$$f_X(x_0) = \frac{1}{N\lambda} \sum_{i=1}^{N} K_\lambda(x_0, x_i), \tag{2.6}$$

kde $f_X(x_0)$ je výsledná odhadovaná funkce, N je počet vzorků, λ je šířka okna a $K_\lambda(x_0, x_i)$ je jádrová funkce.

Jádrová funkce má podobu funkce hustoty pravděpodobnosti. Z důvodu testování dat na normalitu sleduje jádrová funkce Gaussovo rozdělení hustoty pravděpodobnosti. Je definována jako $K_\lambda = \phi(\frac{x-x_0}{\lambda})$ s tím, že $\mu = 0$ a $\sigma = \lambda$. Výsledná odhadovaná funkce hustoty pravděpodobnosti je průměr vážených příspěvků jádrových funkcí v bodě x_0 a její tvar je

$$f_X(x_0) = \frac{1}{N(2\lambda^2\pi)^{\frac{p}{2}}} \sum_{i=1}^{N} e^{-\frac{1}{2}(\frac{x_i-x_0}{\lambda})^2}. \tag{2.7}$$

Zhodnocení podobnosti odhadu funkce hustoty pravěpodobnosti s funkcí hustoty pravděpodobnosti pro Gaussovo normální rozložení je výsledkem této metody[11][6].

2.2.5 Box plot

Boxplot je množina vizuálních technik, které zobrazují distribuci vzorků v množině. Dále je popsán jeden konkrétní model grafu. Ke grafické reprezentaci potřebuje tři hodnoty: první kvartil, medián a třetí kvartil. Poté ještě definuje tzv outliers - jsou to vzorky, které leží mimo hranice distribuce. Dolní hranice DH je stanovená jako

$$DH = P_{25} - 1.5 * IQR, \tag{2.8}$$

horní hranice HH jako

$$HH = P_{75} + 1.5 * IQR. \tag{2.9}$$

Všechny zobrazované hodnoty jsou umístěny na vertikální nebo horizontální ose. Box znázorňuje pozici IQR. Předností boxplotu je možnost zjištění počtu outlierů[11][14][9].

3 KORELAČNÍ ANALÝZA

3.1 Korelace

Korelace je metoda pro stanovení velikosti pravděpodobnosti, že existuje lineární vztah mezi dvěma měřenými množinami.

3.1.1 Pearsonův korelační koeficient

Pearsonův korelační koeficient informuje o intenzitě vzájeného vztahu mezi dvěma náhodnými proměnnými X a Y. Implementuje v sobě kovarianci c_{xy}, která sděluje, jak velký je vzájemný lineární vztah mezi dvěma náhodnými proměnnými. Korelační koeficient je definován jako

$$\rho_{XY} = \frac{kov(X,Y)}{\sqrt{s^2(X)}\sqrt{s^2(Y)}} = \frac{c_{xy}}{\sigma_x \sigma_y}, \tag{3.1}$$

kde fce $s^2(X)$ a $s^2(Y)$ jsou hodnoty rozptylu pro náhodné proměnné X a Y a σ_x, σ_y jsou z nich vypočítané střední odchylky. Hodnota ρ_{XY} se pohybuje v rozmezí

$$-1 \leq \rho_{XY} \leq 1. \tag{3.2}$$

Hodnota $\rho_{XY} = 0$ stanovuje, že X a Y nemají spolu žádný lineární vztah. Naopak čím více se hodnota $|\rho_{XY}|$ blíží 1, tím více jsou X a Y mezi sebou korelovány. Znaménko ρ pak udává směr korelace[10][4][11].

3.1.2 Spearmanův koeficient pořadové korelace

Spearmanův koeficient nepočítá se skutečnými hodnotami proměnných, jako Pearsonův koeficient, ale pouze s jejich pořadovými čísly. Dále vyžaduje, aby vzájemná závislost pořadových čísel pro proměnné X a Y měla monototónní průběh. Při jeho výpočtu je nejdříve každé hodnotě x_i a y_i z náhodných proměnných X a Y přiřazeno pořadové číslo dle její velikosti vzhledem k ostatním v dané skupině. Poté platí, že

$$r_s = 1 - \frac{6 \sum (d_i)^2}{n(n^2-1)}, \tag{3.3}$$

kde d_i je rozdíl zpárovaných pořadových čísel pro x_i a y_i a n je celkový počet párů. Spearmanův koeficient potvrzuje nebo vyvrací tezi, že existuje dokonalá korelace mezi veličinami X a Y. Naprostá shoda se nachází na hodnotách ± 1, kde znaménko odpovídá směru korelace. Nulová shoda pak platí pro $r_s = 0$([4][9]).

Ke každé hodnotě korelačního koeficientu nebo Spearmanova koeficientu je počítána hladina významnosti p. Ta stanovuje s jakou pravděpodobností se může nulová

hypotéza pro daný koeficient vyskytnout. Pro oba koeficienty lze stanovit, že nulová hypotéza má znění: Neexistuje korelace mezi proměnnými X a Y. Mezní hodnoty pravděpodobnosti p se liší, například v literatuře[9] je uvedeno, že do hodnoty $p = 0.02$ lze nulovou hypotézu bez obav zamítnout. Naopak pro hodnoty $p > 0.15$ již musí být nulová hypotéza brána v potaz[9][13].

3.2 Regrese 1. řádu

Regrese 1.řádu nachází rovnici regresní přímky k, která nejlépe popisuje vztah mezi dvěma proměnnými X a Y. K nalezení rovnice k se používá metoda nejmenších čtverců, která hledá nejmenší hodnotu sumy

$$\sum (y - y')^2, \tag{3.4}$$

kde y je hodnota proměnné Y a y' je pak ekvivalentní bod na odhadované přímce k. Tvar k má podobu:

$$k = b_0 + b_1 x, \tag{3.5}$$

kde b_0 je průsečík s osou y a b_1 je směrnice přímky k. Přímka k vždy prochází bodem c o souřadnicích (a, b), který se nazývá centroid. Výpočet konstant b_0 a b_1 a souřadnic pro centroid c lze najít v literatuře [9].

Vykreslená přímka k s určitou směrnicí informuje o lineárním vztahu mezi proměnnými X a Y. V případě, že $b_1 = 0$, bude mít k tvar $k = b_0$, což naznačuje, že proměnná Y je nezávislá na proměnné X. Směrnice b_1 pak představuje průměrný vzrůst(pokles) hodnoty proměnné Y v závislosti na vzrůstu proměnné X o jednu jednotku[8].

3.2.1 Korelační diagram

Koleračí diagram je grafická metoda, která zobrazuje dvě množiny hodnot X a Y ze stejného zdroje dat. Je zvyk tyto data matematicky vyjádřit jako páry (X,Y), kde X je vstupní proměnná, která bývá zpravidla seřazená a Y je výstupní proměnná. Hodnoty jsou pak nazvány párovými, protože pro každou hodnotu X, existuje korespondující hodnota Y[9].

Zkonstruovaný korelační diagram je vizualizací toho, co bylo v 3. kapitole dosud popsáno. Pokud jsou proměnné X a Y v lineárním vztahu, budou jednotlivé body sledovat trend směrnice regresní přímky. Jejich vzdálenost k regresní přímce bude popisovat Pearsonův korelační koeficient - čím více budou body blíže, tím více jsou proměnné X a Y korelovány a tím vyšší je hodnota ρ. Regresní přímka s kladnou

směrnicí bude informovat o přímé závislosti a se zápornou směrnicí o nepřímé závisloti proměnných X a Y. Nakonec hodnota Spearmanova koeficientu sdělí, jak moc vynesené body korelačního diagramu sledují tvar monotónní funkce[10][9].

4 REALIZACE PROGRAMU

Celá realizace programu se dá rozdělit do tří základních částí:

1. Načítání a zpracování dat
2. Vizualizace dat
3. GUI

V této kapitole bude detailně popsána realizace každé části.

4.1 Načítání a zpracování dat

4.1.1 Formát vstupních dat

Program zkoumá numerické data, které jsou uložena ve sloupcích. Každý sloupec odpovídá měření nějakého parametru. Počet zpracovávaných parametrů není omezený. Ke konkrétnímu parametru je přiřazen konkrétní název. Všechny změřené hodnoty musí být označené. A to jak číselně, slovně nebo i kombinovaně. A opět je nutné, aby každý sloupec takových značek měl nějaké jméno. Pojmenování uvedených proměnných je následující:

- `feat_matrix` - matice změřených hodnot
- `feature_labels` - řádkový vektor s názvy parametrů
- `labels` - pole buňek obsahující značky
- `labels_names` - řádkový vektor s názvy značek

Jediný podporovaný datový formát je soubor s příponou `mat`, který je generován programem MATLAB. Požadovaný obsah každého takovéhoto souboru je zobrazen na obrázku 4.1. Program umožní další zpracování dat pouze v případě, že jsou splněny všechny uvedené vlastnosti. Soubor mat musí obsahovat čtyři proměnné s

Vstupní *.mat soubor				
Název	feature_matrix	feature_labels	labels	labels_names
Datový typ 1 Rozměr	double (A,B)	cell (1,B)	cell (A,C)	cell (1,C)
Datový typ 2		char	char,double	char

Obr. 4.1: Informace o formátu vstupních dat.

uvedenými jmény v přesně stejném znění. Není dovoleno, aby kterákoliv proměnná byla prázdná. Například proměnná `feat_matrix` má být maticí s počtem řádků A a s počtem sloupců B s tím, že její datový typ je `double`. Naopak proměnná `labels` musí mít datový typ `cell`. Počet řádků má odpovídat počtu řádků matice

`feat_matrix` a počet sloupců C má být nenulový. Každý sloupec buňek pak může ještě nabývat buď datového typu `double` nebo `char`.

Žádná z proměnných nemůže být prázdná. Například pro matici `feat_matrix` rozměru $(A, 2)$ by ostatní hodnoty měly nabývat těchto hodnot:

- názvy parametrů v proměnné `feature_labels` by měly mít rozměr $(1, 2)$,
- značky v proměnné `labels` - $(A, 1)$,
- název značek v proměnná `labels_names` - $(1, 1)$.

Před vizualizací je potřeba proměnné z `mat` souboru zkontrolovat a případně vhodně upravit pro jednotlivé vizualizační fce. Dále jsou popsány funkce ve vzestupném pořadí, které to mají na starosti.

4.1.2 Funkce `CheckMyData`

Po načtení `mat` souboru do paměti je to první funkce, která pracuje se vstupnímy daty. Vstupní proměnné nemění, ale pouze je kontroluje a případně vyvolá výskyt chyby. U vstupních dat zkoumá:

- dovolené rozměry (viz 4.1),
- dovolený obsah.

Chyba je detekována v případě, že jakákoliv proměnná je prázdná. Dále se kontroluje konzistence proměnné `labels`. V jednom sloupci se můžou vyskytovat proměnné typu `double` nebo `char`. V případě výskytu jejich kombinace je vyhlášena chyba.

Výstupní proměnnou je buňka s názvem `TypeOfLabelsData` 4.2. Jedná se o řádkový vektor, který obsahuje informaci o datových typech sloupců proměnné `labels`.

TypeOfLabelsData					
Number	Number	Char	Number	Number	Number

Obr. 4.2: Obsah buňky `TypeOfLabelsData` v případě, kdy proměnná `labels` má 6 sloupců s danými datovými typy.

4.1.3 Funkce `ControllAllLabels`

Fce kontroluje na základě stavu proměnné `TypeOfLabelsData`, zda se nenachází v `labels` více jak jeden sloupec hodnot s datovým typem `char`. V případě porušení této podmínky, je vyhlášena chyba a je uživateli doporučeno zanechat v `labels` pouze jeden vektor typu `char`.

Poté fce prohledává `labels` a vyjme sloupec s typem `char`. Ten je pak uložen do proměnné `TAGS`. Pokud `labels` neobsahuje sloupcový vektor typu `char`, je do všech řádků `TAGS` zkopírována jedna značka. Obsah `TAGS` je pro další zpracování klíčový a bude vysvětlen dále.

Díky této fci dochází k rozdělení datových typů `char` a `double` do dvou proměnných `TAGS` a `labels` s ekvivalentní úpravou jejich jmen v proměnných `TAGS_NAMES` a `labels_names`.

4.1.4 Funkce `GetLabelIndex`

Často jsou jednotlivé řádky ve sloupci konkrétního parametru postupně vybírány v cyklu for. Jednotlivé značky pro tyto řádky jsou uložené v proměnné `TAGS`. Například prvních deset řádků `feat_matrix` může mít značku XX a zbytek řádků značku YY. Lze tedy vidět, že parametry jsou rozděleny do dvou intervalů XX a YY. To jakou mají značku pak ovlivní jejich grafickou vizualizaci. Fce `GetLabelIndex` proto kontroluje, zda jsou značky v seřazeném pořadí. Pokud nejsou, musí si uživatel sám provést seřazení.

Výstupem fce je proměnná `LabelIndex`. Jedná se o buňku, která má v prvním sloupci názvy značek a v druhém sloupci počáteční indexy jejich výskytu.

LabelIndex	
'XX'	1
'YY'	11

Obr. 4.3: Obsah buňky `LabelIndex` pro dvě značky XX a YY s počátečními indexy 1 a 11.

4.1.5 Funkce `AfterPickFiller4TheRest`

Uživatel má možnost si vybírat, které parametry(sloupce z `feat_matrix`) a značky(sloupce z `labels`) chce vizalizovat. Informace o vybraných datech je uložena v proměnné `buffer`. Proměnná `buffer` je maticí čísel, kde v prvním řádku je informace o typu proměnné a v druhém řádku číslo vybraného sloupce.

Na základě stavu `buffer` pak fce uloží vybrané sloupce do nové proměnné. Zároveň vytvoří záznam jmen vybraných sloupců. Jinými slovy vytvoří nové proměnné `feat_matrix` a `feature_labels`, které odpovídají proměnné `buffer`.

buffer			
1	0	0	1
100	3	1	2100

Obr. 4.4: Ukázka obsahu proměnné `buffer` pro vizualizaci čtyř sloupců dat. V prvním řádku číslo 1 odpovídá proměnné `feat_matrix` a 0 proměnné `labels`. V druhém řádku pak jsou čísla vybraných sloupců.

4.1.6 Funkce `LoadMatrix2Cell`

Jelikož mnohé vizualizační fce pracují se vstupní proměnnou jako s celým sloupcem, bylo nutné najít způsob, jak data vhodně naformátovat k co nejjednoduššímu procházení. To má za úkol fce `LoadMatrix2Cell`. Rozděluje `feat_matrix` do pole buněk. Výstupní proměnná se pak nazývá `data` a má stejný počet sloupců jako `feat_matrix`. Rozdílná je v počtu řádků. Fce rozděluje každý sloupec `feat_matrix` do tolika buňek, kolik existuje unikátních značek.

Je důležité zmínit, že `LoadMatrix2Cell` pracuje na vstupu s „libovolnou“ maticí čísel a `LabelIndex`. Vůbec se nezajímá, zda sloupce matice odpovídají proměnným `feat_matrix` nebo `labels`. To má na starost fce `AfterPickFiller4TheRest`4.1.5, která vhodně naplní matici a `LoadMatrix2Cell` ji pouze rozloží do buněk.

Obrázek 4.1.6 demonstruje stav proměnné `data` po sekvenci všech zmíněných fcí. Ve sloupcích jsou jednotlivé proměnné, které se řídí obsahem `buffer`. A počet řádků je odvozen z obsahu `LabelIndex`. S takto transformovanými daty se pak jednoduše pracuje ve vizualizačních funkcích.

data			
feat_matrix(1:10,100)	labels(1:10,3)	labels(1:10,1)	feat_matrix(1:10,2100)
feat_matrix(11:end,100)	labels(11:end,3)	labels(11:end,1)	feat_matrix(1:end,2100)

Obr. 4.5: Obsah proměnné `data`, který je ovlivněn obsahem proměnných `buffer` a `LabelIndex`.

Právě tato fce nabízí výpočet normalizace pro všechny buňky v `data`. Normalizace je provedena pomocí vzorce:

$$SV = \frac{SV - mean(SV)}{std(SV)}, \tag{4.1}$$

kde SV je sloupcový vektor konkrétní buňky, $mean$ je průměr a std je směrodatná odchylka.

4.1.7 Funkce `AfterPickFiller4Corr`

K vizualizaci korelace byl zvolen jiný formát dat. Proto tato fce nahrazuje dvojici fcí `LoadMatrix2Cell` a `AfterPickFiller4TheRest`. Fce dovoluje zpracovávat data, která se skládají pouze ze sudého počtu sloupců datového typu `double`(viz 4.4). Zároveň dovoluje zpracovávat data, které jsou značeny pouze jednou značkou. Takže v příkladu dat zobrazených na obrázku 4.3 by uživatel volání fce `AfterPickFiller4Corr` neměl vůbec k dispozici.

Výstupní proměnná `scatter_feat_cell` je buňka, která má vždy čtyři řádky. Počet sloupců odpovídá počtu dvojic, které se necházejí v proměnné `buffer`. Jednotlivé řádky mají následující význam:

1. Sloupcový vektor hodnot daného parametru pro osu x.
2. Název parametru pro osu x.
3. Sloupcový vektor hodnot dané značky pro osu y.
4. Název značky pro osu y.

Pokud vstupní data budou obsahovat jen jednu značku(`LabelIndex` pak má jen jeden řádek) a obsah `buffer` bude stejný jako je na obrázku 4.4, pak `scatter_feat_cell` bude mít následující podobu:

scatter_feat_cell	
feat_matrix(:,100)	labels(:,1)
feature_labels{1,100}	labels_names{1,1}
labels(:,3)	feat_matrix(:,2100)
labels_names{1,3}	feature_labels{1,2100}

Obr. 4.6: Obsah `scatter_feat_cell` s daty značenými pouze jednou značkou.

4.2 Vizualizace dat

V této části jsou popsány detailní informace o každé vizualizační funkci. Program umožňuje celkem 6 vizualizací a výpis popisných statistik. Výsledné grafy vizualizací jsou ukázány až v poslední kapitole 5.

Vlastnosti vizualizačních fcí:

- Všechny vizualizace obsahují kontrolu na přítomnost NaN(výsledek není číslo – not a number). Každá fce pak patřičně reaguje na přítomnost NaN.
- K získání různých barev bylo použito externí fce `brewermap`[1] ze stránek [2]. Umožňuje jednoduše získat matici barev RGB z vybrané palety barev. Každá

[1]K fci `brewarmap` existuje i autorova fce `brewermap_view`, kterou lze najít na přiloženém DVD. Umožňuje uživateli si prohlédnout, jaké palety barev brewarmap obsahuje.

z vizualizačních fcí obsahuje vstupní proměnnou pro volbu palety barev. Fce `brewermap` je volána až uvnitř vizualizační fce.

- Všechny grafy obsahují legendu, která patřičně reaguje na uživatelem zvolené sloupce hodnot.
- V případě nutnosti popisky os mění svůj formát k zajištění přehlednosti.
- Všechny použité fce k vizualizaci jsou naprogramovány pomocí základních interních fcí k vykreslování grafů(např.: fce `plot`). Nebo jsou použity externí fce, které byly různě modifikovány, aby vyhovovaly požadavkům této práce.

4.2.1 Funkce `PopisneStatistiky`

Tato funkce provádí výpočet nad celou maticí `feat_matrix`. Nevyužívá tedy fce spojené s `buffer`. Sled použitých fcí je následující:

- `CheckMyData`
- `ControllAllLabels`
- `GetLabelIndex`
- `LoadMatrix2cell`

Fce má na starost vypočítat 11 popisných statistik pro každou buňku z proměnné `data`. Výsledkem je tedy sloupcová matice o 11 řádcích pro každou buňku. Jednotlivé sloupce jsou skládány vedle sebe jak vertikálně, tak i horizontálně. Výstupní matice se pak nazývá `FINAL_1D_MATRIX`. Další výstupní proměnnou je sloupcový vektor buněk `FINAL_1D_TAGS`, ve kterém jsou uloženy názvy pro všechny řádky `FINAL_1D_MATRIX`. Na obrázku 4.7 je ukázána podoba `FINAL_1D_MATRIX` pro čtyři značky a pro matici `feat_matrix` o velikosti (A,6).

FINAL_1D_MATRIX					
vector_1_1(11,1)	vector_1_2(11,1)	vector_1_3(11,1)	vector_1_4(11,1)	vector_1_5(11,1)	vector_1_6(11,1)
vector_2_1(11,1)	vector_2_2(11,1)	vector_2_3(11,1)	vector_2_4(11,1)	vector_2_5(11,1)	vector_2_6(11,1)
vector_3_1(11,1)	vector_3_2(11,1)	vector_3_3(11,1)	vector_3_4(11,1)	vector_3_5(11,1)	vector_3_6(11,1)
vector_4_1(11,1)	vector_4_2(11,1)	vector_4_3(11,1)	vector_4_4(11,1)	vector_4_5(11,1)	vector_4_6(11,1)

Obr. 4.7: Matice popisných statistik pro celou proměnnou `data`.

4.2.2 Funkce `PlotPP`

Fce vykresluje P-P plot a využívá k tomu interní fci MATLAB s názvem `normplot`. Ta splňuje všechny definice P-P plot uvedené v kapitole 2.2.1. Fce `normplot` je volána nad každou buňkou v proměnné `data`. Jednotlivé grafické realizace jsou od sebe rozlišeny barvou. Kvůli individuální preferencím uživatelů je v této fci v paletě

barev pořadí barev navíc náhodně zamícháno. Výsledný graf má tak vždy trochu jinou barevnou podobu při každém novém volání fce.

Fce `normplot` interně nabízí vykreslení křivek, které spojují 1., 4. kvartil a 2.,3. kvartil. Jelikož je P-P plot z definice citlivý v pozici mezi 2. a 3. kvartilem, přebírá pro přehlednost spojnice těchto kvartilů barvu vynesených bodů. Uživatel má volbu tyto křivky vypnout.

4.2.3 Funkce `PlotQQ`

Tato fce vykresluje Q-Q plot pomocí interní fce MATLAB `qqplot`. `PlotQQ` je velmi podobná fci `PlotPP` a pracuje prakticky stejně. Jedinou změnou ve vykreslování jsou kvartilové křivky. Jelikož Q-Q plot je citlivý v oblastech 1. a 4. kvartilu, přebírá právě tato spojnice barvu vynesených bodů.

4.2.4 Funkce `PlotBox`

Funkce `PlotBox` vykresluje graf Box plot. Používá k tomu externí fci `aboxplot`[2] ze stránek [1], která umožňuje naproti interním fcím MATLAB vykreslovat krabicové grafy ve skupinách.

Vstupní proměnná `data` musela být ještě rozdělena podle značek do buněk. Pokud vstupní data budou mít například 2 značky(viz 4.1.6), budou `data` rozdělena do sloupcového vektoru buněk o dvou řádcích. V každém řádku pak budou zástupci všech sloupců pro danou značku. Situace je znázorněna na obrázku 4.8.

BOX				
buňka značky XX	feat_matrix(1:10,100)	labels(1:10,3)	labels(1:10,1)	feat_matrix(1:10,2100)
buňka značky YY	feat_matrix(11:end,100)	labels(11:end,3)	labels(11:end,1)	feat_matrix(1:end,2100)

Obr. 4.8: Obsah proměnné `BOX` pro dvě značky a 5 sloupců v proměnné `buffer`.

Dolní a horní hranice krabicového grafu odpovídají hodnotám specifikovaným v kapitole 2.2.5. Graf také vykresluje outliers, což jsou body které za nachází za zmíněnými hranicemi. V případě vykreslení pouze jednoho sloupce s jednou značkou je výsledná barva krabicového grafu náhodná.

Graf je vhodný k zobrazování velkého počtu sloupců. Fce totiž umožňuje volit mezi numerickými popisky a popisky s názvy pro osu x. Další vlastností fce je možnost zkrácení dlouhých textových popisků.

[2]Fce dále obsahuje autorovy `m` soubory s názvy `colorgrad.m` a `quartile.m`, které potřebuje pro svou funkčnost.

4.2.5 Funkce `PlotHistogram`

Fce vykresluje frekvenční histogram 2.2.3 pomocí externí fce `histf` ze stránek [5]. `PlotHistogram` je volána nad všemi buňkami data podobně jako u grafů P-P plot nebo Q-Q plot. Hrany každého binu reagují na nenormalizovaná a normalizovaná vstupní data. Histogram má zvýrazněné hrany pouze u nenormalizovaných dat. Při zapnuté normalizaci se tak graf stává přehlednějším. Uživatel může volit počet binů histogramu. Defaultně je počet binů odvozen od odmocniny počtu řádků [7]. Pokud je zvolen k vykreslení pouze jeden sloupec dat, výsledná barva histogramu je náhodná.

4.2.6 Funkce `KernelDensityEstimation`

Fce má na starost vykreslení grafu jádrových odhadů. K získání hodnot křivky jádrového odhadu je použita interní fce `ksdensity`. Opět je volána nad všemi buňkami data. V případě vykreslení pouze jednoho sloupce dat je výsledná barva křivky náhodná. Pokud vstupní data obsahují více rozdílných značek, jsou jednotlivé parametry (případně číselné značky) od sebe rozlišeny jiným tipem čáry. Fce nabízí možnost vykreslení vstupních dat pro konkrétní křivku. Činí tak pomocí barevně vykresleného čísla, které odpovídá číslu křivce v legendě.

4.2.7 Funkce `CorrelationPlot`

Tato fce má za úkol vykreslit korelační diagram spolu se všemi dalším informacemi, které se korelace týkají. Počítá Pearsonův korelační koeficient a Spearmanův koeficient, které vyobrazuje nad grafem korelace. Dále umožňuje zobrazit regresní křivku pro danou dvojici vstupních dat, u které lze měnit řád polynomu. Fce rozděluje vykreslené body do intervalů stejné šířky, které jsou barevně rozlišeny a jejichž počet může uživatel měnit. Graf lze vykreslovat ve dvou různých grafických verzích. Korelační graf používá vlastní pevně definovanou paletu barev, která zaručuje dobrou přehlednost grafu. Pokud je počet vstupních dat větší jak 29, je použita paleta barev zvolená uživatelem.

Jelikož `CorrelationPlot` nevyužívá fci `LoadMatrix2Cell`, je normalizace počítána uvnitř této fce. K rozdělení hodnot do intervalů pro osu y se používá externí fce `split_data` ze stránek [3]. Výsledkem `split_data` je logická matice rozměru (A,I), kde A je počet řádků `feat_matrix` a I je počet intervalů. Matice má hodnotu log. 1 u bodu, který patří do konkrétního intervalu a log. 0 u bodu který do intervalu nepatří. Jednoduše pak lze pomocí fce `plot` vykreslit data pouze z daného intervalu.

Vstupní proměnnou fce `CorrelationPlot` je výsledek fce `AfterPickFiller4Corr` - `scatter_feat_cell`. Všechny mezivýsledky pro konkrétní sloupec `scatter_feat_cell`

jsou ukládány do této proměnné na další řádky. Na obrázku 4.9 je naznačen obsah proměnné pro 4 sloupce vstupních dat vybrané uživatelem (viz 4.6)[3] s tím, že počet intervalů je 2.

feat_cell			
scatter_feat_cell	(:,1)	scatter_feat_cell	(:,1)
RGB_pro_intervaly_1	(2,3)	RGB_pro_intervaly_2	(2,3)
Pearsonův_koeficient_1	(1,2)	Pearsonův_koeficient_2	(1,2)
Spearmanův_koeficient_1	(1,2)	Spearmanův_koeficient_2	(1,2)
Regresní_křivka_1	(1,2)	Regresní_křivka_2	(1,2)
Log_matice_intervalů_1	(A,2)	Log_matice_intervalů_2	(A,2)
Intervaly_pro_legendu_1	(1,3)	Intervaly_pro_legendu_2	(1,3)
Název_osy_x_1		Název_osy_x_2	
Název_osy_y_1		Název_osy_y_2	
Informace_o_regresi_1	(4,:)	Informace_o_regresi_2	(4,:)

Obr. 4.9: Obsah proměnné `feat_cell`, se kterou pracuje fce `CorrelationPlot`. Údaje v závorkách odpovídají rozměru dané proměnné.

4.3 Grafické rozhraní - GUI

GUI bylo vytvořeno pomocí vestavěného prostředí pro jeho tvorbu - `guide`. GUI umožňuje uživateli jednoduše volat fce pomocí myši bez interakce s klávesnicí. Na obrázku 4.10 lze vidět stav GUI po jeho prvním spuštění. Pro komunikaci GUI s uživatelem byl zvolen anglický jazyk.

Rozhraní je rozděleno do čtyř oblastí:

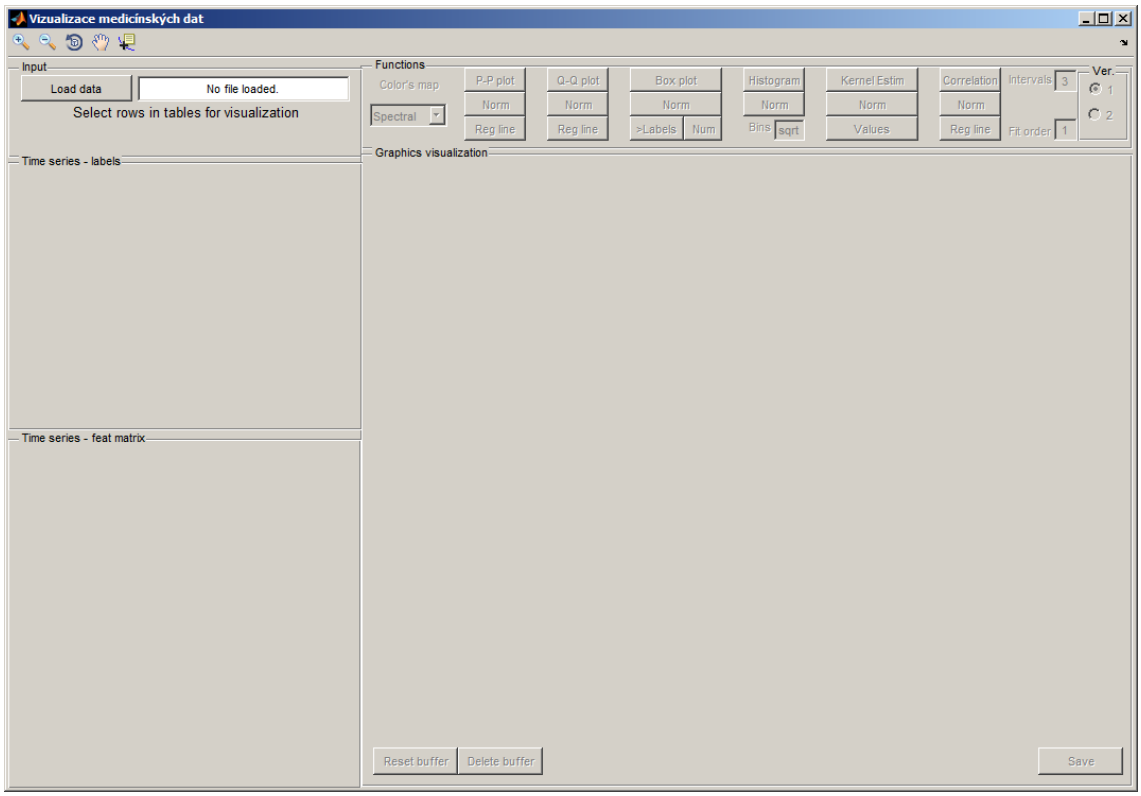
1. oblast načítání dat,
2. oblast popisných statistik,
3. oblast volby vizualizačních fcí a jejich parametrů,
4. oblast vykreslení grafů.

GUI se skládá z funkcí, které se volají na základě uživatelovy činnosti. Spouští se pomocí souboru `GUI.m`.

Stisknutím libovolného tlačítka uvnitř GUI dochází k zavolání tzv.: fce `Callback`, ve které se nachází sekvece dalších funcí připravených k provedení. Po dokončení fce `Callback` je GUI v režimu, ve kterém čeká na další uživatelovu činnost.

Na rozdíl od bežné práce s programem MATLAB je uložným prostorem v případě GUI datová struktura tzv.: `handles`. V této struktuře se nachází ukazatele

[3]Proměnná `scatter_feat_cell` se uvnitř fce nazývá `feat_cell`.



Obr. 4.10: Grafické prostředí po první spuštění.

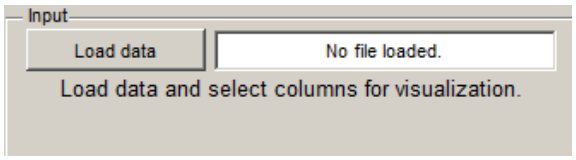
na všechny prvky v GUI a zároveň slouží jako místo pro ukládání. Pokud uvnitř fce `Callback` dojde k nějakému výpočtu a je cílem jeho výsledek použít v jiné fci `Callback`(vyvolané jiným tlačítkem), musí být výsledek uložen v `handles`. Pokud nebude uložen, tak jakmile daná fce `Callback` bude dokončena, všechny ostatní proměnné budou vymazány.

4.3.1 Oblast načítání dat

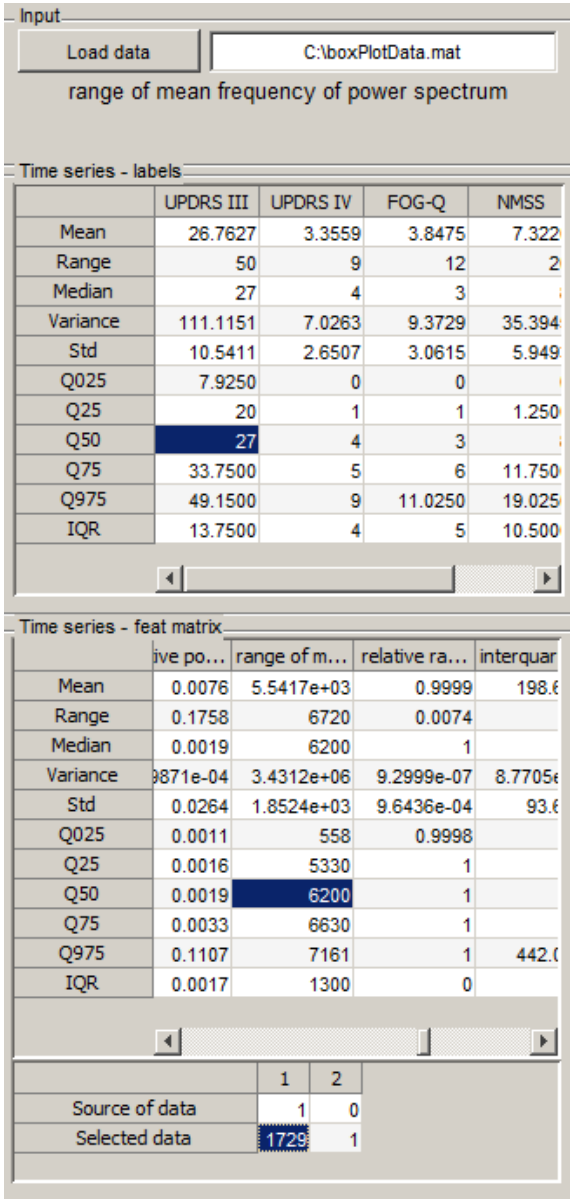
Prvním krokem, který uživatel musí udělat je načíst vstupní data. Po stisknutí tlačítka `Load data` se provede následují sekvence fcí:

- dialogové okno pro výběr souboru,načtení a kontrola souboru,
- fce `CheckMyData`, `ControllAllLabels`, `GetLabelIndex`,
- fce `PopisneStatistiky` pro proměnnou `feat_matrix` a případně `labels`.

Cesta k souboru je pak vypsána v okně na obr. 4.11 s textem „No file loaded“. Pod tlačítem `Load data` se nachází textové pole, které informuje uživatele o aktuálním stavu GUI.



Obr. 4.11: Část GUI pro načítání dat.



Obr. 4.12: Část GUI pro výpis popisných statistik.

4.3.2 Oblast popisných statistik

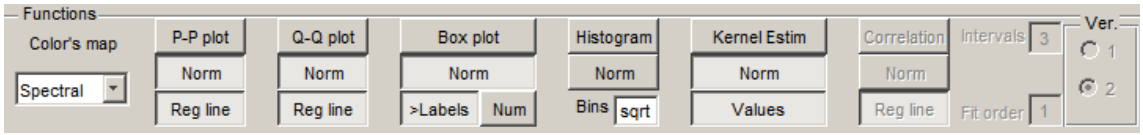
Po načtení dat se vypočítaná matice popisných statistik spolu s vhodnými popisky objeví v tabulce `Time series-feat_matrix`, případně `Time series-labels`. Tyto

tabulky(viz 4.12) jsou interaktivní. Uživatel si kliknutím do kteréhokoli řádku vybere příslušný sloupec, který chce vizualizovat. O svém výběru je informován informačním textem v oblasti načítání dat. A zároveň se objeví aktuální obsah proměnné `buffer` v nejnižší tabulce v této oblasti. Dalším výběrem sloupců se tato tabulka aktualizuje. Tabulka je také interaktivní a kliknutím na příslušné políčko je uživatel informován o tom, co dané číslo představuje. Informace o výběru se zobrazuje opět v oblasti načítání dat.

Kvůli nedostatku místa byla tlačítka pro ovládání přesunuta do oblasti vykreslení grafů. Tlačítko na obr. 4.14 `Reset buffer` vymaže celou proměnnou `buffer` a tlačítko `Delete buffer` vymaže poslední záznam.

4.3.3 Oblast volby vizualizačních fcí a jejich parametrů

Po výběru sloupců uživatelem se tato oblast zviditelní. Její viditelnost je přímo závislá na obsahu a počtu proměnné(tabulce) `buffer` 4.4. V případě sudého počtu sloupců jsou viditelná všechna tlačítka, v případě lichého počtu se viditelnost tlačítek pro korelaci vypne. Chování viditelnosti tlačítek má na starost fce `checkButtons`, která je volána při každém výběru sloupce uživatelem a při každém stisku tlačítek `Reset buffer` a `Delete buffer`. Z obrázku 4.13 je patrné, že tlačítka pro jednot-



Obr. 4.13: Oblast volby vizualizačních fcí.

livé vizualizační fce jsou seskupená. V první řadě se nachází tlačítka pro volání konkrétních vizualizačních fcí. Zbylé tlačítka slouží k volbě konkrétních parametrů vizualizační fce.

Jednotlivá tlačítka mají následující význam:

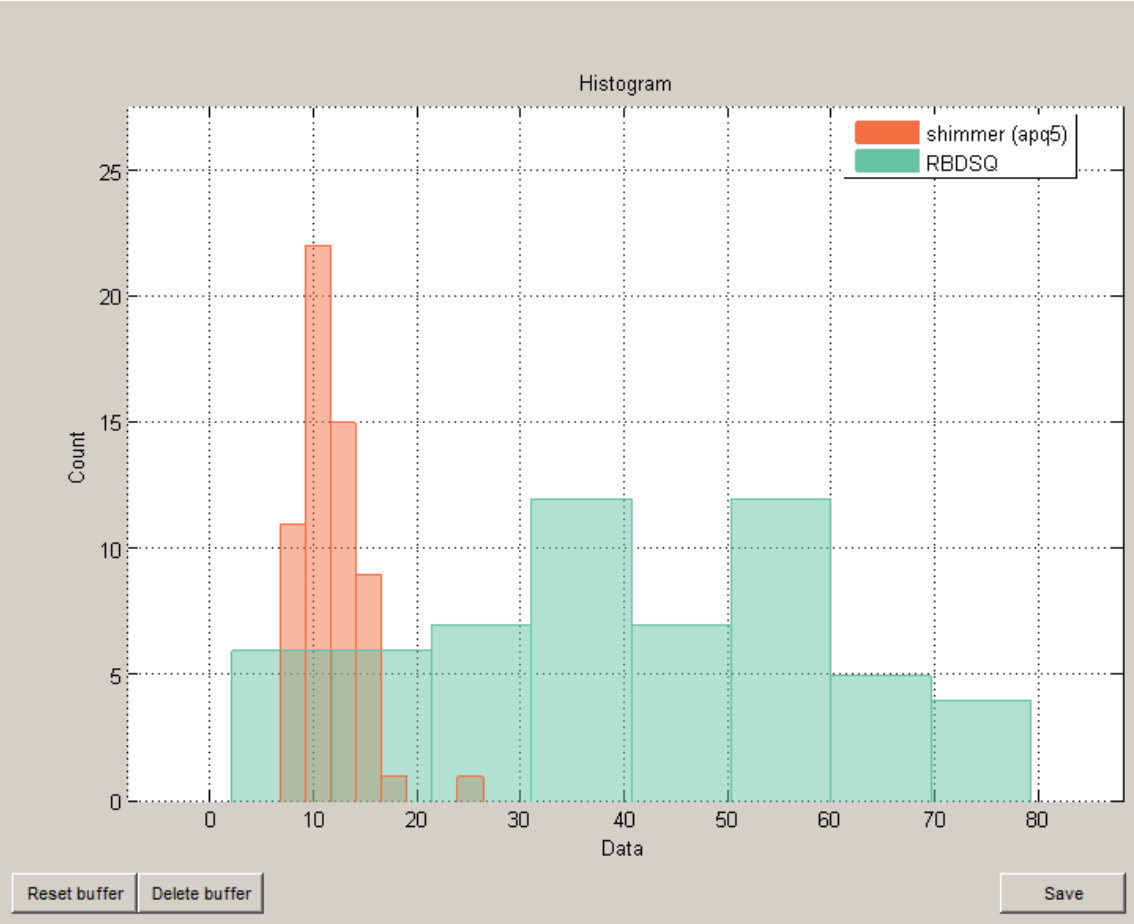
- `Color's map` nabízí uživately výběr různých barevných palet.
- Tlačítko `P-P plot` vykreslí graf P-P plotu s tím, že bere v potaz aktualní stav tlačítek pod sebou. To znamená tlačítek `Norm` pro normalizaci a `Reg line` pro vykreslení regresních křivek. Podobně je tomu i u tlačítka `Q-Q plot`.
- Tlačítko `>Labels` kontroluje zkrácení nebo ponechání popisků osy x pro Box plot. A tlačítko `Num` kontroluje, zda se budou vykreslovat textové nebo číselné popisky osy x.
- U tlačítka `Histogram` je pole za textem `Bins`, které umožňuje zadávat počet binů histogramu. Je zajištěna kontrola zadávaných hodnot.

- Tlačítko `Values` kontroluje vykreslování jednotlivých bodů u grafu jádrových odhadů.
- Pole za texty `Intervals` a `Fit order` představují vstup pro volbu počtu intervalů a řádu regresní přímky. I zde je kontrola vstupních hodnot zajištěna. Tlačítka `1` a `2` slouží k volbě grafické realizace korelačního diagramu.

4.3.4 Oblast vykreslení grafů

V tomto prostoru dochází k vykreslení zvolených grafů. Při každém novém volání vizualizační fce je celý graf vykreslen znova. Při vykreslení velkého počtu vstupních dat dochází k nárůstu velikosti textů os a legend. V takovém případě je lepší výsledný graf uložit pomocí tlačítka `Save`.

Tlačítko `Save` je dostupné kdykoliv je vykreslen nějaký graf. Umožňuje uložit aktuální graf do několika datových typů: `fig`, `jpeg`, `png`, `bmp`, `tiff`, `pdf`, `eps`, `ps`.



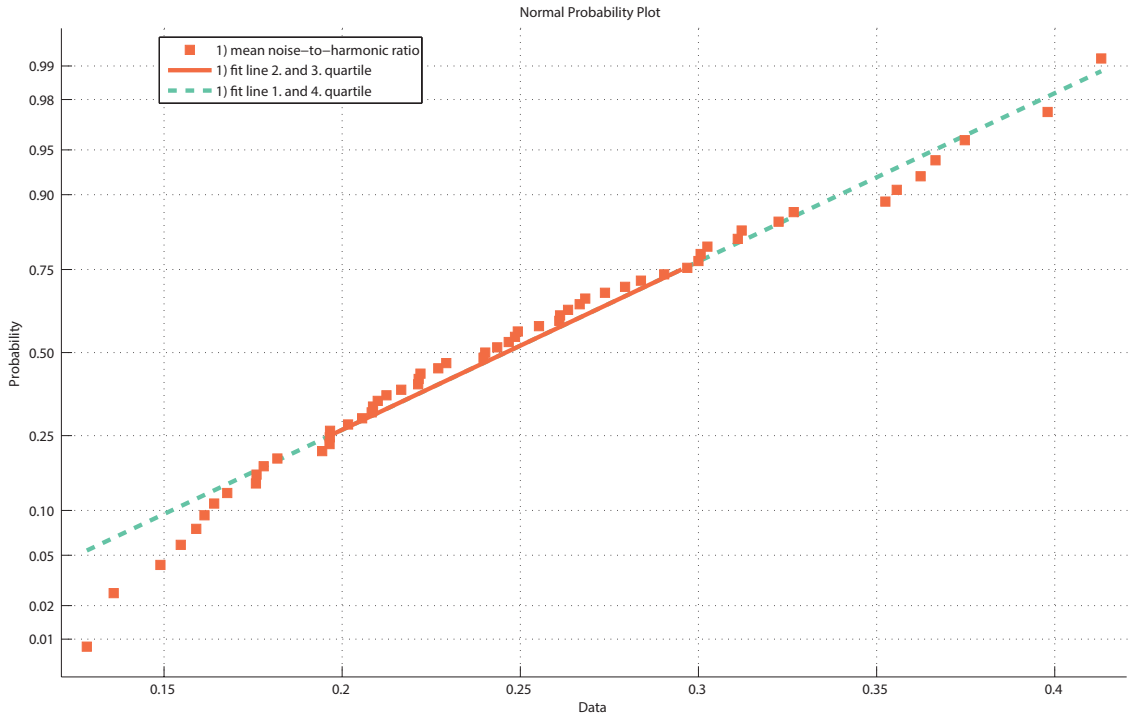
Obr. 4.14: Oblast vykreslení grafů v GUI.

5 VÝSLEDNÉ GRAFY PRO VIZUALIZAČNÍ METODY

V této kapitole jsou ukázány a popsány výsledné grafy. Grafy byly exportovány z GUI pomocí tlačítka `Save` ve formátu `ps`. U každého obrázku grafu jsou informace o použité paletě barev, počtu značek, normalizaci, počtu sloupců vstupních dat a o viditelnosti regresních nebo kvartilových přímek.

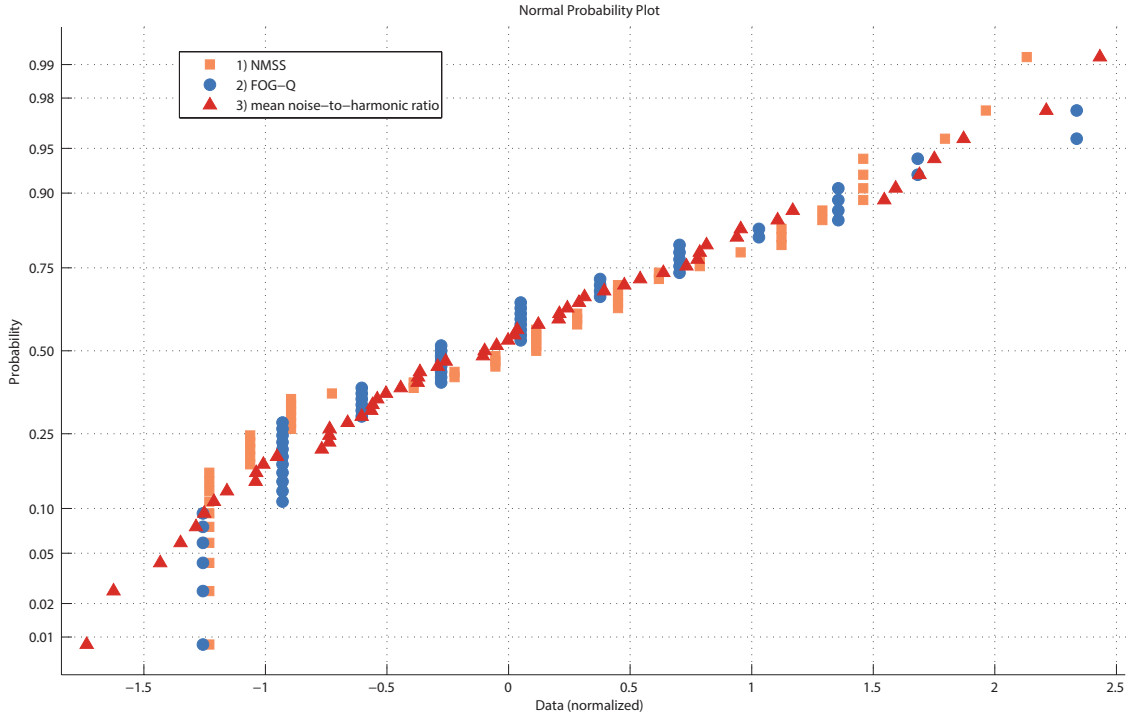
5.1 Metoda P-P plot

Obrázek 5.1 představuje příklad vykreslení P-P plot pro jeden parametr z `feat_matrix`. Body uprostřed grafu věrně sledují kvartilovou přímku a pouze počáteční body se od přímky odchylují. Co se týká sklonu přímky, tak nemá 45 stupňů. Lze odhadnout, že data s určitou chybou sledují normální rozložení dat.

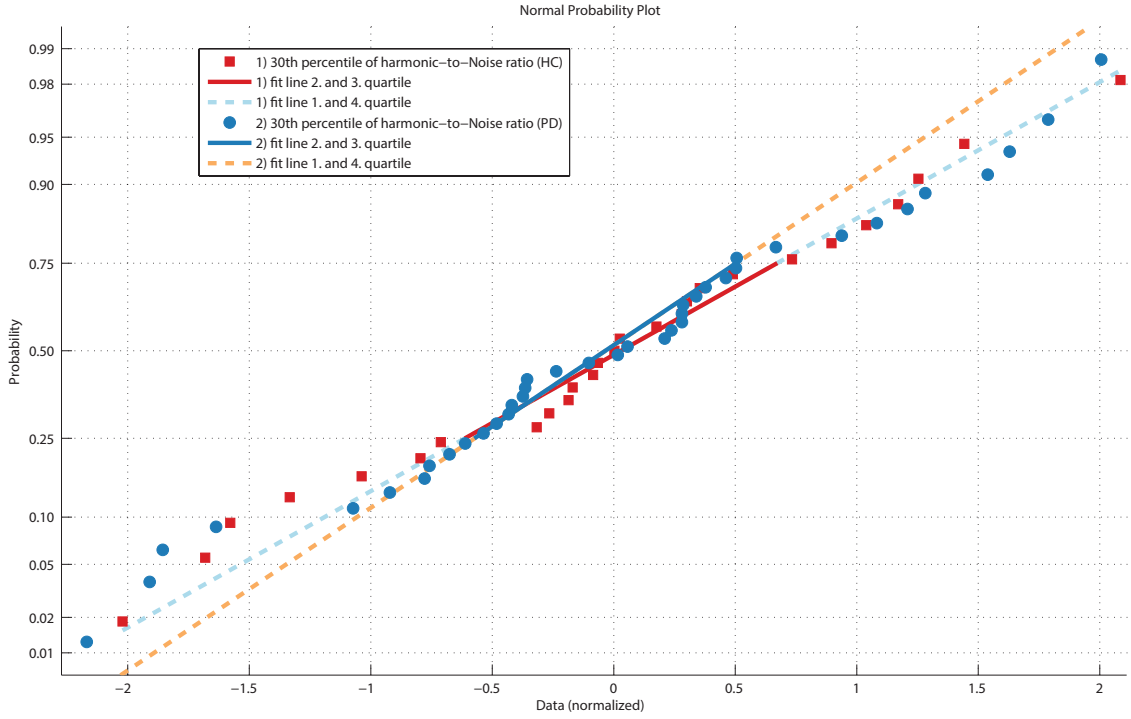


Obr. 5.1: Graf P-P plot: paleta Spectral, jedna značka, nenormalizováno, jeden sloupec vstupních dat, kvartilové přímky jsou vykresleny.

Na dalším obrázku 5.2 lze vydět, že přidané sloupce z `labels` mají velmi odlišné rozdělení dat než původní parametr. Kvartilové přímky byly vypnuty kvůli přehlednosti. O sloupcích z `labels` „NMSS“ a „FOG-Q“ lze říci, že jejich hodnoty nejsou normálně rozložené.



Obr. 5.2: Graf P-P plot: paleta RdYlBu, jedna značka, normalizováno, tři sloupce vstupních dat, kvartilové přímky nejsou vykresleny.

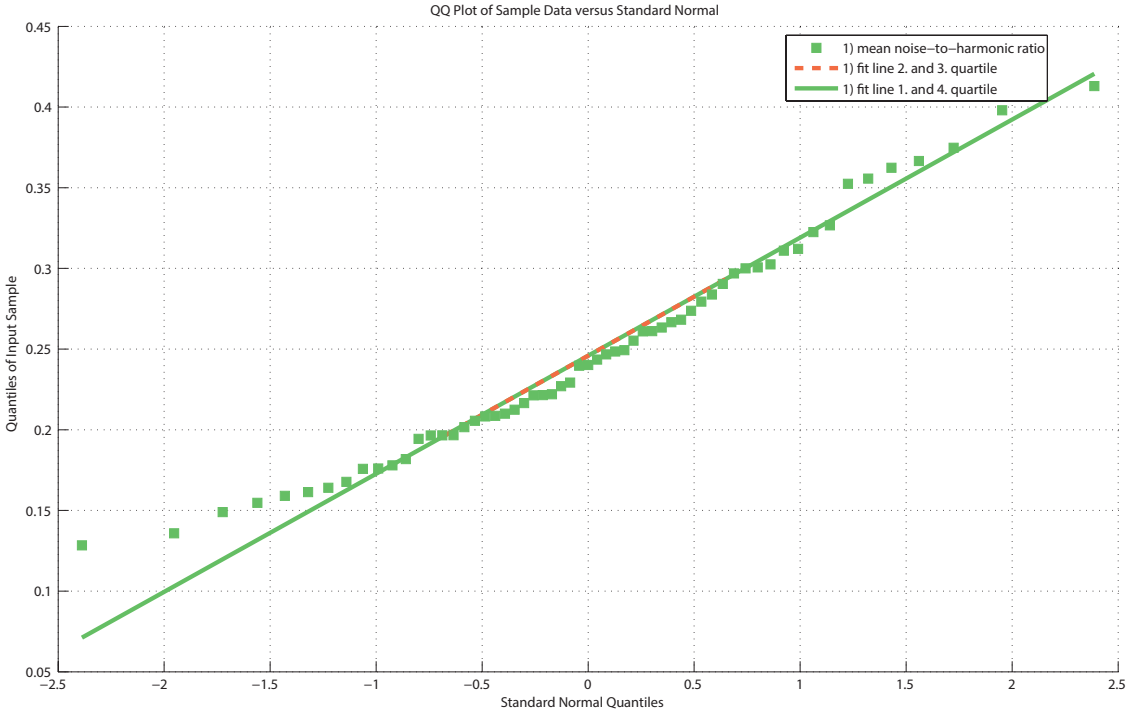


Obr. 5.3: Graf P-P plot: paleta RdYlBu, dvě značky, normalizováno, jeden sloupec vstupních dat, kvartilové přímky jsou vykresleny.

V grafu na obr. 5.3 je vybrán jeden sloupec označený dvěma intervaly značek. V legendě jsou jednotlivé značky vypsány v závorkách za popisem parametru. Body s jistou chybou sledují trend svých kvartilových křivek. Opět lze říci, že jejich data mají normální rozložení hodnot.

5.2 Metoda Q-Q plot

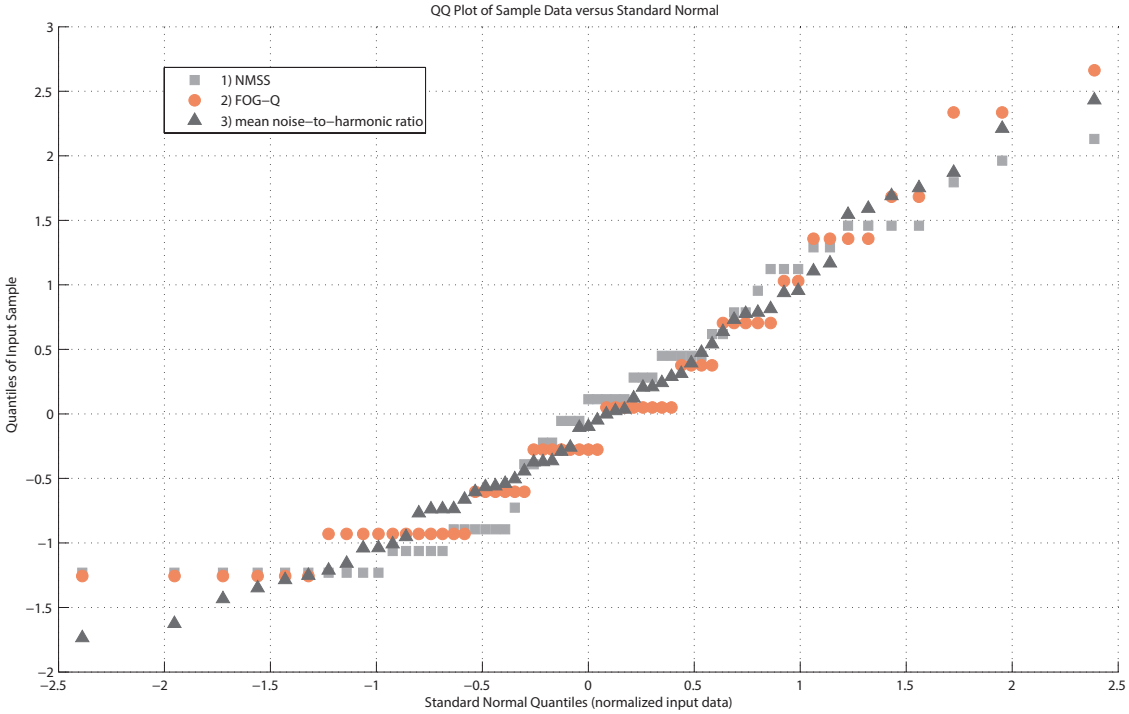
Na obr. 5.4 je zobrazen Q-Q plot pro stejná data jako byla použita na obr. 5.1. Q-Q plot je citlivý na data, která leží na 1. a 4. kvartilu. Výsledný graf potvrzuje, že můžeme s jistou chybou považovat data za normálně rozložená.



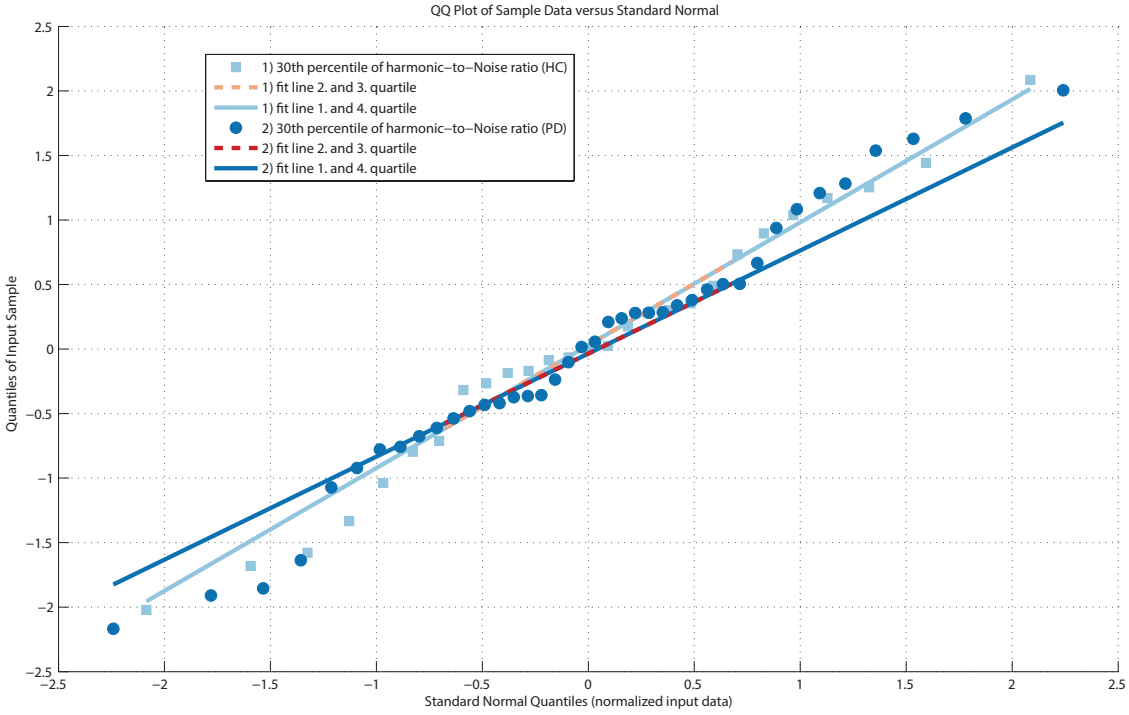
Obr. 5.4: Graf Q-Q plot: paleta RdYlGn, jedna značka, nenormalizováno, jeden sloupec vstupních dat, kvartilové přímky jsou vykresleny.

Další graf na obr. 5.5 dokazuje, podobně jako obr. 5.2, že labels „NMSS“ a „FOG-Q“ nemají normální rozdělení hodnot.

Graf na obr. 5.6 zobrazuje Q-Q plot pro jeden parametr značený dvěma různými intervaly. I přes relativně velké odchylky vykreslených bodů tyto body do jisté míry sledují trend kvartilové přímky. A zároveň se kvartilová přímka blíží ideální přímce se sklonem 45 stupňů. Proto lze tvrdit, že tento parametr má hodnoty s normálním rozložením.



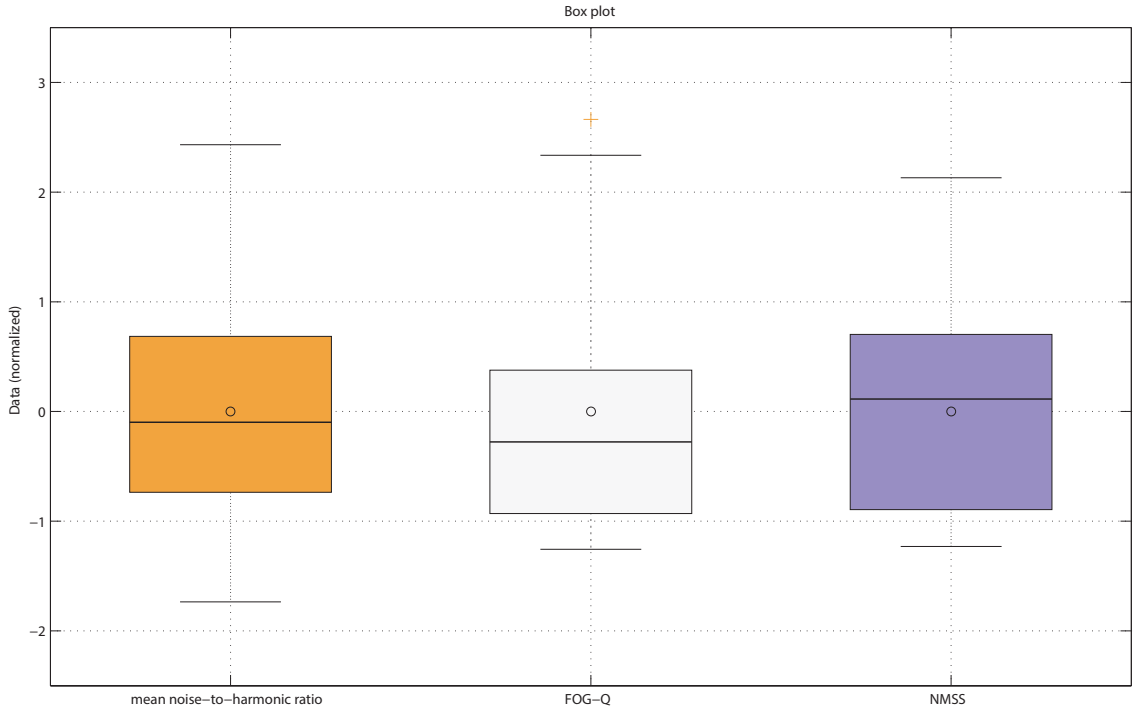
Obr. 5.5: Graf Q-Q plot: paleta RdGy, jedna značka, normalizováno, tři sloupce vstupních dat, kvartilové přímky nejsou vykresleny.



Obr. 5.6: Graf Q-Q plot: paleta RdBu, dvě značky, normalizováno, jeden sloupec vstupních dat, kvartilové přímky jsou vykresleny.

5.3 Metoda box plot

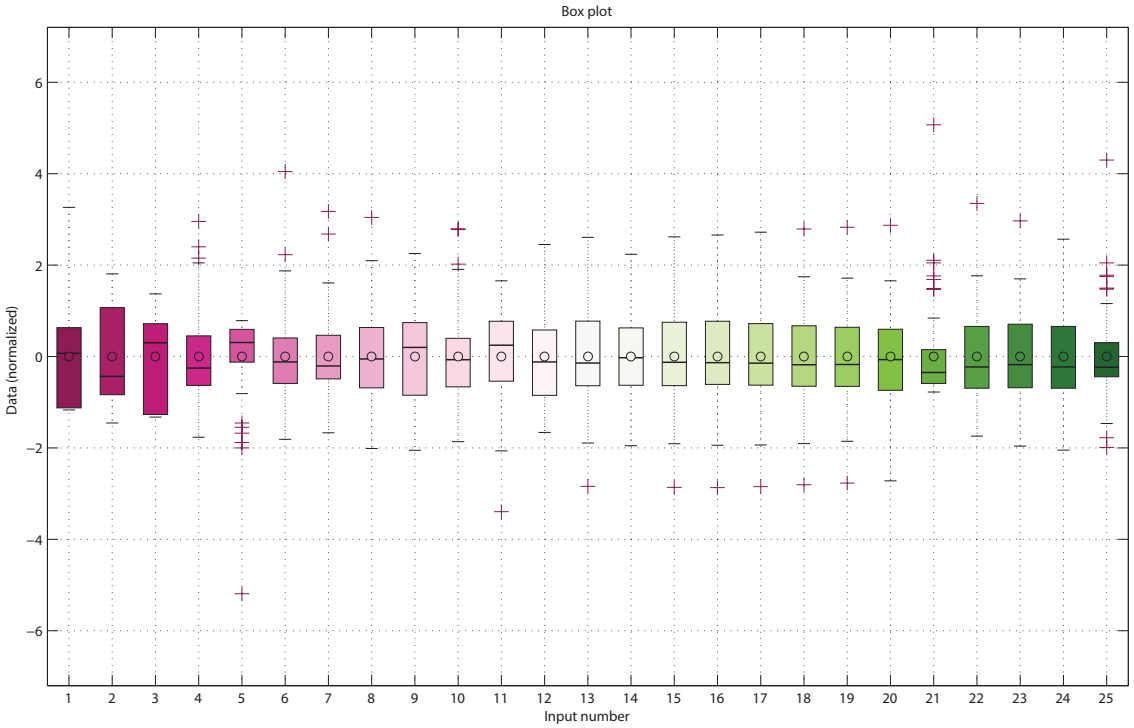
Box plot s normálním rozložením má dolní a horní hranici, která vychází z výsledku $P_{75|25} \pm 1,5 * IQR$ (viz 2.2.5). Na obr. 5.7 je zobrazen Box plot se vstupními daty, které byly použity pro P-P plot nebo Q-Q plot. Lze vidět, že i v tomto případě graf potvrzuje normalitu dat pouze u prvního parametru, který je vykreslen oranžově. Zbylé boxy mají dolní hranice blíže, něž je požadovaných $P_{25} - 1,5 * IQR$. Jelikož jsou data normalizována, značení průměru(kruh) všech dat má hodnotu vždycky 0.



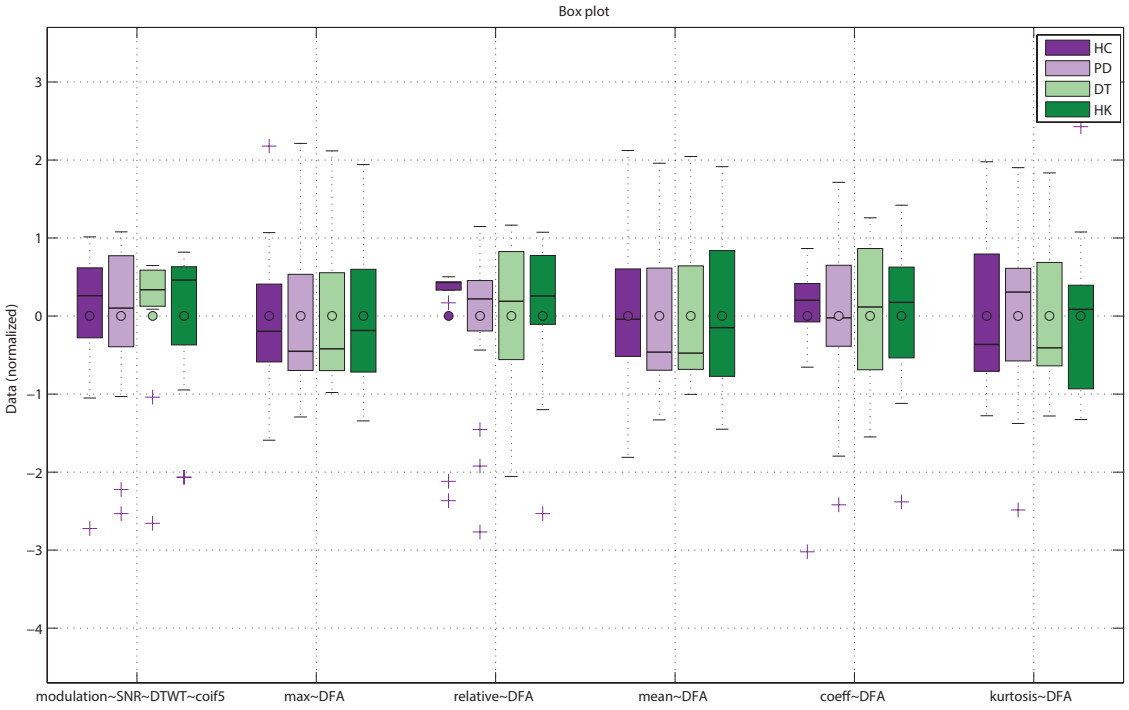
Obr. 5.7: Graf Box plot: paleta PuOr, jedna značka, normalizováno, tři sloupce vstupních dat.

Pro graf na obrázku 5.8 bylo zvoleno 25 vstupních parametrů. Díky metodě box plot lze velmi rychle zjistit, které parametry mají a které nemají normální rozložení dat. V tomto případě určitě parametry č.: 1,2,3,5,21,25 nemají své hodnoty normálně rozloženy.

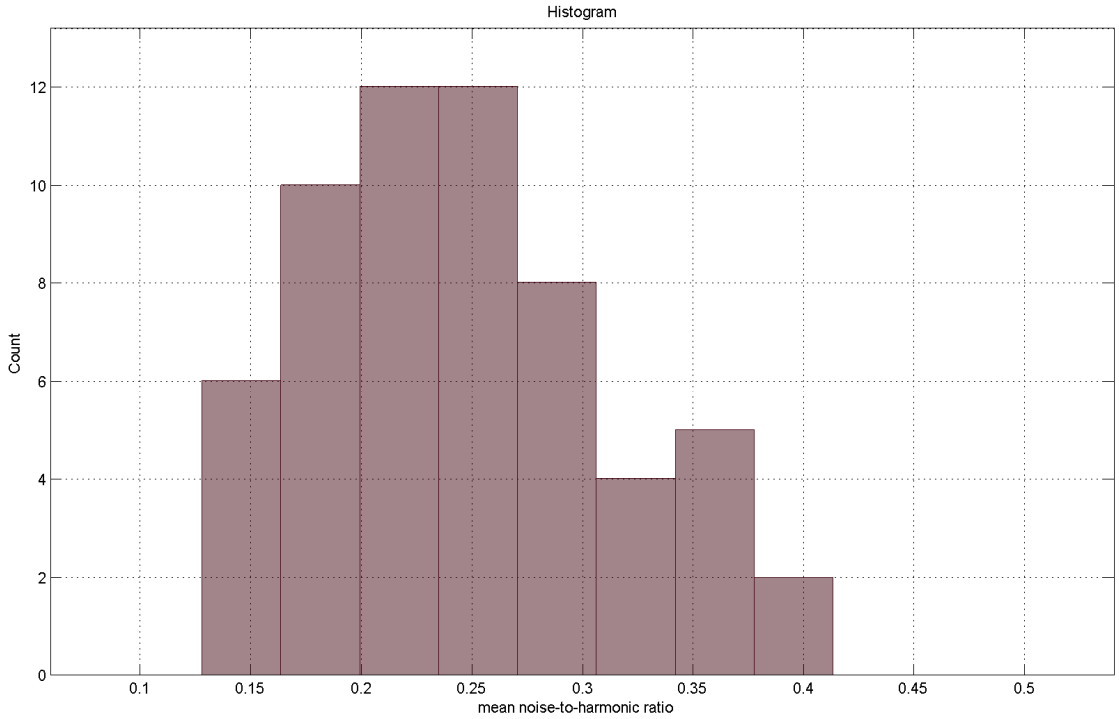
Na posledním obrázku 5.9 je vyobrazena ilustrativní situace se 4 různými intervaly značek. Boxy označených dat pro jednotlivé parametry jsou vždycky u sebe. Intervaly jsou rozlišeny barevně (viz legenda). Popisky na ose x jsou z důvodu úspory místa zkráceny.



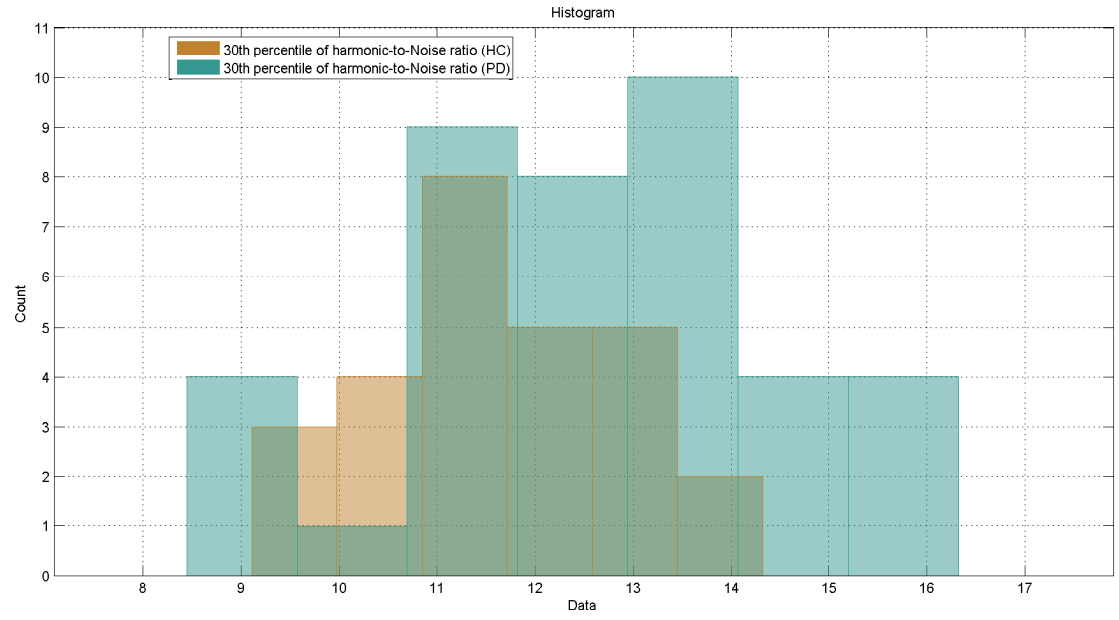
Obr. 5.8: Graf Box plot: paleta PiYG, jedna značka, normalizováno, 25 sloupců vstupních dat.



Obr. 5.9: Graf Box plot: paleta PRGn, 4 značky, normalizováno, 6 sloupců vstupních dat.



Obr. 5.10: Graf Histogram: jedna značka, nenormalizováno, jeden sloupec vstupních dat.



Obr. 5.11: Graf Histogram: paleta BrBG, dvě značky, nenormalizováno, jeden sloupec vstupních dat.

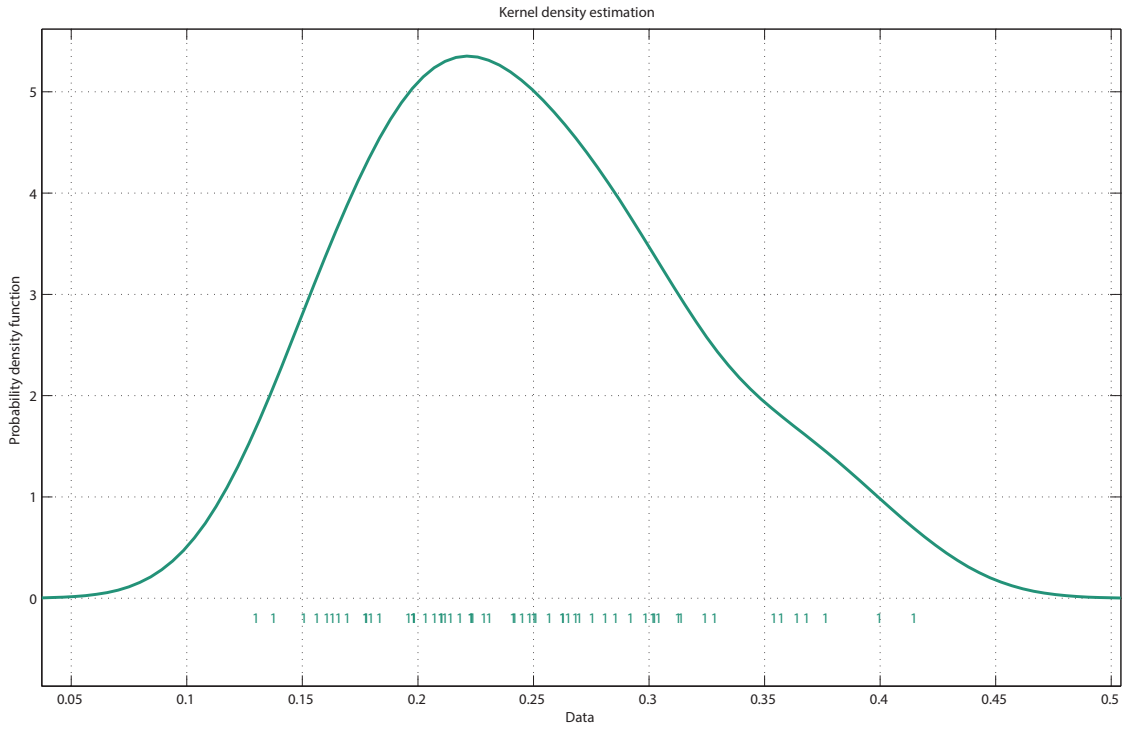
5.4 Metoda histogramů

Na obrázku 5.10 na předchozí straně je vykreslen histogram pro již známý vstupní parametr. Počet sloupců je odvozen pomocí odmocniny z počtu řádků parametru. Lze vidět, že histogram se s jistou chybou podobá normálnímu rozdělení hustoty pravděpodobnosti.

Na dalším obrázku 5.11 je vyobrazena situace pro jeden parametr se dvěma značkami. Jednotlivé histogramy se liší v počtu binů. První se značkou „HC“ má 6 binů a druhý jich má 7. Je to z toho důvodu, že histogram „PD“ je vytvořen z dat, které mají více řádků.

5.5 Metoda jádrových odhadů

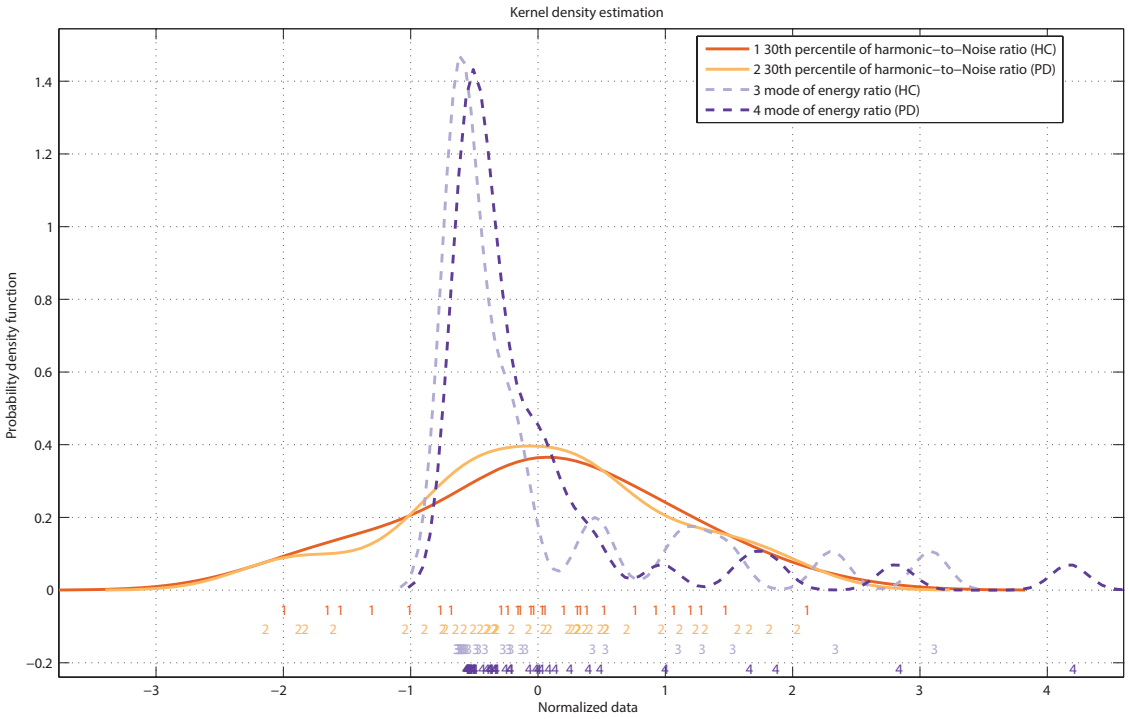
Graf jádrových odhadů na obr. 5.12 vychází ze stejných dat, jako například grafy na obrázcích 5.1,5.4 nebo 5.10. Pod křivkou jádrových odhadů jsou vykreslena původní vstupní data. Křivka s učitou chybou kopíruje ideální křivku normálního rozdělení hustoty pravděpodobnosti.



Obr. 5.12: Graf Jádrové odhady: jedna značka, nenormalizováno, jeden sloupec vstupních dat, symboly vstupních dat jsou vykresleny.

Na následujícím obrázku 5.13 je graf zkonstruován ze dvou sloupců dat, které

jsou označeny dvěma intervaly značek. Jednotlivé sloupce dat jsou od sebe rozlišeny různým typem čáry. Data parametru „mode of energy ratio“ nemají s velkou pravděpodobností normální rozložení.

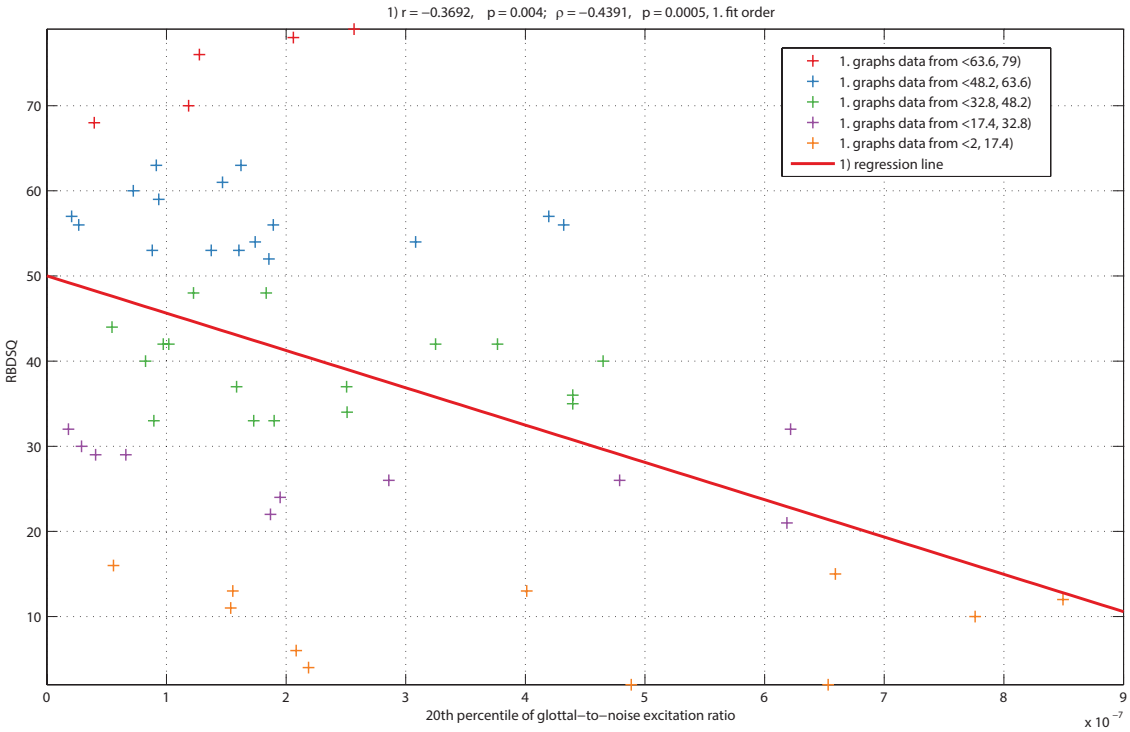


Obr. 5.13: Graf Jádrové odhady: paleta PuOr, dvě značky, normalizováno, dva sloupce vstupních dat, jsou vykreslena vstupní data, symboly vstupních dat jsou vykresleny.

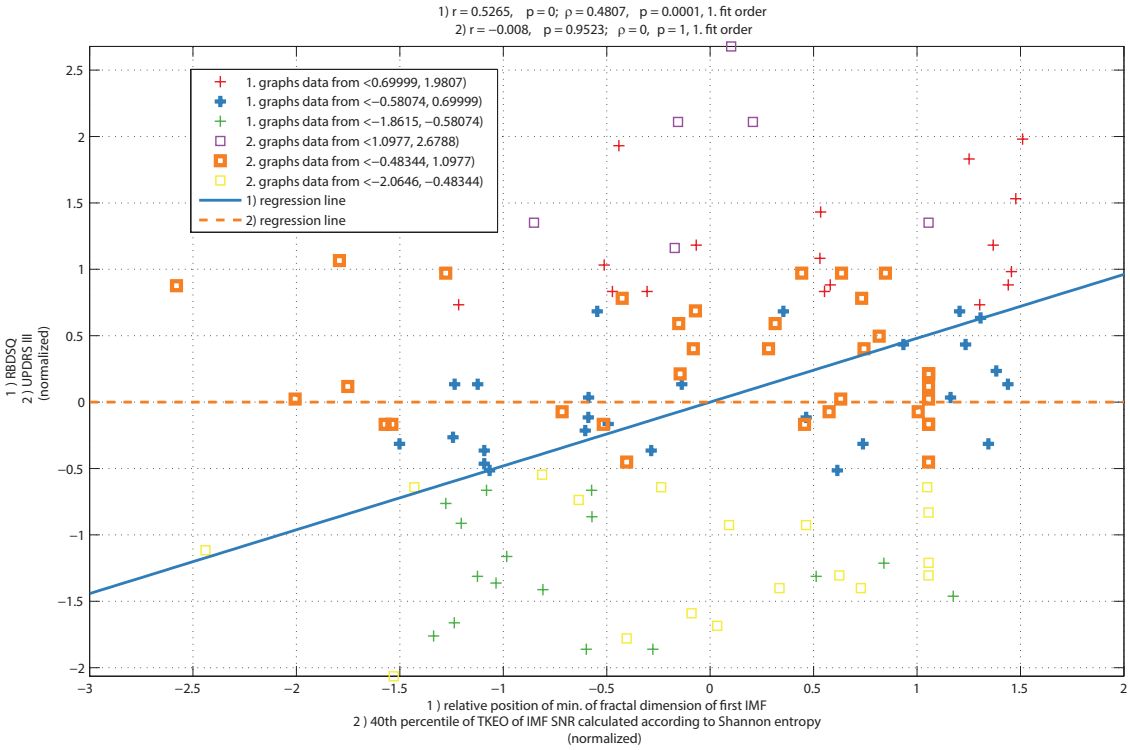
5.6 Metoda korelačních diagramů

Na obr. 5.14 je zobrazen graf korelace se dvěma sloupci vstupních dat. Vykrelené body jsou rozděleny do 5 intervalů podle hodnot osy y. V legendě jsou vypsány koncové hodnoty intervalů. Hodnota Pearsonova koeficientu ρ je -0.4391 s hladinou významnosti $p = 0.0005$. Relativně vysoká hodnota ρ vypovídá o silném záporném lineárním vztahu vstupních dat.

Poslední obrázek 5.15 je zkonstruován ze dvou korelačních grafů. Byly použity 4 různé sloupce dat. První dvojice, podobně jako u situace na předcházejícím grafu 5.14, vykazuje velkou míru lineární závislosti(nyní kladné). Naopak druhá dvojice představuje opačný případ. Hodnota Pearsonova koeficientu ρ je 0 s hladinou významnosti $p = 1$. A zároveň regresivní křivka je téměř rovnoběžná s osou x. Proto lze tvrdit, že vstupní data této dvojice spolu nemají skoro žádný lineární vztah.



Obr. 5.14: Graf korelace: jedna značka, nenormalizováno, dva sloupce vstupních dat, 5 intervalů, verze vykreslení č. 1, je vykreslena regresní přímka.



Obr. 5.15: Graf korelace: jedna značka, normalizováno, 4 sloupce vstupních dat, 3 intervaly, verze vykreslení č. 2, jsou vykresleny regresivní přímky.

6 ZÁVĚR

V práci byly naprogramovány požadované vizualizační fce. V případě externích fcí byly tyto fce dále upraveny tak, aby vhodně vizualizovaly biomedicínská data. Jednotlivé fce nabízí různou variabilitu parametrů pro dosažení požadovaného výsledku uživatelem. Pokud oživatel dodrží formát vstupních dat, tak mu GUI nabízí jednoduchou volbu vstupních data a fcí pro vizualizaci. Před vizualizací GUI vhodně zpracovává data. Díky funkcím pro zpracování dat je GUI dále lehce rozšiřitelné o další statistické metody. Nakonec byly jednotlivé vizualizační fce otestovány na parametrech vypočítaných z řeči poškozené přítomností Parkinsonovy nemoci.

LITERATURA

[1] BIKFALVI, Alex. *Advanced Box Plot for Matlab* [počítačový soubor]. Ver. 1.0, 2012. poslední aktualizace 27. 9. 2012 [cit. 21. 5. 2016]. Dostupné z URL: <http://alex.bikfalvi.com/download/aboxplot.zip>.

[2] COBELDICK, Stephen. *ColorBrewer: Attractive and Distinctive Colormaps* [počítačový soubor]. Ver. 2.2, 2014. poslední aktualizace 19. 3. 2016 [cit. 21. 5. 2016]. Dostupné z URL: <http://www.mathworks.com/matlabcentral/fileexchange>.

[3] GALÁŽ, Zoltán. *Software pro vizualizaci dyzartrické řeči* [počítačový soubor]. Ver. 1.0, 2015. [cit. 21. 5. 2016]. Dostupné z URL: <http://splab.cz/wp-content/uploads/2015/10/dSpeech-visualization.zip>.

[4] GIBBONS, Jean Dickinson a Subhabrata CHAKRABORTI. *Nonparametric Statistical Inference.* 4. vyd. Marcel Dekker, c2003. 677 s. ISBN 0-8247-4052-1.

[5] GREENE, Chad. *histf* [počítačový soubor]. Ver. 1.0, 2014. poslední aktualizace 15. 8. 2016 [cit. 21. 5. 2016]. Dostupné z URL: <http://www.mathworks.com/matlabcentral/fileexchange>.

[6] HASTIE, Trevor, Robert TIBSHIRANI a Jerome FRIEDMAN. *The Elements of Statistical Learning.* 2. vyd. Springer, c2009. 764 s. ISBN 978-0-387-84857-0.

[7] Histogram with a distribution fit. In: *MathWorks* [online]. [cit. 21. 5. 2016]. Dostupné z URL: <http://www.mathworks.com/help/stats/histfit.html>.

[8] JAMES, Gareth, Daniela WITTEN, Trevor HASTIE a Robert TIBSHIRANI. *An Introduction to Statistical Learning.* 4. vyd. Springer, 2014. 440 s. ISBN 978-1-4614-7137-0.

[9] JOHNSON, Roberh a Patricia KUBY. *Elementary statistics.* 11. vyd. Brooks/-Cole, c2012. 833 s. ISBN 0-538-73350-0.

[10] MANOLAKIS, Dimitris G. a Vinay K. INGLE. *Applied Digital Signal Processing.* Cambridge University Press, c2011. 1009 s. ISBN 978-0-521-11002-0.

[11] MARTINEZ, Wendy a Angel MARTINEZ. *Computational Statistics Handbook with MATLAB.* 1. vyd. Chapman & Hall/CRC, c2002. 584 s. ISBN 1-58488-229-8.

[12] MORIN, David. *Combinatorics* [online]. 4. vyd. Harvard University, 2009 [cit. 21. 5. 2016]. Dostupné z URL: <http://www.people.fas.harvard.edu/~djmorin/probability.pdf>.

[13] SMÉKAL, Zdeněk. *Číslicové zpracování signálu: 9. Základy statistiky* [online prezentace]. VUT Brno, 2012 [cit. 21. 5. 2016]. Dostupné z URL: <`http://www.vutbr.cz`>.

[14] SONDEREGGER, Derek. *Introduction to Statistics for Researchers* [online]. Northern Arizona, 2014, poslední aktualizace 18. 4. 2014 [cit. 21. 5. 2016]. Dostupné z URL: <`https://www.github.com/dereksonderegger/STA_570_Book/raw/master/Stat_570.pdf`>.

[15] ŠAFR, Jiří. *Analýza kvantitativních dat I.: Popisné statistiky a explorační jednorozměrná analýza* [online]. Praha, 2009, poslední aktualizace 28. 2. 2015 [cit. 21. 5. 2016]. Dostupné z URL: <`http://metodykv.wz.cz/AKD1_explor_analyza1.ppt`>.

[16] THODE, Henry C. *Testing for normality.* Marcel Dekker, c2002. 368 s. ISBN 0-8247-9613-6.

SEZNAM SYMBOLŮ, VELIČIN A ZKRATEK

GUI	grafické uživatelské rozhraní – graphical user interface
IQR	interkvartilové rozpětí –interquartile range
NaN	výsledek není číslo – not a number
RGB	barevný model červená-zelená-modrá – red-green-blue color model

SEZNAM PŘÍLOH

A DALŠÍ VÝVOJ PROGRAMU

Dosud naprogramovaný program představuje kostru, která může být lehce rozšířena o další funkce. V této kapitole je uvedeno, o jaké fce by mohlo jít. Případně jsou zde informace o částech programu, které by mohly být vylepšeny:

1. Podpora více formátu vstupních dat.
   Program má omezenou podporu vstupních dat pouze na `*.mat`. Další rozšíření o nové podporované formáty(`csv`, `xls`) by umožnilo pohodlnější práci s programem.
2. Kontrola vstupní proměnné `feat_matrix` na `char`.
   Proměnná `feat_matrix` neobsahuje kontrolu na přítomnost znaku, ale pouze na NaN, nebo vhodné rozměry.
3. Zachytávání ošetřených warning hlášení a jejich vhodné předání uživateli.
   Program ve svém běhu průběžně informuje uživatele o aktuálním stavu. Například o přeskočení hodnot NaN, o stavu vstupních dat, apod. Tato informace se zobrazuje pomocí fce `warning` na příkazové řádce programu MATLAB, která je za oknem GUI. Pro lepší přehlednost by měly být tyto hlášení zakomponovány do okna GUI.
4. Podpora více sloupců typu `char` v `labels`.
   Uživatel nemá možnost volby mezi různými sloupci značek. Ve vstupním souboru se může nacházet pouze jeden sloupec značek. Pokud by bylo uživateli dovoleno měnit značení dat, nemusel by mít více vstupních souborů. Mohl by na stejných datech vyzkoužet různé značení dat.
5. Rozšíření popisných statistik a přidání nových vizualizačních funkcí.
6. Překreslování různých grafů přes sebe v jednom okně.
7. Regrese u všech pp a qq kvartilových přímek.
   U kvartilových přímek nejsou uvedeny jejich rovnice.
8. Oprava tisku histogramu do formátu `fig`.
   Při ukládání histogramu dochází z neznámého důvodu k vytvoření špatně naformátovaného souboru `fig`.
9. Tisk normální křivky pro danou vizualizační funkci.
   Každá vizualizační fce by mohla obsahovat volbu pro vykreslení dat z normálového rozložení pro konkrétní zvolená vstupní data (pomocí jejich průměru a směrodatné odchylky).

B OBSAH PŘILOŽENÉHO DVD

Program byl testován v program MATLAB ve verzi 2014a. U dřívějších verzí není zaručena jeho funkčnost.

GUI se spouští pomocí souboru `GUI.m`. Následně je třeba načíst soubor `*.mat` ve správném formátu. Správný formát je popsán v kapitole 4.1, nebo lze využít testovací soubory ve složce `correct_data`. Pak je již GUI připraveno k vizualizaci zvolených dat.

- `/` ..... kořenový adresář přiloženého DVD
  - `xzvonc01_diplom_prace.pdf` ..... elektronická verze práce
  - `GUI.m` ..... soubor pro spuštění GUI
  - `GUI.fig` ..... soubor vygenerovaný pomocí fce `guide`
  - `correct_data` ..... složka se testovacími soubory ve správném formátu
    - `labels_1char_4TAGS_INTHEMIDDLE.mat` ..... data se 4 značkami
    - `labels_1char_1TAG.mat` ..... data s 1 značkou
    - `kernelDensityData.mat` ..... data se 2 značkami
    - `correlationData.mat` .... data s 1 značkou – `labels` neobsahuje double
  - `incorrect_data` ..... složka s testovacími soubory v nesprávném formátu
  - `Data_processing` ..... složka s funkcemi pro zpracování dat
    - `LoadMatrix2Cell.m` ..... transformace `feat_matrix` na buňku `data`
    - `CheckMyData.m` ..... kontrola vstupních dat
    - `GetLabelIndex.m` ..... získání indexů značek
    - `ControllAllLabels.m` ..... kontrola dat uvnitř `labels`
    - `AfterPickFiller4TheRest.m` ... úprava `feat_matrix` na základě `buffer`
    - `AfterPickFiller4Corr.m` ... podobně jako předchozí fce, ale pro korelaci
    - `brewermap_view.m` ..... externí fce - prezentace volitelných palet RGB
    - `brewermap.m` ..... externí fce - generování matice RGB
  - `Box_plot` ..... složka s funkcemi pro graf Box plot
    - `PlotBox.m` ..... vizualizační fce pro graf Box plot
    - `aboxplot.m` ..... externí fce - advanced boxplot
    - `quartile.m` ..... externí fce - součást `aboxplot.m`
    - `colorgrad.m` ..... externí fce - součást `aboxplot.m`
  - `Correlation_plot` ..... složka s funkcemi pro graf korelace
    - `CorrelationPlot.m` ..... vizualizační fce pro graf korelace
    - `split_data.m` . externí fce - získání matice pro rozdělení dat na intervaly
  - `Histogram_plot` ..... složka s funkcemi pro histogram
    - `PlotHistogram.m` ..... vizualizační fce pro histogram
    - `histf.m` ..... externí fce - histf
  - `KernelDensityEstimation_plot` ... složka s funkcemi pro graf jád. odhadů
    - `KernelDensityEstimation.m` vizualizační fce pro graf jádrových odhadů
  - `PP_plot` ..... složka s funkcemi pro P-P plot
    - `PlotPP.m` ..... vizualizační fce pro P-P plot
  - `QQ_plot` ..... složka s funkcemi pro Q-Q plot
    - `plotQQ.m` ..... vizualizační fce pro Q-Q plot
  - `Time_series` ..... složka s funkcemi pro výpočet popisných statistik
    - `PopisneStatistiky.m` ..... fce pro výpočet matice popisných statistik