Západočeská univerzita v Plzni
Fakulta aplikovaných věd

NEPARAMETRICKÝ ODHAD SPOLEHLIVOSTI A ODHAD TRENDOVÉ SLOŽKY

Ing. Tomáš Ťoupal

disertační práce

k získání akademického titulu doktor

v oboru Aplikovaná matematika

Školitel: Doc. Ing. František Vávra, CSc.

Katedra matematiky

Plzeň 2013

University of West Bohemia in Pilsen
Faculty of Applied Sciences

NONPARAMETRIC ESTIMATION OF RELIABILITY AND TREND COMPONENT ESTIMATION

Ing. Tomáš Ťoupal

dissertation thesis in partial fulfilment of requirements

for the degree of Doctor of Philosophy

in specialization Applied Mathematics

Supervisor: Doc. Ing. František Vávra, CSc.

Department of Mathematics

Pilsen 2013

Prohlášení

Prohlašuji, že jsem práci vypracoval samostatně s použitím odborné literatury a pramenů, jejichž přehled je součástí předkládané práce.

V Plzni 15. 4. 2013

……

Ing. Tomáš Ťoupal

Poděkování

Tímto bych chtěl poděkovat svému školiteli doc. Ing. Františku Vávrovi, CSc. za cenné připomínky a rady v průběhu mého doktorského studia.

Dále bych rád poděkoval rodině, přítelkyni a všem přátelům včetně zvířecích za velmi silnou a potřebnou podporu.

Abstrakt

V reálném světě jsou odhady spolehlivosti nebo naopak selhání nejrůznějších systémů často používány v mnoha aplikacích, zejména v technických konceptech. Z tohoto důvodu jsou v literatuře často diskutovány a představovány parametrické odhady hustot a distribučních funkcí pro získání hodnoty spolehlivosti pro známé pravděpodobnostní rozdělení. Problém ale může nastat v případě, kdy nám tato rozdělení nejsou známy nebo dostupné pro efektivní modelování. Modelování spolehlivosti je součástí analýz rizik.

Uvedená práce se zabývá problematikou spolehlivosti pro dvourozměrné náhodné veličiny s pomocí neparametrických jádrových odhadů, které využívají několik možných jádrových funkcí a navrhuje některé závěry a metody. Odvozené metody jsou prezentovány při získání spolehlivosti z vygenerovaného a náhodného souboru dat z platební bilance České republiky. V tomto případě je spolehlivost reprezentována skutečností (pravděpodobností), že celková výše výdajů nepřesáhne celkovou výši příjmů za celé časové období, což vede na odhad trendové složky pomocí nově navržené metody s využitím vygenerovaného ortonormálního systému. V závěru této práce jsou prezentovány některé další otevřené otázky pro budoucí možný výzkum.

Klíčová slova

Spolehlivost, selhání, neparametrický jádrový odhad, jádrová funkce, hustota a distribuční funkce, časová řada, ortonormální systém, skalární součin, trendová složka

Hlavní cíl

Mezi hlavní cíle této práce patří nalezení spolehlivosti u dvourozměrné náhodné veličiny za předpokladu stacionárního charakteru získaných dat. V opačném případě je dalším cílem této práce nalezení odhadu (aproximace) neznámé trendové složky, kterou lze vzhledem k fundamentální podstatě makroekonomických časových řad zde předpokládat.

Struktura práce je popsána následujícími body:

- Stručná prezentace aktuálně používaných metod k odhadu spolehlivosti.
- Prezentace neparametrických jádrových odhadů spolehlivosti v případě dvourozměrné náhodné veličiny a odhad (predikce) trendové složky.
- Analýza vlastností získaných odhadů.
- Ověření na simulovaném i reálném souboru dat.

Abstract

In reality, the reliability or failure estimations of various systems are often used in many applications, especially in engineering concepts. Therefore, the parametric estimation of density and distribution functions of reliability following some known distribution has been discussed extensively in the literature and they are also briefly introduced. However, the problem could occur in particular if these distributions are unknown or available for effective modelling. Reliability modelling is a part of risk analysis.

This work deals with the problem of reliability estimation for two-dimensional random variables by nonparametric estimation using several types of kernel functions and suggests some of conclusions and methods. Derived methods are presented to obtain the reliability of generated and obtained data collections of the balance of payments of the Czech Republic. In this case, the reliability is represented by the fact (probability) that the total amount of expenditures is not greater than the total amount of incomes over the whole time period. Of course there are discussed the problems with real data collection which leads to an estimate of the trend component using the newly proposed method with some generated orthonormal system. In the end of this work, there are also presented some open questions for future research.

Keywords

Reliability, failure, nonparametric kernel estimation, kernel function, density and distribution function, time series, orthonormal system, inner product, trend component

Main Goal

Main goal of this thesis is to find a two-dimensional nonparametric kernel estimation of reliability in the case of stationary character of the observed data. Otherwise, this thesis is focused on the trend component estimation (approximation) of time series which it can be assumed due to the fundamental macroeconomics nature of the obtained data collection.

Structure of this thesis is described by following points:

- Presentation of current state of art.
- Presentation of nonparametric kernel estimation of reliability in the case of two-dimensional random variable and trend component estimation (prediction).
- Analysis of properties of presented estimators.
- Properties verification of the estimators by the simulation and by obtained real data collections.

Obsah

Seznam obrázků

Seznam tabulek

Seznam použitého značení

Čísla v závorce označují stránky, kde je symbol poprvé použit nebo definován.

Římské symboly

n	obvykle počet pozorování (16)
e	$\cong 2.718\ 28 \ldots$ Eulerovo číslo (14)
$P(.)$	pravděpodobnost (6)
$F(.)$	distribuční funkce pravděpodobnosti (6)
$\bar{F}(.)$	funkce přežití (15)
$f(.)$	funkce hustoty pravděpodobnosti (6)
$E\{.\}$	střední hodnota při odvozování (17)
R	spolehlivost (5)
F	selhání (5)
$N(.)$	normální rozdělení pravděpodobnosti (14)
$k(.)$	jádrová funkce hustoty (20)
$K(.)$	jádrová distribuční funkce (26)
$s_n^2\{.\}$	výběrový rozptyl z n pozorování (34)
h	vyhlazovací parametr (parametr měřítka) (19)
ln	přirozený logaritmus (9)
e_k	jednotkový vektor (70)
a_k	$k-t\acute{y}$ koeficient trendové složky (70)
g_k	$k-t\acute{y}$ prvek ortonormálního systému (71)
c	copula funkce hustoty (13)
C	copula distribuční funkce (13)
V_m	Vandermondova matice (68)
V	vektorový prostor (63)
$T = o(h_n)$	označuje $lim_{n\to\infty} \frac{\lvert T\rvert}{h_n} = 0$ (30)
$x(t)$	časová řada pro $t = 1, \ldots, T$ (8)
$X(t)$	předpokládaná a neznámá trendová složka (datový vektor) pro $t = 1, \ldots, T$ (8)
$X_m(t)$	odhad trendové složky z m komponent pro $t = 1, \ldots, T$ (76)
$X_I(t)$	odhad trendové složky z I komponent, vybraných heuristickým postupem pro $t = 1, \ldots, T$ (89)

Řecké symboly

μ	parametr střední hodnoty, již odvozený nebo definovaný (14)
σ^2	parametr rozptylu, již odvozený nebo definovaný (14)
$\sigma^2\{.\}$	parametr rozptylu při odvozování (17)
ε	náhodná složka (8)
π	$\cong 3.14159 \ldots$ Pí (14)
ρ	koeficient korelace (14)
τ	čas do první rizikové události (10)
Φ	distribuční funkce normovaného normálního rozdělení $N(0,1)$ (17)
α	podmíněná pravděpodobnost vyjádřená vztahem (2.12) (10)

Symboly podobné písmenům

$\mathbb{R}$	množina všech reálných čísel (6)
$\mathbb{R}^n$	n-rozměrný Euklidovský prostor, $n \in \mathbb{N}$ (64)
$\mathbb{N}_0$	množina všech přirozených čísel a nula, tj. $\{0,1,2, \dots\}$ (68)
$\mathbb{N}_+$	množina všech přirozených čísel, tj. $\{1,2, \dots\}$ (19)
$\mathbb{Z}$	množina všech celých čísel (60)

Zkratky

i.i.d.	nezávislé a stejně rozdělené náhodné veličiny (16)
Var	rozptyl vyjádřený vztahem (3.22) (28)
Bias	vychýlení vyjádřené vztahem (3.23) (28)
Cov	kovariance (14)
Corr	korelace (14)
ISE	Integrální kvadratická chyba (28)
MSE	Střední kvadratická chyba (28)
MISE	Střední integrální kvadratická chyba (30)
AMISE	Asymptotická střední integrální kvadratická chyba (29)
proj	operátor projekce (70)
det	determinant (67)
■	zakončení důkazu (31)

Zpřehledňující zkratky

$F_{X,Y}(x,y)$	dvourozměrná distribuční funkce pro náhodné veličiny X, Y (6)
$f_{X,Y}(x,y)$	dvourozměrná funkce hustoty pro náhodné veličiny X, Y (6)
ε_{X_t}	náhodná složka pro náhodný proces X v čase t (8)
$\hat{f}(x;h)$	neparametrický jádrový odhad funkce hustoty (19)
$\hat{F}(x;h)$	neparametrický jádrový odhad distribuční funkce (26)
$E_{\hat{f}(x;h)}\{X\}$	odhad střední hodnoty pro náhodnou veličinu X s použitím modelové funkce hustoty, definováno ve vztahu (3.5) (22)
$\sigma^2_{\hat{f}(x;h)}\{X\}$	odhad rozptylu pro náhodnou veličinu X s použitím modelové funkce hustoty, definováno ve vztahu (3.7) (22)
$E\{\hat{F}(x;h)\}$	odhad střední hodnoty pro náhodnou veličinu $\hat{F}(x;h)$, který je pro každé x náhodnou veličinou (37)
$\sigma^2\{\hat{F}(x;h)\}$	odhad rozptylu pro náhodnou veličinu $\hat{F}(x;h)$, který je pro každé x náhodnou veličinou (39)
$\hat{F}_m(x;h)$	neparametrický jádrový odhad distribuční funkce součtu m nezávislých náhodných veličin (44)
$\hat{f}_m(x;h)$	neparametrický jádrový odhad funkce hustoty součtu m nezávislých náhodných veličin (45)
$E_{\hat{f}_m(x;h)}\{Z_{m+1}\}$	odhad střední hodnoty náhodné veličiny Z_{m+1}, vzniklé ze součtu $(m+1)$ nezávislých náhodných veličin s využitím $\hat{f}_m(x;h)$ (45)
$\sigma^2_{\hat{f}_m(x;h)}\{Z_m\}$	odhad rozptylu náhodné veličiny Z_m, vzniklé ze součtu m nezávislých náhodných veličin s využitím $\hat{f}_m(x;h)$ (46)
$E_{k_x}\{X\}$	odhad střední hodnoty komponenty $\frac{1}{h_x}k_x\left(\frac{x-x_i}{h_x}\right)$ pro náhodnou veličinu X (51)
$\sigma^2_{k_x}\{X\}$	odhad rozptylu komponenty $\frac{1}{h_x}k_x\left(\frac{x-x_i}{h_x}\right)$ pro náhodnou veličinu X (51)

Kapitola 1

Úvod

Jeden z možných a primárních cílů statistické analýzy obecně na získaném souboru reálných dat je odhadnout „vhodný“ statistický model známého (již popsaného) pravděpodobnostního rozdělení, který lze následně použít pro další navazující a rozšiřující statistické zkoumání, v našem případě pro získání neznámé hodnoty spolehlivosti nebo selhání uvažovaného systému (objektu, …). K získání potřebné hodnoty lze použít více možných přístupů. Jeden možný a nikoliv jediný, který lze použít k „odhadnutí“ neznámého tvaru hustoty nebo distribuční funkce pravděpodobnostního rozdělení na souboru pozorování, je jejich vykreslení s pomocí dobře známého schématu, které se ve statistické literatuře nazývá histogram[1], a které nám poskytuje prvotní názornou představu o důležitých souvislostech.

Použití tohoto přístupu může být chápáno i jako odhad hustoty ze získaných náhodných vzorků s využitím tzv. jádrového odhadu hustoty, popřípadě jádrového odhadu distribuční funkce. Uvedený postup pro získání odhadů s pomocí různých typů jádrových funkcí je v literatuře označován jako neparametrické jádrové odhady hustoty nebo neparametrické jádrové odhady distribuční funkce neznámého pravděpodobnostního rozdělení z náhodného souboru získaných pozorování.

Získané výsledky (tvary a vlastnosti pravděpodobnostního rozdělení) mohou být následně použity pro uskutečnění „spolehlivostní analýzy (analýzy rizik)“ prostřednictvím modelu, který reprezentuje reálnou situaci[2] (situaci z běžného života), bez předem známého pravděpodobnostního rozdělení (např. často zmiňované sdružené dvourozměrné normální rozdělení pravděpodobnosti). Pokud se tedy zaměříme na možné reálné situace, potom tento přístup může být používán v mnoha vzniklých situacích, zejména v nejrůznějších technických konceptech (mechanické konstrukce, statika budov a mostů, pevnost tlakových nádob, …), v lékařství, v řízení kvality ve firemním sektoru, armádní využití nebo i ve finančních operacích (v platební bilanci) k následným a opodstatněným rozhodováním, zda zde mohou nastat významná „rizika[3]“

[1] Statistický přístup pro odhad funkce hustoty nebo distribuční funkce pomocí sloupcového grafu se sloupci stejné šířky, jejichž výška je určena počtem vyskytujících se hodnot v každém sloupci.

[2] Tento pojem poprvé prezentoval William I. Thomas větou: „Jestliže je určitá situace lidmi definovaná jako reálná, pak je reálná ve svých důsledcích“ Thomas (1923).

[3] Existují různé definice pojmu „rizika“, který je často používán bez exaktnějšího pohledu. Obecně lze riziko označit za pravděpodobnost výskytu nežádoucí události (s nežádoucími účinky). Upřesňující význam pro účely této práce je detailně vysvětlen v Kapitole 2.

jako např. částečné nebo celkové nebezpečí poškození funkčnosti zařízení, selhání systému, vzplanutí nebo exploze, vznik zranění, vzniklé finanční škody, insolvence atd.

Z těchto často závažných a důležitých rozhodovacích situací si uvedená práce klade za svůj hlavní cíl navržení a následné verifikování možných (vytvořených) modelů k získání neznámé hodnoty pravděpodobnosti ať už spolehlivosti nebo selhání ze souboru reálných dat a dále na jejich podkladech formulovat i interpretovat opodstatněné závěry a komentáře pro případné posouzení co možná nejširšího spektra rizik.

Celková koncepce uvedené práce je založena na třech hlavních obsahových částí, kde se každá z nich podrobněji zaměřuje na detailní řešení odlišné části vzniklého problému vzhledem k celkově provázané problematice.

V první části se hlavní pozornost zaměřuje na co největší přiblížení teoretické úrovně této problematiky za pomoci upřesňujících definic či vět a poté následuje odhad neznámé hodnoty spolehlivosti[4] pro jednorozměrnou i dvourozměrnou náhodnou veličinu s použitím několika možných typů jádrových funkcí pro odhad (aproximaci) křivky hustoty a distribuční funkce pro nás apriorně neznámého avšak v literatuře již dostatečně popsaného pravděpodobnostního rozdělení. V druhé části této práce navazuje hlavní pozornost na již vzniklé problémy z předchozí části a navrhuje následné možné řešení zakládající se na odhadu (aproximaci) předpokládané a opět apriorně neznámé trendové[5] složky v uvažované makroekonomické časové řadě[6] s využitím vygenerovaného libovolného systému s pomocí různých numerických metod (Gram-Schmidtův ortonormalizační proces, $Q - R$ rozklad, …). To je provedeno v případech, kdy vznikne podezření na nestacionární charakter[7] získaného reálného souboru náhodných pozorování (pokud uvažujeme makroekonomický soubor reálných dat, potom lze primárně předpokládat nestacionárních charakter u většiny datových zdrojů).

Následuje tvorba „vhodného“ modelu k predikci vývoje prostřednictvím trendové křivky, který je založen na navrhnutých postupech a využívá předchozí znalosti a modely. Zakončení této práce spočívá v propojení těchto přístupů a znalostí pro výslednou spolehlivost na získaném souboru reálných dat, kterou lze interpretovat i za podmínky počáteční nestacionarity.

Nejprve v případě podrobnějšího zaměření na předpokládané a neznámé hodnoty spolehlivosti (selhání) zjistíme, že jeden z možných a zároveň jednodušších přístupů jak

[4] Pojem spolehlivost (opak selhání) představuje neznámou a odhadovanou pravděpodobnost intuitivní situace, která bude podrobněji popsána v následující kapitole.

[5] Všeobecně používaný pojem, představující možný dlouhodobý vývoj sledované veličiny, jehož problematika a zejména nejednoznačnost je popsána dále.

[6] Intuitivně představuje tento pojem časově srovnatelná a uspořádaná pozorování dat, vykazující možné souvislosti vzhledem ke svému časovému vývoji.

[7] Nestacionární charakter reálných dat ze statistického úhlu pohledu je zde uvažován v předpokladu, že marginální rozdělení jednotlivých pozorování jsou (funkčně) nezávislá na čase realizace.

získat požadovanou hodnotu spolehlivosti spočívá v použití přístupu pomocí tzv. relativní frekvence[8]. Následný možný přístup k získání neznámé hodnoty spolehlivosti je založen na použití již známého a častokrát odvozovaného modelu (např. dvourozměrné normální rozdělení) nebo v transformaci získaného souboru náhodných pozorování na uvažovaný a známý sdružený dvourozměrný model[9].

Pokud tyto přístupy selžou (ať již z důvodu (ne)stacionarity, neznalosti vhodného modelu, náročnosti na čas a výpočty, …), potom k odhadu lze použít dále popsaný přístup, který je založen na využití zmiňovaného neparametrického jádrového odhadu právě bez použití některého explicitního modelu. Detailní popis tohoto přístupu a jeho vlastnosti jsou jednou z hlavních náplní této disertační práce, která si ve stručnosti klade za hlavní cíl právě návrh různých modelů pro odhad neznámé hodnoty.

Z hlediska věcného popisu jsou jednotlivé kapitoly systematicky členěny podle svého obsahu, kde navazuje po úvodní části druhá kapitola, ve které jsou podrobně popsány a vysvětleny nezbytné základní definice a vztahy potřebné nejen k prvotnímu popisu dané problematiky, ale i k následné tvorbě navrhovaných modelů (vztahů) pro získání požadovaných hodnot spolehlivosti. Dále jsou zde publikovány tři obecné přístupy k tvorbě dvourozměrných pravděpodobnostních modelů a dva „odlišné“ modely z pohledu (subjektivní) náročnosti včetně odvození základních vztahů pro tvorbu neparametrického jádrového modelu k odhadu spolehlivosti.

Ve třetí kapitole jsou diskutovány modely jednorozměrného a dvourozměrných neparametrických jádrových odhadů hustot a distribučních funkcí, které při svých odhadech používají tzv. jádrové[10] funkce, u kterých je zároveň interpretováno několik intenzivněji používaných typů těchto funkcí. Následně jsou dva typy vybrány (Parzenova a Gaussova jádrová funkce) a použity vzhledem k jejich požadovaným vlastnostem. Dále je zde uvedena problematika „výběru“ vyhlazovacího parametru jednak pro odhad hustoty, ale i modifikace pro odhady distribučních funkcí, plynoucí z dílčí znalosti i neznalosti „skutečného“ pravděpodobnostního rozdělení náhodné veličiny. Závěr kapitoly se zabývá využitím předchozí problematiky (včetně Kapitoly 2) pro neparametrické jádrové odhady distribučních funkcí součtů a použitím pro získání neznámých hodnot (pravděpodobností) spolehlivosti (selhání).

Ve čtvrté kapitole je využití modelů v aplikační sféře. V úvodu je popsán soubor získaných pozorování z platební bilance České republiky, včetně charakteru dat, kde v případě nestacionárního charakteru získaných pozorování je navržen model pro možnou aproximaci neznámé „trendové složky“ z předpokládaného aditivního

[8] Relativní frekvence představuje podíl počtu výskytu události ku počtu pokusů celkem a je blíže vysvětlena v Kapitole 5 .

[9] Zde je myšleno zejména sdružené dvourozměrné normální rozdělení vzhledem k jeho vlastnostem.

[10] Pod tímto pojmem si lze představit libovolnou funkci na oboru reálných čísel, splňující dále uvedené předpoklady.

heuristického modelu časové řady. Uvedené poznatky jsou založeny na skutečnosti, že u velkého počtu získaných souborů makroekonomických dat nelze z jejich fundamentální podstaty předpokládat právě stacionární charakter získaných pozorování (tj. předpoklad nestacionárního charakteru). Nalezená „trendová složka“ ze získané časové řady je detailně popsána (včetně vzniklých problémů) a odhadována na základě vygenerovaného ortonormálního systému (Gram-Schmidtův ortonormalizační proces, $Q - R$ rozklad) z vygenerovaných indexových[11] čar. Dále kapitola pojednává o tzv. „retrospektivním“ a „prediktivním“ trendu. Retrospektivní trend charakterizuje oddělení systematické časově proměnné složky a složky náhodné (vzniklé jádrovým odhadem) ze získaných pozorování, které jsou podrobeny statistické inferenci. Prediktivní trend je propojen s vybraným krátkodobým predikčním modelem. Závěr kapitoly je zaměřen na nezápornost platební bilance modelované spolehlivostí a na platební bilanci obecně, jak v aktuálních hodnotách, tak i v kumulacích.

Pátá kapitola se zaměřuje na konkrétní prezentace a ověření navrhovaných vztahů a modelů. Je zde prostor pro vlastní realizaci popisované problematiky, která v sobě zahrnuje použití navrhovaných vztahů včetně k tomu vytvořených algoritmů. Výsledky a z nich odvozené závěry jsou prezentovány jak v aktuálních hodnotách, tak v kumulacích (platební bilance). Závěr kapitoly popisuje celkové shrnutí provedené práce na souboru reálných (makroekonomických) dat a popis ověřovacího systému, tj. vytvořeného uživatelského programu v software MATLAB 2010a.

V závěrečné šesté kapitole jsou prezentovány některé další otevřené problematické otázky i návrhy pro budoucí rozšiřující výzkum a práce je zakončena stručným celkovým závěrem shrnujícím použité metody a získané výsledky z pohledu autora.

K práci je také připojena jedna příloha, jejíž obsah nebyl vzhledem ke svému rozsahu zařazen do hlavního obsahu práce. Jsou zde detailněji popsány dva náročnější (časově i plošně) důkazy vlastností neparametrického jádrového odhadu, asymptotická nestrannost a vydatnost.

[11] Lze použít libovolnou vhodnou posloupnost funkcí např. úrokové křivky, které jsou použity a detailně popsány v příslušné kapitole dále.

Kapitola 2

Spolehlivost

V této části práce jsou prezentovány a diskutovány vybrané pojmy z teorie zabývající se spolehlivostí, které jsou nezbytné pro účely této práce. Rozbory, odvozené důkazy i vztahy (vazby, modely) a více podrobnějších informací, týkající se tohoto tématu, mohou být nalezeny ve velkém množství nejrůznějších prací prolínajících se různými vědními obory. K dispozici lze shlédnout mnoho knih a článků na uvedené téma, které popisují prezentovanou problematiku z různých úhlů pohledu, jako jsou články např. od Nadarajah (2005), Hangal (1996), Nadarajah a Kotz (2005) nebo v knize Gupta a Subramanian (1998) atd.

2.1 Různá pojetí spolehlivosti a jejich souvislosti

Cílem disertační práce je odhad neznámých parametrů z teorie spolehlivosti, která sama o sobě zahrnuje nejrůzněji se vyskytující rizikové události a z toho plynoucí rizika i následky pro rozhodování. Z tohoto důvodu se celá tato práce zabývá pojmem „spolehlivost“ (nebo též pojmem opačným tzv. „selhání“) včetně související teorie, a proto zde bude nejprve uveden jeden možný popis pro upřesnění významu tohoto slova. V odborné literatuře se lze setkat s mnoha uvedenými popisy a jeden z nich prezentuje např. M. Todinov (2005) ve své knize v následujícím znění:

Spolehlivost je schopnost daného subjektu vykonávat požadovanou funkci za určených podmínek na uvažovaném časovém intervalu. V matematickém významu je spolehlivost měřena pravděpodobností, že celý systém nebo jen jeho komponenta bude pracovat bez poruchy během určitého časového intervalu za daných provozních podmínek a prostředí.[12]

(citace přeložena z Todinov (2005), s. 1)

Následuje popis postupu k získání (kvantifikace) neznámé hodnoty spolehlivosti (často označované písmenem R z anglického překladu „reliability“) nebo naopak postup k získání neznámé hodnoty s opačným významem, která je charakterizována jako „selhání“ systému (a je často označována písmenem F opět z anglického „failure“).

[12] *The reliability is the ability of an entity to perform a required function under given conditions for a given time interval. In mathematical sense, reliability is measured by the probability that a system or a component will work without failure during a specified time interval under given operating conditions and environment.*

Neznámou hodnotu (pravděpodobnost) spolehlivosti lze v případě dvourozměrné náhodné veličiny obecně formulovat jako $R = P(X < Y)$. Tento přístup je možné stručně popsat jako pravděpodobnost takové situace, že vybraná náhodná veličina (vnější zátěžová síla, výdaje, …) nepřekročí další působící náhodnou veličinu (pevnost materiálu, příjmy, …).

Obecnější interpretace spolehlivosti může být formulována jako pravděpodobnost toho, že uvažovaná náhodná veličina označovaná jako X bude nabývat hodnot nižších, než jiná vzájemně působící náhodná veličina označovaná jako Y. Pro přenesení této problematiky do reálné situace uvažujeme náhodnou veličinu Y, která reprezentuje např. pevnost uvažované komponenty (subjektu) podléhající jiné náhodné veličině reprezentující na ní působící náhodnou sílu X (např. působení větru).

V celé této práci bude nadále předpokládána dvourozměrná spojitá náhodná veličina označovaná jako (X, Y), která je zároveň charakterizována sdruženou dvourozměrnou funkcí hustoty $f_{X,Y}(x, y)$ a distribuční funkcí $F_{X,Y}(x, y)$.

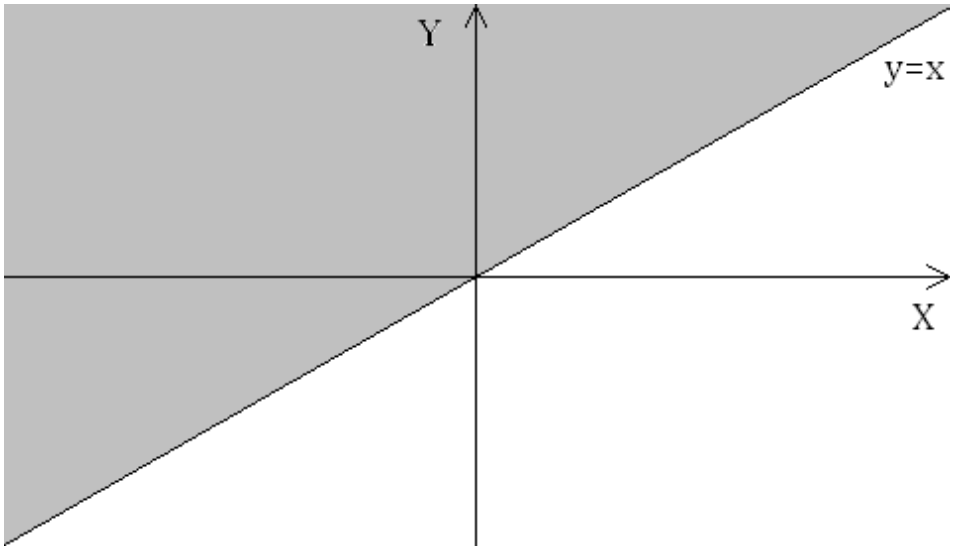
Nyní již lze definovat spolehlivost pomocí sdružené funkce hustoty pravděpodobnosti na základě výše uvedené formulace, kterou ve svých publikacích uvedli např. Nadarajah (2005), Hangal (1996) nebo Nadarajah a Kotz (2005).

Definice 2.1 *Nechť X a Y jsou spojité náhodné veličiny na celém definičním oboru se sdruženou funkcí hustoty $f_{X,Y}(x, y)$ a nechť $x \in \mathbb{R}$. Potom spolehlivost R je*

$$R = P(X < Y) = \int_{-\infty}^{+\infty} \int_{x}^{+\infty} f_{X,Y}(x, y) dy dx. \tag{2.1}$$

(převzato z Nadarajah a Kotz (2005), s. 2)

Opakem spolehlivosti je tzv. nespolehlivost, která je často definována pravděpodobností selhání uvažovaného systému (komponenty) a je označována jako možný rizikový faktor na stejném intervalu nebo prostoru tj. v jednorozměrném nebo dvojrozměrném případě. Grafické znázornění „spolehlivostního prostoru“ (oblasti vyhovující výše uvedeným požadavkům v definici) v uvažovaném dvourozměrném případě je ukázáno na níže uvedeném Obrázku 1 pomocí šedé barvy.



Obrázek 1: *Spolehlivostní prostor*

2.1.1 Různé typy spolehlivosti

Dostupná literatura obsahově zaměřená na teorii spolehlivosti se rozrostla v několika posledních letech. Tento fakt je samozřejmě spojen s rychlým vývojem výpočetních technologií a samotný výraz spolehlivost včetně různých odvození a postupů k následnému odhadu je často používáno v mnoha známých i vyvíjejících se oborech a specializacích. Navíc může být spolehlivost sama o sobě (a často se lze s tímto přístupem setkat) subjektivně kategorizována do několika významově odlišných skupin.

Možné skupiny jsou roztříděny a uvedeny např. v Nadarajah (2005), (2006) podle obsahu či použití do následujících skupin:

- *Ve financích: spolehlivost R představuje pravděpodobnost kladného hospodářského výsledku (zisku), kde náhodná veličina X reprezentuje celkové náklady a náhodná veličina Y reprezentuje celkové výnosy hospodařícího subjektu.*
- *V armádě: spolehlivost R může být interpretována jako pravděpodobnost, že předpokládaná dávka munice zasáhne svůj předem určený cíl.*
- *V lékařství: spolehlivost R reprezentuje pravděpodobnost, že zkoumaný lidský orgán (oko, ucho nebo jiný libovolný orgán) je schopný nadále vykonávat svou funkci.*
- *Ve stavebnictví: spolehlivost R představuje pevnost budov (mostů, …), kde náhodné veličiny X a Y jsou budoucí pozorování mající vliv na stabilitu celé konstrukce.*
- *V inženýrství: spolehlivost R charakterizuje reálné opotřebení navržených komponent nebo změny tlaku ve skladovacích komorách.*

(převzato z Nadarajah (2005) nebo (2006), s. 1-2)

Poznámka 2.1 *V prezentované práci je používán soubor reálných experimentálních dat z platební bilance České republiky pro testovací část. Spolehlivost je zde reprezentována pravděpodobností kladné obchodní bilance za celé uvažované období (tj. schopnost splácet závazky ze zdrojů daného období), respektive selhání je reprezentováno pravděpodobností záporné obchodní bilance za celé uvažované období (tj. neschopnost splácet závazky ze zdrojů daného období).*

2.1.2 Rizikové události, statistické pojetí

Pojem spolehlivost v sobě samozřejmě skrývá pojem rizikové události (obecně rizika), který je v praxi často používán bez kritického a exaktnějšího pohledu. Tento efekt mizí, když se dostaneme do oblasti kvantifikace (kolik, kdy, …) a srovnávání (větší nebo menší riziko). V odborné literatuře již existují pokusy o exaktním zavedení právě tohoto pojmu, které mají velký význam, ale doposud jsou na počátku svého vývoje. Existují i některé dopracované metodiky „měření rizika“ např. Introduction to RiskMetrics™ (1995), ty ale v kontextu dalších událostí (tzv. „finanční krize“) týkajících se majitelů práv k těmto metodikám působí poněkud rozpačitě, a proto je předmětem zájmu jeden specifický pojem a „riziková událost“ při porovnávání dvou náhodných procesů.

Riziková událost je tedy jev, který pokud nastane, lze ho charakterizovat následující trojicí údajů. Časem kdy nastal (měřeno od nějakého, vztažného, počátečního okamžiku), efektem (náklady rizikové události, které mohou být charakterizování jedním nebo více čísel) a pravděpodobností toho, že nebude detekována (časté v technických aplikacích, např. odhalitelnost materiálové vady).

Uvažovaným modelem rizikové události může být tzv. bilanční vztah, kdy jedna náhodná veličina převýší druhou. V technických aplikacích lze zmínit (výše uvedený) vztah, kdy zátěžová síla převýší pevnost materiálu a v ekonomických aplikacích lze uvažovat vztah, kdy např. výdaje převýší příjmy (tzv. likvidní default) nebo kdy náklady převýší výnosy o známou mez danou vlastním jměním, atd.

Vzhledem k charakteru použitých dat lze dále předpokládat dva náhodné procesy[13] , označované jako $x(t)$ a $y(t)$. Potom se za rizikovou událost bude považovat jev, kdy nastane nerovnost $x(t) > y(t)$. Nechť dále každý dílčí náhodný proces obsahuje jednak systematickou složku[14] („trend“, sezónnost, …) označovanou jako $X(t)$, $Y(t)$ a zároveň složku náhodnou s označením ε_{X_t}, ε_{Y_t}. Potom pravděpodobnost rizikové události, v pojetí této práce, lze vyjádřit jako

$$P(x(t) > y(t)). \tag{2.2}$$

[13] Náhodný nebo též stochastický proces si lze představit jako systém náhodných veličin $\{X_t : t \in T\}$ pro $t = 1, \dots, T$ na stejném pravděpodobnostním prostoru, blíže viz Cipra (2008).

[14] Přesné definice a diskuze k jednotlivým pojmům je uvedena v kapitole 4 “Využití modelů v aplikační sféře”.

Samotné propojení obou složek může mít některou z následujících povah:

- Aditivní dekompozice

$$x(t) = X(t) + \varepsilon_{X_t},$$
$$y(t) = Y(t) + \varepsilon_{Y_t},$$
$$P(x(t) > y(t)) = P\left(\varepsilon_{X_t} - \varepsilon_{Y_t} > Y(t) - X(t)\right). \tag{2.3}$$

- Multiplikativní dekompozice

$$x(t) = X(t)\left(1 + \varepsilon_{X_t}\right), \;\; X(t) > 0, \varepsilon_{X_t} > -1,$$
$$y(t) = Y(t)\left(1 + \varepsilon_{Y_t}\right), \;\; Y(t) > 0, \varepsilon_{Y_t} > -1,$$
$$P(x(t) > y(t)) = P\left(\frac{1 + \varepsilon_{X_t}}{1 + \varepsilon_{Y_t}} > \frac{Y(t)}{X(t)}\right) = P\left(ln\left(\frac{1 + \varepsilon_{X_t}}{1 + \varepsilon_{Y_t}}\right) > ln\left(\frac{Y(t)}{X(t)}\right)\right). \tag{2.4}$$

- Smíšené dekompozice, např.

$$x(t) = X(t)\left(1 + \varepsilon_{X_t}\right), \;\; X(t) > 0, \varepsilon_{X_t} > -1,$$
$$y(t) = Y(t) + \varepsilon_{Y_t},$$
$$P(x(t) > y(t)) = P\left(X(t)\varepsilon_{X_t} - \varepsilon_{Y_t} > Y(t) - X(t)\right). \tag{2.5}$$

Všechny zde uvedené modely (samozřejmě i jiné) lze převést do pravděpodobnostního vyjádření ve tvaru

$$P\left(\varepsilon_{Z_t} > Z(t)\right), \tag{2.6}$$

kde u aditivního modelu jsou jednotlivé proměnné vyjádřeny jako

$$\varepsilon_{Z_t} = \varepsilon_{X_t} - \varepsilon_{Y_t}, \qquad Z(t) = Y(t) - X(t). \tag{2.7}$$

U multiplikativního modelu

$$\varepsilon_{Z_t} = ln\left(\frac{1 + \varepsilon_{X_t}}{1 + \varepsilon_{Y_t}}\right), \qquad Z(t) = ln\left(\frac{Y(t)}{X(t)}\right). \tag{2.8}$$

A u uvedeného příkladu smíšeného modelu

$$\varepsilon_{Z_t} = X(t)\varepsilon_{X_t} - \varepsilon_{Y_t}, \qquad Z(t) = Y(t) - X(t). \tag{2.9}$$

Poznámka 2.2 *Zde uvedená „oddělení“ u náhodné a systematické složky nejsou jediné možné.*

Dále je pracováno s modelem pravděpodobnosti výskytu rizikové události (2.6) a vstupujícím časem τ prvního dosažení rizikové oblasti $\varepsilon_{Z_t} > Z(t)$ za předpokladu, že ve vztaženém počátečním čase t_0 platila nerovnost $\varepsilon_{Z_{t_0}} < Z(t_0)$. To lze chápat jako soustavu v nerizikovém stavu na počátku sledování (samozřejmě lze modelovat situaci analogicky opačným vztahem). Z uvedeného textu je zřejmé, že platí následující vztah

$$\left(\varepsilon_{Z_t} > Z(t)\right) \subset (\tau < t). \tag{2.10}$$

Uvedený vztah v sobě skrývá vazbu mezi časem do výskytu první rizikové události τ a chováním obou srovnávaných „komponent“, přesněji vazbu mezi jejich agregovanou náhodnou složkou ε_{Z_t} a agregovanou systematickou složkou $Z(t)$. Potom pravděpodobnost toho, že se náhodná složka bude nacházet v rizikové „zóně“ (tj. bude setrvávat nebo se bude opakovat riziková událost) pokud se tam alespoň jednou dostala, lze vyjádřit jako

$$P\left(\varepsilon_{Z_t} > Z(t) | \tau < t\right) = \frac{P\left(\varepsilon_{Z_t} > Z(t); \tau < t\right)}{P(\tau < t)} = \frac{P\left(\varepsilon_{Z_t} > Z(t)\right)}{P(\tau < t)}. \tag{2.11}$$

Dále jsou uvažovány jen takové soustavy, u nichž platí následující vztah (2.12). Tento přístup je motivován např. Rozanov (1979) a pojmem „náhodná procházka“.

Uvažujeme tedy jen soustavy ve tvaru

$$\frac{P\left(\varepsilon_{Z_t} > Z(t)\right)}{P(\tau < t)} = \alpha, \quad 0 < \alpha < 1. \tag{2.12}$$

Podmíněná pravděpodobnost α charakterizuje situaci, kdy náhodná složka ε_{Z_t} převýší složku systematickou $Z(t)$ (riziková „zóna“, tj. nastala nebo nastala opakovaně riziková událost), pokud se alespoň jednou předtím již realizovala riziková událost. Za podmíněnou pravděpodobnost bývá obvykle uváděna „učebnicová“ hodnota $\alpha = \frac{1}{2}$. V takovém případě se jedná o „symetrickou“ náhodnou složku kolem systematické složky a lze to charakterizovat tak, že je stejná hypotetická pravděpodobnost výskytu v rizikové a nerizikové „zóně“ za podmínky, že již bylo alespoň jednou do rizikové „zóny“ vstoupeno. Podmínku (2.12) splňuje poměrně široká skupina modelů náhodné složky ε_{Z_t} ve vztahu ke složce systematické (procesy s nezávislými přírůstky, jejich limitní verze, …), a proto se nejedná o významné omezení.

Ze vztahu (2.12) bezprostředně plyne rovnost

$$P(\tau < t) = \frac{1}{\alpha} P\left(\varepsilon_{Z_t} > Z(t)\right), \quad 0 < \alpha < 1. \tag{2.13}$$

Tento vztah propojuje za uvedených podmínek dvě častá pojetí modelování rizik, tj. času do selhání a vztahu (2.1).

Nyní pro další úpravy označíme distribuční funkce:

$$\begin{aligned} F_{\tau}(t) &= P(\tau < t), \\ F_{\varepsilon_{Z_t}}(t;\, x) &= P\left(\varepsilon_{Z_t} < x\right), \end{aligned} \tag{2.14}$$

kde $F_{\tau}(t)$ označuje distribuční funkci náhodné proměnné prvního času dosažení rizikové oblasti a $F_{\varepsilon_{Z_t}}(t;\, x)$ distribuční funkci náhodné proměnné ε_{Z_t} v čase t. Tj. přepsáním výrazu (2.13) je získán tvar

$$F_{\tau}(t) = \frac{1}{\alpha}\left(1 - F_{\varepsilon_{Z_t}}\left(t; Z(t)\right)\right), \quad 0 < \alpha < 1. \tag{2.15}$$

Duálně lze k tomu lze vytknout

$$F_{\varepsilon_{Z_t}}\left(t; Z(t)\right) = 1 - \alpha F_{\tau}(t), \quad 0 < \alpha < 1. \tag{2.16}$$

Nepřímým důsledkem vztahu (2.12) je skutečnost, že budeme pracovat s náhodnou složkou, jejíž pravděpodobnostní model je „translačně invariantní“, tj.

$$F_{\varepsilon_{Z_t} - \varepsilon_{Z_0}}(t; x) = F_{\varepsilon_{Z_t}}\left(t - t_0; x - \varepsilon_{Z_0}\right), \quad \varepsilon_{Z_0} < Z(t_0). \tag{2.17}$$

Stručně lze uvedenou skutečnost formulovat tak, že pro měření (zjišťování) jsou podstatné časové a hodnotové přírůstky. To znamená, že sledujeme a vyhodnocujeme běžící, nikoliv jednorázová dění.

Za daných výchozích podmínek jsou vztahy (2.15) a ekvivalentně (2.16) tuhou vazbou mezi časem dosažení „rizikové oblasti“ a popis vztahu náhodné složky ke složce systematické jednoznačně určuje čas prvního dosažení „rizikové oblasti“ a naopak.

Prezentovanou problematiku lze rozšířit tak, že vzájemně srovnávané procesy $x(t)$, $y(t)$ mohou mít povahu v čase aktuálních hodnot, kumulací od počátku pozorování (náhodné procházky, procesy s nezávislými přírůstky, jejich limitní verze, …) nebo kumulací za uvažované období. Tomu pak musí odpovídat uvedený model (2.14) a z modelované reality dále vyplyne, zda se jedná o (alespoň potenciálně) pokračující procesy (po dosažení rizikové zóny) nebo se jedná o procesy, které po dosažení rizikové zóny končí (úlohy o „stopping time“, úlohy obnovy,…) a následně dochází k opětovnému „restartu“ (oprava, náprava, …). V případě takových procesů se bude parametr α zjišťovat jinou metodikou (např. metodou maximální věrohodnosti), než u soustav, kde jsou pozorovatelné „pobyty“ v rizikové zóně a mimo ni.

Vztah (2.12) určuje, že nenáhodná (systematická) složka $Y(t) - X(t)$ musí být odhadována (zde je běžný a známý problém jak oddělit systematickou a náhodnou složku) jako kvantilová čára (přesněji dvě kvantilové čáry pro oba procesy $x(t)$ a $y(t)$) z pozorování po prvním dosažení rizikové oblasti (nebo alespoň potenciálně). Jedna z možných metodik pro takové postupy je uvedena v následujících kapitolách této práce.

Poznámka 2.3 *Obecně je v navrhovaném modelu rizikové události* $x(t) > y(t)$ *zahrnuta i situace „s posunutím“ tj.* $x(t) > y(t) + c$*, kde* c *je nějaká konstanta. V konkrétním případě je však zapotřebí respektovat její specifika. Tento případ je poměrně častý a lze se s ním setkat např. v situaci, kde* $x(t)$ *jsou náklady,* $y(t)$ *jsou výnosy a* c *jsou disponibilní, dodatečné, zdroje (vlastní kapitál, …).*

2.2 Některá klasická pravděpodobnostní pojetí

Bude uveden přehled „základních“ přístupů pro tvorbu již existujících modelů v oblasti dvourozměrných pravděpodobnostních modelů k získání neznámé hodnoty spolehlivosti dle (2.1). Z velkého počtu dostupných modelů jsou zde uvedeny jen dva konkrétní modely jako příklady. Důkazy a detailněji popsaná problematika včetně navazujících teoretických oblastí může být nalezena ve velkém množství dosud publikovaných článků a knih obsahující rozbory a postřehy od různých autorů, např. Nadarajah (2005), Hangal (1996), Nadarajah a Kotz (2005) nebo jen Kotz a kol. (2000) atd.

Nejprve jsou zde stručně uvedeny dvourozměrné modely, které jsou založeny na publikovaných[15] dvourozměrných pravděpodobnostních rozděleních a poté „více“ propracovanější a náročnější modely, které již implicitně nebo explicitně používají „teorii copul[16]“. Celou problematikou se již detailně zabývali zejména Kotz a kol. (2000), kteří diskutovali několik možných postupů týkajících se právě konstrukce různých typů modelů. Cílem kapitoly je informace o složitosti a s tím spojených problémů vzniklých s používáním právě uvedených modelů.

2.2.1 Některé příklady existujících modelů

Obecně lze říci, že výsledné modely mohou být získány pomocí tří přístupů, jak uvedl Johnson a kol. (1999). První přístup konstrukce dvourozměrných modelů spočívá v transformaci souboru získaných (naměřených) pozorování na známé a již publikované sdružené dvourozměrné rozdělení pravděpodobnosti např. Normální, Gama, Exponenciální a Paretovo rozdělení pravděpodobnosti viz Kotz a kol. (2000), Gupta a Subramanian (1998), Nadarajah (2005), Hangal (1996), Nadarajah a Kotz (2005) atd. Druhý přístup spočívá v dále uvedeném modelovém vztahu pro tvorbu sdružené dvourozměrné funkce hustoty a distribuční funkce (oddělení marginálních rozdělení a vzájemné vazby náhodných veličin). Základní vztah pro tvorbu požadovaného modelu, konkrétně sdružené distribuční funkce, lze formulovat ve tvaru

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)[1 + \delta\{1 - F_X(x)\}\{1 - F_Y(y)\}], \tag{2.18}$$

[15] Pod tímto výrazem jsou zahrnuty pravděpodobnostní modely, které byly doposud publikovány v přístupné literatuře.
[16] Podrobnější popis funkce je uveden dále.

kde $F_X(x), F_Y(y)$ jsou marginální distribuční funkce (zde je u obou předpokládáno stejné pravděpodobnostní rozdělení) a parametr $|\delta| \leq 1$ značí "korelaci[17]" mezi oběma komponentami. Použitá komponenta $C = xy\big(1 + \delta(1-x)(1-y)\big)$ se nazývá copula[18] (jedna z možných) a odráží model závislosti mezi jednorozměrnými komponentami (pravděpodobnostními rozděleními, náhodnými veličinami). Samozřejmě to není jediný možný přístup.

Poslední přístup tvorby sdružených modelů spočívá v kombinaci předchozích dvou, blíže viz Johnson a kol. (1999). Podrobný popis prezentované problematiky není cílem práce, a proto zde již nebude nadále detailněji popisován a rozebírán.

Dále lze stručně zmínit funkce copul, které charakterizují možný přístup pro tvorbu sdruženého rozdělení pravděpodobnosti. Tímto způsobem lze obecně reprezentovat různé typy závislosti. Vzhledem k frekventovanému používání těchto funkcí (jak distribučních, tak i hustot) je vhodné formulovat základní definici publikovanou např. v Skiadas (2007) nebo Nelsen (2006).

Definice 2.2 *Distribuční funkce copul $C(u, v)$ je vícerozměrná distribuční funkce definovaná na n-dimenzionální jednotkové krychli $[0,1]^n$ taková, že každé marginální rozdělení je rovnoměrné na intervalu $[0,1]$, shrnuto:*

$$C\colon [0,1]^n \to [0,1]\colon (u_1, \dots, u_n) \mapsto C(u_1, \dots, u_n). \tag{2.19}$$

(převzato ze Skiadas (2007), s. 4)

Poznámka 2.4 *V uvedené problematice je vhodné rozlišovat mezi dvěma funkcemi copul. Funkce označované jako $C(u, v)$ jsou funkce copul pro sdruženou distribuční funkci a funkce označované jako $c(u, v)$ jsou funkce copul pro sdruženou funkci hustoty.*

2.2.2 První příklad

Subjektivním výběrem možného modelu, který lze přiřadit do třídy[19] „jednodušších" dvourozměrných modelů pro odhad spolehlivosti, byl vybrán model sdruženého dvourozměrného normálního rozdělení pravděpodobnosti. Výhody daného rozdělení jsou založeny na empirickém faktu o intenzivním používání v různých vědeckých oblastech. Uvedené pravděpodobnostní rozdělení je velmi zajímavé i ze statistického úhlu pohledu. Představuje přijatelný model pro praktické použití a zároveň reprezentuje sdružené rozdělení pravděpodobnosti pro dvě (nebo více) náhodné veličiny, které

[17] Korelace zde (např. i klasický korelační koeficient) reprezentuje určitou míru závislosti ze široké třídy různě definovaných vztahů týkající se závislostí, přesněji míru pravděpodobnostní závislosti dvou náhodných veličin X a Y.

[18] Jedna z možných knih, kde lze nalézt detailní interpretaci pojmu "copula", je např. Nelsen (2006).

[19] Třídou je myšleno možné (nikoliv standardizované) rozdělení vybraných modelů podle tvaru a náročnosti odhadů jednotlivých parametrů.

v našem případě mohou představovat např. navzájem působící síly (tzv. "stress-strength" model).

Nechť tedy pro dvourozměrnou náhodnou veličinu (X, Y) předpokládáme sdružený model dvourozměrného normálního rozdělení, který tvoří dvě související, normálně rozdělené spojité náhodné veličiny (marginály) označované jako $X \sim N(\mu_X, \sigma_Y^2)$ a $Y \sim N(\mu_Y, \sigma_Y^2)$. Potom vztah pro sdruženou funkci hustoty lze popsat jako

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y(1-\rho^2)} e^{\left(-\frac{z}{2(1-\rho^2)}\right)}, \tag{2.20}$$

kde π, e značí (běžně známé a zavedené) konstanty popsané v seznamu použitého značení v úvodu práce, σ_X a σ_Y směrodatné odchylky[20] příslušných náhodných veličin a proměnná z vyjadřuje výraz

$$z = \left(\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right), \tag{2.21}$$

kde μ_X, μ_Y značí střední hodnoty[21] příslušných náhodných proměnných a koeficient

$$\rho = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X\sigma_Y} \tag{2.22}$$

je koeficient korelace (Pearsonův korelační koeficient „lineární závislosti") mezi náhodnými veličinami X a Y (Kenney a Keeping (1951), s. 92 a 202–205 nebo Kotz, Balakrishnan a Johnson (2000), s. 251-252) a $cov(X,Y)$ označuje kovarianci[22] mezi dvěma náhodnými veličinami, jak je uvedeno např. v Hátle a Likeš (1974).

Jednotlivé marginální hustoty jsou dány pomocí jednorozměrných funkcí hustot normálního rozdělení, přesněji

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}} \tag{2.23}$$

pro obě náhodné veličiny X, Y a sdružená dvourozměrná distribuční funkce je popsána s využitím hustoty $f_{X,Y}(x,y)$ jako

$$F_{X,Y}(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} \frac{1}{2\pi\sigma_X\sigma_Y(1-\rho^2)} e^{\left(-\frac{z}{2(1-\rho^2)}\right)} dx\, dy. \tag{2.24}$$

[20] Směrodatná odchylka neboli druhá odmocnina z rozptylu, více Hátle a Likeš(1974).

[21] Střední (očekávaná) hodnota náhodné veličiny X charakterizuje jistou charakteristiku polohy tzv. první obecný moment, více Hátle a Likeš (1974).

[22] Kovariance značí jistou míru vazby mezi dvěma náhodnými veličinami, přesněji střední hodnotu součinu odchylek obou náhodných veličin od jejich středních hodnot.

Lze použít pro reprezentaci spolehlivosti a obvykle jsou hodnoty vypočítávány numericky. Důležitá vlastnost použitého pravděpodobnostního rozdělení spočívá v „náročnosti odhadu koeficientů“.

2.2.3 Další příklad

Příklad „náročnějšího“ modelu a odhady jeho parametrů jsou zde prezentovány modelem s názvem „Hougaardovo sdružené exponenciální rozdělení“. Výběr byl proveden opět ze subjektivního úhlu pohledu, aby názorně demonstroval potenciální problémy s výběrem „složitějšího“ modelu. Mezi často uváděné modely lze dále zařadit např. sdružené dvourozměrné Paretovo, Gama nebo Exponenciální rozdělení a jejich různé modifikace. Problém spočívá nejprve v určení „vhodného“ modelu pro získání spolehlivosti ze souboru reálných pozorování a následuje odhadem použitých parametrů (a samozřejmé testování přijatelnosti modelu). Uvedené skutečnosti jsou hlavní příčinou pro výzkum neparametrických jádrových odhadů spolehlivosti.

Nechť X a Y jsou spojité náhodné veličiny a nechť sdružená náhodná veličina (X, Y) se řídí upraveným exponenciálním rozdělením pravděpodobnosti s tzv. „funkcí přežití[23]“ $\bar{F}$, kterou lze popsat vztahem

$$\bar{F}_{X,Y}(x,y) = P(X > x, Y > y) = e^{\left(-\left\{\left(\frac{x}{\theta}\right)^r + \left(\frac{y}{\varphi}\right)^r\right\}^{\frac{1}{r}}\right)}. \tag{2.25}$$

Potom sdružená funkce hustoty s parametry θ, φ a upraveným koeficientem korelace je

$$f_{X,Y}(x,y) = \frac{(xy)^{r-1}}{(\theta\varphi)^r}\left\{\left(\frac{x}{\theta}\right)^r + \left(\frac{y}{\varphi}\right)^r\right\}^{\frac{1}{r}-2}\left[r - 1 + \left\{\left(\frac{x}{\theta}\right)^r + \left(\frac{y}{\varphi}\right)^r\right\}^{\frac{1}{r}}\right]\bar{F}_{X,Y}(x,y), \tag{2.26}$$

pro $x, y, r > 0$ a $\theta, \varphi \geq 0$. Marginální funkce náhodných proměnných X a Y se řídí exponenciálním rozdělením pravděpodobnosti s parametry $\frac{1}{\theta}, \frac{1}{\varphi}$. Pro přehlednost lze provést úpravu, která spočívá v transformaci náhodných veličin $(U, V) = \left(\frac{X}{\theta}, \frac{Y}{\varphi}\right)$. Výsledkem je modifikovaná funkce $\bar{F}$ pro transformované náhodné veličiny (U, V)

$$\bar{F}_{U,V}(u,v) \quad = \quad e^{-(u^r + v^r)^{\frac{1}{r}}} \tag{2.27}$$

a sdružená funkce hustoty pro dvourozměrnou náhodnou veličinu (U, V)

$$f_{U,V}(u,v) = (uv)^{r-1}(u^r + v^r)^{\frac{1}{r}-2}\left[r - 1 + (u^r + v^r)^{\frac{1}{r}}\right]\bar{F}_{U,V}(u,v). \tag{2.28}$$

Výslednou spolehlivost lze získat jako

[23] Předklad z anglického “survival function“.

$$R = P\left(\frac{\theta U}{\varphi} < V\right)$$
$$= \int_0^{\infty} u^{r-1} \int_{\frac{\theta u}{\varphi}}^{\infty} v^{r-1}(u^r+v^r)^{\frac{1}{r}-2}\left\{r-1+(u^r+v^r)^{\frac{1}{r}}\right\}\bar{F}_{U,V}(u,v)dvdu\,. \quad (2.29)$$

Ucelený postup včetně podrobného popisu a odvození je publikován v Nadarajah a Kotz (2005, s. 202 – 205) nebo Hougaard (1986, s. 671 - 678).

2.3 Neparametrický model zvoleného pojetí spolehlivosti

V návaznosti na předchozí kapitoly je důležité získat použitelný vztah mezi uvažovanými náhodnými veličinami, který ve své podstatě charakterizuje vztah zatím blíže nespecifikovaných veličin (např. vstupující fyzikální síly, …) do uvažovaného modelu dále rozvinutého pro získání spolehlivosti.

2.3.1 Odvození základních vztahů

Nechť $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ je náhodný výběr (i.i.d.) získaných párovaných pozorování dvou náhodných veličin X, Y rozsahu n, které se řídí sdruženou funkcí hustoty a distribuční funkcí, označované jako $f_{X,Y}(x, y)$ a $F_{X,Y}(x, y)$, a splňující požadované předpoklady[24]. Předmětem zájmu je hodnota pravděpodobnosti, kterou lze popsat pomocí vztahu $P(X < Y)$. Formální definice tohoto výrazu je detailněji vysvětlena v úvodu této kapitoly a popisovaná událost selhání (spolehlivost) může být chápána jako selhání systému, deformace (povrchová, vratná či nevratná…) materiálu, pravděpodobnost insolvence ve finančním sektoru a další. Vzhledem k charakteru použitých dat je zde důležitý zejména pojem insolvence[25] (nebo též bankrot[26]) z finančního sektoru, který bude využíván v dalších částech.

Insolvence (přesněji pravděpodobnost insolvence) sama o sobě může být popsána pravděpodobností $P(X < Y)$, kde náhodná veličina X označuje (celkové, kumulované, …) příjmy a náhodná veličina Y označuje model výdajů vybrané společnosti nebo státu. Naopak doplněk této pravděpodobnosti, který lze popsat vztahem $1 - P(X < Y)$, může být prezentován jako pravděpodobnost přijatelné finanční situace vybrané společnosti, běžným provozním stavem atd.

Po úvodním seznámení bude uveden jeden možný přístup, který je založen na odvození vztahu mezi zvolenými náhodnými veličinami řídící se sdruženým dvourozměrným normálním rozdělením pravděpodobnosti.

[24] Zde spojité marginální a sdružená distribuční funkce na celém definičním oboru, tj. tam, kde $f_{X,Y}(x, y) > 0$.
[25] Insolvence neboli platební neschopnost je neschopnost daného subjektu dostát svým závazkům. I zde může mít platební neschopnost další stupně.
[26] Ne v právním slova smyslu, který vyžaduje delší dobu uvažovaného stavu.

Věta 2.1 *Nechť X a Y jsou dvě spojité náhodné veličiny řídící se sdruženým dvourozměrným normálním rozdělením pravděpodobnosti (tedy dvourozměrná náhodná veličina) pak*

$$\begin{aligned} Cov(X,Y) &= \rho\sigma_X\sigma_Y, \\ E\{X-Y\} &= \mu_X - \mu_Y, \\ \sigma^2\{X-Y\} &= \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y, \\ X-Y &\sim Normálním\ rozdělením\ pravděpodobnosti, \end{aligned} \tag{2.30}$$

kde μ_X a μ_Y jsou střední hodnoty náhodných veličin X a Y, σ_X^2 a σ_Y^2 jsou jejich rozptyly[27] *a ρ značí koeficient korelace mezi oběma veličinami.*

Důkaz. Triviální.

Předpoklad sdruženého normálního rozdělení pravděpodobnosti je nezbytný pro získání normality u rozdělení náhodné veličiny vzniklé rozdílem jednotlivých náhodných veličin $(X-Y)$. V tomto případě nestačí mít pouze předpoklad normálního rozdělení marginálních náhodných veličin, ale je zde nezbytné právě sdružené normální rozdělení.

Pokud se tedy uvažované náhodné veličiny X a Y řídí sdruženým dvourozměrným normálním rozdělením pravděpodobnosti, potom nově vzniklá náhodná veličina rozdílu $(X-Y)$ se také řídí normálním rozdělením pravděpodobnosti. Důkaz tohoto lze nalézt v Rényi (1972).

Věta 2.2 *Nechť X a Y jsou spojité náhodné veličiny řídící se sdruženým dvourozměrným normálním rozdělením pravděpodobnosti, potom distribuční funkci náhodné veličiny vzniklé z rozdílu obou náhodných veličin $(X-Y)$ lze popsat jako*

$$F_{X-Y}(x) = \Phi\left(\frac{x-(\mu_X-\mu_Y)}{\sqrt{\sigma_X^2+\sigma_Y^2-2\rho\sigma_X\sigma_Y}}\right), \tag{2.31}$$

kde Φ označuje distribuční funkci normovaného[28] *normálního rozdělení pravděpodobnosti $N(0,1)$.*

Důkaz. Podstata důkazu je popsána v následujícím postupu.

[27] Rozptyl (druhý centrální moment) náhodné veličiny ve statistice charakterizuje variabilitu rozdělení kolem jeho střední hodnoty, přesněji střední hodnota kvadrátů odchylek od střední hodnoty náhodné veličiny, více Hátle a Likeš(1974).
[28] Více např. Hátle a Likeš (1974).

$$
\begin{aligned}
F_{X-Y}(x) &= P(X-Y<x) \\
&= P((X-Y)-(\mu_X-\mu_Y)<x-(\mu_X-\mu_Y)) \\
&= P\left(\frac{(X-Y)-(\mu_x-\mu_y)}{\sqrt{\sigma_X^2+\sigma_Y^2-2\rho\sigma_X\sigma_Y}}<\frac{x-(\mu_x-\mu_y)}{\sqrt{\sigma_X^2+\sigma_Y^2-2\rho\sigma_X\sigma_Y}}\right).
\end{aligned}
\tag{2.32}
$$

Nyní je získána „nová“ náhodná veličina rozdílu, která se řídí opět normálním rozdělením pravděpodobnosti. Normovaná náhodná veličina je vyjádřena vztahem

$$
\frac{(X-Y)-(\mu_X-\mu_Y)}{\sqrt{\sigma_X^2+\sigma_Y^2-2\rho\sigma_X\sigma_Y}}. \tag{2.33}
$$

Důsledek 2.1 *Nechť X a Y jsou spojité náhodné veličiny, které se řídí sdruženým dvourozměrným normálním rozdělením pravděpodobnosti se sdruženou funkcí hustoty a distribuční funkcí, potom vztah pro výpočet (pravděpodobnosti) spolehlivosti, definované jako $P(X<Y)$ z uvažovaných náhodných veličin, může být přehledně upraven jako pravděpodobnost rozdílu náhodných veličin v nulové hodnotě tj. $P(X-Y<0)$ a vyjádřen jako*

$$
\begin{aligned}
R &= F_{X-Y}(0) \\
&= P(X-Y<0) \\
&= P(X<Y) \\
&= \Phi\left(\frac{\mu_Y-\mu_X}{\sqrt{\sigma_X^2+\sigma_Y^2-2\rho\sigma_X\sigma_Y}}\right).
\end{aligned}
\tag{2.34}
$$

Navíc v případě dvou spojitých a nezávislých náhodných veličin, které se řídí normálním rozdělením pravděpodobnosti, lze vztah pro výpočet spolehlivosti vyjádřit jako v níže uvedeném důsledku, který popisuje jen formální zjednodušení předchozího vztahu.

Důsledek 2.2 *Nechť X a Y jsou spojité a **nezávislé** náhodné veličiny, které se řídí sdruženým dvourozměrným normálním rozdělením pravděpodobnosti, potom vztah pro modelovanou spolehlivost $P(X<Y)$ může být vyjádřen*

$$
\begin{aligned}
R &= P(X<Y) \\
&= \Phi\left(\frac{\mu_Y-\mu_X}{\sqrt{\sigma_X^2+\sigma_Y^2}}\right).
\end{aligned}
\tag{2.35}
$$

Zde je nezbytné upozornit na skutečnost, že všechny výše formulované závěry jsou platné pro případ, kdy jsou splněny výše popsané předpoklady. Tedy, praktické a přímé využití těchto vztahů je velmi omezené. Budou však využity jako dílčí prostředek v dalším textu.

Kapitola 3

Neparametrické jádrové odhady

V kapitole jsou nejprve podrobně diskutovány a popsány neparametrické jádrové odhady (aproximace) hustot i distribučních funkcí (jednorozměrné i dvourozměrné) s použitím různých typů jádrových[29] funkcí. Následně je věnován prostor pro popis problematiky vyhlazovacího parametru jak u hustoty, tak i u distribuční funkce a jádrovému odhadu distribuční funkce součtů náhodných veličin. Závěr kapitoly popisuje využití neparametrických modelů za účelem získání spolehlivosti.

3.1 Neparametrický odhad „jednorozměrné“ hustoty, vlastnosti

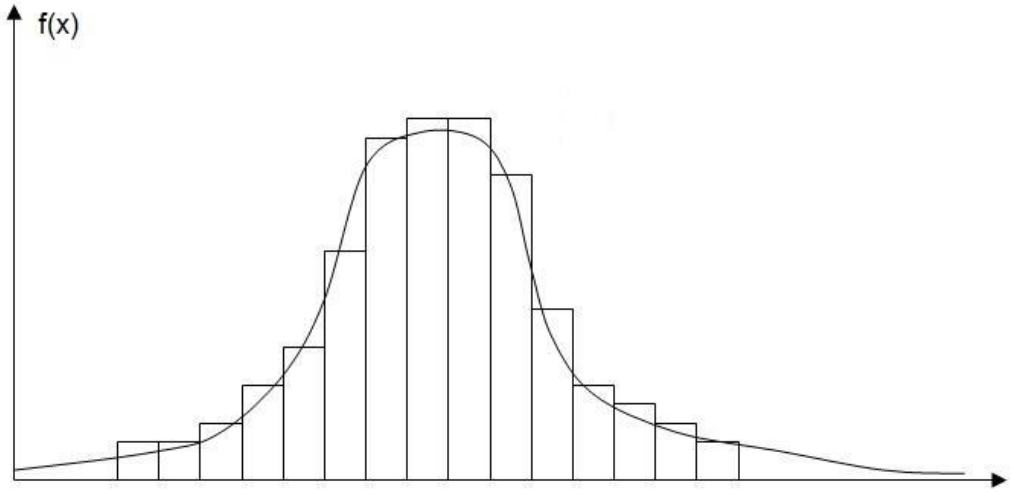
Vzhledem k zamýšlené tvorbě jednorozměrných a dvourozměrných modelů se hlavní pozornost zaměří nejprve na jednorozměrné neparametrické jádrové odhady hustoty. V této problematice je již publikováno velké množství literatury týkající se neparametrického jádrového odhadu hustoty pro soubory nezávislých a stejně rozdělených pozorování náhodné veličiny pevného rozsahu n, které se řídí některým popsaným modelem pravděpodobnostního rozdělení (např. normálním). Cílem je aproximovat „vhodným tvarem“ neznámé funkce hustoty označované jako $f(x)$ pomocí statistického přístupu, který je ve statistice znám pod názvem „neparametrické jádrové odhady“.

Definice neparametrických jádrových odhadů neznámé hustoty bude zaměřena na modely pro jádrové odhady hustoty, které nejdříve představili Rosenblatt (1956) a Parzen (1962). Obecně lze říci, že neparametrické jádrové odhady hustot a distribučních funkcí zobecňují odhady pomocí tzv. histogramů, které jsou charakterizovány „schodovitou funkcí“ a patří mezi jednodušší a často používané metody k prvotnímu náhledu na pravděpodobnostní rozdělení, s použitím alternativních jádrových funkcí. Tento přístup dokládá následující model tvorby histogramu, který je pro spojitou náhodnou veličinu X ve tvaru

$$\hat{f}(x;h) = \frac{n(x)}{nh}, \qquad h > 0, \quad n \in \mathbb{N}_+, \quad x \in \mathbb{R}, \tag{3.1}$$

[29] Jádrové funkce představují třídu funkcí s jistými požadovanými vlastnostmi, které budou popsány dále.

kde parametr h je prozatím zvolená šířka dělícího intervalu (parametr měřítka detailně popsaný dále), n je počet pozorování (realizace náhodného výběru pevného rozsahu) a $n(x)$ je počet pozorování, které padnou do zvoleného intervalu obsahujícího bod x.



Obrázek 2: *Relativní histogram*

Takto získaný model pro tvorbu histogramu lze dále upravit s použitím tzv. jádrových funkcí, jak je popsáno v následující definici.

Definice 3.1 Nechť $\{x_1, \dots, x_n\}$ *je i.i.d. výběr náhodné veličiny* X *pevného rozsahu* n *a nechť* $x \in \mathbb{R}$*, potom neparametrické jádrové odhady neznámé hustoty jsou definovány vztahem*

$$\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right), \qquad h > 0 \qquad n \in \mathbb{N}_+, \tag{3.2}$$

kde h *označuje parametr měřítka (často označován v anglické literatuře jako tzv. scale nebo bandwidth parameter), který také bývá často označován jako vyhlazovací parametr.*

Odhad a vliv této komponenty je diskutován v následujících kapitolách, kde jsou uvedeny dva možné (nikoliv jediné) přístupy vedoucí k jeho „možnému“ odhadu. Pokud je tento parametr na něčem funkčně závislý, pak pouze na počtu pozorování n nebo někdy i na pozorovaných hodnotách x_i. Dále se zde vyskytuje kladná funkce označovaná jako $k(x)$, která představuje jádrovou funkci hustoty v oboru reálných čísel, u které se předpokládá, že je integrovatelná a zároveň splňuje následující podmínky:

$$\begin{aligned} \int_{-\infty}^{+\infty} k(x)dx &= 1 \\ \int_{-\infty}^{+\infty} xk(x)dx &= 0 \\ \int_{-\infty}^{+\infty} x^2k(x)dx &= 1 \end{aligned} \tag{3.3}$$

a $k(x) \geq 0$ pro $\forall x \in \mathbb{R}$.

Poznámka 3.1 *V obecné teorii neparametrických jádrových odhadů není kladnost jádrové funkce nezbytná a dokonce jsou někdy v této teorii uvedeny i další požadavky, mezi kterými lze nalézt např.* $\int_{-\infty}^{+\infty} x^2k(x)dx \neq 1$ *společně s nulovou střední hodnotou.*

Z výše uvedeného je evidentní, že jsou zde požadovány tři triviální podmínky. První podmínka znamená, že vybraná jádrová funkce je funkcí hustoty. Druhá podmínka říká, že jádrová funkce je centrovaná kolem střední hodnoty a ve třetí podmínce je zahrnut předpoklad, že jsou tyto jádrové funkce normovány s jednotkovým rozptylem (samozřejmě za platnosti druhé podmínky).

3.1.1 Základní vlastnosti neparametrického jádrového odhadu hustoty

Vybraný neparametrický jádrový model k aproximaci neznámé funkce hustoty lze nyní podrobit detailnější analýze základních vlastností. Pozornost se nejprve zaměří na aproximovanou funkci hustoty, zda se opravdu jedná o funkci hustoty neznámého pravděpodobnostního rozdělení či nikoliv.

Věta 3.1 *Neparametrický jádrový odhad* $\hat{f}(x;h)$ *je nějakou hustotou.*

Důkaz. Podstata důkazu plyne z postupu níže.

$$\begin{aligned} \int_{-\infty}^{+\infty} \hat{f}(x;h)dx &= \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} k\left(\frac{x-x_i}{h}\right)dx \\ &=_a \frac{1}{nh}\sum_{i=1}^{n} h\int_{-\infty}^{+\infty} k(z)\,dz \\ &= \frac{1}{nh}\sum_{i=1}^{n} h = 1. \end{aligned} \tag{3.4}$$

Použitím symbolu $=_a$ při výpočtu je označeno místo, kde je použita substituce $z = \frac{x-x_i}{h}$. Uvedená substituce je dále používána i v následujících výpočtech rovnic (3.5) a (3.7).

Věta 3.2 *Odhad střední hodnoty na základě získané aproximace funkce hustoty $\hat{f}(x;h)$ je roven hodnotě aritmetického průměru ze souboru získaných pozorování.*

Důkaz. Podstata důkazu je založena na uvedeném postupu.

$$\begin{aligned} E_{\hat{f}(x;h)}\{X\} &= \int_{-\infty}^{+\infty} x\hat{f}(x;h)dx \\ &= \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} xk\left(\frac{x-x_i}{h}\right)dx \\ &=_a \frac{1}{nh}\sum_{i=1}^{n} h\int_{-\infty}^{+\infty}(hz+x_i)k(z)\,dz \\ &= \frac{1}{n}\sum_{i=1}^{n} h\int_{-\infty}^{+\infty} zk(z)dz + \frac{1}{n}\sum_{i=1}^{n} x_i\int_{-\infty}^{+\infty} k(z)dz \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i\,. \end{aligned} \tag{3.5}$$

Poznámka 3.2 *Odhad střední hodnoty náhodné veličiny X, který je získán s použitím aproximované funkce hustoty $\hat{f}(x;h)$, a který lze zapsat s použitím notace $E_{\hat{f}(x;h)}\{X\}$, bude pro přehlednost a zjednodušení v následujících odvození dále označován jako μ_a .*

Důsledek 3.1 *Hodnota rozptylu $\sigma^2_{\hat{f}(x;h)}\{X\}$ náhodné veličiny X, který je získán s použitím aproximované funkce hustoty $\hat{f}(x;h)$ je roven*

$$\sigma^2_{\hat{f}(x;h)}\{X\} = h^2 + \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_a)^2. \tag{3.6}$$

Důkaz. Podstata důkazu plyne z uvedeného postupu.

$$\begin{aligned} \sigma^2_{\hat{f}(x;h)}\{X\} &= \int_{-\infty}^{+\infty}(x-\mu_a)^2\hat{f}(x;h)dx \\ &= \frac{1}{nh}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}(x-\mu_a)^2 k\left(\frac{x-x_i}{h}\right)dx \\ &=_a \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}(hz+x_i-\mu_a)^2 k(z)\,dz \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\int_{-\infty}^{+\infty}(hz)^2k(z)dz + \int_{-\infty}^{+\infty}(x_i-\mu_a)^2k(z)dz + \int_{-\infty}^{+\infty}2hz(x_i-\mu_a)k(z)dz\right] \\ &= h^2 + \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_a)^2\,. \end{aligned} \tag{3.7}$$

Zde je důležité upozornit na skutečnost, že k získání výsledné hodnoty rozptylu jsou využívány všechny dříve uvedené předpoklady (3.3).

Po získání odhadu střední hodnoty a rozptylu se lze dále zaměřit i na další významné statistiky. Přesněji řečeno, zda je vybraný model pro aproximaci neznámé funkce hustoty $\hat{f}(x;h)$ asymptoticky nestranný a vydatný? Odpověď je pozitivní. To je uvedeno a dokázáno v příloze této práce.

3.2 Vybrané typy jádrových funkcí

Obecně lze jádrovou funkci považovat za funkci centrovanou kolem nuly s hodnotou integrálu rovné jedné. V této části práce je představeno několik různých (možných nikoliv jediných) typů jádrových funkcí, u kterých jsou uvedené konstanty $a, b \in \mathbb{R}$ voleny tak, aby byly splněny výše popsané podmínky. Mezi některé používané a „běžné“ jádrové funkce mohou patřit např.:

1. Parzenova (obdélníková) jádrová funkce $k(x)$

$$\begin{aligned} k(x) &= \frac{1}{2a} \Leftrightarrow -a \leq x \leq a, \quad a > 0, \\ &= 0, \quad jinde. \end{aligned} \tag{3.8}$$

Ve všech uvedených jádrových funkcí vycházíme z podmínek (3.3). V tomto případě je první podmínka zřejmá. Druhá podmínka je též zřejmá v souvislosti se symetrií a ze třetí podmínky lze odvodit hodnotu neznámé konstanty a nálsedovně:

$$\frac{1}{2a}\int_{-a}^{+a} x^2\,dx = \frac{a^2}{3} = 1 \Rightarrow a = \sqrt{3}. \tag{3.9}$$

2. Epanechnikova jádrová funkce $k(x)$

$$\begin{aligned} k(x) &= b\left(1 - \left(\frac{x}{a}\right)^2\right) \Leftrightarrow -a \leq x \leq a, \quad a, b > 0, \\ &= 0, \quad jinde. \end{aligned} \tag{3.10}$$

Vyjádření neznámé konstanty b může být odvozeno jako

$$b\int_{-a}^{+a}\left(1 - \left(\frac{x}{a}\right)^2\right)dx = \frac{4}{3}ab = 1 \Rightarrow b = \frac{3}{4a}. \tag{3.11}$$

Druhá podmínka je též zřejmá v souvislosti se symetrií a ze třetí podmínky je možné získat hodnoty obou neznámých parametrů a, b:

$$b\int_{-a}^{+a}\left(x^2-\frac{x^4}{a^2}\right)dx=\frac{4a^3b}{15}=1\Rightarrow a=\sqrt{5},\ b=\frac{3}{4\sqrt{5}}. \tag{3.12}$$

3. Trojúhelníková jádrová funkce

$$\begin{aligned} k(x)&=b\left(1-\left|\frac{x}{a}\right|\right)\Leftrightarrow -a\le x\le a,\quad a,b>0,\\ &=0,\quad jinde. \end{aligned} \tag{3.13}$$

Hodnota neznámé konstanty b:

$$2\int_{0}^{+a} b\left(1-\left(\frac{x}{a}\right)\right)dx=ab=1\Rightarrow b=\frac{1}{a}. \tag{3.14}$$

Druhá podmínka je též zřejmá v souvislosti se symetrií a ze třetí podmínky je možné získat hodnoty obou neznámých parametrů a, b:

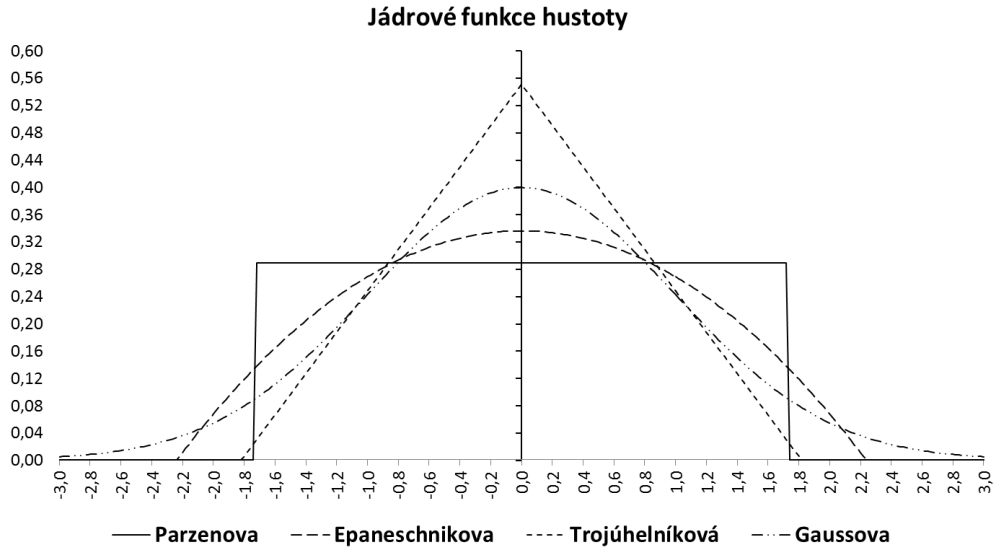
$$2b\int_{0}^{+a} x^2\left(1-\left(\frac{x}{a}\right)\right)dx=\frac{a^3b}{6}=1\Rightarrow a=\sqrt[2]{6}, b=\frac{1}{\sqrt[2]{6}}. \tag{3.15}$$

4. Gaussova jádrová funkce

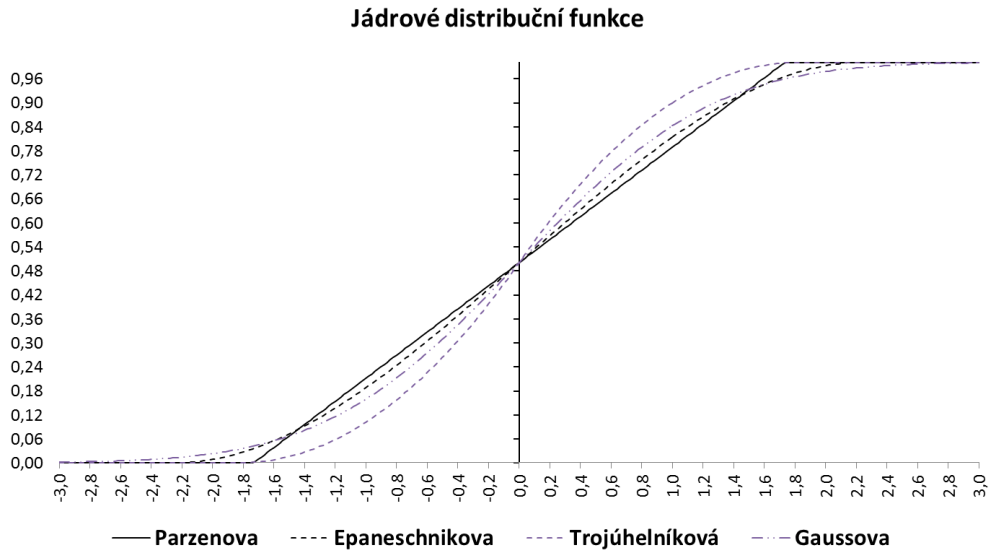
$$k(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}. \tag{3.16}$$

Všechny tři podmínky jsou zřejmé. Funkce je v literatuře často označována jako „běžná“ jádrová funkce, a proto i v této práci dochází k jejímu častému použití.

Grafické interpretace tvarů hustot i distribučních funkcí u popsaných typů jádrových funkcí jsou prezentovány na následujícím Obrázku 3 a 4 pro možné vzájemné porovnání, kde odlišné tvary korespondují s jejich názvy.



Obrázek 3: *Vybrané typy jádrových funkcí hustoty* $k(x)$



Obrázek 4: *Vybrané typy distribučních jádrových funkcí (3.18)*

Všechny uvedené jádrové funkce jsou pouze nepatrným výčtem ze všech možných jádrových funkcí a jsou vhodné pro uvažovaný neparametrický jádrový odhad spolehlivosti z důvodu jejich „jednoduchosti“.

Vzhledem k obsáhlosti experimentů by bylo velmi obtížné použít všechny uvedené funkce, proto byly subjektivně vybrány pouze dvě jádrové funkce, které jsou použity pro odvozování postupů a testování výsledků na získaných experimentálních datech v Kapitole 5. Jako první byla vybrána Parzenova jádrová funkce a jako druhá Gaussova jádrová funkce, kde rozdíly mezi oběma vybranými funkcemi jsou ukázány v jednotlivých popisech a na dvou výše uvedených obrázcích. Výběr těchto „vhodných“ jádrových funkcí je založen na skutečnosti, že první z nich má „jednoduší“ charakter nejen pro

představivost, ale i pro všeobecnou práci a odvození zamýšlených postupů. Výběr druhé jádrové funkce je založen na jejím obsahu v téměř každém statistickém software a na více „vyhlazeném“ tvaru distribuční funkce i hustoty. Všechny tyto výběrové předpoklady budou podrobněji diskutovány v testové části práce, kde nejsou patrné zásadní změny (rozdíly) mezi odhadovanými distribučními funkcemi náhodné veličiny v závislosti na typu jádrové funkce. Jak je zřejmé, pro takovou volbu mluví numerická jednoduchost nebo dostupnost. Není to ale jediným důvodem. Z literatury např. Devroye a Gyorfi (1985) je známo, že volba „tvaru“ jádrové funkce ovlivňuje vlastnosti výsledné aproximace $\hat{f}(x;h)$ daleko méně než volba „vyhlazovacího parametru *h“.*

3.3 Neparametrický odhad „jednorozměrné“ distribuční funkce, vlastnosti

V návaznosti na předchozí část, která se detailně zabývala popisem modelu pro odhad neznámé funkce hustoty, lze nyní definovat model pro neparametrický jádrový odhad (aproximaci) neznámé distribuční funkce s označením $\hat{F}(x;h)$. Získaná aproximace (funkce) představuje odhad distribuční funkce, která je charakterizována více vyhlazeným tvarem, než u výše uvedeného tvaru neparametrického jádrového odhadu hustoty.

Definice 3.2 Nechť $\{x_1, \dots, x_n\}$ *je i.i.d. výběr náhodné veličiny* X *pevného rozsahu* n *a nechť* $x \in \mathbb{R}$*, potom model pro neparametrický jádrový odhad distribuční funkce je definován vztahem*

$$\hat{F}(x;h) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \qquad h > 0 \qquad n \in \mathbb{N}_+, \tag{3.17}$$

kde $K(x)$ *je jádrová distribuční funkce, která je získána z jádrové funkce hustoty*

$$K(x) = \int_{-\infty}^{x} k(z)dz. \tag{3.18}$$

Vztah uvedený v této definici vychází z definice distribuční funkce a je odvozen za použití integrace dříve uvedeného modelu pro odhad neznámé funkce hustoty (3.2) na oboru reálných čísel do požadované hodnoty x:

$$\begin{aligned}\hat{F}(x;h) &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{x}\frac{1}{h}k\left(\frac{z-x_i}{h}\right)dz \\ &=_a \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\frac{x-x_i}{h}}k(y)\,dy \\ &= \frac{1}{n}\sum_{i=1}^{n}K\left(\frac{z-x_i}{h}\right),\end{aligned} \tag{3.19}$$

kde je opět pomocí symbolu $=_a$ označena použitá substituce $y = \frac{z-x_i}{h}$.

3.4 Problematika vyhlazovacího parametru u odhadu hustoty

Kvalita popisovaného odhadu je závislá nejen na volbě jádrové funkce (slaběji, viz výše), ale i na volbě parametru h, a proto je nezbytné se uvedenou problematikou zabývat. Výběr odpovídající hodnoty tohoto parametru je velmi důležitý, protože vliv parametru ve formě vysokých hodnot vede k tzv. „přehlazení“, zatímco vliv malých hodnot parametru vede na tzv. „podhlazení“ nebo též „členitý“ odhad. Obecně se dostáváme do situace, kdy v prvním případě dochází při aproximaci ke ztrátě detailů a ve druhém případě získaná aproximace příliš „osciluje“. Obě situace budou názorně prezentovány dále.

Pro hodnocení kvality celého modelu je nezbytné nejprve vybrat „vhodnou“ míru charakterizující rozdíl mezi skutečnou funkcí hustoty $f(x)$ a popsaným modelem pro její odhad $\hat{f}(x;h)$. Podstata a význam tohoto problému je diskutován v mnoha doposud zveřejněných publikacích, jako např. Turlach (1993), kde je publikována výhoda volby míry v L^2normách z důvodu jednodušší analýzy než zkoumání dané problematiky v jiné normě (např. norma L^1). Následná analýza výběru hodnoty vyhlazovacího parametru může být založena na následujících postupech, které předpokládají znalost „skutečného“ pravděpodobnostního rozdělení náhodné veličiny (tj. apriorní informaci).

3.4.1 Odhad vyhlazovacího parametru při znalosti funkce hustoty

Celá tato část je motivována Turlach (1993), Řezáč (2007), Nováková (2009). Postup je založen na znalosti *„Střední kvadratické chyby*[30]*“*, kterou lze v literatuře nalézt pod zkratkou MSE z anglického překladu „Mean Squared Error“. Uvedené kritérium lze použít nejen u parametrických odhadů, ale i u neparametrických odhadů pro měření jejich „kvality“ v konkrétních bodech x a je interpretováno jako

$$MSE\left(\hat{f}(x;h)\right) = E\left\{\left(\hat{f}(x;h) - f(x)\right)^2\right\}. \tag{3.20}$$

[30] Podrobnější popis střední kvadratické chyby lze nalézt např. v Hátle a Likeš (1974) nebo Cipra (2008).

Vhodnými úpravami lze přepsat vybrané kritérium do tvaru

$$\begin{aligned} MSE\left(\hat{f}(x;h)\right) &= E\left\{\left(\hat{f}(x;h)-f(x)\right)^2\right\} \\ &= E\{f^2(x)-2f(x)\hat{f}(x;h)+\hat{f}^2(x;h)\} \\ &= f^2(x)-2f(x)E\{\hat{f}(x;h)\}+E\{\hat{f}^2(x;h)\}+E^2\{\hat{f}(x;h)\}-E^2\{\hat{f}(x;h)\} \\ &= E\{\hat{f}^2(x;h)\}-E^2\{\hat{f}(x;h)\}+E^2\{\hat{f}(x;h)\}-2f(x)E\{\hat{f}(x;h)\}+f^2(x) \\ &= E\{\hat{f}^2(x;h)\}-E^2\{\hat{f}(x;h)\}+\left(E\{\hat{f}(x;h)\}-f(x)\right)^2. \end{aligned} \tag{3.21}$$

Tím lze interpretovat MSE pomocí dvou členů, kde první z nich značí rozptyl odhadu a druhý tzv. kvadrát vychýlení, což lze zapsat pro přehlednost pomocí zkratek jako

$$Var\{\hat{f}(x;h)\} = E\left\{\hat{f}^2(x;h)\right\}-E^2\{\hat{f}(x;h)\} \tag{3.22}$$

a

$$Bias\left(\hat{f}(x;h)\right) = E\{\hat{f}(x;h)\}-f(x). \tag{3.23}$$

Potom lze přepsat MSE s využitím obou členů vztahem

$$MSE\left(\hat{f}(x;h)\right) = Var\{\hat{f}(x;h)\}+\left(Bias\left(\hat{f}(x;h)\right)\right)^2. \tag{3.24}$$

Nevýhoda kritéria je v hodnocení rozdílů pro jednotlivá pozorování, a proto je nahrazeno kritériem pro měření chyby na celém intervalu ISE, tj. *„Integrální kvadratická chyba“* z anglického překladu „Integrated Squared Error“. To udává vzdálenost mezi (skutečnou, známou) funkcí hustoty $f(x)$ a jejím odhadem $\hat{f}(x;h)$ na celém intervalu.

Uvedená míra představuje jeden z jednodušších možných nástrojů k hodnocení kvality odhadu a její interpretace je

$$ISE\left(\hat{f}(x\,;h)\right) = \int\left(\hat{f}(x;h)-f(x)\right)^2 dx. \tag{3.25}$$

Opět nevýhoda uvedeného kritéria ISE spočívá ve skutečnosti, že jeho výsledky lze považovat ze statistického úhlu pohledu do (jisté míry) za náhodné a tedy i nevhodné. Proto lze jeho tvar (do)upravit a (do)dodefinovat s využitím střední hodnoty, což je v literatuře nazýváno jako tzv. *„Střední integrální kvadratická chyba (MISE)“*, opět z anglického překladu *„Mean Integrated Squared Error“.* Interpretaci upraveného kritéria lze popsat jako

$$MISE\left(\hat{f}(x\,;h)\right) = E\left\{ISE\left(\hat{f}(x\,;h)\right)\right\} = E\left\{\int\left(\hat{f}(x;h)-f(x)\right)^2 dx\right\}, \tag{3.26}$$

kde integrační oblast představuje obor reálných čísel a symbol E označuje zmíněnou střední (nebo také očekávanou) hodnotu. Popsané kritérium je již mnohem více používáno z důvodu jeho matematické jednoduchosti a statistické interpretovatelnosti. Dále je možné kritérium upravit s využitím střední kvadratické chyby (MSE) a záměny pořadí při integraci

$$\begin{aligned} MISE\left(\hat{f}(x\,;h)\right) &= \int E\{\hat{f}(x;h) - f(x)\}^2 dx \\ &= \int MSE(\hat{f}(x;h))dx\,. \end{aligned} \tag{3.27}$$

Nevýhoda uvedeného kritéria spočívá v „dosti komplikované“ závislosti na neznámém vyhlazovacím parametru h, a proto se používá jiné známé kritérium tzv. „asymptotická aproximace“ MISE, která umožňuje přehlednější vyjádření závislosti chyby odhadu na volbě vyhlazovacího parametru. Pro odvození tohoto kritéria lze použít předchozí tvar MSE, kde je použit rozptyl odhadu a kvadrát vychýlení. Asymptotický případ *MISE* je odvozen a ukázán např. v Turlach (1993, s. 6 – 7), kde definice tzv. *„Asymptotické střední integrální kvadratické chyby (AMISE)“* z překladu *„Asymptotic Mean Squared Error“* je

$$AMISE(h) = \frac{R(k)}{nh} + h^{2s}\left(\frac{u_s(k)}{s!}\right)^2 R\big(f^{(s)}\big), \tag{3.28}$$

kde

$R(\cdot)$ značí pro nějakou integrovatelnou funkci L funkcionál $R(L) = \int L^2(x)dx$,

$u_j(\cdot), j\epsilon\mathbb{N}$ opět značí pro nějakou funkci L funkcionál $u_j(L) = \int x^j L(x)dx$,

f je známá nebo “pravá” funkce hustoty pravděpodobnosti a

$f^{(j)}, j\epsilon\mathbb{N}$ je $j-t$á derivace známé funkce hustoty f.

Odvození uvedeného tvaru není zcela triviální a je založeno na několika dalších předpokladech:

- Uvažovaná hustota $f(x)$ má spojité derivace řádu nejméně $(s+2)$ a jádrová funkce hustoty k je právě řádu[31] s.

- Šířka vyhlazovacího parametru $h = \{h_n\}_{n=1}^{\infty}$ je posloupnost nenáhodných kladných čísel, pro kterou platí, že h konverguje k 0 pomaleji než $\frac{1}{n}$, tedy
$$\lim_{n\to\infty} h = 0, \lim_{n\to\infty} nh = \infty.$$

[31] K pojmu řád jádrové funkce viz např. Turlach (1993, s. 6 – 7) nebo Řezáč (2007, s. 15).

Použitý řád jádrové funkce hustoty je podrobně popsán v uvedené literatuře (viz poznámka pod čarou), a protože se nejedná o cíl práce, není zde již dále popisován.

Poznámka 3.3 *Závislost ISE a MISE na vyhlazovacím parametru h se odráží v těchto výpočtech, ale ne závislost na vybrané jádrové funkci k.*

Jeden možný příklad o tomto kritériu je publikován od Marron a Wand (1992). Zde byly studovány *MISE* a *AMISE* pro případ, když je funkce hustoty mixem normálních hustot a k je Gaussova jádrová funkce ($s = 2$).

Nechť jsou tedy splněny předchozí předpoklady, potom je možné psát hodnotu vychýlení v bodě x jako

$$Bias\left(\hat{f}(x;h)\right) = E\{\hat{f}(x;h)\} - f(x) = \frac{1}{2}h^2 u_2(k) f''(x) + o(h^2), \tag{3.29}$$

kde $o(\cdot)$ symbolizuje vyjádření asymptotické chyby uvedeného rozvoje (aproximace) a je popsána níže (a v seznamu značení).

Důkaz. Nejprve se zaměříme na výpočet střední hodnoty odhadu[32] funkce hustoty $\hat{f}(x\,;h)$ v bodě $x \epsilon \mathbb{R}$:

$$\begin{aligned} E\{\hat{f}(x;h)\} &= \frac{1}{n}\sum_{i=1}^{n} E\left\{\frac{1}{h}k\left(\frac{x-x_i}{h}\right)\right\} = \frac{1}{h}E\left\{k\left(\frac{x-X}{h}\right)\right\} \\ &= \frac{1}{h}\int k\left(\frac{x-y}{h}\right) f(y)dy\,, \end{aligned} \tag{3.30}$$

kde X charakterizuje náhodnou veličinu, x_i její realizace a ostatní parametry jsou již popsány v předchozí části. Po zavedení substituce $z = \frac{x-y}{h}$ je tento výraz upraven do tvaru

$$E\{\hat{f}(x;h)\} = \int k(z) f(x - hz)dz. \tag{3.31}$$

Nyní lze provést Taylorův rozvoj funkce $f(x - hz)$ v bodě x (zde řádu 2 pro zjednodušení) s výslednou hodnotou

$$f(x - hz) = f(x) - hzf'(x) + \frac{1}{2}h^2 z^2 f''(x) + o(h^2), \tag{3.32}$$

kde pomocí symboliky $o(h^2)$ je vyjádřena asymptotická chyba daného výrazu v závislosti na hodnotě parametru h. Dosazením odvozeného rozvoje do předchozího výrazu (3.31) je získán tvar výpočtu střední hodnoty aproximované funkce v bodě x

[32] Podrobněji provedený postup tohoto výpočtu je proveden v příloze A dokazující odhad asymptotické nestrannosti a vydatnosti.

$$E\{\hat{f}(x;h)\} = f(x)\int k(z)dz - hf'(x)\int zk(z)dz + \frac{1}{2}h^2 f''(x)\int z^2 k(z)dz + o(h^2). \tag{3.33}$$

Výraz lze dále upravit s použitím dříve prezentovaných momentových podmínek pro zvolenou jádrovou funkci, kdy je získána střední hodnota odhadu (aproximace) v bodě x ve zjednodušeném tvaru

$$E\{\hat{f}(x;h)\} = f(x) + \frac{1}{2}h^2 f''(x)u_2(k) + o(h^2) \tag{3.34}$$

a zároveň vychýlení, které je vyjádřeno s použitím známé funkce hustoty $f(x)$ jako

$$E\{\hat{f}(x;h)\} - f(x) = \frac{1}{2}h^2 f''(x)u_2(k) + o(h^2). \tag{3.35}$$

■

Poznámka 3.4 *Na základě výše uvedeného vztahu je vhodné zmínit fakt, že vychýlení je funkčně závislé na hodnotě parametru h^2. Pokud budeme dále předpokládat, že počet pozorování $n \to \infty$, potom lze považovat odhad funkce hustoty za asymptoticky nevychýlený. Uvedená skutečnost je dokázána v příloze této práce, kde je názorně proveden důkaz asymptotické vydatnosti a nestrannosti získaného odhadu neznámé funkce hustoty.*

Poznámka 3.5 *Uvedenou symboliku „o“ lze zjednodušeně označit za „asymptotickou chybu“ provedeného odhadu, tedy nechť φ je reálná funkce definovaná v okolí bodu a, nechť ω je funkce kladná v prstencovém okolí bodu a. Potom symbol*

$$\varphi(x) = o(\omega(x)) \tag{3.36}$$

pro $x \to a$ značí, že

$$\lim_{x\to a}\frac{|\varphi(x)|}{\omega(x)} = 0. \tag{3.37}$$

Zde se vyskytuje symbol o ve tvaru $\varphi(h) = o(h^k)$ pro $h \to 0$, což lze přepsat tak, že

$$\lim_{h\to 0}\frac{|\varphi(h)|}{h^k} = \lim_{h\to 0}\frac{|o(h^k)|}{h^k} = 0. \tag{3.38}$$

(převzato z Nováková (2009))

Více informací, důkazů a podrobného rozpracování týkající se této problematiky lze nalézt např. v Řezáč (2007), Nováková (2009), Orava (2008).

Nechť jsou tedy splněny předchozí předpoklady, potom vztah (3.22) pro rozptyl odhadu $\hat{f}(x;h)$ v bodě x lze přepsat s využitím předchozích postupů jako

$$Var\{\hat{f}(x;h)\} = E\{\hat{f}(x;h)^2\} - E^2\{\hat{f}(x;h)\} = \frac{1}{nh}R(k)f(x) + o\left(\frac{1}{nh}\right). \tag{3.39}$$

Důkaz. Pro odvození tohoto tvaru použijeme předchozí Taylorův rozvoj a níže uvedené zjednodušení

$$\hat{f}(x;h) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}k\left(\frac{x-x_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n}k_h(x-x_i). \tag{3.40}$$

Uvedený výraz lze dále přepsat s použitím *o* notace (asymptotické chyby odhadu) na tvar

$$k_h(x-x_i) = f(x) + o(1), \tag{3.41}$$

kde $o(1)$ lze chápat jako funkci jdoucí limitně k 0. Výsledné odvození lze následně provést pomocí následujícího postupu:

$$\begin{aligned}
Var\{\hat{f}(x;h)\} &= Var\left\{\frac{1}{n}\sum_{i=1}^{n}k_h(x-x_i)\right\} \\
&= \frac{1}{n}Var\{k_h(x-X)\} \\
&= \frac{1}{n}E\{k_h(x-X)^2\} - \frac{1}{n}(E\{k_h(x-X)\})^2 \\
&= \frac{1}{n}\int k_h(x-y)^2 f(y)dy - \frac{1}{n}\left(\int k_h(x-y)f(y)dy\right)^2 \\
&= \frac{1}{nh}\int k(z)^2 f(x-hz)dz - \frac{1}{n}\left(\int k(z)f(x-hz)dz\right)^2 \\
&= \frac{1}{nh}\int k(z)^2(f(x)+o(1))dz - \frac{1}{n}\left(\int k(z)(f(x)+o(1))dz\right)^2 \\
&= \frac{1}{nh}f(x)\int k(z)^2dz + o\left(\frac{1}{nh}\right), \\
&= \frac{1}{nh}f(x)R(k) + o\left(\frac{1}{nh}\right).
\end{aligned} \tag{3.42}$$

■

Poznámka 3.6 *Zde je vhodné upozornit na skutečnost, že z výše uvedených vlastností vyhlazovacího parametru konverguje součin $nh \to \infty$, rozptyl je tedy nepřímo úměrný tomuto součinu a konverguje k 0 pro $n \to \infty$.*

Nyní již lze přepsat MSE pomocí asymptotického vyjádření střední hodnoty a rozptylu

$$\begin{aligned}
MSE\left(\hat{f}(x;h)\right) &= Var\{\hat{f}(x;h)\} + \left(Bias\left(\hat{f}(x;h)\right)\right)^2 \\
&= \frac{1}{nh}R(k)f(x) + \frac{1}{4}h^4\left(f''(x)\right)^2\left(u_2(k)\right)^2 + o\left(\frac{1}{nh} + h^4\right)
\end{aligned} \tag{3.43}$$

a za splnění uvedených předpokladů lze přepsat MISE do tvaru

$$MISE\left(\hat{f}(x\,;h)\right) = AMISE\left(\hat{f}(x\,;h)\right) + o\left(\frac{1}{nh} + h^4\right), \tag{3.44}$$

kde

$$AMISE\left(\hat{f}(x\,;h)\right) = \frac{1}{nh}R(k) + \frac{1}{4}h^4 R\left(f^{''}\right)\left(u_2(k)\right)^2. \tag{3.45}$$

Důkaz. Odvození výrazu lze provést, pokud f je integrovatelná funkce a jsou zároveň splněny všechny uvedené předpoklady.

$$\begin{aligned} MISE\left(\hat{f}(x\,;h)\right) &= \int MSE(\hat{f}(x;h))dx \\ &= \frac{1}{nh}\int R(k)f(x)\,dx + \frac{1}{4}h^4\int \left(u_2(k)\right)^2\left(f^{''}(x)\right)^2 dx + o\left(\frac{1}{nh} + h^4\right) \\ &= \frac{1}{nh}R(k) + \frac{1}{4}h^4 R\left(f^{''}\right)\left(u_2(k)\right)^2 + o\left(\frac{1}{nh} + h^4\right). \end{aligned} \tag{3.46}$$

■

Poznámka 3.7 *Výpočet integrálu druhé mocniny vychýlení je přímo úměrný hodnotě h^4, a proto hodnotu vychýlení lze snižovat menším vyhlazovacím parametrem. Na druhou stranu menší hodnota vyhlazovacího parametru způsobí nárůst hodnoty integrálu rozptylu vzhledem k uvedené nepřímé závislosti. Závěr z uvedeného odvození ukazuje, že je vhodné volit kompromis pro „optimální[33]“ hodnotu vyhlazovacího parametru.*

Získané postupy lze následně využít pro odvození „vhodného“ vyhlazovacího parametru h, pro který nabývá kvalitativní kritérium $AMISE$ minimální hodnoty, tedy

$$\hat{h} = \left(\frac{R(k)}{\left(u_2(k)\right)^2 R(f^{''})n}\right)^{\frac{1}{5}}. \tag{3.47}$$

Důkaz. Přístup k získání optimální hodnoty vyhlazovacího parametru h je založen na hledání minima s použitím první parciální derivace odvozeného výrazu podle hledaného parametru

$$\frac{\partial}{\partial h}\left(\frac{1}{nh}R(k) + \frac{1}{4}h^4 R\left(f^{''}\right)\left(u_2(k)\right)^2\right) = -\frac{1}{nh^2}R(k) + h^3 R\left(f^{''}\right)\left(u_2(k)\right)^2, \tag{3.48}$$

kterou k získání stacionárních bodů položíme rovnou 0 a dostaneme hodnotu výsledného tvaru (3.47):

[33] Optimální ve smyslu odhadu, který není „přehlazený“ ani „podhlazený“, tedy podle uvedeného kritéria je optimální hodnota rovna minimální hodnotě kritéria.

$$
\begin{aligned}
0 &= \frac{1}{nh^2} R(k) + h^3 R\left(f^{\prime\prime}\right)\left(u_2(k)\right)^2 \\
\hat{h}^5 &= \frac{R(k)}{R(f^{\prime\prime})\left(u_2(k)\right)^2 n} \\
\hat{h} &= \left(\frac{R(k)}{R(f^{\prime\prime})\left(u_2(k)\right)^2 n}\right)^{\frac{1}{5}}.
\end{aligned}
\tag{3.49}
$$

■

Poznámka 3.8 *Vyhlazovací parametr odhadovaný tímto způsobem je závislý na apriorní znalosti jádrové funkce, počtu získaných pozorování a funkci hustoty rozdělení, která většinou není známa na souboru reálných dat, a proto využití tohoto pravidla není v praxi téměř možné.*

Uvedená problematika včetně navazujících různých testovacích kritérií již nebude dále detailně vyšetřována, protože je zde jeden velmi důležitý předpoklad, který je uveden v poznámce výše, a to že apriorně neznáme „skutečnou“ a známou funkci hustoty.

3.4.2 Odhad vyhlazovacího parametru při neznalosti funkce hustoty

Předpoklad apriorní znalosti funkce hustoty představuje hlavní důvod, proč je v celé této práci nadále používán jen odhad parametru h založený na odvození z rovnosti mezi hodnotou výběrového rozptylu a získaným rozptylem neparametrického jádrového modelu pro aproximaci hustoty.

Odvozený vztah pro získání rozptylu (3.6) u neparametrickoho jádrového modelu je již znám a vztah pro odhad výběrového rozptylu, označovaného jako s_n^2, lze vyjádřit jako

$$
s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_a)^2. \tag{3.50}
$$

Zde je obsažen odhad střední hodnoty modelu $\hat{f}(x;h)$, kde je značen notací μ_a. Potom rozptyl vybraného modelu pro neparametrický jádrový odhad hustoty (3.6) může být upraven s použitím předchozího vztahu pro odhad výběrového rozptylu

$$
\begin{aligned}
\sigma^2_{\hat{f}(x;h)}\{X\} &= h^2 + \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_a)^2 \\
&=_a h^2 + \frac{n-1}{n} s_n^2 \\
&= h^2 + \left(1 - \frac{1}{n}\right) s_n^2.
\end{aligned}
\tag{3.51}
$$

Použité značení $=_a$ označuje místo, kde je použito vyjádření

$$\sum_{i=1}^{n}(x_i - \mu_a)^2 = (n-1)s_n^2. \tag{3.52}$$

Z tohoto důvodu je možné stanovit přibližnou rovnost mezi oběma rozptyly $s_n^2 \cong \sigma_a^2$ a výsledný odhad neznámého parametru vyhlazení h může být vyjádřen v následujícím tvaru.

Důsledek 3.2 *Nechť $\{x_1, \dots, x_n\}$ je i.i.d. výběr náhodné veličiny X pevného rozsahu n a nechť s_n^2 je výběrový rozptyl, potom hodnota neznámého vyhlazovacího parametru h může být vyjádřena jako*

$$h = \sqrt{\frac{s_n^2}{n}}, \quad n \in \mathbb{N}_+. \tag{3.53}$$

Důkaz. Jako důkaz pro vyjádření uvedeného vztahu slouží výše uvedený postup.

Poznámka 3.9 *Zde je důležité upozornit na vzniklou skutečnost, že při použití výše uvedeného postupu pro odhad vyhlazovacího parametru je předpokládáno, že neznámý vyhlazovací parametr h je určen reálným číslem (jedná se o deterministickou hodnotu), ale v tomto případě má hodnota tohoto parametru charakter náhodné veličiny, což je obecně problém.*

Další modely pro odhad „vhodného“ vyhlazovacího parametru jsou velmi často diskutovány v odborné literatuře (např. střední integrální absolutní chyba „Mean Integrated Absolute Error“, atd.), kde jsou také doporučeny případné výpočetní postupy včetně předpokladů na ně kladených. Jiná cesta pro odhad tohoto parametru je založena na vybraném numerickém postupu a jeho rostoucí hodnotě až do “přijatelného tvaru”.

3.5 Problematika vyhlazovacího parametru u distribuční funkce

Doposud byl hlavním předmětem této práce odhad (aproximace) neznámé funkce hustoty podle neparametrických jádrových modelů. Nyní se lze podívat na uvedenou problematiku i z jiného úhlu pohledu a zaměřit se na jádrovou aproximaci distribuční funkce pro získaný soubor reálných pozorování, u kterého není známa dostatečná apriorní informace o tvaru pravděpodobnostního rozdělení.

3.5.1 Popis vzniklé situace

Primárním cílem práce je neznámá hodnota (pravděpodobnosti) spolehlivosti nebo selhání, kterou lze získat za předpokladu znalosti (odhadu) distribuční funkce ze získaných pozorování. V předchozích kapitolách byly tyto odhady prezentovány pomocí neparametrických jádrových odhadů, tedy s využitím různých typů jádrových funkcí, které splňují předem dané předpoklady (3.3) a zároveň zde byly popsány nedostatky

ovlivňující kvalitu odhadu tj. přístupy k volbě vyhlazovacího parametru (parametru měřítka). Samotný vyhlazovací parametr významně ovlivňuje tvar (kvalitu) aproximovaných funkcí a v případě nevhodného odhadu jsou výsledné hodnoty zkreslené.

3.5.2 Neparametrický jádrový odhad distribuční funkce

Obecný model neparametrického jádrového odhadu distribuční funkce používající různé typy jádrových funkcí je definován modelem (3.17). Pro účely zde použité aproximace je vhodným typem distribuční jádrové funkce $K(x)$ tzv. Parzenova jádrová funkce popsaná vztahem (3.8), která je vhodná vzhledem k následně vyhovujícím vlastnostem.

Poznámka 3.10 *Z důvodu přehlednosti v níže uvedených postupech je vhodné připomenout rozdíly mezi značením jádrových funkcí, kde $k(x)$ značí jádrovou funkci hustoty a $K(x)$ je integrál této jádrové funkce tj. distribuční jádrová funkce.*

Dále budeme předpokládat, že jádrová funkce hustoty opět splňuje výše uvedené předpoklady (3.3) tj. spojitá centrovaná a normovaná funkce hustoty kolem středí hodnoty a její jádrová distribuční funkce $K(x)$ zároveň splňuje pro $a \in \mathbb{R}$:

1. $K(x)$ je spojitá distribuční funkce na celé reálné ose,
2. $\exists a > 0$ takové, že $K(x) = 0$ pro $x \leq -a$ a zároveň $K(x) = 1$ pro $x \geq a$

a vyhlazovací parametr h je funkčně závislý nanejvýš na hodnotě rozsahu získaného souboru dat, tedy na hodnotě n.

Poznámka 3.11 *Pokud se použije model (3.17) pro odhad distribuční funkce s použitím Parzenovy jádrové funkce a vyhlazovacím parametrem $h \to 0$, pak pro takový jednotkový skok lze hovořit o empirické distribuční funkci[34] (EDF).*

Nechť $\{x_1, \dots, x_n\}$ je i.i.d. výběr náhodné veličiny X pevného rozsahu n, potom sdružená funkce hustoty

$$f(x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i), \qquad n \in \mathbb{N}_+. \tag{3.54}$$

Tento vztah lze použít k odhadu střední hodnoty aproximované (neznámé distribuční funkce z uvažovaného náhodného výběru $E\{\hat{F}(x;h)\}$.

Věta 3.3 *Nechť $\{x_1, \dots, x_n\}$ je i.i.d. výběr náhodné veličiny X pevného rozsahu n a nechť existuje neparametrický jádrový odhad distribuční funkce $\hat{F}(x;h)$, potom odhad střední*

[34] Empirická distribuční funkce náhodné veličiny X zjednodušeně představuje příslušné kumulativní četnosti k vzestupně seřazeným pozorováním (k pořádkovým statistikám nad uvažovaným výběrem), více např. Rényi (1972).

hodnoty aproximované distribuční funkce uvažované náhodné veličiny je možné vyjádřit jako

$$E\{\hat{F}(x;h)\} = F(x - h\xi), \qquad h > 0, \qquad \xi \in \langle -a, a\rangle, \qquad n \in \mathbb{N}_+. \tag{3.55}$$

Poznámka 3.12 *V přesném slova smyslu je myšlen odhad střední hodnoty z náhodné veličiny neparametrického jádrového odhadu distribuční funkce $\hat{F}(x;h)$, který je pro každé x náhodnou proměnnou (tj. deterministickou funkcí pozorování z náhodného výběru).*

Důkaz. Postup pro vyjádření uvedeného vztahu je založen na výrazu (3.54) a předpisu pro odhad střední hodnoty. Níže uvedený postup v sobě odráží smysl náhodné veličiny i vícerozměrnou funkci hustoty uvažovaného pravděpodobnostního rozdělení, tedy

$$\begin{aligned} E\{\hat{F}(x;h)\} &= \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \hat{F}(x;h) \left[\prod_{i=1}^{n} f(x_i)\right] dx_1 \dots dx_n \\ &= \frac{1}{n}\sum_{j=1}^{n} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} K\left(\frac{x - x_j}{h}\right) \left[\prod_{i=1}^{n} f(x_i)\right] dx_1 \dots dx_n \\ &= \frac{1}{n}\sum_{j=1}^{n} \int_{-\infty}^{+\infty} K\left(\frac{x - x_j}{h}\right) f(x_j) dx_j \\ &= \int_{-\infty}^{+\infty} K\left(\frac{x - z}{h}\right) f(z) dz, \qquad h > 0, \qquad n \in \mathbb{N}_+, \end{aligned} \tag{3.56}$$

kde ve třetím řádku dostáváme aritmetický průměr ze shodných jádrových funkcí a díky tomu lze odstranit závislost na rozsahu n. Následuje úprava výsledku pro získání střední hodnoty za použití substituce $\frac{x-z}{h} = y \Longrightarrow z = x - hy \Longrightarrow dz = -hdy$, kterou je získán vztah

$$E\{\hat{F}(x;h)\} = h \int_{-\infty}^{+\infty} K(y)\, f(x - hy) dy. \tag{3.57}$$

Vzhledem k výše uvedené vlastnosti č. 2 je možné uvedený výraz dále modifikovat do tvaru obsahujícího součet dvou komponent

$$h \int_{-\infty}^{+\infty} K(y)\, f(x - hy) dy = h \int_{-a}^{+a} K(y)\, f(x - hy) dy + h \int_{+a}^{+\infty} f(x - hy)\, dy. \tag{3.58}$$

Oba dva členy na pravé straně z výše uvedeného součtu lze dále upravovat. Nejprve je pozornost zaměřena na druhý člen součtu, který za použití zpětné předchozí substituce $z = x - hy \Longrightarrow \frac{x-z}{h} = y \Longrightarrow \frac{-dz}{h} = dy$ je převoditelný do tvaru

$$
\begin{aligned}
h\int_{+a}^{+\infty} f(x-hy)\,dy &= -\int_{x-ha}^{-\infty} f(z)\,dz \\
&= \int_{-\infty}^{x-ha} f(z)\,dz \\
&= F(x-ha), \qquad h>0.
\end{aligned}
\tag{3.59}
$$

Potom je pozornost zaměřena na úpravu prvního členu výše uvedeného součtu, která je proveditelná podle druhé věty o střední hodnotě integrálního počtu popsané níže.

Věta 3.4 *Nechť existují konstanty $a, b \in \mathbb{R}$ takové, že $a < b$, nechť dále $f(x)$ je reálná funkce, u které existuje integrál od a do b, a nechť $g(x)$ je monotónní na intervalu $\langle a, b\rangle$. Potom existuje číslo $\xi \in \langle a, b\rangle$ takové, že*

$$
\int_a^b f(x)g(x)\,dx = g(a)\int_a^{\xi} f(x)\,dx + g(b)\int_{\xi}^{b} f(x)\,dx. \tag{3.60}
$$

(převzato z Jarník (1954, s. 241)

Důkaz. Postup celého důkazu věty (3.4) je detailně popsán v Jarník (1954, s. 241-246) a je proveden za daných a velmi obecných předpokladů.

Nyní již nic nebrání k provedení zamýšlené úpravy u první části uvedeného součtu s využitím výše uvedené věty o střední hodnotě integrálního počtu následovně:

$$
\begin{aligned}
h\int_{-a}^{+a} K(y)\,f(x-hy)dy &= h\left(K(-a)\int_{-a}^{\xi} f(x-hy)\,dy + K(a)\int_{\xi}^{+a} f(x-hy)\,dy\right) \\
&= h\left(\int_{\xi}^{+a} f(x-hy)\,dy\right) \\
&=_a h\left(\frac{1}{h}\int_{x-ah}^{x-\xi h} f(z)\,dz\right) \\
&= F(x-h\xi) - F(x-ha), \qquad h>0,
\end{aligned}
\tag{3.61}
$$

kde je pomocí symboliky $=_a$ použita předchozí substituce $z = x - hy$. Na závěr lze shrnout uvedený poustup tak, že $\exists \xi \in \langle -a, a\rangle$ takové, že střední hodnota z aproximace distribuční funkce v bodě $x \in \mathbb{R}$ je rovna

$$
\begin{aligned}
E\{\hat{F}(x;h)\} &= h\int_{-a}^{+a} K(y)\,f(x-hy)dy + h\int_{+a}^{+\infty} f(x-hy)\,dy \\
&= F(x-h\xi) - F(x-ha) + F(x-ha) \\
&= F(x-h\xi), \qquad h>0, \qquad \xi \in \langle -a, a\rangle.
\end{aligned}
\tag{3.62}
$$

■

Zjednodušeně se získaný výsledek může interpretovat tak, že střední hodnota neparametrického jádrového odhadu neznámé distribuční funkce náhodné veličiny X je pro každé reálné x hledaná (neznámá, vzorová) distribuční funkce posunutá právě o hodnotu „$h\xi$“.

Dále je vhodné odhadnout rozptyl uvažovaného odhadu, kde náhodnou veličinu představuje odhad (aproximace) neznámé distribuční funkce $\hat{F}(x;h)$.

Věta 3.5 *Nechť $\{x_1, \dots, x_n\}$ je i.i.d. výběr náhodné veličiny X pevného rozsahu n a nechť existuje neparametrický jádrový odhad distribuční funkce $\hat{F}(x;h)$ a konstanty $\xi, \eta \in \langle -a, a\rangle$, potom odhad rozptylu aproximované distribuční funkce uvažované náhodné veličiny je možné vyjádřit jako*

$$\sigma^2\{\hat{F}(x;h)\} = \frac{F(x-\eta h) - F^2(x-\xi h)}{n}, \tag{3.63}$$

kde $h > 0$ a $n \in \mathbb{N}_+$.

Důkaz. Podstata důkazu je založena na několika postupných krocích, kdy v prvním kroku lze nejprve odvodit kvadrát uvažované náhodné veličiny, tedy

$$\begin{aligned} \hat{F}^2(x;h) &= \frac{1}{n^2}\left[\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)\right]^2 \\ &= \frac{1}{n^2}\left[\sum_{i=1}^{n} K^2\left(\frac{x-x_i}{h}\right) + \sum_{i\neq j} K\left(\frac{x-x_i}{h}\right) K\left(\frac{x-x_j}{h}\right)\right] \end{aligned} \tag{3.64}$$

pro $h > 0$ a $n \in \mathbb{N}_+$. Takto získaný kvadrát uvažované náhodné veličiny je následně používán k odvození střední hodnoty, která je nezbytná k doplnění do vztahu k výpočtu rozptylu

$$\begin{aligned} E\{\hat{F}^2(x;h)\} &= \frac{1}{n^2}\left[\sum_{i=1}^{n} \int_{-\infty}^{+\infty} K^2\left(\frac{x-x_i}{h}\right) f(x_i)dx_i + \right. \\ &\quad \left. + \sum_{i\neq j} \int_{-\infty}^{+\infty} K\left(\frac{x-x_i}{h}\right) f(x_i)dx_i \int_{-\infty}^{+\infty} K\left(\frac{x-x_j}{h}\right) f(x_j)dx_j\right] \\ &= \frac{1}{n^2}\left[n \int_{-\infty}^{+\infty} K^2\left(\frac{x-z}{h}\right) f(z)dz + (n^2-n)F^2(x-h\xi)\right] \\ &= \frac{1}{n}\left[\int_{-\infty}^{+\infty} K^2\left(\frac{x-z}{h}\right) f(z)dz + (n-1)F^2(x-h\xi)\right] \end{aligned} \tag{3.65}$$

pro $h > 0$, $n \in \mathbb{N}_+$ a $\xi \in \langle -a, a\rangle$. Dále lze použít shodných postupů jako u výpočtu výše uvedené střední hodnoty včetně použitých substitucí (3.58), (3.59), (3.60) a (3.61). Jinými slovy, první člen součtu ve vztahu (3.65) lze rozepsat do následujícího tvaru

$$\begin{aligned}\int_{-\infty}^{+\infty} K^2\left(\frac{x-z}{h}\right) f(z)dz &= h\int_{-a}^{+a} K^2(y)\, f(x-hy)dy + h\int_{+a}^{+\infty} f(x-hy)\,dy \\ &= h\int_{-a}^{+a} K^2(y)\, f(x-hy)dy + F(x-ah) \\ &= h\int_{\eta}^{+a} f(x-hy)dy + F(x-ah), \qquad h>0,\end{aligned} \tag{3.66}$$

kde $\eta \in \langle -a, a\rangle$ a je obecně různé od hodnoty ξ. Odtud je možné dále upravit opět první člen součtu ze získaného vztahu (3.66) na hodnotu

$$\begin{aligned}h\int_{\eta}^{+a} f(x-hy)dy + F(x-ah) &= h\left(\frac{1}{h}\int_{x-ah}^{x-\eta h} f(z)\,dz\right) + F(x-ah) \\ &= F(x-\eta h) - F(x-ah) + F(x-ah) \\ &= F(x-\eta h), \qquad h>0, \qquad \eta \in \langle -a, a\rangle.\end{aligned} \tag{3.67}$$

Výše uvedeným postupem je získána výsledná hodnota odhadu střední hodnoty kvadrátu aproximované distribuční funkce v bodě x (která má v danou chvíli charakter náhodné veličiny) ve tvaru

$$\begin{aligned}E\{\hat{F}^2(x;h)\} &= \frac{1}{n}[F(x-\eta h) + (n-1)F^2(x-\xi h)] \\ &= F^2(x-\xi h) + \frac{1}{n}[F(x-\eta h) - F^2(x-\xi h)],\end{aligned} \tag{3.68}$$

kde $h>0$, $n \in \mathbb{N}_+$ a $\xi, \eta \in \langle -a, a\rangle$, ale jsou obecně navzájem různá. Nyní je již možné odvodit hodnotu rozptylu uvažované náhodné veličiny $\hat{F}(x;h)$ podle dobře známého vztahu

$$\sigma^2\{\hat{F}(x;h)\} = E\{\hat{F}^2(x;h)\} - E\{\hat{F}(x;h)\}^2, \qquad h>0, \qquad n \in \mathbb{N}_+, \tag{3.69}$$

s využitím odvozeného vztahu (3.68) a následujícího postupu za podmínek uvedených výše:

$$\begin{aligned}\sigma^2\{\hat{F}(x;h)\} &= F^2(x-\xi h) + \frac{1}{n}[F(x-\eta h) - F^2(x-\xi h)] - F^2(x-\xi h) \\ &= \frac{F(x-\eta h) - F^2(x-\xi h)}{n}.\end{aligned} \tag{3.70}$$

■

Uvedený postup lze shrnout tak, že pokud existují $\xi, \eta \in \langle -a, a\rangle$ potom odhad rozptylu z náhodné veličiny $\hat{F}(x;h)$ je $\sigma^2\{\hat{F}(x;h)\} = \frac{F(x-\eta h) - F^2(x-\xi h)}{n}$.

Poznámka 3.13 *K tomu, aby bylo předchozí odvození korektní a dávalo smysl, je vhodné „ověřit podmínky existence" pro použitou střední hodnotu* $E\{\hat{F}(x;h)\}$ *a* $E\{\hat{F}^2(x;h)\}$ *na jejichž základě lze dále odvozovat i neznámou hodnotu rozptylu.*

Nejprve existence střední hodnoty odhadu distribuční funkce v bodě $x \in \mathbb{R}$, která je charakterizována samotnou existencí vztahu $E\{\hat{F}(x;h)\} = \int_{-\infty}^{+\infty} K\left(\frac{x-z}{h}\right) f(z)dz$ a $K(x)$ je nezáporná funkce zdola omezená nulou a shora omezená jednotkou, tj. funkce hustoty náhodné veličiny $f(z)$ je její majorantou s omezením maximální hodnoty rovné 1 (omezení pomoci majoranty). Dále pokud existuje uvedená střední hodnota, pak je i hodnota

$$E\{\hat{F}^2(x;h)\} = \frac{1}{n^2}\left[n \int_{-\infty}^{+\infty} K^2\left(\frac{x-z}{h}\right) f(z)dz + (n^2-n)F^2(x-h\xi)\right] \tag{3.71}$$

a $K^2(x)$ je opět nezáporná funkce zdola omezená nulou a shora omezená jednotkou.

Tvrzení 3.1 *Nechť* $\hat{F}(x;h)$ *je modelová distribuční funkce a* $F(x)$ *je předpokládaná a neznámá distribuční funkce ze získaného souboru reálných dat pevného rozsahu* $n \in \mathbb{N}_+$*, nechť dále hodnota* $a \in \mathbb{R}$ *je odvozená hodnota uvažované distribuční jádrové funkce,* $h > 0$ *je přidružený vyhlazovací parametr a v bodě* $x \in \mathbb{R}$ *existuje střední hodnota odhadované distribuční funkce* $E\{\hat{F}(x;h)\}$*, potom platí následující nerovnosti:*

1. $F(x-ah) \le E\{\hat{F}(x;h)\} \le F(x+ah)$,
2. $\sigma^2\{\hat{F}(x;h)\} \le \frac{F(x+ah)-F^2(x-ah)}{n} \le \frac{1}{n}$. (3.72)

Důkaz. Podstata důkazu je založena na vlastnostech uvažovaných funkcí, kde předpokládané funkce $F(x)$ a $F^2(x)$ jsou spojité neklesající distribuční funkce definované na celém definičním oboru a zároveň tedy platí nerovnost $F^2(x) \le F(x)$. Odtud první nerovnost lze dokázat nahrazením libovolného čísla $\xi \in \langle -a, a\rangle$ číslem $(-a)$. Pro druhou nerovnost platí, že $F(x+ah) \le 1$ a tedy i $F^2(x-ah) \in \langle 0,1\rangle$, což lze označit za omezenost konstantou dle definice.

Tvrzení 3.2 *Pokud pro* $n \in \mathbb{N}_+$ *a* $h > 0$ *platí* $n \to \infty \Longrightarrow h \to 0$*, pak je daný odhad za výše uvedených podmínek asymptoticky nestranný a asymptoticky vydatný.*

Důkaz. Přímý důsledek předchozího tvrzení 3.1.

3.5.3 Algoritmus odvozeného modelu

Pomocí předchozího odvození byl získán model pro odhad (aproximaci) neznámé hodnoty distribuční funkce a nyní je vhodná doba pro detailní popis postupu (algoritmu) k získání modelové distribuční funkce na získaném souboru reálných pozorování.

Nechť je k dispozici získaný soubor náhodných pozorování $\{x_1, \dots, x_n\}$ pevného rozsahu n spojité náhodné veličiny X s rozdělením pravděpodobnosti popsané funkcí hustoty i distribuční funkcí a tomuto náhodnému výběru odpovídá „seřazený“ náhodný výběr

$$\{x_{(1)}, x_{(2)}, \dots, x_{(n)}\}: x_{(i)} \leq x_{(i+1)}, \qquad i = 1, \dots, n-1. \tag{3.73}$$

Potom pro takto seřazený soubor v souvislosti s neseřazeným platí

$$\hat{F}(x;h) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - x_{(i)}}{h}\right), \qquad h > 0, \qquad n \in \mathbb{N}_+, \tag{3.74}$$

protože se jedná o konečný součet naměřených pozorování, kde lze přehazovat dle libosti vzhledem k hodnotám na reálné ose x. Dále je možné tento model přepsat do tvaru

$$\hat{F}(x;h) = \frac{1}{n}\left[D(x) + \sum_{i=D(x)+1}^{H(x)-1} K\left(\frac{x - x_{(i)}}{h}\right)\right], \qquad h > 0, \qquad n \in \mathbb{N}_+, \tag{3.75}$$

kde $D(x)$ a $H(x)$ jsou definovány jako

$$D(x) = max\{i: x_{(i)} \leq x - ah; 0\}, \tag{3.76}$$

tj. započtu všechny hodnoty až do uvažované hranice $(x - ah)$ a

$$H(x) = min\{i: x_{(i)} \geq x + ah; n + 1\}, \tag{3.77}$$

tj. všechny větší než je hranice $(x + ah)$ nezapočtu, nebo započtu nulovou hodnotu.

Důkaz. Hlavní podstatou důkazu jsou následující odvození, která nepotřebují bližší popis, tedy

$$\left(\frac{x - x_{(i)}}{h} \leq -a \Longrightarrow K\left(\frac{x - x_{(i)}}{h}\right) = 0\right), \tag{3.78}$$

ale

$$\left(\frac{x - x_{(i)}}{h} \leq -a\right) \Leftrightarrow \left(x_{(i)} \geq x + ah\right) = 0 \tag{3.79}$$

a

$$\left(\frac{x - x_{(i)}}{h} \geq +a \Longrightarrow K\left(\frac{x - x_{(i)}}{h}\right) = +1\right), \tag{3.80}$$

ale

$$\left(\frac{x - x_{(i)}}{h} \geq +a\right) \Leftrightarrow \left(x_{(i)} \leq x - ah\right) = 0. \tag{3.81}$$

■

Poznámka 3.14 *K pochopení celého algoritmu je nejvhodnější názorná ukázka, která obsahuje použité značení a uvažované „meze“, které hrají v tomto algoritmu velmi důležitou roli.*

1	2	3	$D(x)+1$	$D(x)+2$	$D(x)+3$	$n-2$	$n-1$	n
$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(D(x)+1)}$	$x_{(D(x)+2)}$	$x_{(D(x)+3)}$	$x_{(n-2)}$	$x_{(n-1)}$	$x_{(n)}$
$K\left(\frac{x-x_{(i)}}{h}\right)=+1$			$K\left(\frac{x-x_{(i)}}{h}\right)$			$K\left(\frac{x-x_{(i)}}{h}\right)=0$		
			$x-ah$			$x+ah$		

Tabulka 1: *Aproximační algoritmus distribuční funkce*

V uvedené ukázce jsou patrné původní předpoklady na jádrovou distribuční funkci, které jsou v případě výběru Parzenovy jádrové funkce splněny. Pří bližším prozkoumání tohoto algoritmu lze dále zjistit, že uvedený model k odhadu neznámé distribuční funkce není nic jiného, než „zobecněná empirická distribuční funkce“, jak naznačuje slovní popis pro aproximaci neznámé hodnoty $F(x)$.

Stručně řečeno, aproximovanou hodnotu předpokládané a neznámé distribuční funkce $F(x)$ v libovolném bodě $x \in \mathbb{R}$ získáme tak, že jsou nejprve načítány přírůstky $1/n$ až do hodnoty dolní meze $(x-ah)$ za každou menší hodnotu pozorování v seřazeném souboru, jako při tvorbě EDF. Následně jsou pro všechna pozorování v intervalu $(x-ah, a+ah)$ dopočítány hodnoty vybrané jádrové distribuční funkce. Pro pozorování vyšší než je horní mez uvažovaného intervalu $(x+ah)$ již nejsou načítány další hodnoty jádrové distribuční funkce a výsledná hodnota je získána tak, že k součtu přírůstků je přičten celkový součet hodnot v intervalu vydělený počtem pozorování n.

3.6 Neparametrický jádrový odhad hustoty a distribuční funkce součtů

Doposud prezentovaná podrobná analýza neparametrických jádrových modelů je dále rozšiřitelná na jádrové odhady součtů, přesněji na aproximaci distribuční funkce součtů. Myšlenka u jádrového odhadu součtů je inspirována skutečností, že předmětem zájmu není stavová veličina (tj. spolehlivost v daném časovém okamžiku, např. měsíci), ale kumulace veličiny za uvažované období (tj. spolehlivost za celé uvažované období), což může být podmíněno např. ekonomickým významem aplikace.

Nechť je předpokládáno, že máme k dispozici dvě nezávislé stejně rozdělené spojité náhodné veličiny X a Y, které mají funkce hustoty $f_X(x), f_Y(x)$ a distribuční funkce $F_X(x), F_Y(x)$. Dále budeme předpokládat, že existuje rozdělení jejich součtu, tedy

kumulovaná náhodná veličina $Z = (X + Y)$ mající distribuční funkci s využitím konvoluce[35]

$$F_Z(x) = \int_{-\infty}^{+\infty} F_X(x - z) f_Y(z) dz\,, \quad x, z \in \mathbb{R}. \tag{3.82}$$

Poznámka 3.15 *Pravděpodobnostní rozdělení s distribuční funkcí $F_Z(x)$ se nazývá konvoluce rozdělení s distribučními funkcemi $F_X(x), F_Y(x)$.*

Celá problematika konvolucí pravděpodobnostních rozdělení včetně důkazů a vlastností (komutativnost, asociativnost) je podrobně rozepsána v Rényi (1972, s. 180). Potom odhad distribuční funkce součtu m nezávislých stejně rozdělených náhodných veličin bude nadále označován jako $\hat{F}_m(x; h)$. Nechť je dále předpokládán pro náhodný výběr model (3.2) pro neparametrický jádrový odhad hustoty (včetně jeho předpokladů) s použitím Parzenovy jádrové funkce (3.8).

Potom inspirace pro aproximaci neznámé distribuční funkce součtu je založena na postupu, kde je nejprve vyjádřen odhad distribuční funkce pro součet $(m + 1)$ náhodných veličin

$$\hat{F}_{m+1}(x; h) = \int_{-\infty}^{+\infty} \hat{F}_m(x - z; h) \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{z - x_i}{h}\right) dz\,, \quad h > 0, \quad n \in \mathbb{N}_+. \tag{3.83}$$

Zde je vhodné použít úpravu vzhledem k použité jádrové funkci hustoty a její definici, tj.

$$\begin{gathered} -a < \frac{z - x_i}{h} < a \\ -ah < z - x_i < ah \\ x_i - ah < z < x_i + ah, \end{gathered} \tag{3.84}$$

A předchozí výraz lze zjednodušit na tvar

$$\hat{F}_{m+1}(x; h) = \frac{1}{2anh} \sum_{i=1}^{n} \int_{x_i - ah}^{x_i + ah} \hat{F}_m(x - z; h) dz\,, \quad h > 0, \quad n \in \mathbb{N}_+. \tag{3.85}$$

Věta 3.6 *Nechť $\{x_1, \dots, x_{m+1}\}$ jsou realizace nezávislých a stejně rozdělených náhodných veličin $\{X_1, \dots, X_{m+1}\}$, potom neparametrický jádrový odhad (aproximace) distribuční funkce $\hat{F}_{m+1}(x; h)$ pro součet $(m + 1)$ náhodných veličin je ve tvaru*

$$\hat{F}_{m+1}(x; h) \cong \frac{1}{n} \sum_{i=1}^{n} \hat{F}_m(x - x_i; h)\,, \quad n \in \mathbb{N}_+ \tag{3.86}$$

[35] Konvoluce (skládání funkcí) představuje zjednodušeně matematický operátor (operace) zpracovávající dvě funkce.

a neparametrický jádrový odhad funkce hustoty $\hat{f}_{m+1}(x;h)$ *pro stejný počet prvků* $(m+1)$ *je*

$$\hat{f}_{m+1}(x;h) \cong \frac{1}{n}\sum_{i=1}^{n} \hat{f}_m(x-x_i;h)\,, \quad n \in \mathbb{N}_+. \tag{3.87}$$

To jsou poněkud jednodušší výrazy než vlastní konvoluce.

Důkaz. Celý postup je založen na výrazu (3.85) a jeho úpravě s použitím tzv. „lichoběžníkového pravidla[36]“ pro výpočet určitého integrálu u libovolné spojité funkce $g(x)$, které lze vyjádřit jako

$$\begin{aligned}\int_{x_1}^{x_2} g(x)dx &= (x_2-x_1)g(x_1) + \frac{(x_2-x_1)\big(g(x_2)-g(x_1)\big)}{2} \\ &= (x_2-x_1)\left[g(x_1) + \frac{1}{2}g(x_2) - \frac{1}{2}g(x_1)\right] \\ &= \frac{1}{2}(x_2-x_1)\big(g(x_2)+g(x_1)\big).\end{aligned} \tag{3.88}$$

Aplikací pravidla na modelové hodnoty distribuční funkce, kde $g(x_2) = \hat{F}_m(x-x_i-ah;h)$ a $g(x_1) = \hat{F}_m(x-x_i+ah;h)$ je po následném dosazení do uvedeného výrazu získána „přibližná hodnota“ součtu distribuční funkce. Pojem přibližná hodnota je použit z důvodu aplikace lichoběžníkového pravidla.

$$\begin{aligned}\hat{F}_{m+1}(x;h) &\cong \frac{1}{2anh}\sum_{i=1}^{n} ah\left(\hat{F}_m(x-x_i-ah;h) + \hat{F}_m(x-x_i+ah;h)\right) \\ &= \frac{1}{2n}\sum_{i=1}^{n}\left(\hat{F}_m(x-x_i-ah;h) + \hat{F}_m(x-x_i+ah;h)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\left(\hat{F}_m(x-x_i-ah;h) + \hat{F}_m(x-x_i+ah;h)\right)}{2} \\ &= \frac{1}{n}\sum_{i=1}^{n}\hat{F}_m(x-x_i;h)\,, \qquad h>0, \qquad n \in \mathbb{N}_+.\end{aligned} \tag{3.89}$$

■

Věta 3.7 *Nechť* $\{x_1,\dots,x_{m+1}\}$ *jsou realizace nezávislých a stejně rozdělených náhodných veličin* $\{X_1,\dots,X_{m+1}\}$*, potom odhad střední hodnoty* $E\{Z_{m+1}\}$ *pro náhodnou veličinu* Z *tvořenou součtem* $(m+1)$ *náhodných veličin je ve tvaru*

$$E_{\hat{f}_m(x;h)}\{Z_{m+1}\} = E_{\hat{f}_{m-1}(x;h)}\{Z_m\} + \bar{x}, \tag{3.90}$$

kde $\bar{x}$ značí aritmetický průměr.

[36] Lichoběžníkové pravidlo je pravidlo pro „přibližný“ numerický výpočet určitého integrálu u spojitých funkcí (aproximace pomocí součtu obsahů lichoběžníků).

Důkaz. Nechť $g(x)$ představuje hustotu pravděpodobnosti uvažované náhodné veličiny, potom platí vztah

$$\begin{aligned}\int_{-\infty}^{+\infty} xg(x-x_i)\,dx &=_a \int_{-\infty}^{+\infty} (y+x_i)g(y)\,dy \\ &= E\{X\}+x_i.\end{aligned} \tag{3.91}$$

Pomocí symbolu $=_a$ je označeno místo, kde je použita substituce $y = x - x_i$ pro zjednodušení výpočtu hodnoty integrálu. Odtud lze postupně odvodit střední hodnotu celého součtu

$$\begin{aligned}E_{\hat{f}_m(x;h)}\{Z_{m+1}\} &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} x\hat{f}_m(x-x_i;h)\,dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(E_{\hat{f}_{m-1}(x;h)}\{Z_m\}+x_i\right) \\ &= E_{\hat{f}_{m-1}(x;h)}\{Z_m\}+\bar{x}.\end{aligned} \tag{3.92}$$

Důsledek 3.3 *Z uvedeného odvození je zřejmé, že za podmínky* $E_{\hat{f}(x;h)}\{X\} = \overline{x}$ *platí vztah* $E_{\hat{f}_{m-1}(x;h)}\{Z_m\} = m\overline{x}$.

Důkaz. *Přímý důsledek předchozí věty 3.7.*

Věta 3.8 *Nechť* $\{x_1, \dots, x_{m+1}\}$ *jsou realizace nezávislých a stejně rozdělených náhodných veličin* $\{X_1, \dots, X_{m+1}\}$*, potom odhad hodnoty rozptylu* $\sigma^2\{Z_{m+1}\}$ *pro náhodnou veličinu Z tvořenou součtem* $(m+1)$ *náhodných veličin je ve tvaru*

$$\sigma^2_{\hat{f}_{m+1}(x;h)}\{Z_{m+1}\} = \sigma^2_{\hat{f}_m(x;h)}\{Z_m\} + \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2, \tag{3.93}$$

kde $\bar{x}$ *značí aritmetický průměr.*

Důkaz. Nechť $g(x)$ představuje hustotu pravděpodobnosti uvažované náhodné veličiny, potom platí vztah

$$\begin{aligned}\int_{-\infty}^{+\infty}(x-\mu)^2 g(x-x_i)\,dx &=_a \int_{-\infty}^{+\infty}\left((y-\mu)+x_i\right)^2 g(y)\,dy \\ &= \int_{-\infty}^{+\infty}[(y-\mu)^2 g(y)+2x_i(y-\mu)g(y)dy+x_i^2]g(y)\,dy \\ &= \sigma^2+x_i^2.\end{aligned} \tag{3.94}$$

Odtud lze pomocí následujícího postupu dostat požadovanou hodnotu pro rozptyl

$$\begin{aligned}\sigma^2_{\hat{f}_{m+1}(x;h)}\{Z_{m+1}\} &= \int_{-\infty}^{+\infty}(x-(m+1)\bar{x})^2\hat{f}_{m+1}(x;h)\,dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}(x-(m+1)\bar{x})^2\hat{f}_m(x-x_i;h)\,dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}[(y-m\bar{x})^2+2(y-m\bar{x})(x-x_i)+(x-x_i)^2]\hat{f}_m(y;h)\,dy \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\sigma^2_{\hat{f}_m(x;h)}\{Z_m\}-(x_i-\bar{x})^2\right] \\ &= \sigma^2_{\hat{f}_m(x;h)}\{Z_m\}+\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2.\end{aligned} \tag{3.95}$$

■

Důsledek 3.4 *Z této situace mohou nastat dvě různé varianty. Pokud použijeme pro odhad rozptylu první náhodné veličiny* $\sigma^2_{\hat{f}(x;h)}\{X\}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$*, potom platí*

$$\sigma^2_{\hat{f}_m(x;h)}\{Z_m\}=m\sigma^2_{\hat{f}(x;h)}\{X\}, \tag{3.96}$$

ale pokud je odhad rozptylu první náhodné veličiny $\sigma^2_{\hat{f}(x;h)}\{X\}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$*, potom je výsledná hodnota*

$$\sigma^2_{\hat{f}_{m+1}(x;h)}\{Z_{m+1}\}=\sigma^2_{\hat{f}_m(x;h)}\{Z_m\}+\left(\frac{n-1}{n}\right)\sigma^2_{\hat{f}(x;h)}\{X\}. \tag{3.97}$$

Výraz (3.97) je možné upravit na výraz s použitím $\sigma^2_{\hat{f}(x;h)}\{X\}$. Postup odvození je

$$\begin{aligned}\sigma^2_{\hat{f}_m(x;h)}\{Z_m\} &= \sigma^2_{\hat{f}(x;h)}\{X\}+(m-1)\left(\frac{n-1}{n}\right)\sigma^2_{\hat{f}(x;h)}\{X\} \\ &= \sigma^2_{\hat{f}(x;h)}\{X\}\left[1+\frac{(m-1)(n-1)}{n}\right] \\ &= m\sigma^2_{\hat{f}(x;h)}\{X\}\left(1-\frac{1}{n}\left(1-\frac{1}{m}\right)\right),\end{aligned} \tag{3.98}$$

kde prakticky pro dost velké n opět platí

$$\sigma^2_{\hat{f}_m(x;h)}\{Z_m\}=m\sigma^2_{\hat{f}(x;h)}\{X\}. \tag{3.99}$$

Tvrzení 3.3 *Pro* $n\in\mathbb{N}_+$ *a* $h>0$ *je daný odhad distribuční funkce součtů* $\hat{F}_Z(x)$ *za výše uvedených podmínek asymptoticky nestranný a vydatný.*

Důkaz. Nechť X je náhodná veličina se známou distribuční funkcí $F_X(x)$ a nechť Y je náhodná veličina s neznámou distribuční funkcí $F_Y(x)$ a dostupným náhodným výběrem $\{y_1,\dots,y_n\}$. Potom pro náhodnou veličinu $Z=(X+Y)$ vzniklou součtem obou náhodných veličin lze odhadnout distribuční funkci ve tvaru

$$\hat{F}_Z(x) = \frac{1}{n}\sum_{i=1}^{n} F_X(x - y_i). \tag{3.100}$$

A střední hodnotu odhadu distribuční funkce lze vyjádřit

$$\begin{aligned} E_{f_Y(y)}\{\hat{F}_Z(x)\} &= \frac{1}{n}\sum_{i=1}^{n} E\{F_X(x - y_i)\} \\ &= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{+\infty} F_X(x - y_i) f_Y(y_i)\, dy_i \\ &= \int_{-\infty}^{+\infty} F_X(x - y) f_Y(y)\, dy. \end{aligned} \tag{3.101}$$

Stručně řečeno $\hat{F}_Z(x)$ je za výše uvedených předpokladů nestranným odhadem distribuční funkce součtu dvou náhodných veličin $(X + Y)$, bodově pro $\forall x \in \mathbb{R}$.

Nechť dále předpokládáme nezávislý a stejně rozdělený náhodný výběr $\{y_1, \dots, y_n\}$, potom jsou nezávislé a stejně rozdělené i $F_X(x - y_1), \dots, F_X(x - y_2)$, pro $\forall x \in \mathbb{R}$. Dále dle předchozích odvození platí

$$\sigma^2_{\sum_{i=1}^{n} F_X(x - y_i)} = n\sigma^2_{F_X(x - y_i)}, \tag{3.102}$$

a

$$\sigma^2_{\hat{F}_Z(x)} = \frac{1}{n}\sigma^2_{\hat{F}_X(x - y_i)}, \quad i = 1, \dots, n. \tag{3.103}$$

Tedy $\hat{F}_Z(x)$ je i asymptoticky vydatným odhadem distribuční funkce $(X + Y)$.

3.7 Odhad „dvourozměrné“ funkce hustoty a distribuční funkce

Další oblastí neparametrických jádrových odhadů distribučních funkcí a hustot jsou vícerozměrné neparametrické jádrové odhady. Pro uvažované neparametrické jádrové modely k získání spolehlivosti (selhání) je ale nezbytné používat dvou nebo vícerozměrné modely z důvodu simulace vzájemného působení různých sil, jak je již popsáno v úvodu práce. Proto by bylo nyní vhodné nejprve definovat použitelný model a poté ho aplikovat na vícerozměrný neparametrický odhad distribuční funkce a funkce hustoty. Možná definice již byla publikována např. Hardle (1991, s. 79 - 82).

Definice 3.3 *Nechť $\{x_1, \dots, x_n\}$ je i.i.d. výběr jedné náhodné veličiny X_i z r-rozměrné náhodné veličiny X (zde je uvažován r −rozměrný náhodný vektor $X = (X_1, \dots, X_r)^T$, kde X_i je jednorozměrná náhodná veličina s náhodným výběrem právě rozsahu n), potom vícerozměrné modely pro neparametrické jádrové odhady hustoty a distribuční funkce jsou definovány jako*

$$\hat{f}(x_1, \dots, x_r; h_1, \dots, h_r) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{r}\frac{1}{h_j}k_j\left(\frac{x_j - x_{j,i}}{h_j}\right), \quad h_j > 0, \quad n \in \mathbb{N}_+ \tag{3.104}$$

a

$$\hat{F}(x_1, \dots, x_r; h_1, \dots, h_r) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{r}K_j\left(\frac{x_j - x_{j,i}}{h_j}\right), \quad h_j > 0, \quad n \in \mathbb{N}_+, \tag{3.105}$$

kde

$k_j(x_j)$ *je jádrová funkce hustoty pro* $j-tou$ *komponentu (veličinu)* x_j,

$K_j(x_j)$ *je jádrová distribuční funkce pro* $j-tou$ *komponentu (veličinu)* x_j,

h_j *je vyhlazovací parametr jádrové funkce v* $j-$*té komponentě (veličině), který má vliv na rozptyl jádrové funkce,*

$x_{j,i}$ *je* $i-$*té pozorování (realizace) z* $j-$*té komponenty (veličiny) a*

n *je počet pozorovaných dat v každé* $r-$*rozměrné komponentě (veličině), který je pro všechny komponenty stejný.*

Poznámka 3.16 *Vyjádření těchto definic vícerozměrných modelů se značně zjednoduší v případě volby jen jednoho vyhlazovacího parametru, jak uvedl např. Hardle a kol. (2004).*

Použití uvedeného vícerozměrného neparametrického jádrového modelu je z praktického pohledu omezeno na počet dimenzí $r = 2$. Pro vyšší číslo dimenze se ztrácí efekty neparametrického modelování a i v případě dvourozměrného modelu se jedná pouze o teoretický základ pro odvození dílčích semi-parametrických výsledků. Další negativní stránka vícerozměrných neparametrických modelů spočívá ve výpočetní složitosti, která roste s počtem dimenzí r. Někdy je výše uvedené tvrzení zesilováno. Uvedené postupy nejsou běžně aplikovány na vícedimenzionální data, jejichž dimenze je větší než 5, Hardle a kol. (2004). Opět se ale jedná o empirii, jako i u výroku na počátku tohoto odstavce.

Upravená definice pro dvourozměrné neparametrické jádrové modely odhadu neznámé hustoty a distribuční funkce mohou být vyjádřeny pomocí následujících vztahů, které představují zjednodušení předchozích modelů (3.104) a (3.105).

Definice 3.4 *Nechť* $\{x_1, \dots, x_n\}$ *je i.i.d. výběr náhodné veličiny* X *rozsahu* n *a nechť* $\{y_1, \dots, y_n\}$ *je i.i.d. výběr druhé náhodné veličiny* Y *opět rozsahu* n. *Pozorování v obou výběrech jsou párovány, tj. pozorujeme dvojice* (x_i, y_i). *Nechť dále* h_x *a* h_y *jsou jejich vyhlazovací parametry, potom dvourozměrný neparametrický jádrový model odhadu neznámé funkce hustoty a distribuční funkce je*

$$\hat{f}(x,y;h_x,h_y)=\frac{1}{nh_xh_y}\sum_{i=1}^{n}k_x\left(\frac{x-x_i}{h_x}\right)k_y\left(\frac{y-y_i}{h_y}\right),\quad h_x,h_y>0,\qquad n\in\mathbb{N}_+ \tag{3.106}$$

a

$$\hat{F}(x,y;h_x,h_y)=\frac{1}{n}\sum_{i=1}^{n}K_x\left(\frac{x-x_i}{h_x}\right)K_y\left(\frac{y-y_i}{h_y}\right),\quad h_x,h_y>0,\qquad n\in\mathbb{N}_{+.} \tag{3.107}$$

Ze vztahů uvedených v předchozích definicích je zcela evidentní, že obě aproximace jsou „konzistentní“ vůči operaci přechodu na jejich marginální hustoty a distribuční funkce.

Věta 3.9 *Dvourozměrná aproximace hustoty a distribuční funkce je konzistentní vůči operaci přechodu na jejich marginální hustoty a distribuční funkce.*

Důkaz. Konzistence mezi dvourozměrným a jednorozměrným neparametrickým jádrovým modelem pro odhad neznámé hustoty může být ukázána pomocí následujícího postupu:

$$\begin{aligned}\hat{f}(x;h_x)&=\int_{-\infty}^{+\infty}\hat{f}(x,y;h_x,h_y)dy\\&=\frac{1}{nh_xh_y}\sum_{i=1}^{n}k_x\left(\frac{x-x_i}{h_x}\right)\int_{-\infty}^{+\infty}k_y\left(\frac{y-y_i}{h_y}\right)dy\\&=\frac{1}{nh_x}\sum_{i=1}^{n}k_x\left(\frac{x-x_i}{h_x}\right),\end{aligned} \tag{3.108}$$

kde jsou všechny postupy zcela evidentní.

3.8 Využití neparametrických odhadů pro spolehlivost

Modely pro neparametrické jádrové odhady spolehlivosti jsou založeny na výše odvozených vztazích, zejména vztah pro případ nezávislosti náhodných veličin (2.35). Nadále bude předpokládáno, že je k dispozici n párů pozorování (x_i,y_i), tedy náhodný výběr párovaných pozorování pevného rozsahu n. Cílem této části je získání neznámých hodnot (pravděpodobností) spolehlivosti nebo selhání s použitím dvou vybraných typů jádrových funkcí. První typ jádrové funkce je opět Parzenova (obdélníková) jádrová funkce (3.8) a druhý typ Gaussova jádrová funkce (3.16).

3.8.1 Odhad spolehlivosti s použitím Gaussovy jádrové funkce

Nejprve je vytvořen model neparametrického jádrového odhadu spolehlivosti s použitím Gaussovy jádrové funkce (3.16) pro náhodný výběr párovaných pozorování (x_i,y_i) pevného rozsahu n, který je odvozen na základě dvourozměrného neparametrického jádrového odhadu hustoty. Dále je zde nezbytné upozornit na vlastnosti jednotlivých

„komponent“ (částí jádrového odhadu, které vstupují do výsledného výpočtu pro aritmetický průměr), kdy pro každou takovou uvažovanou „komponentu“ marginální funkce hustoty platí následující věta.

Věta 3.10 Nechť $\{x_1, \dots, x_n\}$ *je i.i.d. výběr náhodné veličiny* X *s její distribuční funkcí a funkcí hustoty, potom pro každou uvažovanou komponentu definovanou vztahem* $\frac{1}{h_x} k_x\left(\frac{x-x_i}{h_x}\right)$ *pro* $h_x > 0$ *z celkového součtu komponent (tj. sčítanec) vstupujících do neparametrického jádrového odhadu marginální funkce hustoty dostaneme jí odpovídající střední hodnotu a rozptyl:*

$$\begin{aligned} E_{k_x}\{X\} &= x_i, \\ E_{k_x}\{X^2\} &= x_i^2 + h_x^2, \\ \sigma_{k_x}^2\{X\} &= h_x^2. \end{aligned} \tag{3.109}$$

Důkaz. Zdůvodnění výše uvedených vztahů spočívá v postupném odvození příslušných výsledků s použitím substituce $\frac{x-x_i}{h_x} = y$ a předpokladem $h_x > 0$.

$$\begin{aligned} E_{k_x}\{X\} &= \int_{-\infty}^{+\infty} \frac{1}{h_x} x k_x\left(\frac{x-x_i}{h_x}\right) dx = x_i\,, \\ E_{k_x}\{X^2\} &= \int_{-\infty}^{+\infty} \frac{1}{h_x} x^2 k_x\left(\frac{x-x_i}{h_x}\right) dx = x_i^2 + h_x^2\,, \\ \sigma_{k_x}^2\{X\} &= h_x^2. \end{aligned} \tag{3.110}$$

■

Stručně lze říci, že střední (očekávaná) hodnota uvedené jádrové funkce (komponenty) je získané pozorování a rozptyl této funkce je kvadrát hodnoty vyhlazovacího parametru. Tyto odvozené vztahy lze nyní použít pro odhad neznámé hodnoty spolehlivosti.

Věta 3.11 *Nechť* X *a* Y *jsou spojité náhodné veličiny a nechť máme k dispozici náhodný výběr párovaných pozorování* $\{(x_1, y_1), \dots, (x_n, y_n)\}$ *pevného rozsahu* n*, potom výslednou hodnotu neparametrického jádrového odhadu spolehlivosti s použitím Gaussovy jádrové funkce a s využitím vlastností* (3.110) *pro dvourozměrné rozdělení pravděpodobnosti v případě dvou nezávislých náhodných veličin* (2.35) *lze vyjádřit jako*

$$R = P(X < Y) = \frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{y_i - x_i}{\sqrt{h_x^2 + h_y^2}}\right), \quad h_x, h_y > 0, \qquad n \in \mathbb{N}_+, \tag{3.111}$$

kde y_i *a* x_i *značí* i *−té párové pozorování,* n *je celkový počet párovaných pozorování,* h_x *a* h_y *jsou hodnoty vyhlazovacích parametrů (parametry měřítka).*

Důkaz. Neznámá hodnota spolehlivosti je vyjádřena jako $P(X < Y)$, jestliže dále použijeme odvozený vztah (2.35) pro dvourozměrné normální rozdělení pravděpodobnosti dvou nezávislých náhodných veličin a využijeme vlastnosti každé komponenty (3.110), potom výsledný tvar může být odvozen postupem:

$$\begin{aligned} R &= P(X < Y) \\ &= \iint\limits_{x<y} \hat{f}(x, y; h_x, h_y)\,dxdy \\ &= \frac{1}{n}\sum_{i=1}^{n} \iint\limits_{x<y} \frac{1}{h_x} k_x\left(\frac{x - x_i}{h_x}\right) \frac{1}{h_y} k_y\left(\frac{y - y_i}{h_y}\right) dxdy \\ &= \frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{y_i - x_i}{\sqrt{h_x^2 + h_y^2}}\right), \quad h_x, h_y > 0, \qquad n \in \mathbb{N}_+. \end{aligned} \tag{3.112}$$

Poznámka 3.17 *Uvedený postup pro získání neznámé hodnoty spolehlivosti je používán u dvourozměrného normálního rozdělení pravděpodobnosti s předpokladem nezávislosti náhodných veličin. Celý postup lze využít ale i v obecném případě (kde jsou předpokládány závislosti), protože dvourozměrný neparametrický jádrový odhad je tvořen z násobků jádrových funkcí, které mají nezávislý charakter a z toho důvodu odhadovaná hodnota spolehlivosti nepotřebuje mít funkci hustoty a distribuční funkci dvourozměrného normálního rozdělení pravděpodobnosti.*

3.8.2 Odhad spolehlivosti s použitím Parzenovy jádrové funkce

Použití Parzenovy jádrové funkce je založeno na stejném principu jako v předchozím případě. Opět je předpokládán náhodný výběr párovaných pozorování (x_i, y_i) pevného rozsahu n dvou náhodných veličin X, Y a cílem je odhadnout neznámou hodnotu spolehlivosti s použitím Parzenovy jádrové funkce (3.8). Uvedený přístup je technicky více náročný, a proto je zde vysvětlen podrobněji.

Věta 3.12 *Nechť X a Y jsou spojité náhodné veličiny s hustotou i distribuční funkcí a nechť máme k dispozici náhodný výběr párovaných pozorování $\{(x_1, y_1), \dots, (x_n, y_n)\}$ pevného rozsahu n, potom výslednou hodnotu neparametrického jádrového odhadu spolehlivosti s použitím Parzenovy jádrové funkce (3.8) a dvourozměrného jádrového modelu pro odhad hustoty (3.106) lze získat jako*

$$R = P(X < Y) = \frac{1}{n 4 a^2 h_x h_y} \sum_{i=1}^{n} I_i\,, \quad h_x, h_y > 0, \qquad n \in \mathbb{N}_+, \tag{3.113}$$

kde a značí parametr Parzenovy jádrové funkce, n je celkový počet párovaných pozorování, h_x a h_y jsou hodnoty vyhlazovacích parametrů a faktor I_i charakterizuje odpovídající prostor. Ten lze vyjádřit jako

$$I_i = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \chi(x<y)\chi(x_i - ah_x < x < x_i + ah_x)\chi\big(y_i - ah_y < y < y_i + ah_y\big)dxdy, \tag{3.114}$$

kde je pomocí symbolu $\chi(*)$ označena charakteristická funkce výroku definovaná na uvažované množině nebo matematický vztah, který označuje členství daného prvku v uvažované podmnožině.

Důkaz. K odvození modelu pro odhad neznámé hodnoty spolehlivosti je použita $P(X<Y)$ s dvourozměrným neparametrickým jádrovým modelem pro odhad hustoty (3.106) a vztah (2.35). Potom výsledný tvar může být odvozen následovně:

$$\begin{aligned} R &= P(X<Y) \\ &= \iint_{x<y} \hat{f}\big(x,y;h_x,h_y\big)dxdy \\ &= \frac{1}{n}\sum_{i=1}^{n}\iint_{X<Y}\frac{1}{h_x}k_x\left(\frac{x-x_i}{h_x}\right)\frac{1}{h_y}k_y\left(\frac{y-y_i}{h_y}\right)dxdy \\ &= \frac{1}{n4a^2h_xh_y}\sum_{i=1}^{n} I_i\,, \quad h_x,h_y>0, \qquad n\in\mathbb{N}_+. \end{aligned} \tag{3.115}$$

Použitý faktor I_i využívá charakteristickou funkci $\chi(*)$, která je rovna 1 nebo 0, pokud matematický výraz uvnitř je pravdivý nebo nepravdivý a celková plocha označovaná pomocí písmene A ve tvaru čtverce nebo obdélníku je odvozena ze vztahu

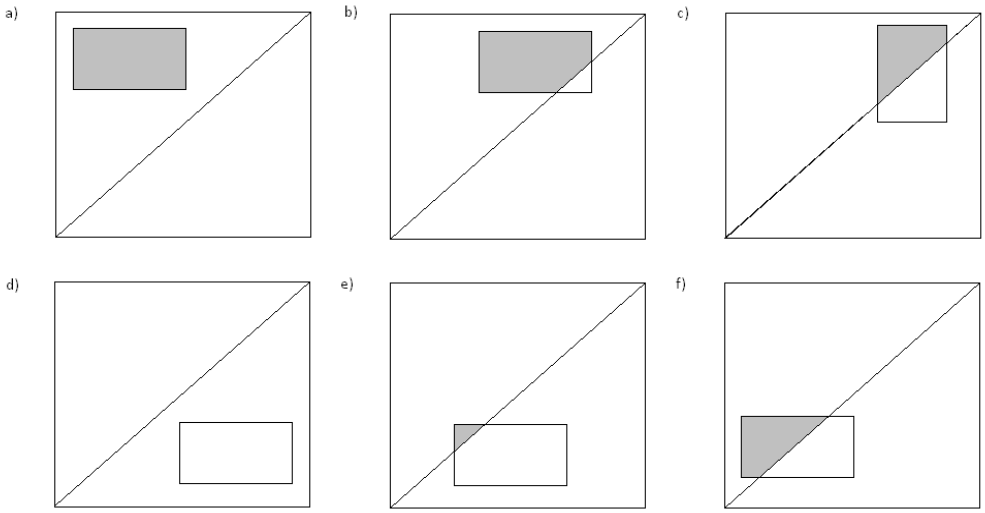
$$A = (x_i - ah_x, x_i + ah_x) * \big(y_i - ah_y, y_i + ah_y\big) = 4a^2h_xh_y. \tag{3.116}$$

Poznámka 3.18 *Uvedený parametr (konstanta) a představuje parametr Parzenovy (obdélníkové) jádrové funkce odvozený vztahem* (3.9).

Vzhledem k uvedenému postupu je vhodné se nyní zaměřit na přesnou interpretaci výpočtu neznámé hodnoty spolehlivosti $i-t\acute{e}ho$ páru pozorování (nebo $i-t\acute{e}$ komponenty). K předchozímu odvození obsahu celkové plochy pro každé párované pozorování je nutné získat obsah části plochy (faktor I_i), která splňuje podmínky charakteristických funkcí uvažovaného faktoru. Získaná hodnota faktoru je tedy plocha, která je dána vztahem (3.114) a při následném dosazení této plochy do podílu s celkovým obsahem plochy (3.116) je získána hodnota spolehlivosti (respektive její pravděpodobnost). Nelze opomenout fakt, že odvozené obsahy obou ploch jsou závislé na „mírách“ (vyhlazovacích parametrech) obou náhodných veličin. Nyní již zbývá získat výslednou hodnotu spolehlivosti z celého rozsahu získaných pozorování, kterou lze

odvodit jako podíl součtu všech hodnot (pravděpodobností) jednotlivých komponent a celkovým počtem získaných pozorování n.

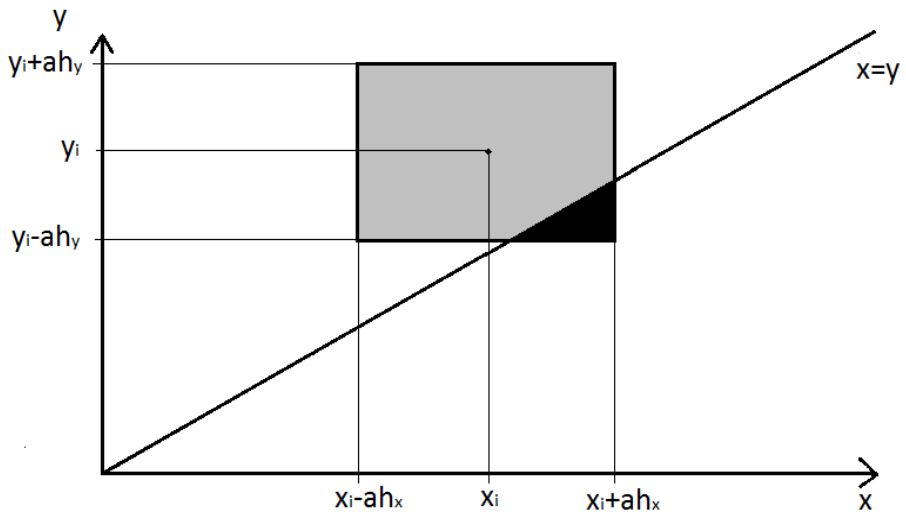
Uvedený postup lze stručně shrnout tak, že hodnota spolehlivosti (selhání) je tedy aritmetický průměr z hodnot jednotlivě „odhadnutých pravděpodobností“ a získané faktory I_i pro $i = 1, \dots, n$ charakterizují plochy označené šedou barvou na Obrázku 5, kde jsou zároveň pro lepší představivost vykresleny všechny možné uspořádání (pozice), které mohou pro každý pár pozorování nastat.



Obrázek 5: *Možné uspořádání (pozice) odhadovaných ploch pro odhad spolehlivosti s použitím Parzenovy jádrové funkce*

Vzhledem k důležitosti a možné nepřehlednosti tohoto obrázku v souvislosti s předchozím popisem je nyní vhodné se zaměřit na jeho obsah detailněji. Subjektivně je proto zvolena a vybrána jedna možná konfigurace (pozice) z Obrázku 5, na které jsou zobrazeny uvedené důležité souřadnice z používaných ploch při odhadu pro intuitivnější pochopení celé této problematiky a zároveň je vykreslena celková plocha označovaná jako A ze vztahu (3.116).

Dále je vhodné připomenout hodnotu uvedené konstanty $a = \sqrt{3}$ ze vztahu (3.9) a možný použitelný odhad neznámého parametru měřítka, který je odvozen vztahem (3.53). Výsledné souřadnice vybrané konfigurace jsou znázorněny na Obrázku 6.



Obrázek 6: *Souřadnice jedné vybrané konfigurace (komponenty)*

Poznámka 3.19 *Obrázek 6 ukazuje vybranou komponentu z postupu k získání neznámé hodnoty spolehlivosti pro i – tý pár pozorování, kde plocha označená šedou barvou splňuje všechny podmínky uvedené v* (3.114) *a pravděpodobnost této komponenty je získána podílem této plochy s hodnotou celkové plochy* (3.116)*, která je označena společně šedou a černou barvou.*

Kapitola 4

Využití modelů v aplikační sféře

Kapitola popisuje využití navrhovaných modelů v aplikační sféře. V úvodu je popsán zdrojový soubor dat z platební bilance České republiky, kde v případě nestacionárního (očekávaného) charakteru získaných pozorování je navržen model pro možnou aproximaci neznámé „trendové složky“ z předpokládaného aditivního heuristického modelu časové řady na základě vygenerovaného ortonormálního systému „indexových“ „čar“. Dále kapitola pojednává o tzv. „retrospektivním“ a „prediktivním“ trendu. Závěr kapitoly je zaměřen na pojmy nezápornosti platební bilance modelované spolehlivostí a na platební bilanci obecně, jak v aktuálních hodnotách, tak v kumulacích.

4.1 Platební bilance, stručný ekonomický význam

Cíl práce je zaměřen na oblast z aplikované matematiky, kterou charakterizuje konkrétní soubor reálných pozorování získaný z platební bilance České republiky, přesněji na hodnoty z její části nazývané obchodní bilance. Zde je spolehlivost reprezentována skutečností, že celková výše (např.) výdajů (za dovoz) není větší než celková výše příjmů (z vývozu) a naopak, selhání reprezentuje skutečnost, že celková výše výdajů je větší než celková výše příjmů za celé uvažované období od 1. 1. 2003 do 31. 5. 2012. Samozřejmě se jedná o dílčí pohled, ekonomicky důležitým jevem je nezáporná (záporná) celková platební bilance (např. za některé období).

V platební bilanci dané země (platební bilanci zahraničního obchodu) je zachycen mezinárodní pohyb statků, služeb, výrobních faktorů, pohledávek a závazků se zahraničím (v peněžním vyjádření). Představuje statistický výkaz, který systematickým způsobem zachycuje ekonomické transakce se zahraničím za určité časové období. Platební bilance má mnoho definic, např. z metodického listu ČNB[37] nebo v knize od autorů Neuman, Lamberský a Jiránková (2010).

Definice 4.1 *Platební bilance státu je statistický záznam (účetní výkaz) ekonomických transakcí subjektů dané země (rezidentů) s ekonomickými subjekty ze zbytku světa (nerezidenty) za určité období, zpravidla jeden rok.*

(převzato z Neumann, Žamberský a Jiránková (2010, s. 94)).

[37] Česká národní banka.

Samotná struktura platební bilance je v různých zemích a institucích nejednotná, proto zde popíšeme přístup ČNB, který vychází z „Příručky k sestavení platební bilance MMF[38] (5. vydání, 1993)“. Zde je platební bilance rozdělena do horizontální nebo vertikální struktury. K účelu této práce byla použita pozorování z horizontální struktury, proto je pozornost zaměřena na její stručný popis, respektive jen na běžný účet. Další informace a podrobné popisy lze nalézt přímo na internetových stránkách České národní banky, nebo ve výše uvedené literatuře.

Horizontální struktura platební bilance je rozdělena do následujících částí:

I. Běžný účet
   a. *Obchodní bilance*
   b. *Bilance služeb*
   c. *Bilance výnosů*
   d. *Běžné převody*
II. Kapitálový účet
III. Finanční účet
   a. *Přímé investice*
   b. *Portfoliové investice*
   c. *Finanční deriváty*
   d. *Ostatní investice*
IV. Rezervy
V. Chyby a opomenutí, kurzové rozdíly

Obchodní bilance (bilance zboží) zahrnuje hodnoty dovozu (import) a hodnoty vývozu (export) zboží do země a ze země ve vztahu: *obchodní bilance = vývoz zboží – dovoz zboží.* Výsledek bilance může být vyrovnaný, kladný (aktivní saldo, přebytek) a záporný (pasivní saldo, schodek). V případě záporného salda běžného účtu je nutné (do)financovat vzniklý deficit v rámci zbylých účtů (kde hlavní roli hraje finanční účet) nebo rezervy, které obsahují i půjčky z ciziny nebo od obyvatel ČR pomocí dluhopisů nebo pokladničních poukázek.

Bilance služeb zahrnuje vývoz a dovoz služeb ze země a do země a často je označována jako tzv. neviditelný obchod, který zahrnuje zejména dopravu včetně tranzitních poplatků, příjmy z cestovního ruchu, různé stavební a montážní práce, licence, patenty atd.

Poznámka 4.1 *Obchodní bilance spolu s bilancí služeb bývá často označována jako tzv. bilance výkonů dané země.*

Předposlední částí je bilance výnosů (důchody z výrobních faktorů), která zahrnuje např. příjmy zahraničních sezónních pracovníků, zisky, dividendy, úroky atd. a poslední část

[38] Původní znění této příručky lze nalézt: http://www.imf.org/external/np/sta/bop/bopcg.pdf

běžného účtu představují tzv. běžné platby, které zahrnují sociální transfery, převody pracovních příjmů cizinců, příspěvky mezinárodním organizacím atd.

Zdrojem informací v zahraničním obchodě jsou podklady získané celními orgány (nejen, viz dále Generální ředitelství cel) a následné zpracování, kontrolu a uveřejnění provádí Český statistický úřad.

Poznámka 4.2 *Platební bilance České republiky vedená v korunách používá pro kursový přepočet průměrný kurz za vykazované období dle údajů ČNB a je sestavována z podkladů obchodních bank, podnikové sféry, centrálních úřadů a institucí.*

Poznámka 4.3 *Platební bilance je vedena na základě podvojného účetnictví, proto musí být z pohledu účetního celku vždy vyrovnaná a k tomu slouží tzv. devizové rezervy. Pokud ve sledovaném období nedojde k poklesu ani k růstu devizových rezerv*[39] *centrální banky, potom je platební bilance v rovnováze (při nulové změně devizových rezerv je běžný účet v rovnováze s finančním účtem platební bilance, tj. pohyb měnového kurzu způsobuje vyrovnávání schodku běžného účtu s přebytkem finančního účtu).*

Získané soubory reálných dat z platební bilance byly převzaty ze statistických výkazů ČNB, kdy banka oficiálně zveřejňuje tyto údaje od roku 1999. Platební bilance je zveřejňována v uvedeném členění čtvrtletně v Kč[40], EUR[41], USD[42] od roku 1993 jako jednotlivá čtvrtletí nebo kumulovaná a vzhledem k požadavkům na co nejobsáhlejší soubor reálných dat byla zvolena jednotlivá měsíční pozorování, která jsou zveřejňována až od roku 2003 v milionech Kč.

4.2 Úvod do modelování časových řad

Výše uvedená metodika předpokládala i.i.d pozorování, tj. stacionární charakter získaných dat ze statistického úhlu pohledu. Stacionarita reálných dat je zde uvažována v předpokladu, že marginální rozdělení jednotlivých pozorování (časových odečtů) jsou (funkčně nebo parametricky) nezávislá na čase realizace (čase zjištění). Jestliže tento předpoklad není splněn (což představuje obvyklý jev v reálných (makro)ekonomických souborech dat), potom je nestacionarita modelem systematických, případně náhodných cenových a objemových změn. Tento jev si lze představit jako situaci, kdy existují alespoň dva časové odečty, u nichž nelze předpoklad stejného rozdělení přijmout. Proto jsou nejprve popsány základní metody analýzy časových řad včetně podrobnějšího popisu „trendové složky“ (komponenty) nebo často skloňovaného pojmu „trend“, protože bude předpokládán model obsahující jen trendovou a náhodnou složku.

[39] Devizové rezervy představují zásobu zahraničních měn, cenných papírů, zlata atd., které lze použít pro intervenci na měnovém trhu.
[40] Zkratka interpretuje měnu České republiky.
[41] EUR představuje zkratku pro měnu Evropské unie s názvem Euro.
[42] USD představuje zkratku pro měnu Spojených států americký s názvem americký dolar.

Časové řady jsou ve své podstatě velmi užitečné a důležité nejen pro statistické analýzy, ale i v oblasti ekonomické teorie pro mikroekonomické i makroekonomické analýzy a predikce. Při bližším zaměření na problematiku časových řad se hlavní zájem soustředí na vybranou „trendovou složku“ z důvodu častého nestacionárního charakteru získaného souboru pozorování. V uvedené teoretické oblasti již bylo prezentováno mnoho definic a jedna možná interpretace je definována v Brockwell a Davis (2006).

Definice 4.2 *Časová řada je takový soubor pozorování x_i, kde je každé pozorování zaznamenáváno ve specifickém čase t. Diskrétní časová řada je taková řada, kde množina realizací T_0, ve kterých byly provedeny jednotlivé pozorování, je diskrétní množina, jako například, když jsou jednotlivá pozorování provedena v konstantních časových intervalech. “Spojité časové řady” jsou získávány, když jsou pozorování zaznamenávány spojitě v určitém časovém intervalu, např. když $T_0 = [0,1]$.*[43]

(citace přeložena z Brockwell a Davis (2006), s. 1)

Poznámka 4.4 *V této práci používané označení $x(t)$ charakterizuje diskrétní časovou řadu, ve které jsou jednotlivá pozorování zaznamenávány v diskrétních časových okamžicích pro $t = 1, 2, \dots, n$. Uvedená skutečnost je dána charakterem souboru získaných pozorování z ČNB, kde je interval mezi jednotlivými naměřenými hodnotami jeden měsíc.*

Dále mohou být časové řady klasifikovány mezně jako stochastického nebo deterministického (nestochastického) typu, kde závisí na obsahu náhodné složky.

Poznámka 4.5 *Časové řady používané a testované v této práci jsou stochastické a obecně nestacionární z důvodu vzniklých „nákladů“ na peníze (inflační projevy, nejen), které se mění v čase. Jinými slovy, je zde zahrnuta inflace*[44]*, která z časového pohledu snižuje hodnotu (kupní schopnost nominální hodnoty) peněz.*

Časové řady se obecně „těší velké oblibě“ a lze se s nimi setkat v mnoha vědeckých disciplínách např. statistice, zpracování signálu, ekonometrii a finanční matematice atd. V současné době jsou k dispozici různé přístupy k analýze časových řad, které mohou být nalezeny v dostupné literatuře, jako např. Kvasnička a Vašíček (2001). Autoři zde publikují několik verzí možných analýz použitelných pro získané časové řady:

- *Expertní (kvalitativní) metody*
  Jsou vhodným nástrojem všude tam, kde není možné rozumně kvantifikovat sledované veličiny nebo vlivy, které působí na jejich vývoj.

[43] A time series is a collection of observations x_i, each one being recorded at a specified time t. A discrete-time series is one in which the set T_0 of times at which observations are made is a discrete collection, as is the case for example when observations are made at fixed time intervals. “Continuous-time series” are obtained when observations are recorded continuously over some time interval, e.g. when $T_0 = [0,1]$.

[44] Inflace představuje časové znehodnocení absolutní hodnoty peněz, tedy růst cenové hladiny v uvažovaném časovém období, což způsobuje snížení kupní síly peněz.

- *Grafická analýza*
  Nejjednodušší způsob analýzy časové řady, která je reprezentována výběrem vhodného typu grafu.
- *Dekompozice (rozklad) časových řad*
  Vychází z předpokladu, že náhodný proces, který generuje časovou řadu, je závislý pouze na čase. Dále předpokládá, že časovou řadu je možné rozdělit na několik nezávislých složek.
- *Box-Jenkinsova analýza*
  Tato metoda je založena na analýze náhodné složky, která může být tvořena závislými veličinami.
- *Spektrální analýza*
  Vychází z předpokladu, že časová řada je nekonečnou směsí sinusových a cosinusových křivek s různými frekvencemi a amplitudami.
- *Lineární dynamické (ekonometrické) modely*
  Jedná se o kauzální modely, kde je vysvětlovaná proměnná vysvětlována pomocí jedné nebo více vysvětlujících proměnných.
- *Atd.*

(převzato z Kvasnička a Vašíček (2001), s. 13)

Uvedené přístupy jsou podrobněji vysvětleny a popsány v knihách Kvasnička a Vašíček (2001) nebo Cipra (2008), kde jsou obsaženy i konkrétní příklady s použitím reálných dat. V práci je používána metoda dekompozice časových řad, která je založena na dekompozici časových řad do několika „hlavních“ komponent, které jsou charakterizovány určitými vlastnostmi a způsobem chování. Další nespornou výhodou této metody je následná schopnost monitorovat a predikovat vývoj celé časové řady nebo jednotlivých „základních“ komponent, jak je využíváno v závěru kapitoly.

Poznámka 4.6 *Nechť $x(t)$ je časová řada získaná z diskrétních časových pozorování pro $t\epsilon\mathbb{Z}$ a $t_i < t_{i+1}$, potom za hlavní komponenty metody dekompozice časových řad mohou být označeny:*

- *Trend $X(t)$,*
- *Cyklická složka $C(t)$,*
- *Sezónní složka $S(t)$,*
- *Reziduální (náhodná, zbytková, …) složka ε_t.*

(převzato a upraveno z Cipra (2008))

Trendová složka $X(t)$ charakterizuje dlouhodobé průměrné změny vybraného statistického ukazatele a vzhledem ke skutečnosti, že se jedná o jeden z hlavních pojmů v této práci, bude pojem „trend“ více diskutován v následující podkapitole. *Sezónní složku* $S(t)$ lze charakterizovat jako periodicky se opakující odchylky z trendové složky,

ke kterým dochází pravidelně za uvažované období, nejčastěji za každý rok (ekonomická data). Zde jsou uvažovány tzv. sezónní výkyvy jako např. vztahy mezi sezónní prací a nezaměstnaností. *Cyklická složka* $C(t)$ představuje nejspornější a na odhad nejnáročnější složku v časových řadách a vyznačuje se opakovanými, ale neperiodickými výkyvy, které jsou často nepravidelné cykly s proměnnými periodami (tím jsou myšleny délky jednotlivých period) a amplitudami (tím jsou myšleny různé „výšky“). A poslední z výše uvedených složek je *náhodná složka* ε_t , která je tvořena náhodnými pohyby (fluktuacemi) a často je nazývána ve statistice jako „bílý šum“ (za běžných předpokladů na něj kladených).

Poznámka 4.7 *Při použití metody dekompozice časových řad nemusí uvažovaná časová řada obsahovat všechny výše uvedené složky, ale jen některé z nich.*

Definice 4.3 *Při dekompozici se vychází ze tří různých typových modelů časové řady: aditivního, multiplikativního a smíšeného. Tyto modely specifikují, jakým způsobem jsou jednotlivé složky časové řady „skloubeny“ dohromady.*

1) *Aditivní model předpokládá, že výsledná časová řada je součtem jednotlivých složek. Model této řady má tvar*

$$x(t) = X(t) + S(t) + C(t) + \varepsilon_t. \tag{4.1}$$

2) *Multiplikativní model na druhou stranu předpokládá, že výsledná časová řada je spíše součinem jednotlivých složek*

$$x(t) = X(t) \times S(t) \times C(t) \times \varepsilon_t. \tag{4.2}$$

3) *Smíšený model je vlastně „jen“ kombinací obou předchozích přístupů. Některé složky mohou být v součtu, jiné v součinu. Typickým příkladem může být třeba takovýto model časové řady*

$$x(t) = X(t) \times S(t) \times C(t) + \varepsilon_t. \tag{4.3}$$

(převzato z Kvasnička a Vašíček (2001), s. 55)

Poznámka 4.8 *Předpokládaný aditivní model pro účely této práce (viz Kapitola 2) obsahuje trendovou (systematickou) a náhodnou (nesystematickou) složku:*

$$x(t) = X(t) + \varepsilon_t. \tag{4.4}$$

4.3 Problematika a potíže trendu

„Trend“ je populární a frekventované slovo na celém světě. Tento pojem je všude-přítomný kolem nás např. v makroekonomii, mikroekonomii a ve všech aplikovaných podoblastí ekonomie, statistiky, sociologie, financí, obchodu a samozřejmě hraje neméně podstatnou roli mezi novináři, u kterých je používání tohoto slova nebezpečné.

V současné době je možné obstarat dostatek literatury pojednávající o popisované problematice, protože se jedná o důležitý pojem v řadě výzkumů a tvrzení.

Problematika pojmu je v jeho (ne)jednoznačnosti a přirozenosti. Většina lidí toto slovo zná a dokáže si pod ním „něco“ představit, ale jaká je jeho jedinečná a správná definice nebo co přesně právě tento pojem znamená a jaká je jeho správná interpretace již tak jednoznačné není. Při analýze časové řady libovolné veličiny jsou předmětem zájmu informace, jaký byl dosavadní vývoj, jaká hodnota je dnes ve vztahu k minulosti a jaká hodnota uvažované časové řady by mohla být v budoucím (následujícím a neznámém) vývoji. Odpovědi na tyto otázky mohou být interpretovány různými trendovými křivkami, které uspokojují možné požadavky a které nám umožňují provádět „informované a odůvodněné“ rozhodnutí o budoucím vývoji. Slova „informované a odůvodněné“ jsou uvedeny v uvozovkách, protože budoucnost a budoucí vývoj je často (pokud neuvažujeme deterministický vývoj) přesně nepředpověditelný a obsahuje mnoho náhodných vlivů. Vzhledem k těmto nejednoznačnostem je jedna možná definice publikována v Phillips (2010).

Definice 4.4 *Trendová křivka sumarizuje, kde jsme doposud byli, kde jsme nyní ve vztahu k minulosti a ze všeho nejvíce ukazuje náznak, jakým směrem bychom se mohli v budoucnu odebírat. Jednotlivé křivky, které máme v našich myslích jako ty, které jsou namalované na papíře nebo jsou vytvořeny ekonometrickými metodami, jsou typicky vyhlazené a provedená derivace charakterizuje směrový vektor pro budoucí možný vývoj. Křivky vyhlazující naměřená data ukazují vlastnosti jako je dlouhodobá tendence přes naměřené časové hodnoty, cyklické vzory nebo body zvratu, které mohou být spojeny se známými událostmi a tím přispět k posílení jejich hodnot pro nás.*[45]

(citace přeložena z Phillips (2010), s. 3)

Nebo lze použít jinak formulovanou definici, která je publikována v Cipra (2008).

Definice 4.5 *Trend je křivka odrážející dlouhodobé změny v průměrném chování časové řady jako je dlouhodobý růst nebo pokles.*

(citováno z Cipry (2008), s. 16)

V současné době jsou jedním z velmi diskutovaných témat klimatické změny na zemi. Jestliže je tento problém analyzován do detailu, potom někteří vědci (lidé) chápou tyto změny jako dlouhodobé (například v zemědělství) a jiní jako krátkodobé změny (např.

[45] A trend line summarizes where we have been, shows where we are now in relation to the past, and, most of all, reveals a hint of where we are going. The lines we draw in our minds like those we draw on paper or fit by econometric methods are typically smooth and the derivative is a direction vector for the future. Lines through the data reveal features like a long run tendency to increase over time, a cyclical pattern, or turning points that can be associated with known events, thereby helping to reinforce their value to us.

v geologii). Tedy charakter použitého významu může být pro různé skupiny lidí odlišný, a proto je vhodné si zde význam přesněji specifikovat.

Poznámka 4.9 *Pojem "trend" použitý v této práci lze interpretovat jako „vhodnou" vyhlazovací (spojitou) křivku reflektující dlouhodobé změny v pohybu získaných reálných pozorování. Odhadnutá křivka má za cíl sumarizovat, kde se získané hodnoty doposud pohybovali, kde se tyto hodnoty nacházejí nyní ve vztahu k minulosti a kde se mohou s návaznou predikcí nacházet v budoucnosti.*

Jiné možné použití tohoto slova je v oblasti prodejních objemů, technologických změn ve výrobě, změnách v počtu obyvatel, změnách v průměrných mzdách, cen akcií atd.

4.4 Aditivní heuristický model časové řady

Motivace následující tvorby je založena na nestacionárním charakteru získaného souboru reálných pozorování. Získané hodnoty spolehlivosti mohou být ovlivněny „povahou" takových dat a mohou nabývat zkreslených hodnot z důvodu inflace, nejrůznějších změn v objemu produkce, kurzových změn nebo dalších ovlivňujících faktorů.

Oblast dalšího zájmu spočívá v eliminaci nestacionarity, tj. v aproximaci (modelování) neznámé „trendové křivky" ze získaného souboru pozorování s využitím vygenerované soustavy dat, který tvoří ortonormální systém (vektorů) v prostoru se skalárním součinem, jak definoval např. Deutsch (2001).

Definice 4.6 *Nechť V je lineární (vektorový) prostor nad tělesem reálných čísel $\mathbb{R}$, potom tento lineární prostor[46] V lze označit za prostor se skalárním součinem, jestliže pro každou dvojici jeho prvků (vektorů) $u, v \in V$ je zde definován skalární součin $\langle u, v \rangle$ mající následující vlastnosti (pro každé $u, v, z \in V$ a $\alpha \in \mathbb{R}$):*

1. $\langle u, u \rangle \geq 0$,
2. $\langle u, u \rangle = 0$ *tehdy a jen tehdy* $u = 0$,
3. $\langle u, v \rangle = \langle v, u \rangle$,
4. $\alpha \langle u, v \rangle = \langle \alpha u, v \rangle$,
5. $\langle u + v, z \rangle = \langle u, z \rangle + \langle v, z \rangle$.

(převzato z Deutsch (2001), s. 2-3)

Použitá symbolika $\langle ., . \rangle$ označuje skalární součin v lineárním vektorovém prostoru V.

[46] Definice vektorového prostoru V je uvedena např. ve Williams (2008, s.210), kde *"vektorový prostor je taková množina prvků V, které se nazývají vektory, mající operace sčítání a skalární násobení na ní definované"*.

Poznámka 4.10 *Lineární vektorové prostory nad* $\mathbb{R}$ *se skalárním součinem jsou nazývány jako Euklidovské prostory a značeny včetně dimenze* $\mathbb{R}^n$ *nebo „pre-Hilbertovy" prostory, jak uvádí (a podrobně popisuje) např. Deutsch (2001).*

V prostoru s takto zavedeným skalárním součinem lze (nejen) zavést normu, ze které je možné dále odvodit metriku. Popis a definice Euklidovské normy a metriky je opět prezentována ve velkém množství literatury, a proto je vybrána jedna definice z Williams (2008).

Definice 4.7 *Euklidovská norma, která v Euklidovském prostoru* $\mathbb{R}^n$ *značí velikost (délku) vektoru* $u = (u_1, \dots, u_n)$ *je označována jako* $\|u\|$ *a definována vztahem*

$$\|u\| = \sqrt{u_1^2 + \cdots + u_n^2}. \tag{4.5}$$

(převzato z Williams (2008), s. 46)

Poznámka 4.11 *Vektorová norma může být přepsána pomocí výše definovaného skalárního součinu jako*

$$\|u\| = \sqrt{\langle u, u \rangle}. \tag{4.6}$$

Definice 4.8 *Nechť* $u = (u_1, \dots, u_n)$ *a* $v = (v_1, \dots, v_n)$ *jsou dva body v* $\mathbb{R}^n$*. Rozdíl (vzdálenost) mezi uvedenými body* u *a* v *je označován jako* $d(u, v)$ *a je definován vztahem*

$$d(u, v) = \|u - v\| = \sqrt{(u_1 - v_1)^2 + \cdots + (u_n - v_n)^2}. \tag{4.7}$$

(převzato z Williams (2008), s. 51)

Nechť je k dispozici výše definovaný lineární vektorový prostor se skalárním součinem, normou a metrikou, potom lze definovat Hilbertův prostor s použitím tzv. úplnosti vzhledem k normě. Úplnost je vyjádřena pomocí Cauchyova kritéria pro posloupnosti.

Definice 4.9 *Hilbertův prostor je prostor se skalárním součinem na oboru reálných čísel, ve kterém každá Cauchyovská posloupnost konverguje k nějakému prvku z tohoto prostoru.* [47]

(citace přeložena z Cornwell (1997), s. 286 – 287)

Euklidovské prostory lze označit za Hilbertovy prostory konečné dimenze. Uvedenou definici Hilbertova prostoru včetně detailního popisu vlastností lze nalézt např. v Young

[47] A Hilbert space is a real inner product space in which every Cauchy sequence converges to an element of the space.

(1988) nebo Kufner (1973). V celé následující práci bude používán Euklidovský prostor se skalárním součinem.

4.4.1 Model odhadu trendové složky

Odvození aditivního heuristického modelu k odhadu neznámé trendové složky je založeno na základních znalostech makroekonomické „teorie“. Tato teorie předpokládá dostupnost různých makroekonomických časových řad. Některé z makroekonomických peněžních časových řad mohou být popsány následujícím, složkovým, indexovým, modelem

$$X(t) = X_0 \sum_{i=0}^{m} a_i (1 + r_i)^t + \varepsilon_t, \qquad t = 1, 2, \dots, T, \qquad r_i > -1, \tag{4.8}$$

kde je pomocí faktoru $a_i(1 + r_i)^t$ označován konkrétní růst nebo pokles v některé složce, např. objemu, ceně, měnovém kurzu a ostatní změny počátečního objemu označovaného jako X_0 a pomocí indexů r_i jsou označovány meziroční přírůstky v případě, že $a_i > 0$ nebo meziroční ztráty (úbytky) či stagnace v případě, že $a_i \leq 0$. Dále je zde používán parametr t charakterizující čas a parametr m, který označuje počet dílčích složek.

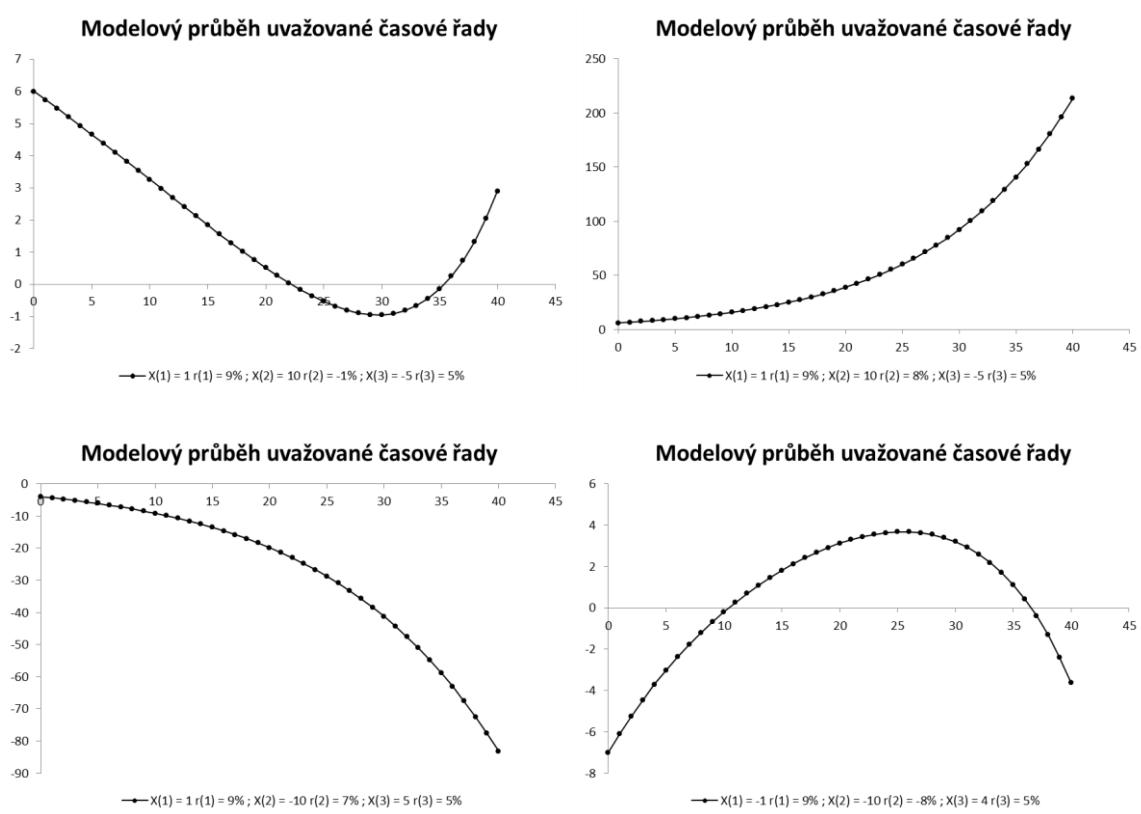
Velikost části, která podléhá vlivu modelovaného indexu r_i je dána hodnotou parametru $|a_i|$ a poslední proměnná z tohoto vztahu je časová řada náhodné složky ε_t, kde požadavky na její vlastnosti budou doplněny v průběhu následujícího textu.

Předchozí model vyjádřený vztahem (4.8) lze dále přepsat do nového upraveného tvaru, ve kterém je použita nová konvence v označování.

$$\begin{aligned} X(t) &= X_0 \sum_{i=0}^{m} a_i (1 + r_i)^t + \varepsilon_t \\ &= \sum_{i=1}^{m} X_i (1 + r_i)^t + \varepsilon_t, \qquad t = 1, 2, \dots, T, \qquad r_i > -1, \end{aligned} \tag{4.9}$$

kde použitá konvence je $X_i = X_0 a_i$. Uvedený vztah (popis) lze považovat za jeden možný z několika modelů finančních časových řad a může být nadále používán k odhadu neznámé a předpokládané trendové složky. Jinými slovy lze říci, že se jedná o aditivní heuristický model časové řady, tedy typový ekonometrický model.

Pro představu o obecnosti uvažovaného modelu jsou uvedeny na následujícím obrázku některé možné vývoje popisované časové řady (jejího „trendu“).



Obrázek 7: *Modelové průběhy aditivního heuristického modelu časové řady*

4.4.2 Formulace problému

Obecně jsou k dispozici „typové“ posloupnosti časových řad (průběhů, …) označované jako $f_i(t)$, kde:

1. $i = 0, 1, \dots, m$
2. $t = 1, \dots, T$
3. $m \leq (T - 1)$
4. $\forall i \in \{0, 1, \dots, m\} \exists t \in \{1, \dots, T\}: f_i(t) \neq 0.$

Podstata modelů těchto časových řad je v jejich aplikační oblasti, ze které pocházejí případná řešení a data tohoto problému. Navíc bude o takových posloupnostech předpokládáno, že se jedná o lineárně nezávislou množinu.

Definice 4.10 *Nechť $v_1, \dots, v_m$ jsou vektory v lineárním vektorovém prostoru V, potom množina těchto vektorů je lineárně závislá, jestliže zde existují nenulová čísla $c_1, \dots, c_m \in \mathbb{R}$ z nichž alespoň jedno je nenulové taková, že*

$$c_1 v_1 + \cdots + c_m v_m = 0. \tag{4.10}$$

Množina těchto vektorů je lineárně nezávislá, jestliže

$$c_1 v_1 + \cdots + c_m v_m = 0 \tag{4.11}$$

může být splněno jen pro $c_1 = 0, \dots, c_m = 0$.

(převzato z Williams (2008), s. 225)

Pro účely formulace tohoto problému lze obecně říci, že podstatou detekce (odhadu) trendové složky je její vytvoření ve formě lineární kombinace popsané jako

$$X(t) = \sum_{i=0}^{m} a_i f_i(t), \quad i = 1, \dots, m, \quad t = 1, 2, \dots, T. \tag{4.12}$$

Důkaz. Nechť $f_i(t) = (1 + r_i)^t$ je množina funkcí definovaná na množině $t \in S = \{0,1,2, \dots\}$, kde $-1 < r_1 < r_2 < r_3 < \cdots < r_m$.[48] Dále zde bude předpokládáno, že uvažovaná množina funkcí je lineárně nezávislá na definované množině. Důkaz lineární nezávislosti může být založen na metodě důkazu sporem.

Nadále je tedy předpokládáno, že jsou zde taková čísla $\{a_1, \dots, a_m\} \in \mathbb{R}$, že

$$\sum_{i=1}^{m} a_i (1 + r_i)^t = 0, \quad t \epsilon S. \tag{4.13}$$

Z tohoto předpokladu plyne, že to musí platit i pro $t \in \{0,1,2, \dots, (m-1)\}$. Poté ale soustava rovnic

$$\sum_{i=1}^{m} a_i (1 + r_i)^t = 0, \quad t \in \{0,1,2, \dots, m-1\} \tag{4.14}$$

má nenulové řešení s ohledem na uvedená čísla $\{a_1, \dots, a_m\}$ právě tehdy, když je její determinant roven nule.

Tvrzení 4.1 *Nechť* $A = \{a_{ij}\}$ *je čtvercová matice* $(n \times n)$ *a nechť* A_{ij} *je čtvercová matice* $(n-1) \times (n-1)$ *získaná vynecháním* $i-$*tého řádku a* $j-$*tého sloupce z matice* A. *Potom determinant* A, *označovaný jako* $\det A$ *nebo* $|A|$, *je definován jako*

$$|A| = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} |A_{ij}|, \tag{4.15}$$

kde i, j *může být libovolně vybráno jako celé číslo od* 1 *do* n *aniž by se měnila hodnota determinantu* $|A|$. *Výraz* $(-1)^{i+j} a_{ij} |A_{ij}|$ *se nazývá „cofaktor*[49]*“ prvku* a_{ij}.

(převzato z Amemiya (1994), s. 261)

[48] Omezení $-1 < r_1$ je z ekonomické povahy věci. Budeme předpokládat, že základní veličiny mohou klesnout nanejvýš o 100%. Tento předpoklad má svou váhu v peněžních vyjádřeních.

[49] Definice tohoto výrazu je prezentována např. v Amemiya (1994).

Ale determinant takové soustavy je determinantem čtvercové matice $(1+r_i)^j$ pro $i = 1, \dots, m$ a $j = 0, \dots, (m-1)$, což je tzv. Vandermondův determinant. Vandermondův determinant není v tomto případě ($r_1 < r_2 < r_3 < \cdots < r_m$) roven nule a tento závěr je tedy ve sporu s výše uvedeným předpokladem.

Tvrzení 4.2 *Nechť V_m je Vandermondova matice[50], která je definována následujícím vztahem*

$$V_m = \begin{bmatrix} 1 & 1 & \dots & 1 \\ a_1 & a_2 & \dots & a_n \\ a_1^2 & a_2^2 & \dots & a_n^2 \\ \dots & \dots & \dots & \dots \\ a_1^{n-1} & a_2^{n-1} & \dots & a_n^{n-1} \end{bmatrix}, \tag{4.16}$$

kde $a_i \in \mathbb{R}$. Potom Vandermondův determinant čtvercové Vandermondovy matice V_m definované výše je

$$|V_m| = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} (a_j - a_i), \tag{4.17}$$

kde V_m je regulární (není singulární) právě tehdy, když všechna reálná čísla a_i jsou odlišná.

(založeno na knize Bečvář (1981), s. 35)

4.5 Přirozené soustavy bazických časových průběhů

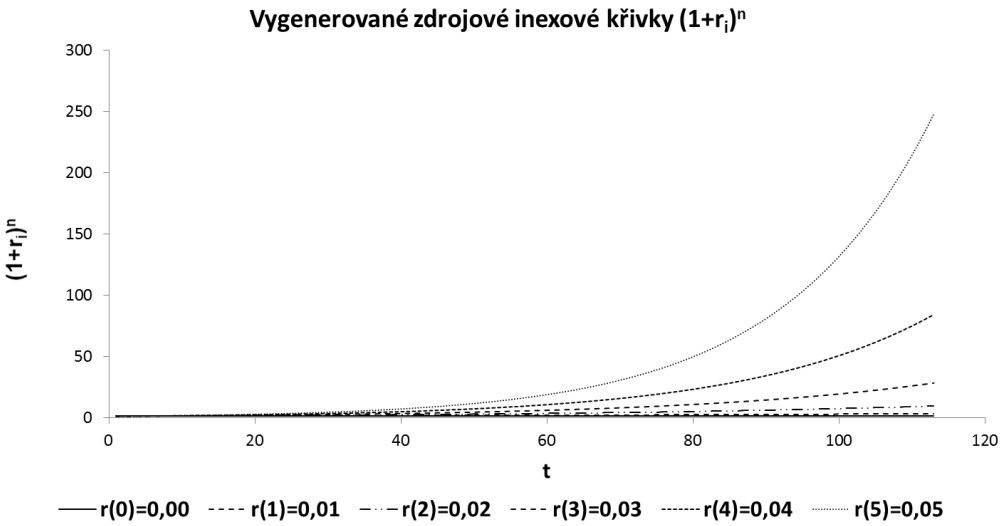
Nyní je vhodná doba pro výběr jedné soustavy (libovolné) vyhovujících posloupností $f_i(t)$, pro $t = 1, \dots, T$ a $i = 1, \dots, m$, ke konstrukci ortonormálního rozvoje, protože tento výběr představuje první krok v dále uvedené metodě pro odhad neznámé trendové složky. Při výběru vhodného lineárně nezávislého systému je obecně k dispozici velký počet možných zdrojových souborů posloupností, ze kterých může být daný výběr proveden. Pro účely této práce zde byly vybrány tzv. indexové čáry (někdy známé jako úrokové čáry), jejichž vyjádření včetně přepisu s využitím přirozeného logaritmu je ve tvaru

$$f_i(t) = (1+r_i)^t = e^{t\, ln(1+r_i)} = e^{t\, S_i}, \tag{4.18}$$

kde pomocí parametrů r_i jsou označeny uvažované indexy (mezi-obdobní) a t pro $t \in \mathbb{N}_0$ označuje čas. Symbol S_i prezentuje provedenou úpravu pro přehlednost při softwarové realizaci $S_i = ln(1+r_i)$. Grafická interpretace vybraného systému je vykreslena na Obrázku 8, kde hodnoty vynesených čísel na obou osách x a y jsou pouze v obecné rovině, tj. reálná čísla na základě zvolené posloupnosti časových řad (zvoleného

[50] Za povšimnutí stojí fakt, že matice V není obecně čtvercová. Čtvercová matice je nezbytná k výpočtu determinantu ve výše uvedeném důkazu.

systému). Časová osa je reprezentována jednotlivými měsíci a první zobrazovaný měsíc je považován za 0.



Obrázek 8: *Vygenerované zdrojové indexové (úrokové) čáry* $(1 + r_i)^t$

4.6 Ortonormální soustavy (ONS) bazických časových průběhů odvozené z přirozených

Nyní je k dispozici lineárně nezávislý systém (množina prvků) $f_i(t)$ pro $i = 0, \dots, m$ a $t = 1, \dots, T$, který může být používán k získání (vygenerování) požadovaného ortonormálního systému vzhledem ke skalárnímu součinu s použitím postupu, který je v literatuře označován jako tzv. Gram-Schmidtův ortonormalizační proces viz Hewitt a Stromberg (1975). Postup lze použít za výše uvedených předpokladů.

Pro popis uvedeného postupu je vhodné dodefinovat základní výrazy a symboly, které jsou vhodné pro tvorbu požadovaného ortonormálního systému. Nejprve je zde popsána definice (standardního) skalárního součinu pro Euklidovský prostor $\mathbb{R}^n$ publikovaná ve Williams (2008).

Definice 4.11 *Nechť* $u = \{u_1, \dots, u_n\}$ *a* $v = \{v_1, \dots, v_n\}$ *jsou vektory v* $\mathbb{R}^n$*. Potom skalární součin těchto vektorů* $\langle u, v \rangle$ *je definován jako*

$$\langle u, v \rangle = u_1 v_1 + \cdots + u_n v_n. \tag{4.19}$$

Výsledkem skalárního součinu je reálné číslo související s úhlem (jeho cosinem) příslušných vektorů.

(převzato z Williams (2008), s. 45)

Dále je nutné pro popis Gram-Schmidtova ortonormalizačního procesu popsat operátor ortogonální projekce v prostoru se skalárním součinem jako např. Cheney a Kincaid (2008).

Definice 4.12 *Nechť* $u = \{u_1, \dots, u_n\}$ *a* $v = \{v_1, \dots, v_n\}$ *jsou vektory v* $\mathbb{R}^n$*. Potom operátor projekce obou vektorů označovaný jako* $proj_u(v)$*, který ortogonálně promítá vektor* v *na vektor* u*, je definován jako*

$$proj_u(v) = \frac{\langle v, u \rangle}{\langle u, u \rangle} u, \tag{4.20}$$

kde $\langle v, u \rangle$ *značí vektorový součin* u *a* v*.*

(převzato z Cheney a Kincaid (2008), s. 722)

Nyní jsou definovány základní pojmy a je tedy možné popsat Gram-Schmidtův ortogonalizační proces, jak publikovali Hewitt a Stromberg (1975).

Definice 4.13 *Nechť* V *je prostor se skalárním součinem a nechť* $\{v_1, \dots, v_n, \dots\}$ *je konečná nebo spočetně nekonečná lineárně nezávislá podmnožina na stejném prostoru* V*, potom Gram-Schmidtův ortogonalizační proces lze definovat vztahem*

$$u_k = v_k - \sum_{j=1}^{k-1} proj_{u_j}(v_k), \qquad e_k = \frac{u_k}{\|u_k\|}, \tag{4.21}$$

kde

k *lze označit jako krok v postupu,*

u_k *představuje* $k-$*tý prvek z vygenerovaných ortogonálních vektorů,*

e_k *je* $k-$*tý normovaný prvek z vygenerovaných ortogonálních vektorů a*

$\|u_k\|$ *je Euklidovská norma.*

(převzato z Hewitt a Stromberg (1975), s. 240)

Poznámka 4.12 *Ortogonalita vektorů charakterizuje (nenulové) vektory takové, „které jsou navzájem kolmé“ (skalární součin těchto vektorů je roven* 0*) a lineárně nezávislé, Zhang (2009).*

Získanou množinu prvků $\{u_1, \dots, u_k\}$ lze označit za orto**gonální** množinu a postup odvození je znám jako Gram-Schmidtova ortogonalizace, zatímco množinu prvků $\{e_1, \dots, e_k\}$ lze označit jako orto**normální** množinu. Vektor označovaný jako e_k je jednotkový vektor, který je popsán v následujícím důsledku.

Důsledek 4.1 *Jednotkový vektor je vektor, jehož norma je rovna jedné. Jestliže* u_k *je nenulový vektor, potom vektor* e_k *získaný jako*

$$e_k = \frac{u_k}{\|u_k\|} \tag{4.22}$$

je jednotkový vektor ve směru vektoru u_k*.*

Gram-Schmidtův ortonormalizační proces popisuje, jak vygenerovat požadovaný ortonormální systém. Nyní je již známa metodika Gram-Schmidtova ortonormalizačního procesu, a proto je dále vhodné uvést podrobnější popis tohoto algoritmu (postupu) pro praktické výpočty.

Důsledek 4.2 *Algoritmus pro vygenerování požadované ortonormální množiny může být interpretován následujícím postupem:*

$$\begin{aligned} g_0(t) &= \frac{f_0(t)}{\|f_0(t)\|}; \qquad i = 0, \qquad t = 1,..,T, \\ h_i(t) &= f_i(t) - \sum_{j=0}^{i-1} \langle f_i, g_j \rangle \, g_j(t); \qquad i = 1,..,m, \qquad t = 1,..,T, \\ g_i(t) &= \frac{h_i(t)}{\|h_i\|}. \end{aligned} \tag{4.23}$$

(převzato z Trefethen & Bau (1997), s. 50 – 58, včetně značení)

Výše uvedený postup a jeho implementace vypadají z procesního hlediska v pořádku, ale problém nastane právě s implementací tohoto systému. Tento proces je většinou „numericky nestabilní“, protože vypočítané vektory nejsou zcela ortogonální díky vyskytujícím se chybám při vnitřním zobrazení čísel. Vzniklá ztráta ortogonality znehodnocuje celý Gram-Schmidtův proces, Trefethen & Bau (1997). Proto lze použít mírnou modifikaci pro numerickou stabilizaci celého procesu a tím i ke zpřesnění ortonormálních prvků uvažovaného systému. Tato jednoduchá úprava pro numerickou stabilitu je nazývána jako modifikovaný Gram-Schmidtův ortonormalizační proces (algoritmus) a je prezentován např. v Trefethen & Bau (1997).

Důsledek 4.3 *Numerická stabilizace Gram-Schmidtova ortonormalizačního procesu je založena na skutečnosti, že místo okamžitého výpočtu vektoru $h_i(t)$ z předchozího postupu je tento vektor odvozen následujícím modifikovaným postupem:*

$$\begin{aligned} g_0(t) &= \frac{f_0(t)}{\|f_0(t)\|}, \qquad i = 0, \qquad t = 1,..,T, \\ h_i(t) &= f_i(t), \qquad i = 1,..,m, \qquad t = 1,..,T, \\ h_i(t) &\leftarrow h_i(t) - \langle h_i, g_j \rangle g_j(t), \qquad j = 0,\dots,i-1, \\ g_i(t) &= \frac{h_i(t)}{\|h_i\|}. \end{aligned} \tag{4.24}$$

(převzato z Trefethen & Bau (1997), s. 50 – 58, včetně značení)

Odvozený postup výše uvedeného algoritmu může být vyjádřen i následujícím postupem:

$$\begin{aligned} g_0(t) &= t_{0,0} f_0(t) & f_0(t) &= s_{0,0} g_0(t) \\ g_1(t) &= t_{1,0} f_0(t) + t_{1,1} f_1(t) & f_1(t) &= s_{1,0} g_0(t) + s_{1,1} g_1(t) \\ \dots & \Rightarrow & \dots & \\ g_m(t) &= \sum_{j=0}^{m} t_{m,j} f_j(t) & f_m(t) &= s_{m,j} g_j(t). \end{aligned} \tag{4.25}$$

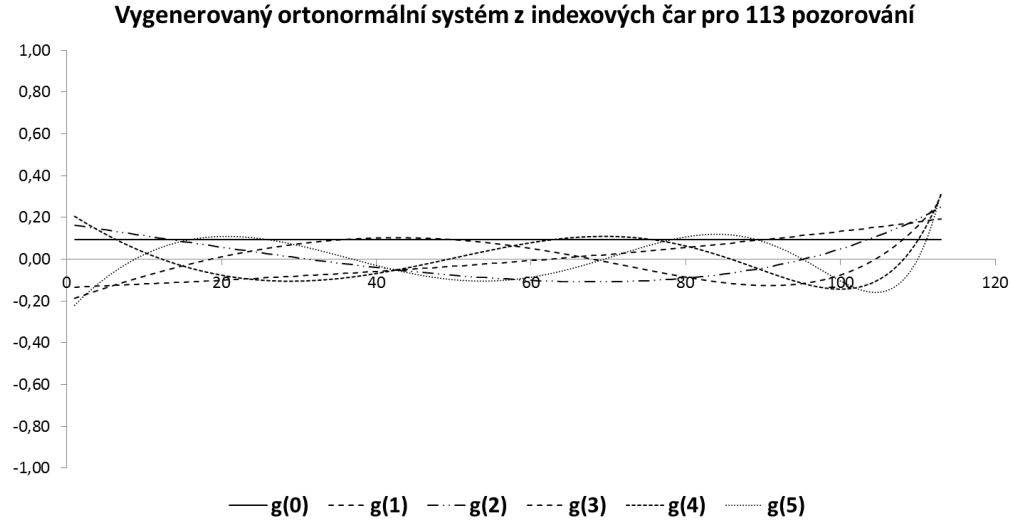
Uvedený postup lze přepsat s použitím matic T a S jako

$$g = T * f \Rightarrow f = S * g. \tag{4.26}$$

Prezentované matice S a T jsou v tomto schématu regulární, vzájemně inverzní a mají dolní trojúhelníkový tvar. Jejich explicitní tvar lze odvodit z výše uvedených ortonormalizačních postupů. Pro praktické použití je (nejen) zajímavá matice S, protože se jedná o matici zpětné (bázové) transformace.

Dále zde může nastat problém pro numerické stanovení $i - tého$ řádku matice S. Jedním (nikoliv jediným) z možných řešení tohoto problému je použití takových algoritmů, jako je např. lineární regresní model bez absolutního členu. Jako druhý vedlejší a užitečný efekt použití lineárního regresního modelu je získání koeficientu determinace, často označovaného jako R^2. Tento koeficient může být interpretován jako podíl rozptylu hodnot závislé proměnné, který se podařilo vysvětlit pomocí odhadovaného regresního modelu. Stručně řečeno, jedná se o jistou míru toho, jak přesně byla posloupnost $g_j(t)$ pro $j = 0, \dots, i$ spočtena.

Vygenerovaný ortonormální systém z množiny indexových čar $g(i)$ je ukázán na následujícím Obrázku 9.



Obrázek 9: *Vygenerovaný ortonormální systém z množiny indexových čar $g(i)$*

4.7 Důvody pro zavedení a využívání ONS

Samotná transformace zdrojových indexových čar (zdrojové soustavy) na ortonormální systém se může zdát jako samoúčelná nebo pouze motivována numerickými důvody, kde pro efektivní a numericky stabilní výpočty je vhodné mít k dispozici dobře podmíněnou bázi.

Výhody vygenerovaného systému jsou ale v jeho vlastnostech při odvozování, které jsou důležité pro výpočty a dokazování zvoleného modelu (tj. je zde vytvořena ortonormální báze uvažovaného podprostoru). Konkrétní popis výhod je proveden postupně v následujícím textu při odvozování vlastností.

4.8 Problematiky generování ONS

V návaznosti na předchozí odvození je z teoretického úhlu pohledu vygenerovaný systém $f_i(t)$ pro $i = 0, \dots, m$ a $t = 1, \dots, T$ lineárně nezávislý a potenciálně rozšiřitelný až do hodnoty $m \leq (T - 1)$. Následně vygenerovaný ortonormální systém s použitím uvedené metody by měl být také lineárně nezávislý a splňující požadavky uvedené dále.

Reálná simulace vybraného systému odhalila možné nedostatky numerického modelování, které nesouhlasily s teoretickým základem požadovaným na uvedená data. Konkrétně při použití software Matlab 2010a, ve kterém byly patřičné simulace prováděny (ve standardní přesnosti „double precision“), se objevila lineární závislost (vlivem vnitřního zobrazení) pro vygenerovaný systém $f_i(t)$, $t = 1, \dots, 113$ přibližně již pro $t > 50$.

Z tohoto důvodu je při generování ortonormálního systému, který je dále použitelný pro predikci trendové křivky, použit vnitřní algoritmus (funkce) uvedeného software pro odhad požadovaných vektorů s názvem „qr“[51]. Podrobný popis a syntaxe jsou uvedeny v příslušném odkazu. Vyvolaná funkce provede ze vstupní matice vygenerovaného systému tzv. QR rozklad $f_i(t)$ na horní trojúhelníkovou matici R a čtvercovou matici Q tak, že $A = Q * R$.

Použitou metodu lze popsat jako způsob, jak rozložit požadovanou matici na součin dvou matic, z nichž je jedna ortogonální nebo má alespoň ortonormální sloupce a druhá je vyjádřena v horním trojúhelníkovém tvaru. Detailní numerické řešení a analýza uvedené problematiky nejsou předmětem této práce, proto zde nejsou detailněji popsány.

[51] Podrobný popis a syntaxe funkce: http://www.mathworks.com/help/matlab/ref/qr.html.

4.9 Retrospektivní trend, oddělení systematické časově proměnné složky a složky náhodné

Hlavní princip požadovaného odhadu je založen na skutečnosti, že bude nalezena „vhodná“ aproximace neznámé a předpokládané trendové křivky označované jako $X(t)$ pro získaný soubor diskrétních pozorování v čase $t = 1, \dots, T$. Pro výběr „vhodné“ aproximace je použita metoda nejmenších čtverců.

Poznámka 4.13 *Metoda nejmenších čtverců může být stručně popsána jako celkové řešení minimalizující součet čtverců chyb provedených v řešení každé rovnice. Detailnější popis, odvození a informace lze nalézt např. v Merriman (2005).*

4.9.1 Metoda nejmenších čtverců pomocí ortonormálního systému

Následná aproximace požadované trendové složky z původní řady $X(t)$, která bude nadále označována jako $X_m(t)$, může být chápána jako vytvořená lineární kombinace vyjádřená pomocí vztahu

$$X_m(t) = \sum_{i=0}^{m} a_i g_i(t), \qquad t = 1, \dots, T, \qquad m \leq (T-1), \tag{4.27}$$

tak, že přístup k řešení problému lineární úlohy pomocí metody nejmenších čtverců může být popsán pomocí následující rovnice

$$\|X_m(t) - X(t)\|^2 = \sum_{t=1}^{T} (X_m(t) - X(t))^2 \underrightarrow{a_0, \dots, a_m} min, \qquad m \leq (T-1), \tag{4.28}$$

kde zápis $\|\cdot\|$ označuje Euklidovskou normu. Následný problém týkající se požadovaného odhadu lze zredukovat na skutečnost, kde je hlavním cílem najít takové neznámé ale předpokládané koeficienty a_i, které minimalizují součet kvadrátů chyb mezi modelem pro odhad neznámého trendu a získaným modelem ze souboru obsahujícího právě T pozorování.

Věta 4.1 *Nechť a_k je $k-tý$ předpokládaný a odhadovaný koeficient, nechť X je datový vektor a nechť g_k je $k-tý$ prvek z vygenerovaného ortonormálního systému, potom odhad neznámého koeficientu a_k může být vyjádřen vztahem*

$$a_k = \langle \mathrm{X}, g_k \rangle, \qquad k = 0, \dots, m, \tag{4.29}$$

kde zápis $\langle .,. \rangle$ označuje skalární součin, X a g_k jsou vektory obsahující celkem T hodnot.

Důkaz. Uvedený vztah je získán na základě postupů hledajících minimum, které představuje požadovaný výsledek. Nejprve lze přepsat metodu nejmenších čtverců do tvaru

$$Q(a_0, \dots, a_m) = \|X_m(t) - X(t)\|^2 = \sum_{t=1}^{T} \left(\sum_{i=0}^{m} a_i g_i(t) - X(t) \right)^2, \quad m \leq (T-1), \tag{4.30}$$

kde použitá funkce Q označuje kvadrát Euklidovský normy rozdílu mezi odhadovanou a získanou (předpokládanou) trendovou křivkou. Potom se využije znalost prvních parciálních derivací, která je velmi užitečná právě při hledání maximálních a minimálních hodnot prezentované funkce Q. S pomocí těchto parciálních derivací funkce Q s ohledem na neznámou hodnotu koeficientu a_k je následně získán požadovaný výsledek pro $k-tý$ koeficient. Pro větší doložení tohoto postupu jsou zde ukázány možné kroky, které umožňují lepší pochopení vztahu (4.29).

$$\begin{aligned} \frac{d}{da_k} Q(a_0, \dots, a_m) &= 2 \sum_{t=1}^{T} g_k(t) \left(\sum_{i=0}^{m} a_i g_i(t) - X(t) \right) \\ &= \sum_{i=0}^{m} a_i \sum_{t=1}^{T} g_i(t) g_k(t) - \sum_{t=1}^{T} X(t) g_k(t) \\ &= \sum_{i=0}^{m} a_i \langle g_i, g_k \rangle - \langle \mathrm{X}, g_k \rangle \\ &= a_k - \langle \mathrm{X}, g_k \rangle, \qquad k = 0, \dots, m. \end{aligned} \tag{4.31}$$

Druhé parciální derivace potvrzují, že funkce Q dosahuje svého lokálního extrému v sedlovém bodu $0 = \frac{d}{da_k} Q(a_0, \dots, a_m)$, tedy za těchto podmínek:

$$\begin{aligned} \frac{d}{da_k da_j} Q(a_0, \dots, a_m) &= 1 \qquad \Leftrightarrow \qquad k = j \\ \frac{d}{da_k da_j} Q(a_0, \dots, a_m) &= 0 \qquad \Leftrightarrow \qquad k \neq j. \end{aligned} \tag{4.32}$$

Pro získání sedlových bodů je první derivace nastavena na hodnotu rovnou nule a následné řešení může být nalezeno s použitím skalárního součinu jako

$$\begin{aligned} 0 &= \frac{d}{da_k} Q(a_0, \dots, a_m) \\ &= a_k - \langle \mathrm{X}, g_k \rangle \\ a_k &= \langle \mathrm{X}, g_k \rangle, \qquad k = 0, \dots, m. \end{aligned} \tag{4.33}$$

■

Z uvedených postupů je získán důsledek, ve kterém je interpretace použitých výrazů publikována v Collatz (1970).

Důsledek 4.4 *Nechť X je datový vektor a nechť g_i je $i-tý$ prvek (vektor) vygenerovaného ortonormálního systému, potom platí*

$$\|X\|^2 \geq \sum_{i=0}^{m} \langle X, g_i \rangle^2 , \qquad m \geq 0. \tag{4.34}$$

(převzato z Collatz (1970), s. 52)

Poznámka 4.14 *Uvedená nerovnost je v literatuře známa jako tzv. Besselova nerovnost*.

Důkaz. Postup důkazu je založen na modifikaci metody nejmenších čtverců. Nejprve je možné přepsat rozdíl mezi aproximovanou trendovou křivkou $X_m(t)$ a předpokládanou trendovou křivkou $X(t)$ s použitím skalárního součinu jako

$$\begin{aligned} X_m(t) - X(t) &= \sum_{i=0}^{m} a_i g_i(t) - X(t) \\ &= \sum_{i=0}^{m} \langle \mathrm{X}, g_i \rangle g_i(t) - X(t), \qquad t = 1, \dots T. \end{aligned} \tag{4.35}$$

Potom funkce $Q(a_0, \dots, a_m)$ obsahuje optimální hodnoty koeficientů $a_i \in \mathbb{R}$ a může být vyjádřena pomocí vztahu s použitím kvadrátu Euklidovské normy, skalárního součinu a vygenerovaného ortonormálního systému:

$$\begin{aligned} Q(a_0, \dots, a_m; optimální) &= \left\| \sum_{i=0}^{m} \langle \mathrm{X}, g_i \rangle g_i(t) - X(t) \right\|^2 \\ &= \langle \sum_{i=0}^{m} \langle \mathrm{X}, g_i \rangle g_i(t) - X(t), \sum_{i=0}^{m} \langle \mathrm{X}, g_i \rangle g_i(t) - X(t) \rangle \\ &= \sum_{i=0}^{m} \langle \mathrm{X}, g_i \rangle^2 - 2 \sum_{i=0}^{m} \langle \mathrm{X}, g_i \rangle^2 + \langle X, X \rangle \\ &= \langle X, X \rangle - \sum_{i=0}^{m} \langle \mathrm{X}, g_i \rangle^2 \\ &= \|X\|^2 - \sum_{i=0}^{m} \langle \mathrm{X}, g_i \rangle^2 \geq 0. \end{aligned} \tag{4.36}$$

■

Tvrzení 4.3 *Vygenerovaný ortonormální systém je lineárně nezávislá množina pro* $t = 1, \dots, T$.

Důkaz. Postup je motivován použitím metody důkazu sporem s úvodním předpokladem. Nechť $g_i(t)$ je vygenerovaný ortonormální systém pro $t = 1, \dots, T$ a $i = 0, \dots, m$ a nechť je zde předpokládána lineární závislost vygenerovaného ortonormálního systému, kterou lze vyjádřit pomocí lineární kombinace jejich prvků

$$\sum_{i=0}^{m} \alpha_i g_i = 0, \qquad \exists i \in \{0, \dots, m\} : \alpha_i \neq 0, \tag{4.37}$$

kde $\alpha_i \epsilon \mathbb{R}$. Potom pro $k = 0, \dots, m$ by měl být skalární součin 0 a g_k roven nule, ale výsledná hodnota je odlišná s předpokladem

$$0 = \langle 0, g_k \rangle = \langle \sum_{i=0}^{m} \alpha_i g_i, g_k \rangle = \alpha_k. \tag{4.38}$$

A to je spor s úvodním předpokladem pro $k = 0, \dots m$.

Důsledek 4.5 *Každý T prvkový ortonormální systém nad $1, \dots, T$ je bází nad $1, \dots, T$.*

Důkaz. Zřejmé.

Věta 4.2 *Nechť X je datový vektor, nechť g_i je $i - tý$ prvek (vektor) vygenerovaného ortonormálního systému a nechť $a_i \in \mathbb{R}$ je odhadovaný koeficient, potom v jakémkoliv konečně-rozměrném prostoru se skalárním součinem platí následující rovnost*

$$\|X\|^2 = \sum_{i=0}^{T-1} \langle X, g_i \rangle^2 = \sum_{i=0}^{T-1} a_i^2, \tag{4.39}$$

kde $\|\cdot\|^2$ značí kvadrát Euklidovské normy a $\langle ., . \rangle$ skalární součin.

Důkaz. Věta 4.2 je triviální důsledek předchozího. Důkaz je založen na výše popsané Besselově nerovnosti (4.34) a na vlastnosti, že vygenerovaný prvek ortonormálního systému $g_i(t)$ pro $i = 0, \dots, m$ a $t = 1, \dots, T$ je bázový. Potom z uvedených vlastností plyne rovnost, kterou lze vyjádřit jako

$$\|X\|^2 = \sum_{i=0}^{T-1} \langle X, g_i \rangle^2. \tag{4.40}$$

Uvedená rovnost je známa pod názvem „Parsevalova identita“ nebo „Parsevalova rovnost“ jak je uvedeno v Bachman & Narici (2000, s. 163). Stručně, Parsevalova identita platí v jakémkoliv konečně-rozměrném prostoru se skalárním součinem a s ortonormální bází.

■

Dále zde bude předpokládáno, že je k dispozici nějaký ortonormální systém označovaný jako $g_i(t)$ pro $i = 0, \dots, m$ a $t = 1, \dots, T$, který je potenciálně rozšiřitelný z důvodu lineární nezávislosti až do hodnoty $m = (T - 1)$ (tj. systém má stejný počet komponent jako počet pozorování časové řady). Potom lze vyjádřit odhad trendové křivky $X_m(t)$ vztahem

$$X_m(t) = \sum_{i=0}^{m} \langle X, g_i \rangle g_i(t), \quad t = 1, \dots, T \tag{4.41}$$

a rozdíl mezi odhadovanou a reálnou trendovou křivkou, který bude nadále označován jako ε_t^m pro $m \leq (T-1)$. Uvažovaný rozdíl mezi prezentovanými hodnotami může být získán jako

$$\varepsilon_t^m = X(t) - X_m(t), \quad t = 1, \dots, T. \tag{4.42}$$

Potom trendová křivka z datového souboru $X(t)$ může být přepsána s použitím předchozího značení a vyjádření pomocí vztahu

$$X(t) = \sum_{i=0}^{m} \langle X, g_i \rangle g_i(t) + \varepsilon_t^m, \quad t = 1, \dots, T. \tag{4.43}$$

Nejedná se o nic jiného než o vyjádření reálných pozorovaných hodnot pomocí vztahu pro odhad trendové křivky a chyby daného odhadu. Také je zde předpokládáno, že je k dispozici „specifický ortonormální systém“, který „obsahuje konstantu“ dále označovanou jako $g_0(t) = g_0 \in \mathbb{R}$ pro $t = 1, \dots, T$.

Věta 4.3 *Nechť G je nějaký ortonormální systém takový, že $g_0(t)$* pro $t = 1, \dots, T$ *je konstanta $g_0(t) = g_0 \in \mathbb{R}$ v konečně-rozměrném prostoru s Euklidovským skalárním součinem $\langle x, y \rangle = \sum_{t=1}^{T} x(t)y(t)$. Potom hodnota konstanty g_0 je dána vztahem*

$$g_0(t) = g_0 = \frac{1}{\sqrt{T}} \tag{4.44}$$

a pro každý prvek $g_i(t)$ vygenerovaného ortonormálního systému pro $i = 1, \dots, m$ platí

$$\sum_{t=1}^{T} g_i(t) = 0. \tag{4.45}$$

Důkaz. Nejprve popis prvního vztahu, který je založen na vlastnostech skalárního součinu a definované normy. Výpočet normy s použitím skalárního součinu lze přepsat:

$$1 = \langle g_0, g_0 \rangle = \sum_{t=1}^{T} g_0^2 = T g_0^2 \Rightarrow g_0 = \frac{1}{\sqrt{T}}, \tag{4.46}$$

kde T značí počet získaných pozorování. Jak je patrné z uvedeného postupu, výsledná hodnota uvažované konstanty ve „specifickém ortonormálním systému“ je rovna hodnotě $\frac{1}{\sqrt{T}}$.

Odvození druhého vztahu je založeno na podobném postupu, tedy na vlastnostech skalárního součinu a vygenerovaného ortonormálního systému:

$$0 = \langle g_0, g_i \rangle = \sum_{t=1}^{T} g_i(t) g_0 = g_0 \sum_{t=1}^{T} g_i(t) \Rightarrow \sum_{t=1}^{T} g_i(t) = 0, \quad i = 1, \dots m. \tag{4.47}$$

■

V následujícím textu je uvažován pouze vygenerovaný „specifický ortonormální systém“ obsahující konstantu v konečně-rozměrném prostoru se skalárním součinem

$$\langle x, y\rangle = \sum_{t=1}^{T} x(t)y(t). \tag{4.48}$$

Věta 4.4 *Nechť G je specifický ortonormální systém v konečně-rozměrném prostoru se skalárním součinem, potom platí vztah mezi datovým vektorem $X(t)$ a odhadem trendové křivky $X_m(t)$*

$$\sum_{t=1}^{T} X(t) = \sum_{t=1}^{T} X_m(t). \tag{4.49}$$

Důkaz. K odvození vztahu je použito vyjádření odhadované trendové složky $X_m(t)$ se skalárním součinem pro $t = 1, \dots, T$ a $i = 0, \dots, m$. Potom uvedený vztah (4.49) může být přepsán následujícím postupem:

$$\begin{aligned}\sum_{t=1}^{T} X_m(t) &= \sum_{t=1}^{T}\sum_{i=0}^{m}\langle X, g_i\rangle g_i(t)\\ &= \sum_{i=0}^{m}\langle X, g_i\rangle \sum_{t=1}^{T} g_i(t)\\ &= T g_0\langle X, g_0\rangle\\ &= T g_0^2 \sum_{t=1}^{T} X(t) = \sum_{t=1}^{T} X(t)\,.\end{aligned} \tag{4.50}$$

■

Z obsahu věty 4.4 přímo plyne následující jednoduchý důsledek 4.6 pro rozdíl mezi odhadovanou a předpokládanou reálnou trendovou křivkou ε_t^m pro $m \leq (T-1)$.

Důsledek 4.6 *Nechť ε_t^m je rozdíl mezi odhadovanou trendovou křivkou a datovým vektorem X, potom pro součet rozdílu platí*

$$\sum_{t=1}^{T} \varepsilon_t^m = 0. \tag{4.51}$$

Důkaz. Dokazování této skutečnosti je založeno na větě 4.4, ze které přímo plyne uvedená rovnost.

Věta 4.5 *Nechť ε_t^m je rozdíl mezi odhadovanou trendovou křivkou a datovým vektorem $X(t)$, nechť $g_k(t)$, $t = 1, \dots, T$ je libovolný vektor z vygenerovaného ortonormálního systému v konečně-rozměrném prostoru se skalárním součinem, potom je součet jejich násobků přes všechny naměřená pozorování T roven nulové hodnotě, exaktněji:*

$$\sum_{t=1}^{T} \varepsilon_t^m g_k(t) = \langle \varepsilon^m, g_k \rangle = 0, \qquad 0 \le k \le m. \tag{4.52}$$

Poznámka 4.15 *Zbytková (reziduální, náhodná, …) složka je ortogonální ke každé komponentě vygenerovaného ortonormálního systému* g_k.

Důkaz. Postup důkazu je založen na vztahu k odhadu předpokládaných koeficientů. Odhady těchto koeficientů jsou vyjádřeny pomocí skalárního součinu hodnot předpokládaného a neznámého trendu ze souboru reálných pozorování a libovolného $k - t\acute{e}ho$ vektoru z vygenerovaného ortonormálního systému. Potom $k - t\acute{y}$ koeficient a_k může být získán za pomoci postupu

$$\begin{aligned} a_k &= \langle X, g_k \rangle = \sum_{t=1}^{T} X(t) g_k(t) \\ &= \sum_{t=1}^{T} \left(\sum_{i=1}^{m} a_i \, g_i(t) + \varepsilon_t^m \right) g_k(t) \\ &= \sum_{i=1}^{m} a_i \sum_{t=1}^{T} g_i(t) g_k(t) + \sum_{t=1}^{T} \varepsilon_t^m g_k(t) \\ &= \sum_{i=1}^{m} a_i \, \langle g_i, g_k \rangle + \sum_{t=1}^{T} \varepsilon_t^m g_k(t) \\ &= a_k + \sum_{t=1}^{T} \varepsilon_t^m g_k(t) \Rightarrow \sum_{t=1}^{T} \varepsilon_t^m g_k(t) = 0. \end{aligned} \tag{4.53}$$

Celkový součet $\sum_{i=1}^{m} a_i \, \langle g_i, g_k \rangle$ má pouze jeden sčítanec odlišný od nulové hodnoty, který nastane pro případ $i = k$, kde je hodnota odhadovaného koeficientu právě a_k. Z tohoto důvodu lze v uvedeném odvození psát, že

$$\sum_{i=1}^{m} a_i \, \langle g_i, g_k \rangle = a_k, \qquad 0 \le k \le m. \tag{4.54}$$

Dále je vhodné poznamenat, že celkový součet násobků $\varepsilon_t^m g_k(t)$ je roven nulové hodnotě pro $t = 1, \dots, T$ jak bylo dokázáno dříve a pozorování ze souboru reálných dat je nahrazeno námi odhadovanou trendovou křivkou a jí přiřazenou náhodnou složkou.

■

Věta 4.6 *Nechť* a_k *je* $k - t\acute{y}$ *odhadovaný koeficient, nechť* $X(t)$ *je datový vektor ze souboru reálných pozorování, nechť* $X_m(t)$ *je odhadovaná trendová křivka a nechť* $g_k(t)$ *je libovolný vektor z vygenerovaného ortonormálního systému, potom v konečně-rozměrném prostoru se skalárním součinem platí rovnost mezi trendovými křivkami při odhadu hledaného koeficientu*

$$a_k = \sum_{t=1}^{T} X(t) g_k(t) = \sum_{t=1}^{T} X_m(t) g_k(t) = \langle X, g_k \rangle = \langle X_m, g_k \rangle, \tag{4.55}$$

pro $0 \leq k \leq m$ *a* $t = 1, \ldots, T$.

Důkaz. Odvození rovnosti je založeno na vztahu pro odhad koeficientu a vyjádření skalárního součinu z definice pomocí celkového součtu násobků jednotlivých prvků. Potom požadovaný koeficient může být vyjádřen s použitím prvků $X(t)$ a $g_k(t)$ jako

$$\begin{aligned} a_k &= \sum_{t=1}^{T} X(t) g_k(t) \\ &= \sum_{t=1}^{T} (X_m(t) + \varepsilon_t^m) g_k(t) \\ &= \sum_{t=1}^{T} X_m(t) g_k(t) + \sum_{t=1}^{T} \varepsilon_t^m g_k(t) \\ &= \sum_{t=1}^{T} X_m(t) g_k(t), \qquad k = 0, \ldots, m. \end{aligned} \tag{4.56}$$

Opět je vhodné poznamenat, že celkový součet násobků $\varepsilon_t^m g_k(t)$ je roven nulové hodnotě pro $t = 1, \ldots, T$ jak bylo dokázáno výše a získaný datový vektor je nahrazen součtem odhadované trendové křivky a jí přiřazené náhodné složky.

4.10 Statistická inference náhodné složky

Vzhledem k předchozím numerickým postupům je nyní vhodné se zaměřit na popisovanou problematiku ze statistického úhlu pohledu. Doposud nebyl rozdíl mezi předpokládanými a odhadovanými koeficienty označován, protože jejich význam intuitivně vycházel z podstaty věci. Nyní jsou pro přehlednost rozdílná značení požadována. Důležitým předpokladem v celé této kapitole je skutečnost, že datový vektor je v realitě popsán modelem (tj. předpokládáme, že ve skutečnosti platí, existuje), který je vyjádřen vztahem

$$X(t) = \sum_{i=0}^{m} a_i g_i(t) + \varepsilon_t, \qquad t = 1, \ldots, T, \qquad m \leq (T-1), \tag{4.57}$$

kde a_i označují reálné (dané), ale neznámé (nedostupné) koeficienty, které jsou odhadovány pomocí koeficientů $\hat{a}_i$ a ε_t je náhodná komponenta, která je obecně označována jako šum (nebo též náhodná složka). Náhodná komponenta je opět nedostupná s některými předem známými vlastnostmi, které budou postupně stanoveny v této části práce a m je daný, ale také předem neznámý počet komponent odhadu trendové křivky.

Pro koeficienty nutné k odhadu trendové složky platí vztah s použitím skalárního součinu a vygenerovaného ortonormálního systému

$$\hat{a}_k = \langle X, g_k \rangle = a_k + \langle \varepsilon, g_k \rangle \;\Rightarrow\; a_k - \hat{a}_k = -\langle \varepsilon, g_k \rangle, \quad k = 0, \dots, m. \tag{4.58}$$

Střední hodnota rozdílu mezi oběma koeficienty $(a_k - \hat{a}_k)$ může být vyjádřena jako

$$\begin{aligned} E\{a_k - \hat{a}_k\} &= E\{-\langle \varepsilon, g_k \rangle\} \\ a_k - E\{\hat{a}_k\} &= -E\left\{\sum_{t=1}^{T} \varepsilon_t g_k(t)\right\} \\ &= -\sum_{t=1}^{T} E\{\varepsilon_t g_k(t)\} \\ &= -\langle E\{\varepsilon\}, g_k \rangle, \quad k = 0, \dots, m, \end{aligned} \tag{4.59}$$

kde E je značení pro střední hodnotu, a_k a $\hat{a}_k$ je $k-tý$ neznámý a jeho odhadovaný koeficient a $\langle .,. \rangle$ označuje skalární součin. Prezentované úpravy mohou být provedeny, protože a_k a g_k jsou deterministické, zatímco $\hat{a}_k$ a ε jsou náhodné veličiny vstupující do uvažovaného modelu. Uvedené skutečnosti jsou následně základem pro větu o vychýlených odhadech neznámého koeficientu a_k.

Věta 4.7 *Nechť $E\{\varepsilon\}$ je střední hodnota náhodné složky ε a nechť g_k je $k-tý$ prvek (vektor) vygenerovaného ortonormálního systému pro $k = 1, \dots, m$ v konečně-rozměrném prostoru se skalárním součinem, potom skalární součin $\langle E\{\varepsilon\}, g_k \rangle$ je roven nule a $\hat{a}_k = \langle X, g_k \rangle$ je nestranný (nevychýlený) odhad neznámého parametru a_k.*

Důkaz. Postup důkazu je odvozen ze vztahu (4.59).

Důsledek 4.7 *Zejména pokud střední hodnota náhodné složky $E\{\varepsilon\}$ je rovna nulové hodnotě, potom získaný odhad koeficientu $\hat{a}_k = \langle X, g_k \rangle$ je nestranný odhad neznámého koeficientu a_k pro $k = 0, \dots, m$.*

Předpokládanou a neznámou trendovou křivku je možné získat ze vztahu

$$X_m(t) = \sum_{i=0}^{m} \hat{a}_i g_i(t) + \varepsilon_t^m, \quad t = 1, \dots, T, \tag{4.60}$$

kde $\hat{a}_i = \langle X, g_i \rangle$ označuje odhady neznámých koeficientů a_i, $g_i(t)$ je $i-tý$ prvek vygenerovaného ortonormálního systému a ε_t^m je reziduální složka. Potom rozdíl mezi předpokládaným modelem $X(t)$ a vytvořeným modelem pro odhad $X_m(t)$, vzniklý odečtem druhé rovnice od první, se rovná nulové hodnotě a celý postup lze popsat jako

$$0 = \sum_{i=1}^{m} (\hat{a}_i - a_i) g_i(t) + (\varepsilon_t^m - \varepsilon_t), \quad t = 1, \dots, T. \tag{4.61}$$

Zároveň může být tento vztah pro rozdíl upraven na vztah využívající skalární součin ve tvaru

$$0 = \sum_{i=0}^{m} (\hat{a}_i - a_i)\langle g_i, g_k\rangle + (\langle \varepsilon^m, g_k\rangle - \langle \varepsilon, g_k\rangle), \quad k = 0, \dots, m \tag{4.62}$$

a odtud

$$(\hat{a}_k - a_k) = \langle \varepsilon, g_k\rangle, \quad k = 0, \dots, m. \tag{4.63}$$

Jinými slovy, rozdíl mezi odhadovaným a reálným (neznámým) koeficientem může být vyjádřen pomocí skalárního součinu náhodné složky ε a prvku vygenerovaného ortonormálního systému. Za předpokladu $\langle E\{\varepsilon\}, g_k\rangle = 0$ je možné přepsat vztah (4.63) jako

$$(\hat{a}_k - a_k) = \langle \varepsilon - E\{\varepsilon\}, g_k\rangle, \quad k = 0, \dots, m. \tag{4.64}$$

Nyní je vhodný podrobnější rozbor předpokladu $\langle E\{\varepsilon\}, g_k\rangle = 0$. Pokud se jedná o náhodnou složku ε s konstantní střední hodnotou, pak je tento předpoklad splněn pro $k = 1, \dots, T-1$, protože

$$\langle E\{\varepsilon\}, g_k\rangle = E\{\varepsilon\} \sum_{t=1}^{T} g_k(t) = E\{\varepsilon\} * 0 = 0, \quad k = 1, \dots, T-1. \tag{4.65}$$

Obecně ale nebude splněn pro $k = 0$

$$\langle E\{\varepsilon\}, g_0\rangle = E\{\varepsilon\} \sum_{t=1}^{T} g_0(t) = E\{\varepsilon\} \frac{T}{\sqrt{T}} = E\{\varepsilon\}\sqrt{T}. \tag{4.66}$$

V tomto případě by byl předpoklad automaticky splněn pokud $E\{\varepsilon\} = 0$. V případě $E\{\varepsilon\} <> 0$ si lze pomoci tím, že budeme předpokládat dále $a_0 g_0(t) + E\{\varepsilon\} = \frac{a_0'}{\sqrt{T}}$ a nadále pracovat s tzv. „licencí“ $a_0' \to a_0$. Případná nenulová konstantní střední hodnota by tedy byla „přesunuta“ do konstantní složky trendu. Proto pro následující úvahy postačí předpoklad konstantní (stacionarita náhodné složky ve střední hodnotě) střední hodnoty náhodné složky nebo je ekvivalentně platný předpoklad $E\{\varepsilon\} = 0$, neboť je spíše filosofickou otázkou, zda „trvalá“ konstantní střední hodnota patří do trendu, či nikoliv. Proto je zde uvažován klasický předpoklad $E\{\varepsilon\} = 0$.

Dále lze ze vztahu (4.63) odvodit korelační matici pro odhady neznámých koeficientů $\hat{a}_k$ za uvedeného předpokladu výše. Tedy

$$(\hat{a}_k - a_k)(\hat{a}_j - a_j) = \langle \varepsilon, g_k\rangle\langle \varepsilon, g_j\rangle = \sum_{t=1}^{T}\sum_{s=1}^{T} \varepsilon_t \varepsilon_s g_k(t) g_j(s) \tag{4.67}$$

a odtud

$$E\{(\hat{a}_k - a_k)(\hat{a}_j - a_j)\} = \sum_{t=1}^{T}\sum_{s=1}^{T} E\{\varepsilon_t \varepsilon_s\} g_k(t) g_j(s). \tag{4.68}$$

Pokud je $E\{\varepsilon_t \varepsilon_s\} = 0$ pro $t <> s$ a $E\{\varepsilon_t \varepsilon_s\} = \sigma_\varepsilon^2$ pro $t = s$, potom se jedná o nekorelovanou a stacionární náhodnou složku v širším smyslu (v prvních a druhých momentech). Dále pokud

$$E\{(\hat{a}_k - a_k)(\hat{a}_j - a_j)\} = \sum_{t=1}^{T} E\{\varepsilon_t^2\} g_k(t) g_j(s) = \sigma_\varepsilon^2 \sum_{t=1}^{T} g_k(t) g_j(s) = \sigma_\varepsilon^2 \langle g_k, g_j \rangle, \tag{4.69}$$

potom $\sigma_\varepsilon^2 \langle g_k, g_j \rangle = 0$ pro $k <> j$ a $\sigma_\varepsilon^2 \langle g_k, g_j \rangle = \sigma_\varepsilon^2$ pro $k = j$. Tj. kovarianční matice je diagonální se stejnými prvky na diagonále. Korelační matice odhadů $\hat{a}_k$ je tedy jednotkovou maticí. Tyto závěry lze shrnout do následujících tvrzení.

Tvrzení 4.4 *Pokud* $E\{\varepsilon\} = 0$ *a* $E\{\varepsilon_t \varepsilon_s\} = 0$ *pro* $t <> s$ a $E\{\varepsilon_t \varepsilon_s\} = \sigma_\varepsilon^2$ pro $t = s$, *pak odhady* $\hat{a}_k$ *jsou nestranné* $(E\{\hat{a}_k\} = a_k)$ *s jednotkovou korelační maticí a s rozptylem* $\sigma_{\hat{a}_k}^2 = \sigma_\varepsilon^2$.

Tvrzení 4.5 *Pokud* $X_m(t) = \sum_{i=0}^{m} \hat{a}_i g_i(t) + \varepsilon_t^m, m = T - 1$, *pak* $\varepsilon_t^m = 0$ *pro* $t = 1, \dots, T$.*To plyne bezprostředně z nezávislosti prvků ortonormálního rozvoje* $g_i(t)$ *pro* $t = 1, \dots, T$ *a* $i = 0, \dots, T - 1$.

Důsledek 4.8 *Triviálně lze získat* $X(t) = \sum_{i=0}^{T-1} \hat{a}_i g_i(t) = \sum_{i=0}^{T-1} a_i g_i(t) + \varepsilon_t$ *a odtud* $\sum_{i=0}^{T-1} (\hat{a}_i - a_i) g_i(t) = \varepsilon_t$.

Věta 4.8 *Nechť* $E\{\varepsilon\}$ *je střední hodnota náhodné složky* ε, *nechť* g_k *je* $k-$*tý vektor vygenerovaného ortonormálního systému v konečně-rozměrném prostoru se skalárním součinem a nechť platí předpoklad* $\langle E\{\varepsilon\}, g_k \rangle = 0$ *pro* $k = 0, \dots, m$. *Potom platí rovnost mezi středními hodnotami náhodných veličin*

$$E\{\varepsilon_t^m\} = E\{\varepsilon_t\}. \tag{4.70}$$

Důkaz. Postup důkazu je založen na vyjádření střední hodnoty vztahu (4.61) a předpokladu nestrannosti odhadu neznámých koeficientů $\langle E\{\varepsilon\}, g_k \rangle = 0$:

$$\begin{aligned} 0 &= E\left\{\sum_{i=1}^{m} (\hat{a}_i - a_i) g_i(t) + (\varepsilon_t^m - \varepsilon_t)\right\} \\ &= \sum_{i=1}^{m} E\{\langle \varepsilon, g_i \rangle g_i(t)\} + E\{\varepsilon_t^m - \varepsilon_t\} \\ &= \sum_{i=1}^{m} \langle E\{\varepsilon\}, g_i \rangle g_i(t) + E\{\varepsilon_t^m\} - E\{\varepsilon_t\} \\ E\{\varepsilon_t^m\} &= E\{\varepsilon_t\}. \end{aligned} \tag{4.71}$$

4.11 Prediktivní trend

Predikce neznámé a předpokládané trendové křivky navazuje na vytvořený model k jejímu odhadu. Nechť je nadále předpokládáno, že jsou k dispozici jednotlivá pozorování $\{x(1), \dots, x(T)\}$ uvažované časové řady na diskrétním intervalu $1, \dots, T$ a vygenerovaný ortonormální systém vůči skalárnímu součinu na intervalu $1, 2, \dots, T, (T+1), \dots, (T+\tau)$. Vygenerovaný ortonormální systém je označován jako $g_0(t), g_1(t), \dots, g_m(t)$ a je potenciálně rozšiřitelný až na hodnoty $g_0(t), g_1(t), \dots, g_{T+\tau-1}(t)$, kde:

1. $t\epsilon\{1, 2, \dots T, (T+1), \dots, (T+\tau)\}$,
2. $0 < m < (T+\tau-1)$,
3. $g_0(t) = \frac{1}{\sqrt{T+\tau}}$.

První vygenerovaný prvek ortonormálního systému $g_0(t)$ je konstantní funkcí. Dále je předpokládáno, že je k dispozici „nějaké“ prodloužení (tím je myšlena možná předpověď, odhad, atd.) uvažované časové řady na intervalu $t\epsilon\{(T+1), \dots, (T+\tau)\}$, které bude označováno jako

$$z(t) = x(t) + \varepsilon_t, \tag{4.72}$$

pro $t\epsilon\{(T+1), \dots, (T+\tau)\}$, kde $x(t)$ je skutečná (budoucí a nedostupná) hodnota časové řady na uvažovaném intervalu a ε_t je chyba (náhodná složka) tohoto prodloužení. Tím je možné získat časovou řadu $\tilde{x}(t)$, pro kterou platí:

$$\begin{aligned} \tilde{x}(t) &= x(t), && t\epsilon\{1, 2, \dots T\}, \\ \tilde{x}(t) &= x(t) + \varepsilon(t), && t\epsilon\{(T+1), \dots, (T+\tau)\}. \end{aligned} \tag{4.73}$$

Nyní se lze zaměřit na množinu vybraných indexů, která je nadále označována jako I. Prozatím se jedná o blíže neurčenou množinu indexů vybraných ortonormálních složek $I \subseteq \{0, 1, \dots, m\}$, pomocí které bude odhadována neznámá trendová složka uvažované časové řady $\tilde{x}(t)$. Ta je nadále označována jako $\widetilde{X_I}(t)$ a lze ji vyjádřit vztahem

$$\widetilde{X_I}(t) = \sum_{i \in I} \tilde{a}_i g_i(t), \tag{4.74}$$

kde dosažitelný odhad neznámých koeficientů $\tilde{a}_i$ je možné rozepsat jako

$$\tilde{a}_i = \sum_{t=1}^{T+\tau} \tilde{x}(t) g_i(t) = \langle \tilde{x}, g_i \rangle = \sum_{t=1}^{T+\tau} x(t) g_i(t) + \sum_{t=T+1}^{T+\tau} \varepsilon_t g_i(t). \tag{4.75}$$

Pokud jsou vyjádřeny nedostupné a „skutečné“ koeficienty a_i jako

$$a_i = \sum_{t=1}^{T+\tau} x(t) g_i(t), \tag{4.76}$$

potom je získána vlivem „nedokonalé“ předpovědi $z(t)$ chyba odhadovaného koeficientu $\tilde{a}_i$ ve tvaru

$$(\tilde{a}_i - a_i) = \sum_{t=T+1}^{T+\tau} \varepsilon(t) g_i(t). \tag{4.77}$$

Potenciálně (ale jen potenciálně, ve skutečnosti je prakticky vždy nedostupná) je možné využít přesnou trendovou křivku časové řady $X_I(t)$, která je vyjádřena vztahem

$$X_I(t) = \sum_{i \in I} a_i g_i(t). \tag{4.78}$$

Poznámka 4.16 *Zde je důležité zdůraznit, že koeficienty $\tilde{a}_i$ jsou potencionálně alespoň dostupné, ale koeficienty a_i nikoliv.*

Pokud je dále k dispozici pro časovou řadu $x(t)$ pro $t\epsilon\{(T+1), \dots, (T+\tau)\}$ nějaký vybraný jednoparametrický trendový odhad (nejen z aplikační oblasti) s označením $z(t,p)$ pro $t\epsilon\{(T+1), \dots, (T+\tau)\}$, může se hodnota neznámého parametru p odhadnout optimalizací „nějakého“ zvoleného kritéria např. z aplikační oblasti. Pokud takové kritérium není k dispozici, pak lze zvolit např.

$$\max_{t=1,\dots,T} \left| x(t) - \sum_{i=0}^{m} \tilde{a}_i g_i(t) \right| \xrightarrow[p]{} min, \tag{4.79}$$

kde

$$\tilde{a}_i = \sum_{t=1}^{T+\tau} \tilde{x}(t) g_i(t) \tag{4.80}$$

a

$$\tilde{x}(t) = z(t,p), \qquad t\epsilon\{(T+1), \dots, (T+\tau)\}. \tag{4.81}$$

Zvolené kritérium obsahuje některé nedostatky, které se projevují zejména v návaznosti na „významně“ odlehlá „pozorování“ a mohou zkreslit požadovaný výsledek. Proto je někdy vhodnější volba jiného kritéria, např. Euklidovská norma

$$\left\| x(t) - \sum_{i=0}^{m} \tilde{a}_i g_i(t) \right\| \xrightarrow[p]{} min, \tag{4.82}$$

která se vyznačuje větší stabilitou vzhledem k výše popsaným vlastnostem normy pro $t = 1, \dots, T$. Uvedený přístup charakterizuje minimalizaci volatility reziduální složky. Celá optimalizace se provádí jen na množině $t\epsilon\{(T+1), \dots, (T+\tau)\}$ proto, aby hodnota parametru p byla „významně“ ovlivněna (určena) již známým průběhem na $t\epsilon\{1, 2, \dots T\}$ a „minimálně“ ovlivňovala $\tilde{a}_i$ (byla ovlivněna odhadem budoucnosti).

Pokud není (tvar) odhad trendu $z(t,p)$ k dispozici, je možné využít tzv. „prodloužení“ přímkou vzhledem ke svým nesporným vlastnostem

$$z(t,p) = x(T) + p(t-T), \qquad t\epsilon\{(T+1), \dots, (T+\tau)\}. \tag{4.83}$$

Nebo lze použít i jiné „prodloužení“, např. pro účely této práce je použito „prodloužení“ ve tvaru

$$z(t,p) = x(T)p^{(t-T)}, \qquad t\epsilon\{(T+1), \dots, (T+\tau)\}, \tag{4.84}$$

tj. vybraný indexový průběh, který je vhodný pro některé ekonomické řady. Takových prodloužení je samozřejmě více. Jejich typ a případně i parametr (parametry) by měl být podstatně ovlivněn aplikační oblastí, odkud pocházejí data.

4.11.1 Výběr množiny indexů ortonormálních složek I

Výběr „vhodné“ množiny indexů ortonormálních složek je založen na heuristickém postupu, který je použitelný i obecně, nejen pro uvedený prediktivní trend. Vybraný postup je inspirován z tzv. Parsevalovy rovnosti (4.40), kterou lze vyjádřit s odhadovanými koeficienty jako

$$\|\tilde{x}\|^2 = \sum_{t=1}^{T+\tau} \tilde{x}^2(t) = \sum_{t=0}^{T+\tau-1} \tilde{a}_i^2. \tag{4.85}$$

Při výběru je nejprve seřazena posloupnost kvadrátů odhadovaných koeficientů $\tilde{a}_i^2$ sestupně, tj.

$$\tilde{a}_{(0)}^2 \geq \tilde{a}_{(1)}^2 \geq \tilde{a}_{(2)}^2 \geq \cdots \geq \tilde{a}_{(m)}^2. \tag{4.86}$$

Dále je důležitý předpoklad, že největší koeficienty (zde ve smyslu $|\tilde{a}_i|$) modelují systematickou složku řady $\tilde{x}(t)$ a zbývající koeficienty náhodnou složku. Náhodná složka mívá navíc tzv. „rovnoměrné spektrum“, kdy od nějakého (l) je „prakticky splněna“ následující rovnost (hodnot) mezi koeficienty

$$\tilde{a}_{(l)}^2 = \tilde{a}_{(l+1)}^2 = \tilde{a}_{(l+2)}^2 = \cdots = \tilde{a}_{(T+\tau-1)}^2 = \tilde{a}^2. \tag{4.87}$$

Z uvedeného plyne, že je potenciálně získána posloupnost ve tvaru

$$\tilde{a}_{(0)}^2 \geq \tilde{a}_{(1)}^2 \geq \tilde{a}_{(2)}^2 \geq \cdots \geq \tilde{a}_{(l)}^2 = \tilde{a}^2 = \tilde{a}^2 = \cdots = \tilde{a}^2 = \tilde{a}_{(T+\tau-1)}^2. \tag{4.88}$$

Nechť tedy existuje množina $L = \{0,1, \dots, (l-1)\}$ a zároveň některá její netriviální nadmnožina J tak, že platí $L \subset J$ a $J \subset \{0, 1, \dots, (T+\tau-1)\}$. Potom je možné označit odhad trendové složky s použitím vybrané množiny indexů J jako $\tilde{X}_J(t)$ a množinu indexů s počtem prvků zvýšeným o jeden jako $J_{+1}(t)$, shrnuto:

$$\begin{aligned}\tilde{X}_J(t) &= \sum_{i\in J} \tilde{a}_i g_i(t),\\ J_{+1}(t) &= J \cup \{k\}, \quad k \in \bar{J} = \{0,1,\dots,T+\tau-1\} - J,\\ \tilde{X}_{J_{+1}}(t) &= \sum_{i\in J_{+1}}^{T} \tilde{a}_i g_i(t).\end{aligned} \tag{4.89}$$

V dalším postupu lze vyjádřit kvadrát Euklidovské normy pro odhadované trendové složky s použitím vybrané množiny indexů jako

$$\begin{aligned}\left\|\tilde{X}_J\right\|^2 &= \sum_{i\in J} \tilde{a}_i^2,\\ \left\|\tilde{X}_{J_{+1}}\right\|^2 &= \sum_{i\in J} \tilde{a}_i^2 + \tilde{a}^2 = \left\|\tilde{X}_J\right\|^2 + \tilde{a}^2.\end{aligned} \tag{4.90}$$

S využitím předpokladu o „rovnosti“ koeficientů lze upravit Parsevalovu rovnost jako

$$\|\tilde{x}\|^2 = \sum_{i\in J} \tilde{a}_i^2 + |\bar{J}|\tilde{a}^2 = \left\|\tilde{X}_J\right\|^2 + |\bar{J}|\tilde{a}^2, \tag{4.91}$$

kde $|\bar{J}|$ značí počet prvků množiny $\bar{J}$, která obsahuje nevybrané indexy k odhadu neznámé trendové složky. Získaný vztah lze využít pro vyjádření $\tilde{a}^2$, tedy

$$\tilde{a}^2 = \frac{\|\tilde{x}\|^2 - \left\|\tilde{X}_J\right\|^2}{|\bar{J}|} \tag{4.92}$$

a odtud je možné dále získat vztah pro vyjádření $\left\|\tilde{X}_{J_{+1}}\right\|^2$ s využitím předchozích poznatků jako

$$\begin{aligned}\left\|\tilde{X}_{J_{+1}}\right\|^2 &= \left\|\tilde{X}_J\right\|^2 + \tilde{a}^2\\ &= \left\|\tilde{X}_J\right\|^2 + \frac{\|\tilde{x}\|^2 - \left\|\tilde{X}_J\right\|^2}{|\bar{J}|}\\ &= \left\|\tilde{X}_J\right\|^2 \left(1 - \frac{1}{|\bar{J}|}\right) + \frac{\|\tilde{x}\|^2}{|\bar{J}|}.\end{aligned} \tag{4.93}$$

Uvedený výsledek je možné při triviální úpravě převést na tvar

$$\frac{\left\|\tilde{X}_{J_{+1}}\right\|^2}{\|\tilde{x}\|^2} = \frac{\left\|\tilde{X}_J\right\|^2}{\|\tilde{x}\|^2}\left(1 - \frac{1}{|\bar{J}|}\right) + \frac{1}{|\bar{J}|}, \tag{4.94}$$

kde takto získaná reálná čísla $\frac{\left\|\tilde{X}_{J_{+1}}\right\|^2}{\|\tilde{x}\|^2}$ a $\frac{\left\|\tilde{X}_J\right\|^2}{\|\tilde{x}\|^2}$ jsou při daném J dostupná. Pro získání neznámé množiny indexů L je vybrána taková množina J, pro kterou je rovnost uvedená výše splněna co nejlépe, např.

$$\left|\left(\frac{\left\|\tilde{X}_{J_{(i)}}\right\|^2}{\|\tilde{x}\|^2}\right)-\left(\frac{\left\|\tilde{X}_{J_{(i-1)}}\right\|^2}{\|\tilde{x}\|^2}\left(1-\frac{1}{|\bar{J}_{(i-1)}|}\right)+\frac{1}{|\bar{J}_{(i-1)}|}\right)\right| \xrightarrow[J]{} min, \tag{4.95}$$

kde $J_{(i-1)}=\{(0),(1),\dots,(i-1)\}$. Přesněji řečeno, cílem této heuristiky je nalézt takový „nejmenší“ index i, při kterém má výraz

$$\left|\left(\frac{\left\|\tilde{X}_{J_{(i)}}\right\|^2}{\|\tilde{x}\|^2}\right)-\left(\frac{\left\|\tilde{X}_{J_{(i-1)}}\right\|^2}{\|\tilde{x}\|^2}\left(1-\frac{1}{|\bar{J}_{(i-1)}|}\right)+\frac{1}{|\bar{J}_{(i-1)}|}\right)\right|, \tag{4.96}$$

„první“ lokální minimum. Tento postup je umožněn tím, že pro reálná data je použit předpoklad

$$\tilde{a}^2_{(l)}=\tilde{a}^2=\tilde{a}^2=\cdots=\tilde{a}^2=\tilde{a}^2_{(T+\tau-1)}, \tag{4.97}$$

který je splněn jen částečně.

4.12 Odhad spolehlivosti s použitím systematické složky

Vzhledem k odvozeným vztahům pro odhad trendové složky je nyní vhodné se vrátit k diskusi týkající se neznámé hodnoty spolehlivosti, jak již bylo rozebíráno dříve. Účel této části spočívá v popisu výsledných změn s použitím získané trendové křivky, které se vyskytnou při výpočtu spolehlivosti nebo naopak selhání. Předpokládaný model pro odhad neznámé trendové křivky je detailně popsán v předchozí části a uvedený vztah lze přepsat jako

$$X(t)=X_m(t)+\varepsilon_{x_t}, \qquad t=1,\dots,T, \tag{4.98}$$

kde $X_m(t)$ označuje odhadovanou trendovou složku s vybranou množinou indexů ortonormálních složek a ε_{x_t} označuje náhodnou složku veličiny X. Stejný vtah je využit i pro druhou veličinu

$$Y(t)=Y_m(t)+\varepsilon_{y_t}, \qquad t=1,\dots,T, \tag{4.99}$$

kde $Y_m(t)$ je odhadovaná trendová křivka s vybranou množinou indexů ortonormálních složek a ε_{y_t} je opět náhodná složka veličiny Y. Potom lze vyjádřit spolehlivost R jako $P(X<Y)$, kde veličiny X a Y jsou nahrazeny výše uvedenými vztahy (4.98) a (4.99) obsahujícími trendovou složku:

$$\begin{aligned} R &= P(X < Y) \\ &= P\{(X_m(t) + \varepsilon_{x_t}) < (Y_m(t) + \varepsilon_{y_t})\} \\ &= P\{(\varepsilon_{x_t} - \varepsilon_{y_t}) < (Y_m(t) - X_m(t))\} \\ &= F_{\varepsilon_x - \varepsilon_y}(Y_m(t) - X_m(t)), \qquad t = 1, \dots, T. \end{aligned} \tag{4.100}$$

Z uvedeného postupu je patrná závislost výsledné hodnoty spolehlivosti na distribuční funkci vzniklé z rozdílu náhodných složek obou veličin$F_{\varepsilon_x - \varepsilon_y}$. Zde je nezbytné upozornit na skutečnost, že pro získání výsledné hodnoty je důležitá znalost celé distribuční funkce a ne pouze znalost její hodnoty v jednom bodě, jak bylo diskutováno v předchozí kapitole. Tam byla doposud předmětem zájmu pravděpodobnost $P(X - Y < 0)$ vyjádřena jako distribuční funkce rozdílu náhodných veličin v bodě $F_{X-Y}(0)$.

4.13 Nezápornost platební bilance modelovaná spolehlivostí

Práce je situována do aplikační sféry, konkrétně na souborech reálných dat získaných z platební bilance České republiky, která je detailně popsána v úvodu kapitoly. Z platební bilance je vybrána bilance obchodní, která v sobě zahrnuje vývoz (import) a dovoz (export) v mil. Kč. Význam spočívá nejen v systematicky seřazených informacích o daném státě, ale i ve vlivu na cenu zboží a služeb, tedy vlivu na vývoz a dovoz a tím na ekonomiku země. Výběr uvedených dat je založen na skutečnosti, že mezinárodní obchod zboží tvoří (obvykle) největší položku v platební bilanci.

Pokud se zaměříme na vzniklý přebytek (nezápornost platební bilance, kladné saldo) nebo schodek (zápornost platební bilance, záporné saldo), potom záporné saldo ukazuje, že peněžní výdaje země do zahraničí jsou vyšší než peněžní příjmy z transakcí s ostatními ekonomikami a země se stává dlužníkem vůči zahraničí. Jinými slovy, zahraniční zadluženost lze popsat skutečností, že běžný účet (obchodní bilance) je dlouhodobě záporný a musí být kompenzován kladným finančním účtem, tj. vypůjčením kapitálu (nejen) v zahraničí. Tato skutečnost má dále vliv na devizový kurz[52] měny (hodnoty vývozu a dovozu). Pokud se obchodní bilance nachází v záporných hodnotách, potom je dovoz vyšší než vývoz tj. ze země „odtéká“ více peněz než do ní „přitéká“ a subjekty poptávají „relativně“ více zahraniční měny (s podstatnými efekty na „zásoby“ cizoměnových, případně i „zlatých“ aktiv), což má vliv na oslabení české měny a naopak. Země s dlouhodobě zápornou hodnotou obchodní bilance se označují jako země s „nestabilní ekonomikou“ mající negativní vliv na vnitřní ekonomické procesy. Proto je téma kladných a záporných obchodních bilancí v posledních několika letech často diskutováno na mezinárodní scéně, zejména s nástupem současné „ekonomické recese“.

[52] Devizový kurz charakterizuje cenu cizí měny vyjádřenou v jednotkách měny domácí.

Důležitý je popis spolehlivosti (zde nezápornosti platební bilance), která je reprezentována pravděpodobností kladného výsledku obchodní bilance, respektive selhání je reprezentováno pravděpodobností záporného výsledku obchodní bilance za uvažované období. Získané výsledky mohou být odhadovány a následně interpretovány jak v krátkodobém časovém horizontu, tak v dlouhodobém a výsledná hodnota spolehlivosti (selhání) v jednotlivých pozorováních může být chápána jako informace s menší váhou (důležitostí) než výsledná hodnota spolehlivosti z celého časového intervalu naměřených pozorování. Např. záporná hodnota platební bilance v jednom měsíci není příliš „zajímavá a důležitá“ pro vládní představitele či „ekonomy“ jako záporná hodnota platební bilance za kalendářní (fiskální) rok.

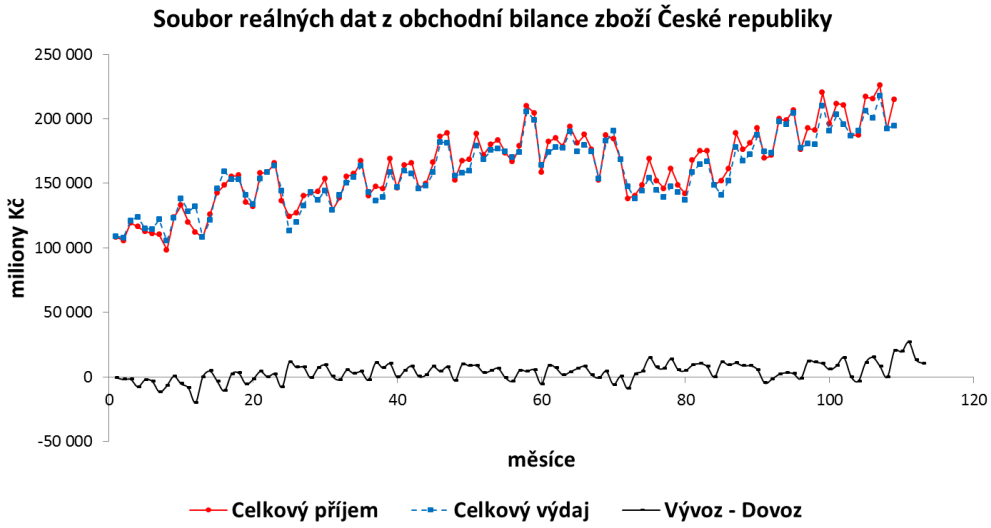
Cílem jednotlivých zemí je udržet vyrovnanou (kladnou, zvyšuje devizové zásoby, obvykle u centrální banky) platební bilanci, tedy uchování měnové stability, což znamená zároveň nepřidávat peníze do oběhu a nezvyšovat tak tlak na inflaci v zemi. Tento přístup může být intuitivně chybný, protože lze utratit jen peníze vytvořené za dané období (účetní fikce). Pokud se pozornost zaměří na popis „uvažovaného období“, potom z podstaty věci (kumulace peněz) je možné používat i peníze z přebytku vytvořeného za minulá období, tj. peníze z rozdílu mezi příjmy a výdaji před aktuálním (posledním) období. Tento přístup lze považovat též za účetní fikci, protože vzniká problém s určení délky uvažovaného období, tj. období, ve kterém byly peníze vydělávány (kumulativní hodnoty). Proto se jeví i určení „skutečného počátku“ za problematické.

V uvedeném popisu vznikají další problémy, např. statistika vzniklých kumulací už je sama o sobě problematická, protože vytvořené modely jsou modely pravděpodobnostní a nikoliv statistické.

Použité vysvětlení je velice stručné, protože hlavní myšlenka těchto komentářů spočívá v aplikaci této práce na získaná pozorování z reálného ekonomického prostředí.

4.13.1 Obchodní bilance v aktuálních hodnotách

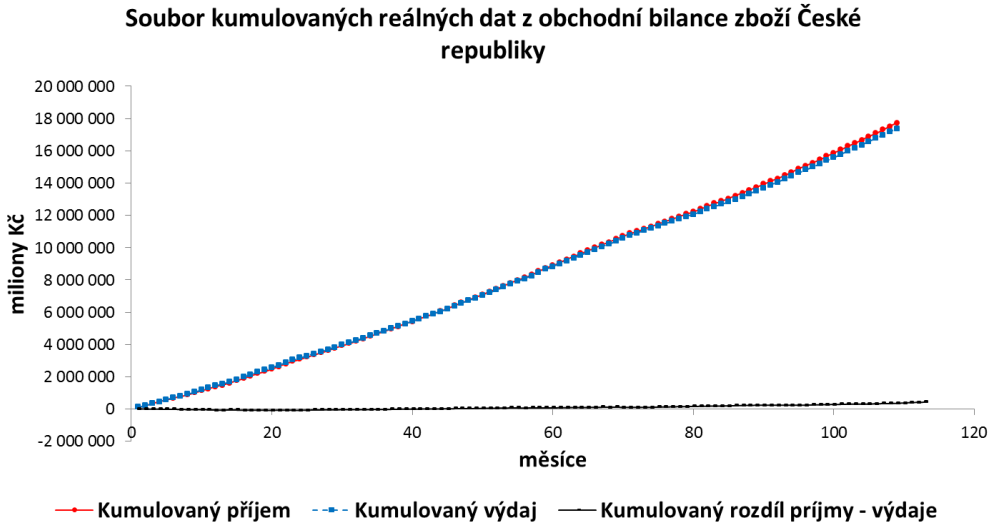
Prezentace získaných souborů pozorování z platební bilance, přesněji obchodní bilance z celkových příjmů za vývoz a celkových výdajů za dovoz v aktuálních hodnotách, je vykreslena na Obrázku 10. Celkové příjmy jsou znázorněny plnou čárou červené barvy a celkové výdaje čárou přerušovanou barvy modré. Rozdíl mezi příjmy a výdaji (obchodní bilance) je vykreslen níže černou barvou, kde si lze povšimnout převahy kladného salda obchodní bilance ve sledovaném období. Hodnoty jsou vykresleny od 1. 1. 2003 do 31. 5. 2012, tj. 113 měsíčních pozorování, vše v milionech Kč.



Obrázek 10: *Získaný soubor reálných dat v aktuálních měsíčních hodnotách z obchodní bilance České republiky*

4.13.2 Obchodní bilance v kumulacích

Prezentace získaných souborů pozorování z obchodní bilance v kumulovaných hodnotách je vykreslena na Obrázku 11. Celkové kumulované příjmy jsou znázorněny plnou čárou červené barvy a celkové kumulované výdaje čárou přerušovanou modré barvy. Kumulovaný rozdíl mezi příjmy a výdaji (kumulovaná bilance) je vykreslen níže, kde si lze potvrdit předchozí tvrzení o převaze kladného salda obchodní bilance ve sledovaném období. Hodnoty jsou vykresleny ve stejném období do 113 pozorování, tj. poslední hodnota značí součet všech 113 hodnot.



Obrázek 11: *Získaný soubor kumulovaných reálných dat z obchodní bilance České republiky*

Kapitola 5

Konkrétní prezentace a ověření

Kapitola popisuje konkrétní prezentace a ověření navrhovaných vztahů a modelů na souboru vygenerovaných dat a získaných (reálných) pozorování. V úvodu je prezentována vlastní tvorba na souboru vygenerovaných dat pro popis vlastností neparametrických jádrových odhadů. Následují získané výsledky a z nich odvozené závěry na souboru reálných dat jak v aktuálních hodnotách, tak v kumulacích (obchodní bilance). Závěr kapitoly popisuje celkové shrnutí provedené práce a popis ověřovacího systému (vytvořeného uživatelského programu) v softwaru MATLAB 2010a.

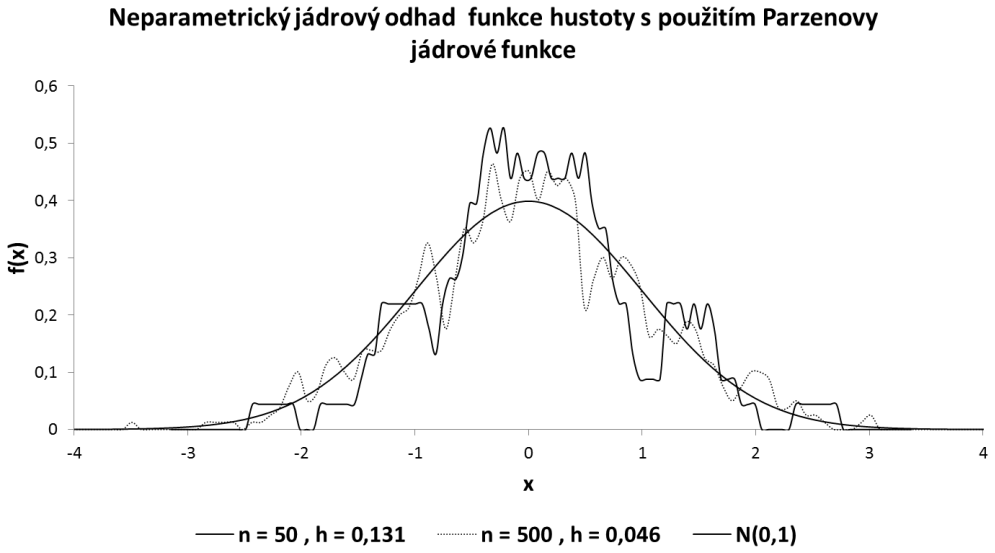
5.1 Vlastní realizace, realizace a využité vztahy

Vlastní realizace popisuje analýzu neparametrických jádrových odhadů a jejich vlastností na vybraném souboru vygenerovaných dat, který se řídí (pro názornost) normálním rozdělením pravděpodobnosti. Předmětem zájmu jsou vlivy použitých proměnných na výsledné hodnoty odhadu (aproximace), dvourozměrný neparametrický jádrový odhad a modelování spolehlivosti (vzhledem k předpokládanému stacionárnímu charakteru dat).

5.1.1 Jednorozměrný soubor vygenerovaných dat

Pozornost je v prvé řadě zaměřena na neparametrický jádrový model pro odhad hustoty a distribuční funkce. Kvalita odhadu je ovlivněna rozsahem pozorování, hodnotou vyhlazovacího parametru a tvarem jádrové funkce. Pokud jsou k dispozici různé rozsahy „náhodně“ vygenerovaných[53] souborů (zde pro názornost 50 a 500), potom si lze povšimnout „menších“ hodnot vychýlení u souboru dat s větším rozsahem. Získané výsledky odhadu „neznámé“ hustoty z vygenerovaných souborů dat jsou uvedeny v následujícím Obrázku 12, kde je intuitivně prezentován empirický předpoklad „přijatelnější“ aproximace hustoty pro rozsáhlejší soubory dat. Použité parametry vyhlazení jsou odvozeny ze vztahu (3.53). Zde stojí za povšimnutí odlišné hodnoty obou vyhlazovacích parametrů h vzhledem k funkční závislosti na pevném rozsahu n.

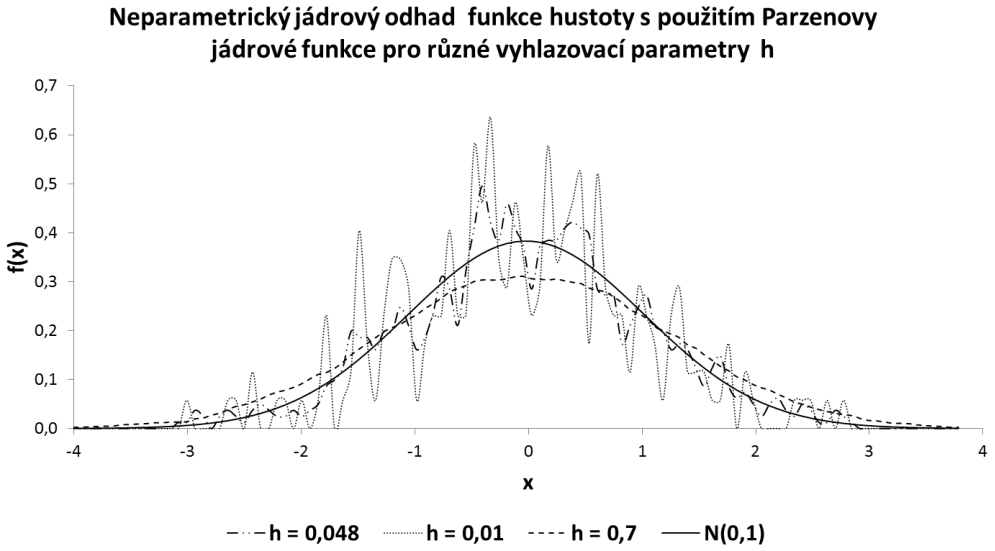
[53] Vzhledem k použitému software se zde používá pojem náhodný, ale ve své podstatě se jedná o generování pseudonáhodných čísel.



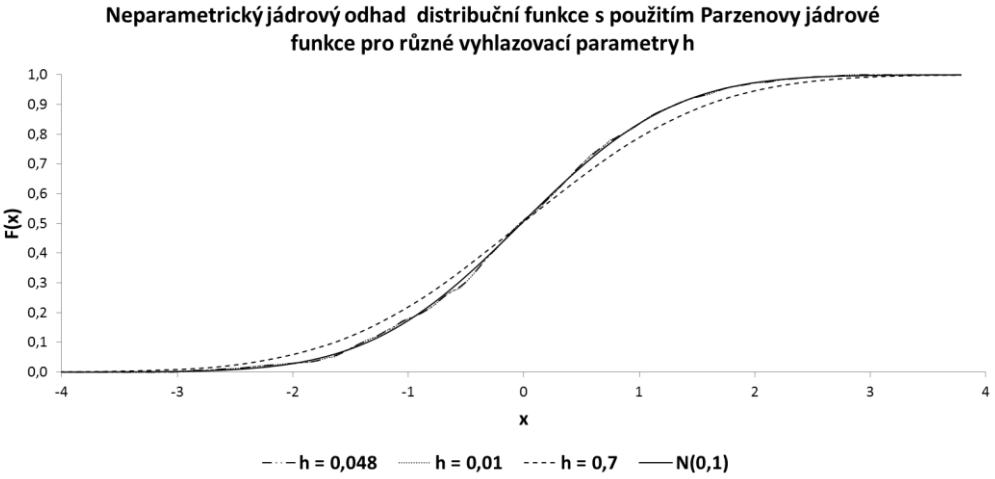
Obrázek 12: *Neparametrický jádrový odhad hustoty používající Parzenovu jádrovou funkci pro vygenerované soubory dat s rozdílným rozsahem* n

Poznámka 5.1 *Všechny zde uvedené obrázky jsou vykresleny v softwaru Microsoft Excel a při použití Parzenovy jádrové funkce mohou být mírně zkreslené. To znamená, že vykreslené křivky jsou oproti „realitě“ mírně vyhlazenější (viz Obrázek 12 výše, kdy není přesně kopírován tvar Parzenovy jádrové funkce).*

Nedostatkem neparametrických jádrových odhadů je vliv vyhlazovacího parametru h, který je široce popisován v předchozí části. Problém spočívá ve volbě „vhodné“ (optimální) hodnoty tohoto parametru, protože v případě „malé“ nebo naopak „velké“ hodnoty nedostaneme „předpokládaný“ tvar hustoty a distribuční funkce (odlišné tvary od tvaru vygenerovaného rozdělení pravděpodobnosti). Důsledky „nevhodné“ volby jsou prezentovány na Obrázku 13 a 14, kde jsou použity 3 odlišné hodnoty parametru. První hodnota je odvozena ze vztahu (3.53) a je použita i v dalších modelech pro získání spolehlivosti (selhání) z důvodu neznalosti „správného (původního)“ rozdělení. Druhá hodnota parametru je úmyslně nižší pro ukázku “nedostatečně vyhlazeného (podhlazeného)” odhadu a třetí hodnota parametru je mnohem vyšší než předchozí dva parametry pro ukázku „přehlazeného“ odhadu.



Obrázek 13: *Neparametrický jádrový odhad hustoty používající Parzenovu jádrovou funkci s různými hodnotami vyhlazovacího parametru h*

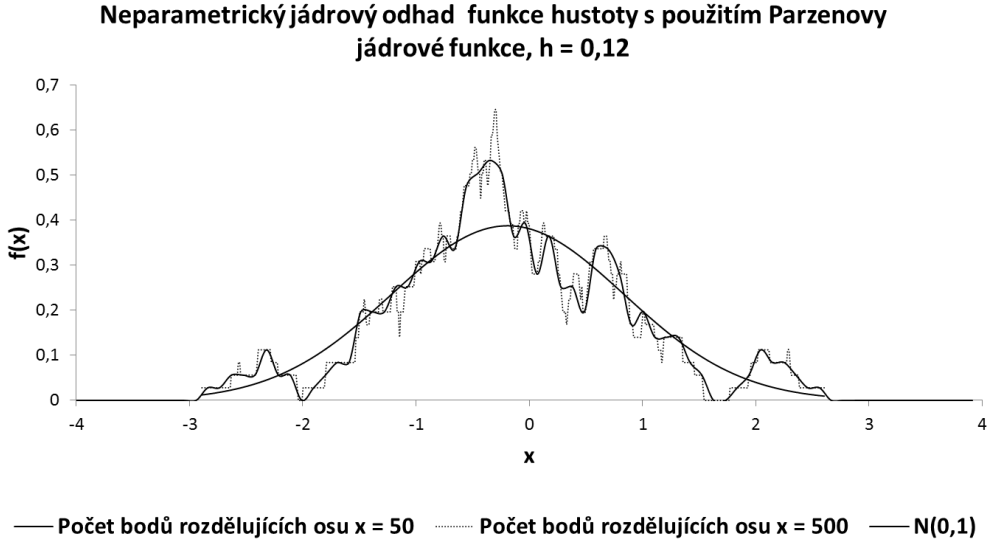


Obrázek 14: *Neparametrický jádrový odhad distribuční funkce používající Parzenovu jádrovou funkci s různými hodnotami vyhlazovacího parametru h*

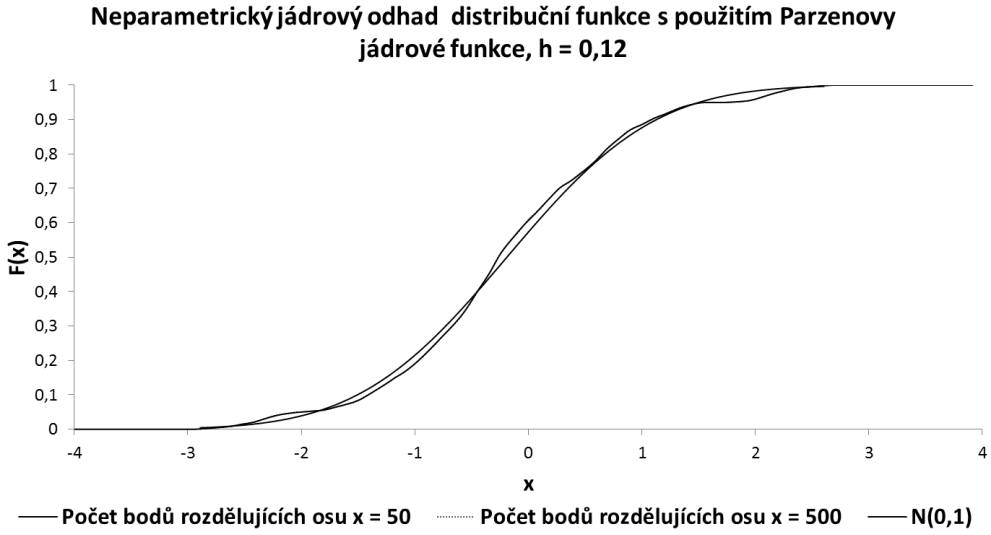
Z uvedených obrázků je patrné, že „kvalita“ odhadu je z největší části ovlivněna právě tímto parametrem. Zároveň je vhodné připomenout předpoklady, že hodnota vyhlazovacího parametru h je určena reálným číslem (nepředpokládá se jeho příslušnost k náhodné veličině) a v případě souboru reálných pozorování není apriorně známo pravděpodobnostní rozdělení. Nelze tedy použít “optimální” hodnoty vyhlazovacího parametru, publikované např. v Devroye a Gyorfi (1985), kde jsou odvozeny postupy a úvahy pro několik vybraných jádrových funkcí.

Problematika vyhlazovacího parametru představuje velmi široké téma a z tohoto důvodu není podrobný rozbor optimálního řešení cílem této disertační práce. Zde zvolený přístup představuje jeden možný návrh z mnoha a detailnější analýza představuje námět

pro následné rozšíření práce. Poslední proměnnou, která má neoddiskutovatelný vliv na tvar a tedy i vzhled odhadované hustoty a distribuční funkce je počet bodů „rozdělujících“ osu x, ve kterých jsou prováděny požadované odhady. Vliv této proměnné je graficky ukázán na Obrázku 15 a 16, kde je pomocí plné čáry označeno rozdělení osy na 50 bodů a pomocí „přerušované“ čáry rozdělení osy na bodů 500.



Obrázek 15: *Neparametrický jádrový odhad hustoty s odlišnými počty bodů na ose x pro $h = 0{,}12$*



Obrázek 16: *Neparametrický jádrový odhad distribuční funkce s odlišnými počty bodů na ose x pro $h = 0{,}12$*

Při konstrukci možných závěrů budeme vycházet z uvedených grafů, na jejichž podkladě se lze domnívat, že zde nejsou žádné „významné změny“ mezi uvažovanými aproximacemi obou hustot a zároveň po bližším prozkoumání distribučních funkcí je

patrné, že jsou „téměř stejné“. Dále je vhodné upozornit na skutečnost, že volba menšího počtu dělících bodů způsobuje více vyhlazenou křivku hustoty a naopak. Z uvedených důvodů je v celé práci osa x roztříděna na 100 hodnot, což má nesporný vliv na výpočetní složitost.

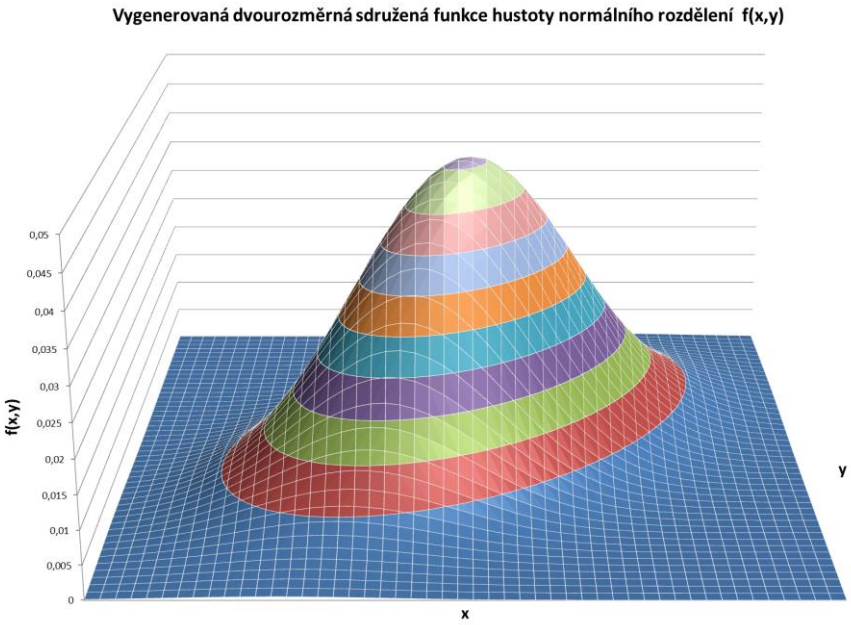
5.1.2 Dvourozměrný soubor vygenerovaných dat

Následuje analýza dvourozměrných neparametrických jádrových modelů a jejich vlastností na souboru vygenerovaných dat, který se řídí sdruženým normálním rozdělením pravděpodobnosti. Obsah kapitoly je podnětný pro analýzu spolehlivosti, protože jsou zde obsaženy dvě náhodné veličiny reprezentující dříve popisovanou reálnou situaci.

Vygenerovaný soubor náhodných[54] pozorování řídící se sdruženým normálním rozdělením pravděpodobnosti obsahuje následující parametry, kde μ_i představuje střední (očekávanou) hodnotu, σ_i je směrodatná odchylka a ρ korelační koeficient (jistá míra vyjadřující lineární závislost). Konkrétní hodnoty parametrů jsou vybrány pouze pro zjednodušení a nabývají hodnot:

μ_1	μ_2	σ_1	σ_1	ρ
0	0	4	4	0,5

Tabulka 2: *Hodnoty parametrů vygenerovaného dvourozměrného sdruženého normálního rozdělení*



Obrázek 17: *Vygenerovaná sdružená funkce hustoty normálního rozdělení* $f(x, y)$

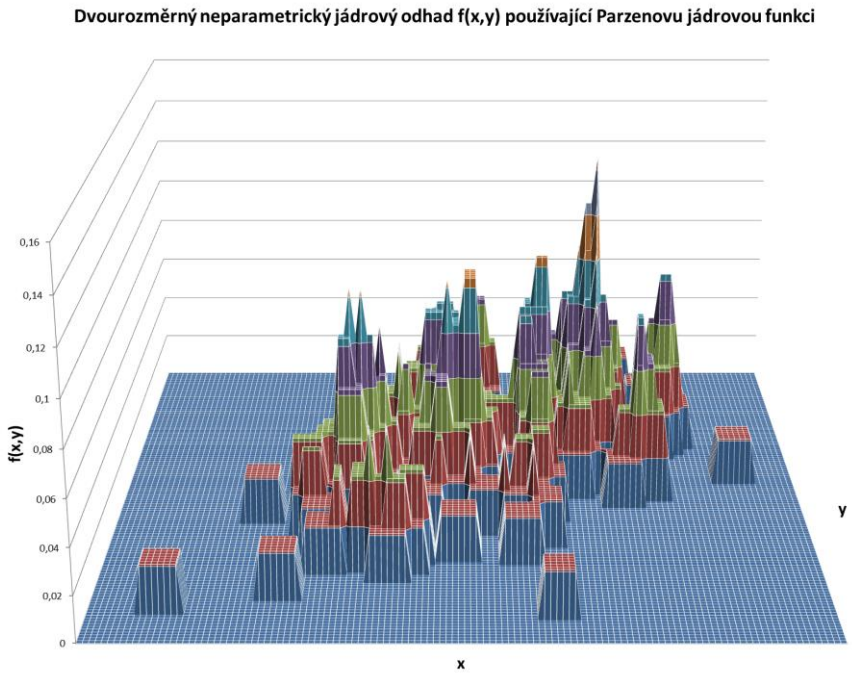
[54] Opět je zde používán soubor vygenerování pseudonáhodných čísel.

Výběr rozdělení je založen na skutečnosti, že se jedná o jedno z běžně používaných dvourozměrných pravděpodobnostních rozdělení. Vzhledem k výše uvedeným vlastnostem jsou použité jádrové funkce opět Parzenova a Gaussova jádrová funkce. Výsledná aproximace hustoty s použitím Parzenovy jádrové funkce je charakterizována „ostrými hranami“, jak je graficky ukázáno dále. Samozřejmě i zde dochází k mírnému zkreslení jako v jednorozměrném případě vzhledem k použitému softwaru. Naopak aproximace s výběrem Gaussovy jádrové funkce je „více vyhlazená“. Obrázek s odhadem používajícím Gaussovu jádrovou funkci je prezentován až na získaném souboru reálných dat.

Počet vygenerovaných náhodných pozorování je opět 100 a hodnoty použitých párových parametrů[55] měřítka jsou nejprve odvozeny ze vztahu (3.53) a potom násobeny (pro názornost) dvěma. Vliv proměnné je ukázán na obrázcích níže, kde je patrný „vyhlazenější“ tvar u aproximace s “vyšší” hodnotou parametru. Konkrétní hodnoty jsou uvedeny v Tabulce 3.

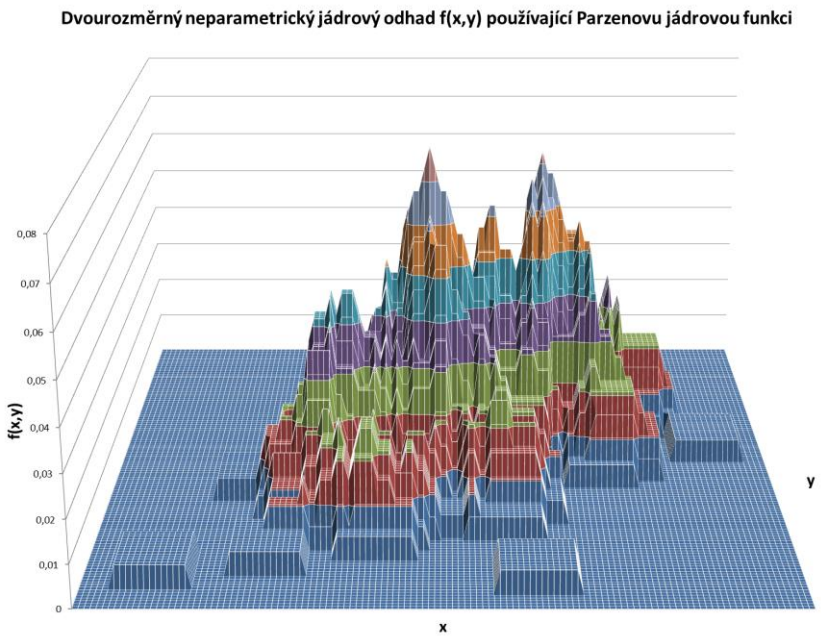
	$\boldsymbol{Odhad\ 1}$	$\boldsymbol{Odhad\ 2}$
$\boldsymbol{h_x}$	0,21	0,42
$\boldsymbol{h_y}$	0,19	0,38

Tabulka 3: *Vyhlazovací parametry dvourozměrného jádrového odhadu funkcí hustot a distribučních funkcí*



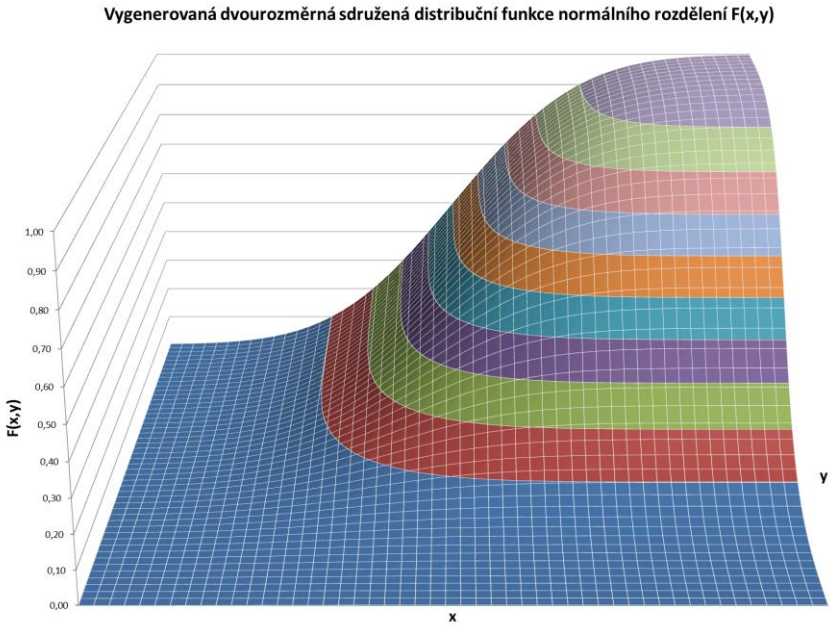
Obrázek 18: *Dvourozměrný neparametrický jádrový odhad hustoty používající Parzenovu jádrovou funkci s odvozenými parametry* h_x, h_y

[55] Dvourozměrný neparametrický jádrový odhad obsahuje dva vyhlazovací parametry.

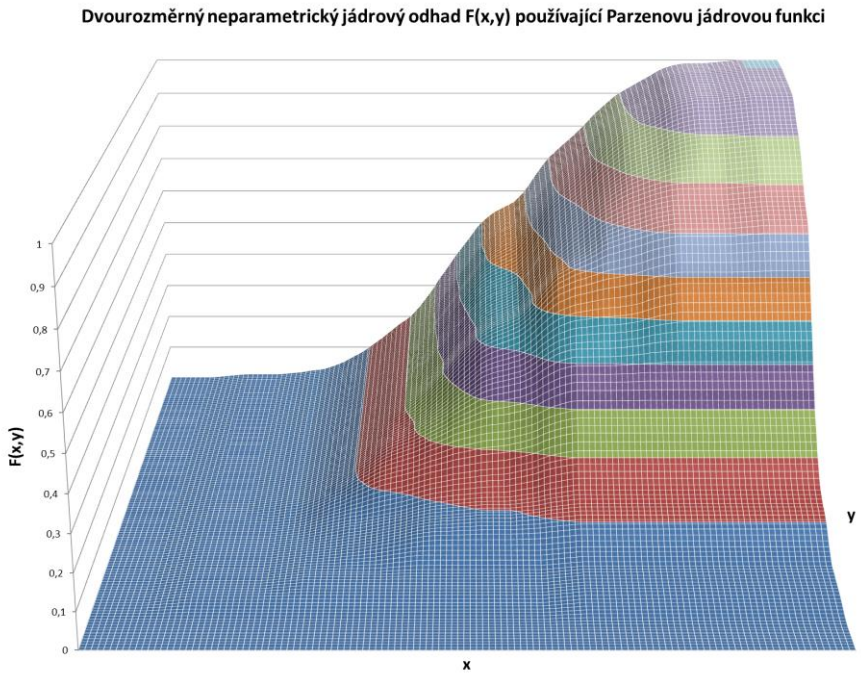


Obrázek 19: *Dvourozměrný neparametrický jádrový odhad hustoty používající Parzenovu jádrovou funkci s „většími“ vyhlazovacími parametry* h_x, h_y

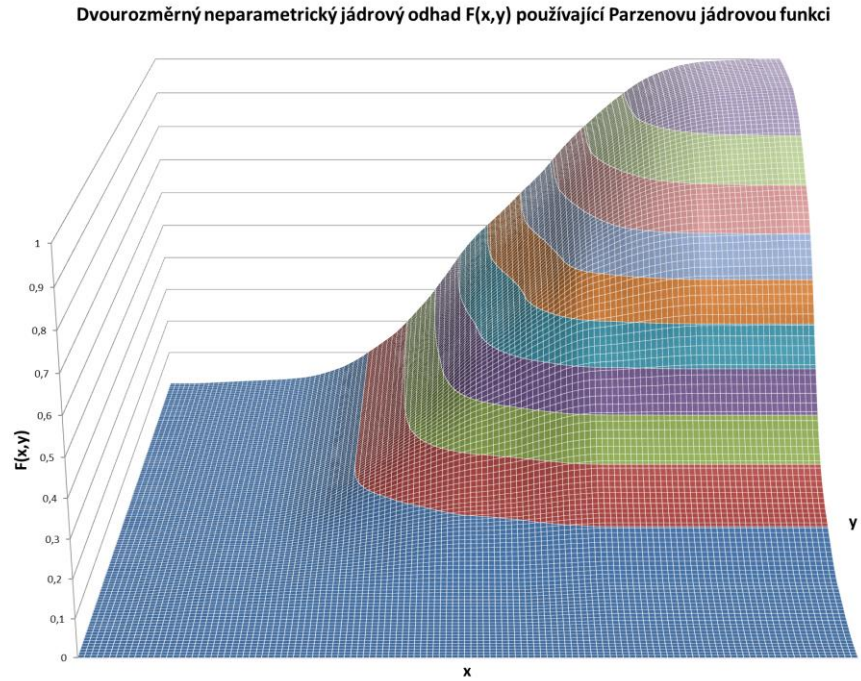
S ohledem na spolehlivost (selhání) jsou zde důležité neparametrické jádrové odhady distribučních funkcí, které jsou prezentovány na Obrázcích 20, 21 a 22. Zde nejsou “téměř žádné změny” mezi 3 vybranými distribučními funkcemi. První distribuční funkce je vygenerovaný sdružený model, který představuje předpokládaný pravděpodobnostní model daného rozdělení. Zbylé dvě funkce jsou odhadované distribuční funkce tohoto modelu s odlišně zvolenými vyhlazovacími parametry (viz výše).



Obrázek 20: Vygenerovaná dvourozměrná sdružená distribuční funkce normálního rozdělení $F(x, y)$



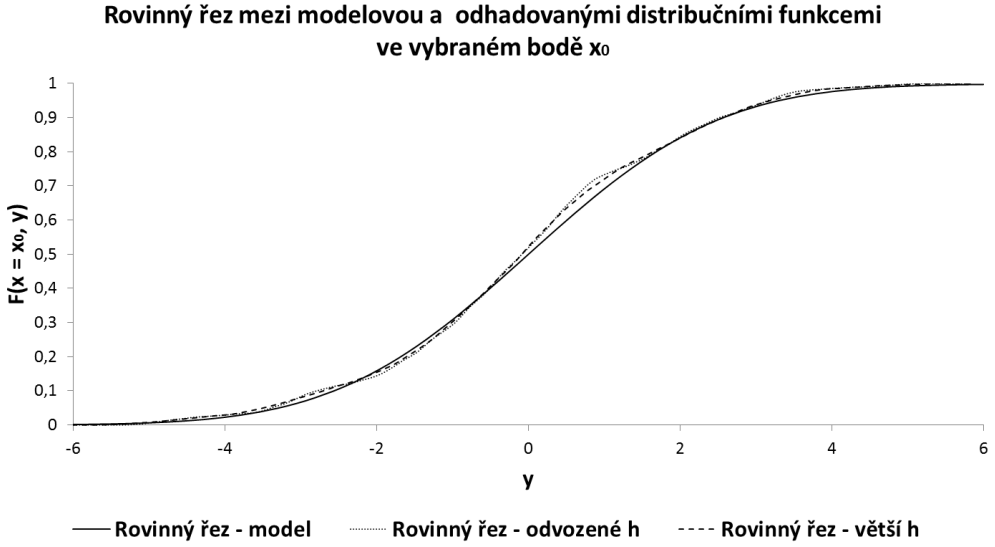
Obrázek 21: *Odhad distribuční funkce sdruženého normálního rozdělení používající Parzenovu jádrovou funkci s odvozenými parametry h_x, h_y*



Obrázek 22: *Odhad distribuční funkce sdruženého normálního rozdělení používající Parzenovu jádrovou funkci s „většími“ vyhlazovacími parametry h_x, h_y*

Dle grafické interpretace získaných výsledků lze usuzovat, že prezentované tvary distribučních funkcí jsou “téměř stejné”, tj. že zde nejsou téměř „žádné rozdíly“ mezi

výslednými pravděpodobnostmi. Při podrobnější analýze vlivu vyhlazovacího parametru je patrný vyhlazenější tvar hustoty pro vyšší hodnoty parametru, ale tvar distribuční funkce je „téměř stejný“. Zdůvodnění může být provedeno rovinným řezem v (libovolně) vybraném bodě na ose x, který je stejný pro všechny uvedené distribuční funkce. Získaný řez ukazuje vzniklé rozdíly mezi modelovou distribuční funkcí a odhadovanými distribučními funkcemi, jak je ukázáno na Obrázku 23.



Obrázek 23: *Rovinný řez mezi modelovanou a odhadovanými distribučními funkcemi*

5.1.3 Relativní frekvence

V předchozích částech jsou navrženy možné metody neparametrického jádrového odhadu k získání spolehlivosti, které budou nyní aplikovány na náhodném souboru vygenerovaných dat řídící se sdruženým normálním rozdělením pravděpodobnosti (viz Tabulka 2). Získané výsledky jsou následně porovnány s jinou možnou metodou, která je publikována pod názvem relativní frekvence.

Relativní frekvence je jedna z možných a jednodušších (časově, výpočetně, …) metod k získání neznámé hodnoty spolehlivosti. Metodu lze popsat jako podíl dvou hodnot, kde v čitateli je počet výskytu nějaké (sledované) náhodné události a ve jmenovateli je celkový počet pozorování. Popsaný výpočet je zde prezentován pro zdůraznění jednoduchosti a bude nadále používán pro závěrečné srovnání získaných výsledků spolehlivosti. Definice relativní frekvence je uvedena např. v Spiegel a Stephens (2008).

Definice 5.1 *Nechť $N(A)$ je počet výskytu vybrané náhodné události A v N opakováních náhodného experimentu, potom relativní frekvence F_A je*

$$F_A = \frac{N(A)}{N}. \tag{5.1}$$

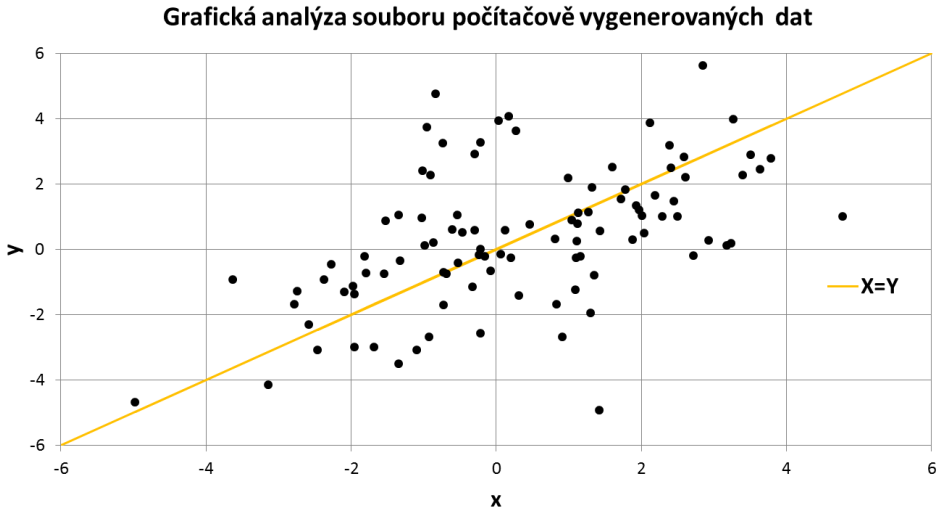
(převzato a přeloženo ze Spiegel a Stephens (2008), s. 32)

Poznámka 5.2 *Odhady pomocí relativní frekvence lze označit za dobré odhady díky její jednoduchosti a zároveň rychlosti celého výpočtu. Problém v jejím použití spočívá ve skutečnosti, že hodnota získaného výsledku je pouze zdánlivá. Dále je zde silný předpoklad, že výběrové náhodné soubory obsahují stacionární*[56] *data, což není v realitě často naplněno (náhodné výběry reálných dat), jak je detailně popsáno v Kapitole 4.*

Použitím relativní frekvence na vygenerovaném souboru dat (předpoklad stacionárního charakteru) jsme získali požadovanou hodnotu:

$$F_A = R = \frac{N(A)}{N} = \frac{51}{100} = 0{,}51. \tag{5.2}$$

Prvotní hodnota spolehlivosti s použitím relativní frekvence je přibližně 51 %. Grafická analýza souboru vygenerovaných dat, včetně přímky $x = y$ zvýrazňující a vymezující oblast pro $x < y$, je ukázána na Obrázku 24.



Obrázek 24: *Vygenerovaný soubor stacionárních dat sdruženého normálního rozdělení pravděpodobnosti,* $n = 100$

5.1.4 Neparametrický jádrový odhad k získání spolehlivosti

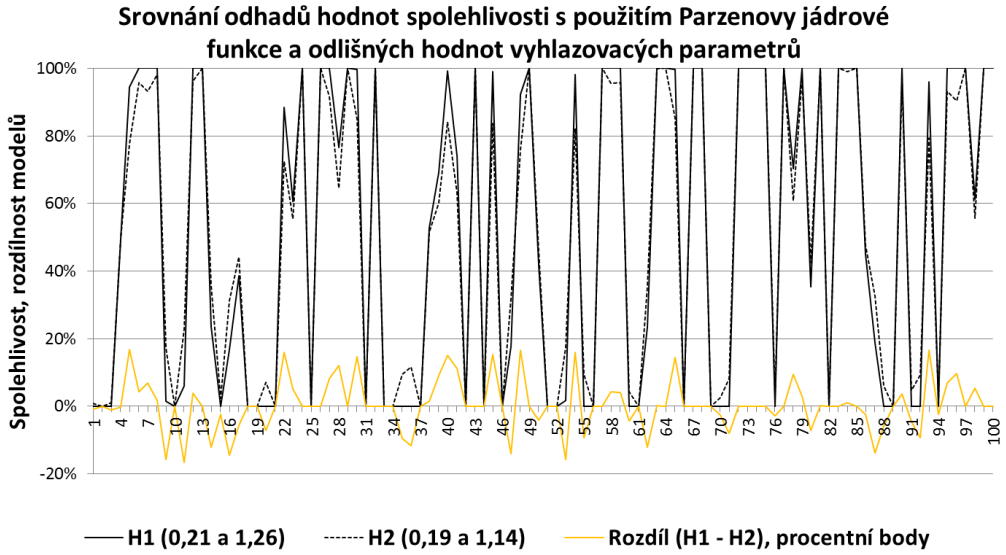
Testovací analýza je založena na modelu s Parzenovou jádrovou funkcí (3.113) pro dvě hodnoty parametrů měřítka h_x, h_y. Tento přístup umožňuje srovnání dvou získaných hodnot spolehlivosti na základě změny uvedených parametrů. První model obsahuje odvozené parametry ze vztahu (3.53) a druhý model obsahuje tyto hodnoty šest krát navýšené (zvolené navýšení je jen subjektivní a dostatečně velké pro demonstraci vlivu změny parametru). Analýza spolehlivosti byla provedena pro více parametrů měřítka, ale nebylo zde dosaženo žádných „významných“ změn ve výsledcích, a proto jsou zde

[56] Stacionarita reálných dat ze statistického pohledu spočívá v předpokladu, že marginální rozdělení jednotlivých pozorování jsou funkčně nezávislá na čase realizace.

prezentovány jen výsledky jedné změny (násobením) parametrů na Obrázku 25. Obrázek přehledně popisuje získané hodnoty spolehlivosti pro každý náhodně vygenerovaný pár ze souboru pozorování (hodnota každé zkoumané komponenty), kde pod zkratkou $H1$ jsou prezentovány parametry měřítka h_x a h_y pro první model a pod zkratkou $H2$ pro druhý model (získaný násobením parametrů).

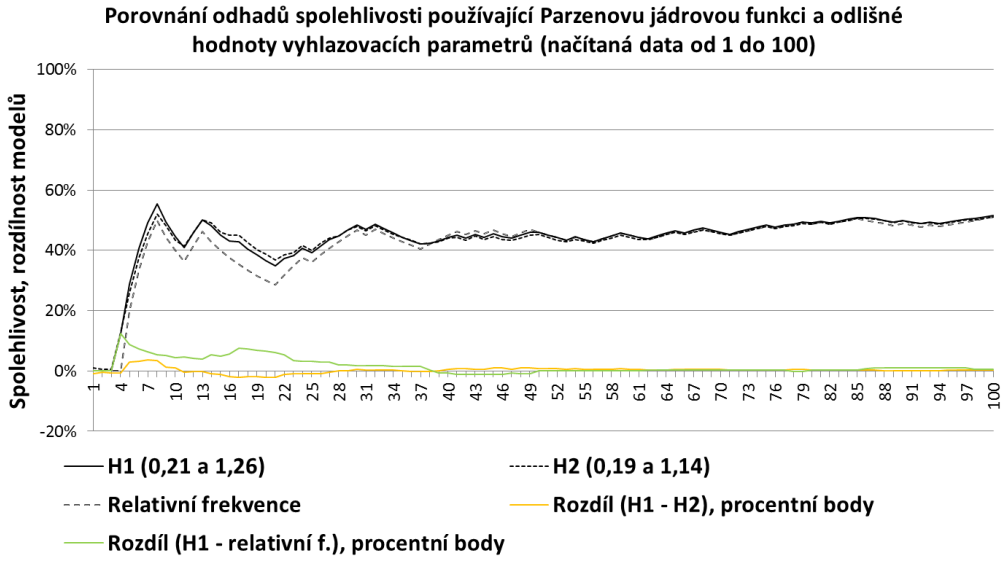
	$Odhad\ H1$	$Odhad\ H2$
h_x	0,21	1,26
h_y	0,19	1,14

Tabulka 4: *Vyhlazovací parametry pro neparametrický jádrový model k získání spolehlivosti*



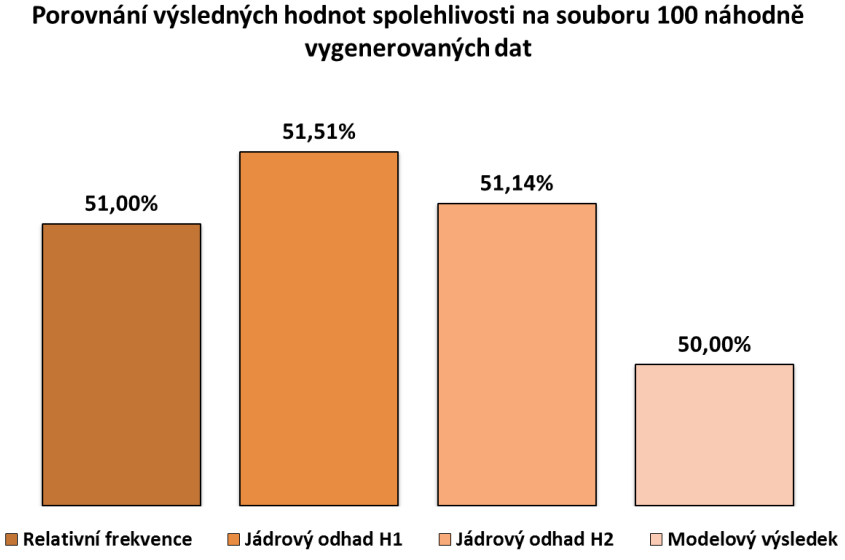
Obrázek 25: *Neparametrický jádrový model k získání spolehlivosti používající Parzenovu jádrovou funkci s odlišnými hodnotami parametrů měřítka pro každou ze 100 vygenerovaných párových hodnot*

Z uvedeného obrázku jsou patrné změny získaných hodnot spolehlivosti v „některých“ vygenerovaných bodech (x_i, y_i) pouze v „jednotkách“ procentních bodů, a proto jsou důležité celkové hodnoty spolehlivosti (pro všechny pozorování, v realitě období). Ty jsou uvedeny na Obrázku 26, kde jsou prezentovány výsledky modelu pro „postupně“ načítaná data. Pojem načítaná data označuje dopočítávané odhady středních hodnot od prvního do posledního vygenerovaného páru dat (tj. v 5 měření dostáváme „průměrnou“ hodnotu spolehlivosti za prvních 5 naměřených párů dat).



Obrázek 26: *Srovnání neparametrického modelu k odhadu spolehlivosti používající Parzenovu jádrovou funkci s rozdílnými hodnotami parametrů měřítka a s výsledky relativní frekvence – načítané hodnoty od 1 do i – tého vygenerovaného bodu (počet pozorování je 100 bodů)*

Získané výsledky slouží jako podklad pro srovnání (analýzu) navrženého neparametrického jádrového modelu spolehlivosti pro odlišné vyhlazovací parametry s hodnotou relativní frekvence. Výsledné hodnoty nejsou popsány ve větším detailu vzhledem k následné duplicitě výzkumu na souboru reálných dat a jsou ukázány na Obrázku 27.



Obrázek 27: *Porovnání získaných hodnot spolehlivosti na souboru 100 náhodně vygenerovaných dat mezi navrženým neparametrickým modelem s různými hodnotami parametrů měřítka a relativní frekvencí*

Na základě získaných hodnot spolehlivosti lze usuzovat, že získané výsledky navrženého modelu používajícího různé parametry měřítka jsou "téměř shodné" s hodnotou relativní frekvence, ale popsané odvození a výpočty jsou časově náročnější, a proto je zde větší pravděpodobnost možných výpočetních chyb. Vliv parametru měřítka není tak „výrazný" a možná cesta k získání spolehlivosti na souboru stacionárních dat je relativní frekvence díky její jednoduchosti a nenáročnosti na výpočet. Z tohoto důvodu se zde nabízí podezření, že závěry získané na základě neparametrických jádrových modelů k ověření spolehlivosti používající různé typy jádrových funkcí nejsou lepší než použití modelu relativní frekvence. Vše za předpokladu stacionárního charakteru dat.

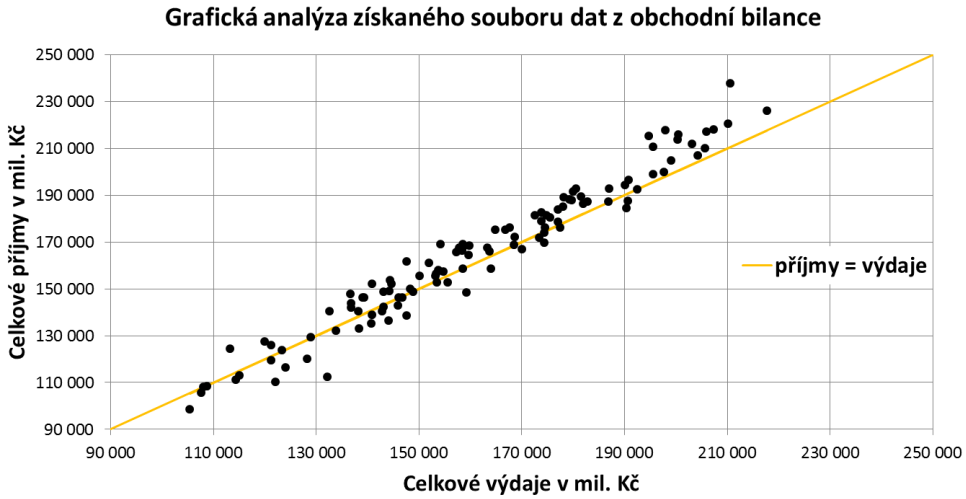
Pokud není tento předpoklad splněn, potom se lze zaměřit na vlastnosti získaného souboru pozorování. Soubor reálných pozorování je získáván z reálně se vyskytujících situací a často se vyznačuje nestacionárním charakterem, např. (makro)ekonomická data (v časových řadách). V uvedeném případě způsobuje obsažená nestacionarita zavádějící výsledky, tj. nepřesnou spolehlivost. Stacionarita reálných dat je zde v předpokladu, že marginální rozdělení jednotlivých pozorování (časových odečtů) jsou (funkčně) nezávislá na čase realizace (odečtu, zjištění). Nestacionarita je tedy taková situace, kdy existují alespoň dva časové odečty, u nichž nelze předpoklad stejného rozdělení přijmout. Obvykle lze takovou situaci ověřit mimo statistiku z fundamentální podstaty dat, nebo z efektů, které nemohou za předpokladu stacionarity dat nastat. Možná cesta ke „stacionarizaci" nestacionárního souboru dat již byla prezentována a podrobná analýza je provedena na získaných pozorování z platební bilance. V tomto případě bude nadále předpokládáno, že získaný výběrový soubor reálných nestacionárních dat může být „transformován" na soubor s charakterem stacionárních dat.

Poznámka 5.3 *Uvedená problematika není tak jednoduchá, jak se může na první pohled zdát z uvedeného textu. Existuje zde velké množství dalších problémů, které nelze bez odůvodněných příčin zanedbat a je třeba je s touto problematikou řešit. Některá možná řešení jsou prezentovány v následujících kapitolách.*

5.2 Výsledky v aktuálních hodnotách

Kapitola se zabývá verifikací navržených metod na získaném souboru reálných dat z platební (obchodní) bilance[57] České republiky v aktuálních (nekumulovaných) hodnotách od 1. 1. 2003 do 31. 5. 2012. Výše popsané metody zde umožní odhadnout jednorozměrné i dvourozměrné hustoty a distribuční funkce, odhalit trendovou složku a následně jsou použity k získání spolehlivosti (selhání). Nejprve je zde získána spolehlivost s použitím modelu relativní frekvence a následuje neparametrický jádrový odhad hustoty a distribuční funkce. K dispozici jsou dva soubory reálných dat, které jsou vykresleny pomocí bodového grafu níže v milionech Kč.

[57] Zdroj: www.cnb.cz



Obrázek 28: *Grafická analýza získaných souborů dat z obchodní bilance ČR*

Vykreslené body reprezentují párovaná pozorování celkových příjmů (osa y) a výdajů (osa x). Přímka vykreslená uvnitř obrázku popisuje situaci $x = y$, kdy jsou obě hodnoty stejné (výdaje = příjmy) ve stejný časový okamžik. Prostor nad vykreslenou přímkou reprezentuje skutečnost, kdy jsou hodnoty celkových příjmů vyšší než celkové hodnoty výdajů a naopak. Potom relativní frekvence pro prvotní pravděpodobnost selhání je

$$F = \frac{N(A)}{N} = \frac{34}{113} = 0{,}3009 = 30{,}09\ \%, \tag{5.3}$$

kde $N(A)$ jsou pozorování pod přímkou a N značí celkový počet získaných pozorování. Zároveň je možné získat hodnotu spolehlivosti

$$R = 1 - F = 1 - 0{,}3009 = 0{,}6991 = 69{,}91\ \%, \tag{5.4}$$

kde R označuje spolehlivost a F selhání.

Poznámka 5.4 *Pojmem spolehlivost je myšlena celková hodnota spolehlivosti za celý uvažovaný časový interval (tedy ze všech naměřených hodnot), tj. pravděpodobnost kladné obchodní bilance za celé měřené období (113 pozorování).*

5.2.1 Jednorozměrné neparametrické jádrové odhady

Před verifikací navržených modelů k získání spolehlivosti je vhodné analyzovat soubory reálných dat s použitím neparametrického jádrového odhadu hustoty a distribuční funkce. Pro tento účel je vykreslen celkový příjem z vývozu s použitím tří typů jádrových funkcí. Zdůvodnění výběru Parzenovy a Gaussovy jádrové funkce je popsáno výše a pro zajímavost je dále přidána kladná jádrová funkce popsaná níže, jak prezentoval Vávra a kol. (2003). Získané tvary hustot a distribučních funkcí se stejným vyhlazovacím parametrem h ze vztahu (3.53) jsou ukázány na Obrázku 29 a 30.

Definice 5.2 *Nechť X je nezáporná náhodná veličina s n nezávislými pozorováními $\{x_1, \dots, x_n\}$ a nechť $a > 0$ je kladné reálné číslo (zde je povolena závislost na počtu pozorování n a na hodnotách pozorování x_i), potom hustota a distribuční funkce mohou být modelovány pomocí vztahů*

$$\hat{f}(x; a, m) = \frac{1}{a}\sum_{i=1}^{n} k\left(\frac{n(x - x_i) + am}{a}\right), \quad a > 0, \quad m \geq 0, \quad n \in \mathbb{N}_+, \tag{5.5}$$

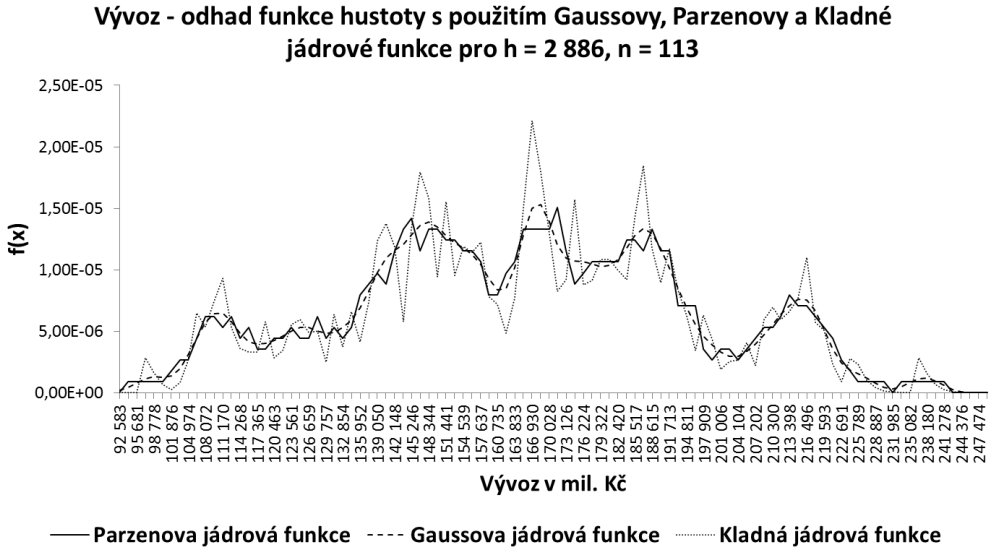
a

$$\hat{F}(x; a, m) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{n(x - x_i) + am}{a}\right), \quad a > 0, \quad m \geq 0, \quad n \in \mathbb{N}_+, \tag{5.6}$$

kde $k(x)$ je kladná[58] jádrová funkce hustoty, $K(x)$ je integrál z jádrové funkce hustoty a m je definováno integrálem $\int_0^\infty \big(1 - K(x)\big)dx$.

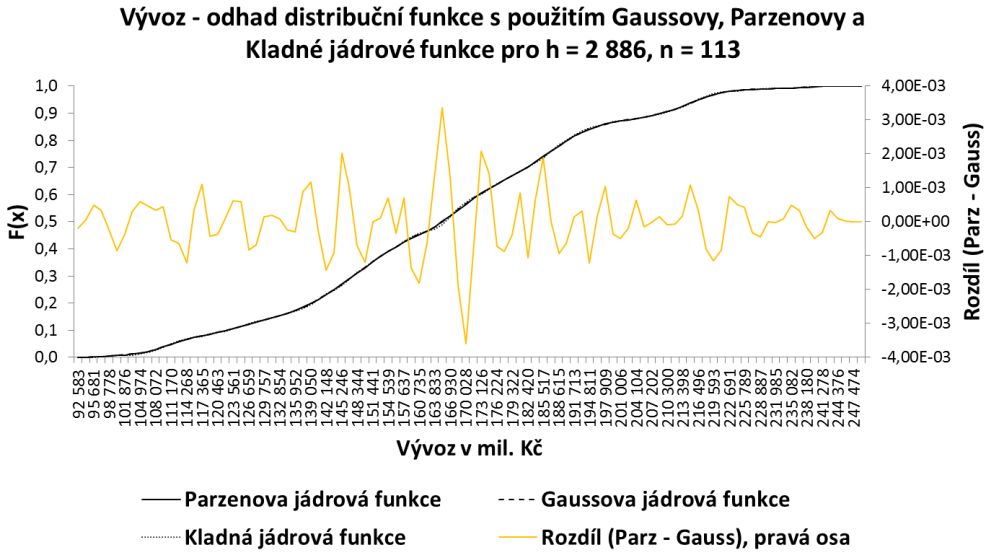
(citováno z Vávra a kol. (2003), s. 341 – 346)

Odhad parametru a a jeho vlastnosti jsou detailně diskutovány v odkazovaném článku.



Obrázek 29: *Neparametrický jádrový odhad hustoty na souboru reálných dat celkového množství příjmů v mil. Kč pro $h = 2\,886$, $n = 113$*

[58] Např. $k(x) = r(1 - e^{-x})^{r-1}e^{-x}$ pro $x > 0$ a $r \geq 1$, jinak 0.



Obrázek 30: *Neparametrický jádrový odhad distribučních funkce na souboru reálných dat celkového množství příjmů v Kč pro* $h = 2\,886$, $n = 113$

Získané hustoty používající různé typy jádrových funkcí mají „mírně odlišné“ tvary, ale odhady distribučních funkcí jsou “téměř stejné”, tj. jsou zde jen „málo“ viditelné změny mezi získanými distribučními funkcemi (viz měřítko změn). Odlišné tvary hustot s použitím Parzenovy a Gaussovy jádrové funkce jsou patrné z uvedených obrázků, kde odhad s použitím Parzenovy jádrové funkce lze považovat za více „kostrbatý“ (včetně mírného zkreslení softwaru Excel). Výběr jádrové funkce má tedy nezanedbatelný vliv na tvar odhadované hustoty, ale „zanedbatelný“ vliv na tvar distribuční funkce. Dále lze usuzovat na závěr vyplývající z výše uvedeného obrázku distribučních funkcí, že výběr jádrové funkce lze provést na základě dalších důvodů, jako je náročnost na počítačové vybavení nebo efektivnost. Zároveň zde musí být poznamenáno, že zde nejsou uvažovány speciální jádrové funkce, jako je např. jádrová funkce ve tvaru „V“.

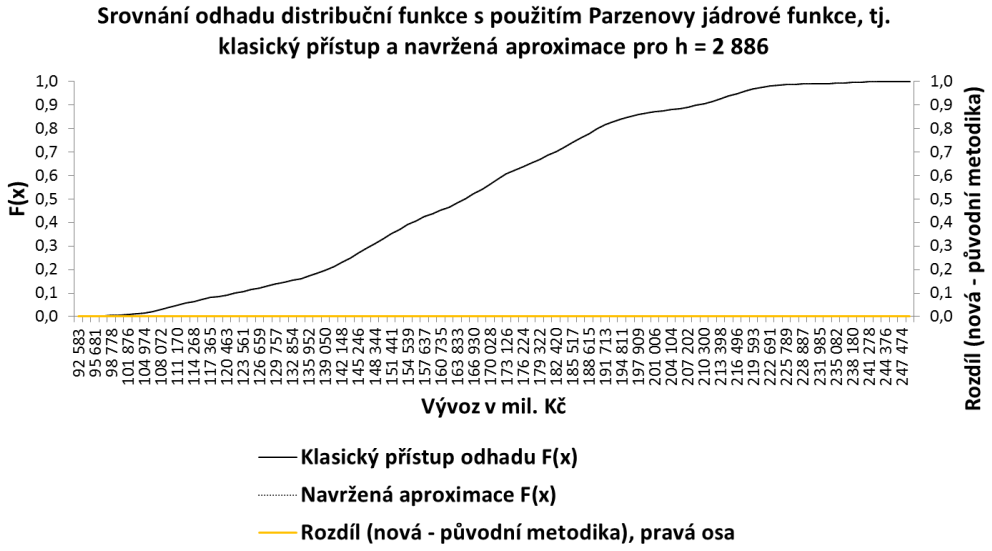
Výběr Parzenovy jádrové funkce je založen na implicitních vlastnostech, kdy se jedná o jednu z „jednodušších“ jádrových funkcí v oblasti výpočetní složitosti a grafické názornosti. Dále lze vypozorovat, že získaný soubor reálných pozorování „se neřídí normálním rozdělením pravděpodobnosti“. Informace o pravděpodobnostním rozdělení nemá vliv na výpočet spolehlivosti (selhání) v následující části práce. Samozřejmě váha tohoto tvrzení by měla být ověřena vhodným testem jako např. χ^2 –kvadrát test dobré shody[59].

5.2.2 Neparametrický jádrový odhad distribuční funkce

Výběr Parzenovy jádrové funkce je založen na implicitních vlastnostech popsaných výše. Proto je vzhledem k náročnosti při odvozování použita i pro aproximaci distribuční

[59] Popis testu včetně vlastností lze nalézt např. v Hátle a Likeš (1974, s. 340 - 348).

funkce, vycházející z empirické distribuční funkce. Výsledný odhad distribuční funkce s použitím Parzenovy jádrové funkce je uveden na obrázku níže, kde je zároveň porovnáván s původním neparametrickým jádrovým odhadem distribuční funkce. Výsledné křivky (tvary) mají ve své podstatě stejný tvar (tj. při volbě stejného typu jádrové funkce a parametru h). Výhoda uvedeného postupu spočívá v nižších nárocích na čas výpočtu, protože oproti klasickému přístupu jsou „prvotní hodnoty“ před dolní mezí $(x - ah)$ načítány stejným způsobem jako při tvorbě empirické distribuční funkce a neprobíhá v nich tzv. neparametrický jádrový odhad distribuční funkce[60].



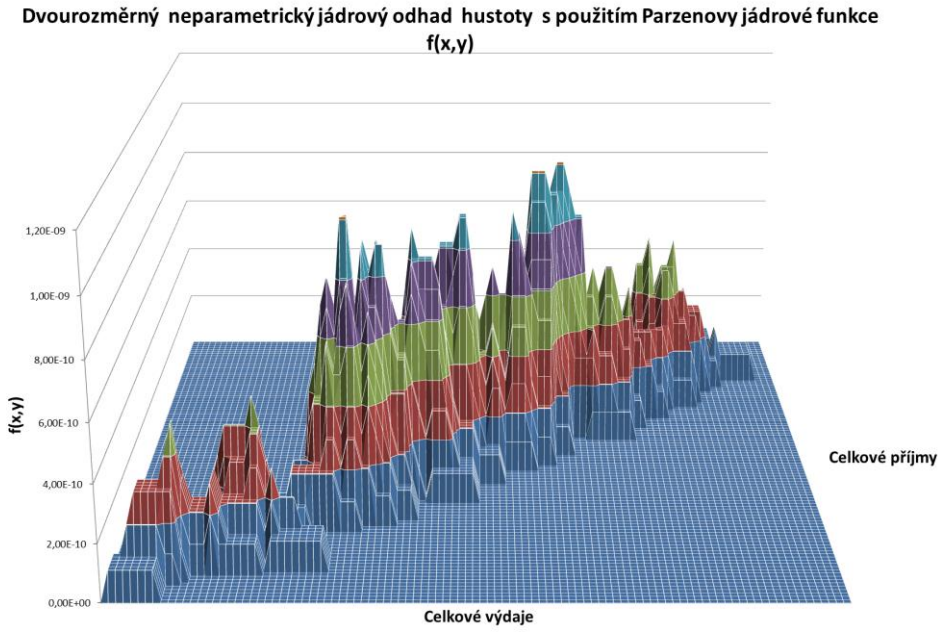
Obrázek 31: *Srovnání odhadu distribuční funkce s použitím Parzenovy jádrové funkce pro klasický neparametrický jádrový odhad a navržené aproximace pro* $h = 2\,886$

5.2.3 Dvourozměrné neparametrické jádrové odhady

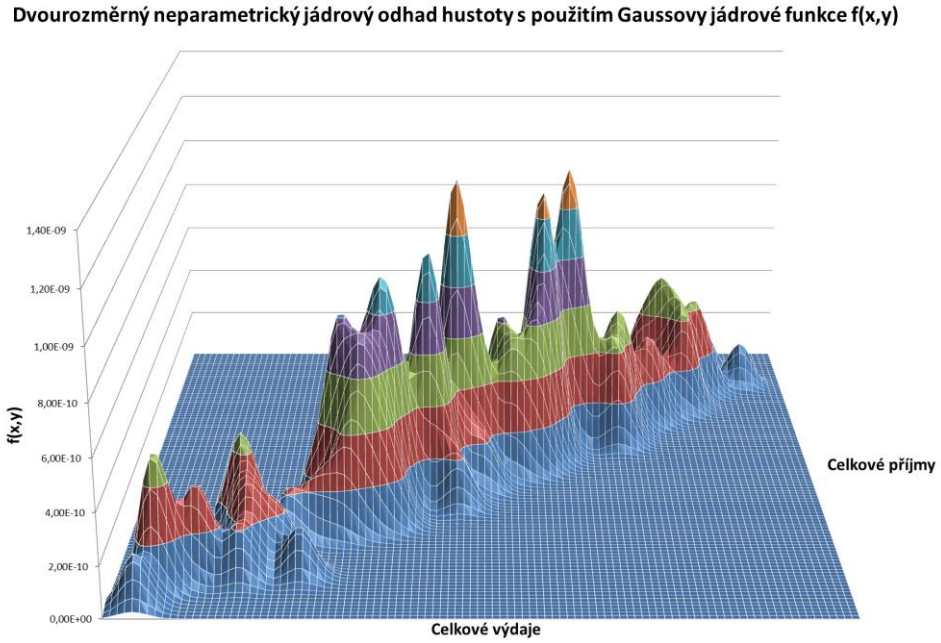
Pravděpodobnost spolehlivosti (selhání) je zde uvažována z dvourozměrné náhodné veličiny, proto jsou nyní předmětem analýzy dvourozměrné neparametrické jádrové odhady hustoty a distribuční funkce ze souboru reálných dat. První náhodná veličina je celková výše příjmů z vývozu a druhá náhodná veličina je celková výše výdajů z dovozu. Neznámá funkce hustoty a distribuční funkce je odhadována s použitím dvou typů jádrových funkcí. První jádrová funkce je Parzenova jádrová funkce a druhá je opět Gaussova jádrová funkce, která již byla charakterizována „vyhlazenějším“ povrchem.

Získané soubory reálných dat mají stejné měřítko (výdaje a příjmy jsou v milionech Kč). Parametry měřítka jsou opět odvozeny ze vztahu (3.53) a pro každou náhodnou veličinu nabývají odlišných hodnot. Výsledné tvary jsou zobrazeny na Obrázku 32, 33 a 34.

[60] Tím je myšlen přístup (3.17) a (3.18).



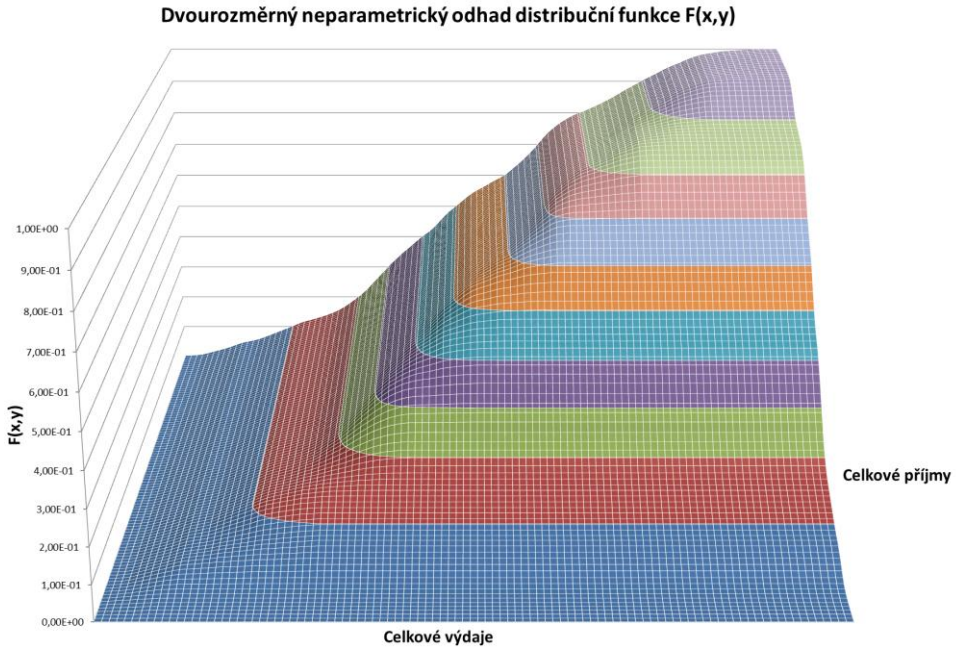
Obrázek 32: *Dvourozměrný neparametrický jádrový odhad hustoty s použitím Parzenovy jádrové funkce na souboru reálných dat z obchodní bilance ČR pro* $h_x = 2\,886$, $h_y = 2\,533$, $n = 113$



Obrázek 33: *Dvourozměrný neparametrický jádrový odhad hustoty s použitím Gaussovy jádrové funkce na souboru reálných dat z obchodní bilance ČR pro* $h_x = 2\,886$, $h_y = 2\,533$, $n = 113$

Stejné experimentální soubory reálných dat jsou použity k aproximaci dvourozměrné distribuční funkce s použitím stejných jádrových funkcí (Parzenova, Gaussova) a shodných párů parametrů měřítka h_x, h_y. Na základě výsledných grafů lze tvrdit, že získané aproximace jsou "téměř stejné" jako v předchozím případě jednorozměrné jádrové aproximace. Opět zde nejsou „významné" rozdíly mezi oběma distribučními

funkcemi, a proto je uveden pouze jeden obrázek výsledné distribuční funkce. Formulace závěru je tedy stejná jako v případě jednorozměrného odhadu.

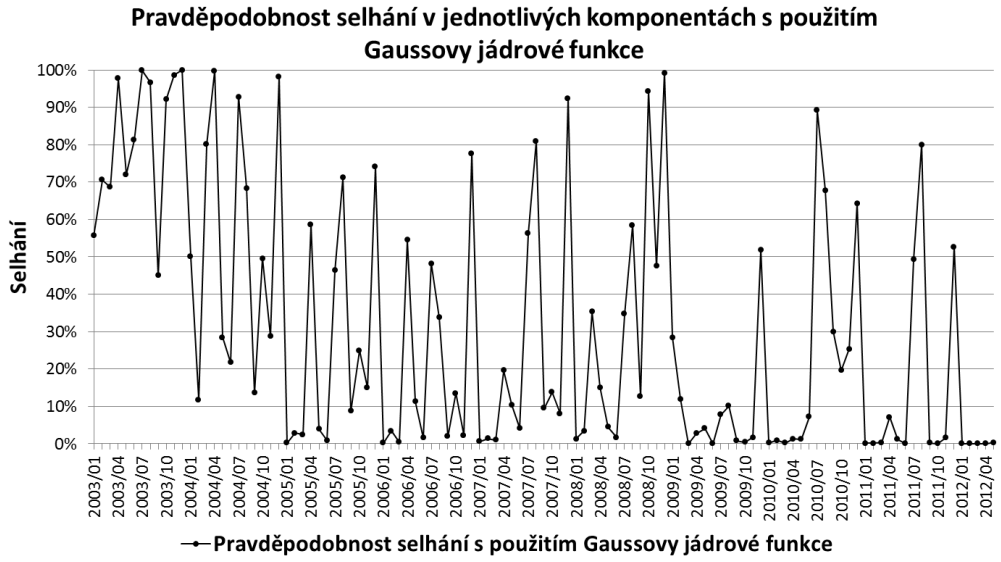


Obrázek 34: *Dvourozměrný neparametrický jádrový odhad distribuční funkce s použitím Parzenovy a Gaussovy jádrové funkce pro* $h_x = 2\,886$, $h_y = 2\,533$, $n = 113$

5.2.4 Analýza spolehlivosti (selhání) pro jednotlivé měsíce

Analýza spolehlivosti (selhání) je provedena na stejných experimentálních souborech dat. Nejprve je spolehlivost (selhání) získána s použitím Gaussovy jádrové funkce a potom s využitím Parzenovy jádrové funkce. Princip výpočtu je založen na postupné analýze párovaných pozorování v $i-tém$ měsíci pro $i = 1, \dots, 113$ a získané dílčí hodnoty (spolehlivosti) jsou následně využity k výsledné pravděpodobnosti překročení celkové výše výdajů nad celkovou výší příjmů nebo naopak (tj. pravděpodobnosti selhání nebo spolehlivosti).

Grafické interpretace získaných hodnot spolehlivosti a selhání s použitím Gaussovy jádrové funkce jsou ukázány na Obrázku 35 a 36 pomocí zvýrazněných bodů pro každý pár pozorování (komponentu). Ten představuje dvojici (x_i, y_i), kde x_i je celková výše výdajů z dovozu a y_i je celková výše příjmů z vývozu v $i-tém$ měsíci. Potom pravděpodobnost selhání $F = P(Y < X)$ je pravděpodobnost, že celková výše příjmů je menší než celková výše výdajů a naopak. Pravděpodobnost spolehlivosti může být vyjádřena pomocí vztahu $R = 1 - F$. Parametry měřítka jsou získány ze vztahu (3.53).

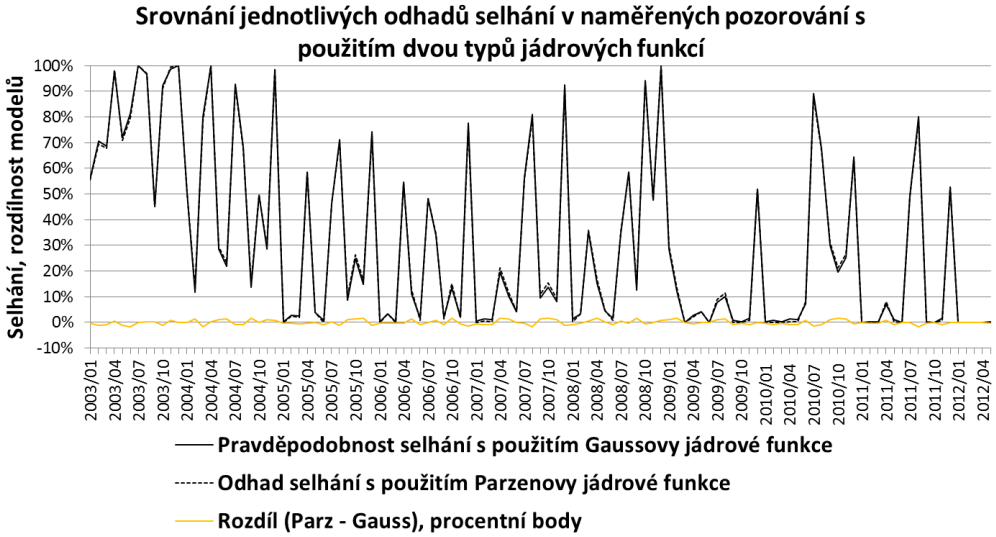


Obrázek 35: *Pravděpodobnost selhání pro každý pár naměřených pozorování (komponentu) s použitím Gaussovy jádrové funkce pro* $h_x = 2\,886$, $h_y = 2\,533$, $n = 113$

Obrázek 36: *Pravděpodobnost spolehlivosti pro každý pár naměřených pozorování s použitím Gaussovy jádrové funkce pro* $h_x = 2\,886$, $h_y = 2\,533$, $n = 113$

Na základě uvedených obrázků lze říci, že vykreslené pravděpodobnostní hodnoty nám umožňují popsat pravděpodobnosti změn záporné (nebo kladné) platební bilance. V tomto konkrétním případě byla pravděpodobnost selhání ze začátku sledovaného období roku 2003 „vysoká“ a během následujícího časového období se postupně snižovala až do inkriminovaného roku 2008, po kterém dochází k opětovnému poklesu až do současných pozorování v roce 2012. Dalším důležitým zjištěním jsou rozdíly mezi dvěma sousedními body, které se ve většině případů jeví jako „významné“, což by mohlo znamenat obsah „významné“ nestacionarity v získaných souborech reálných dat.

Dále lze srovnat hodnoty obou navržených metod s použitím stejných vyhlazovacích parametrů h_x, h_y. Získané výsledky jsou prezentovány na Obrázku 37, kde již nejsou detailně označeny jednotlivé body z důvodu větší přehlednosti.



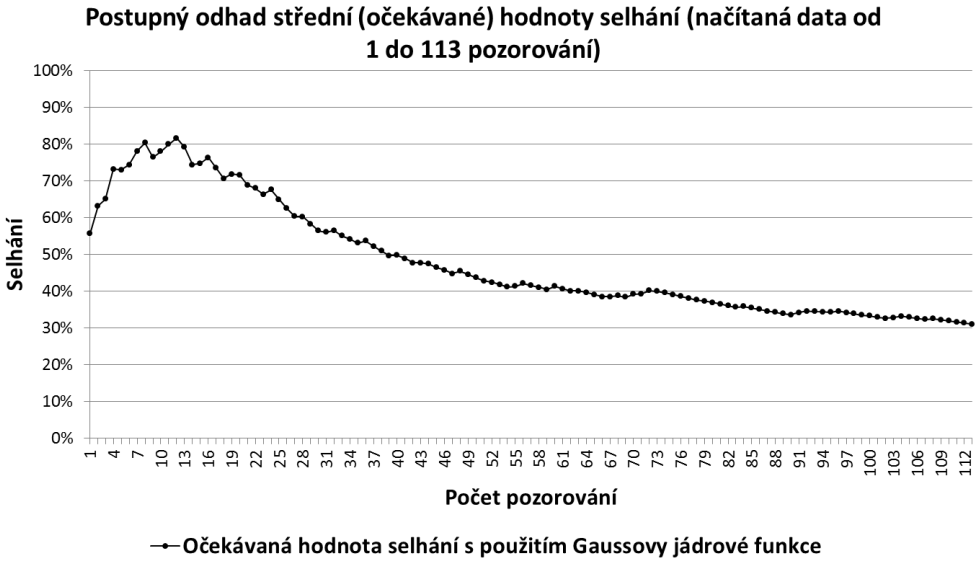
Obrázek 37: *Srovnání navržených modelů k získání pravděpodobnosti selhání pro stejné vyhlazovací parametry* $h_x = 2\,886$, $h_y = 2\,533$, $n = 113$

Ve srovnání obou přístupů nejsou patrné téměř „žádné změny“ (ne v matematickém smyslu slova). Změny jsou pouze v některých pozorováních (naměřených bodech) a jsou pouze v „jednotkách“ procentních bodů. Naopak pohyby mezi dvěma sousedními pozorováními mají ve všech případech stejný směr. Vzniklé rozdíly jsou způsobeny vlastnostmi použitých jádrových funkcí, kdy Gaussova jádrová funkce pokrývá širší oblast než Parzenova jádrová funkce. Z těchto závěrů lze formulovat hlavní otázku, která z uvedených jádrových funkcí je vhodnější pro nalezení spolehlivosti (selhání)? Výsledkem nebo doporučením práce může být formulována odpověď, že jádrová funkce může být vybrána v souladu s výpočetní náročností nebo obtížností při prováděných výpočtech.

5.2.5 Analýza spolehlivosti (selhání) za celé období

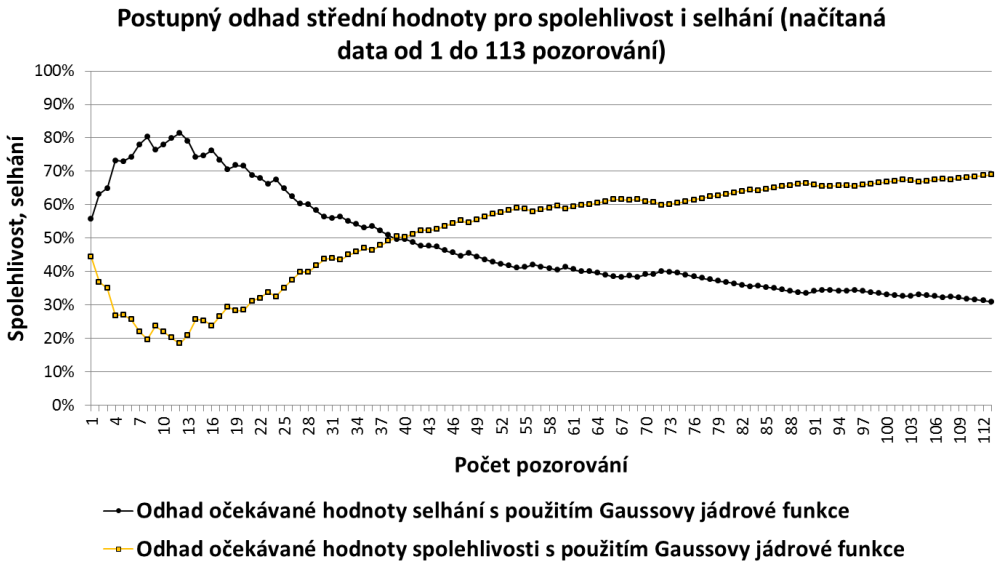
K dispozici jsou známy hodnoty spolehlivosti i selhání v jednotlivých měsících a zbývá získat hodnotu za celé uvažované časové období např. s pomocí aritmetického průměru[61]. Výsledná hodnota je ukázána pomocí postupného výpočtu (vývoje). Obecně jsou zde vypočítány načítané odhady střední hodnoty až do $n-tého$ pozorování. Důležitá (ve smyslu sledovaného období) je poslední získaná hodnota v grafu, která reprezentuje pravděpodobnost spolehlivosti (selhání) pro vybraný typ jádrové funkce.

[61] Aritmetický průměr a jeho popis je velmi dobře znám a je zároveň uveden v každé knize se statistickou tématikou.



Obrázek 38: *Odhad střední hodnoty pro pravděpodobnost selhání s použitím Gaussovy jádrové funkce (načítané data od 1 do 113 pozorování)*

Grafické vyjádření vývoje pravděpodobnosti pro hodnotu spolehlivosti i selhání na stejném souboru reálných dat s použitím Gaussovy jádrové funkce je ukázáno níže.

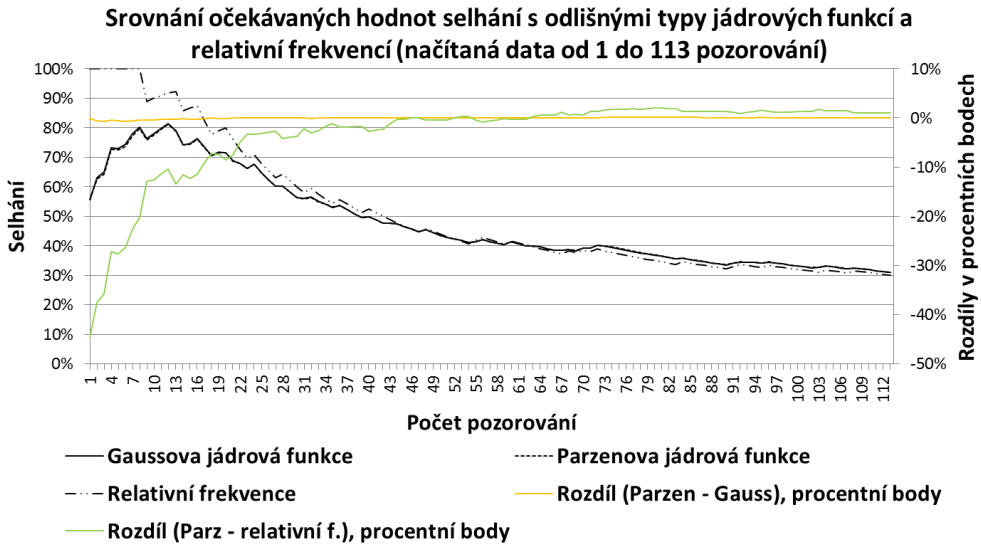


Obrázek 39: *Srovnání odhadů středních hodnot pro pravděpodobnost spolehlivosti a selhání s použitím Gaussovy jádrové funkce*

Takto získaná výsledná hodnota je zajímavá jak ze statistického, tak z ekonomického úhlu pohledu. Prvotně se jedná o požadovanou pravděpodobnost záporné nebo kladné obchodní bilance s vlivem na celou platební bilanci České republiky. Patrná je zde klesající hodnota pravděpodobnosti záporné hodnoty obchodní bilance na „téměř“ celém časovém intervalu, což signalizuje dobré výsledky v bilanci zahraničního obchodu. Vývoj popisované funkce je zajímavý pro analýzu (odhad) možné „trendové“ křivky

a zároveň může charakterizovat ekonomickou situaci v zemi, ale to je již na další diskusi v ekonomickém pojetí a terminologii. Závěr získaný z této analýzy může být interpretován jako možná detekce klesající trendové křivky pravděpodobnosti selhání ve vybrané zemi. Dále je vhodné srovnat získané výsledky s použitím modelu relativní frekvence pro rostoucí počet pozorování n.

Srovnání získaných „křivek“ vývoje pravděpodobnosti selhání relativní frekvence a navržených modelů s použitím Gaussovy a Parzenovy jádrové funkce je ukázáno na Obrázku 40. Zde nejsou patrné „žádné významné změny“ mezi odhady používající odlišné typy jádrových funkcí. Rozdíly jsou mezi výsledky navržených modelů a relativní frekvencí, což může být způsobeno malým počtem pozorování v získaném souboru reálných dat. Z těchto důvodů lze zopakovat předchozí závěr, že výběr jádrové funkce nemá „významný“ vliv na výslednou hodnotu spolehlivosti.

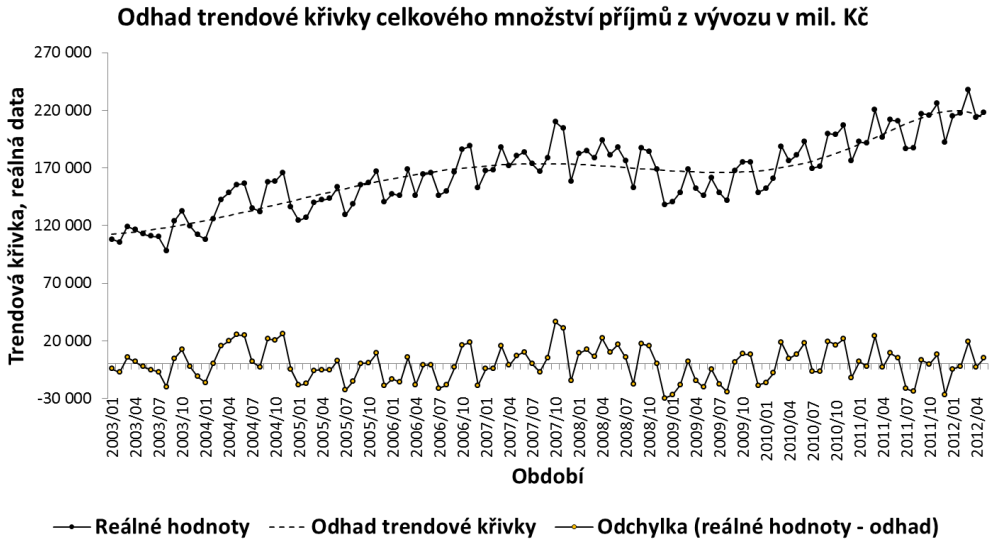


Obrázek 40: *Srovnání mezi odhady středních hodnot selhání používajících odlišné typy jádrových funkcí a relativní frekvencí (načítaná data od 1 do 113 pozorování)*

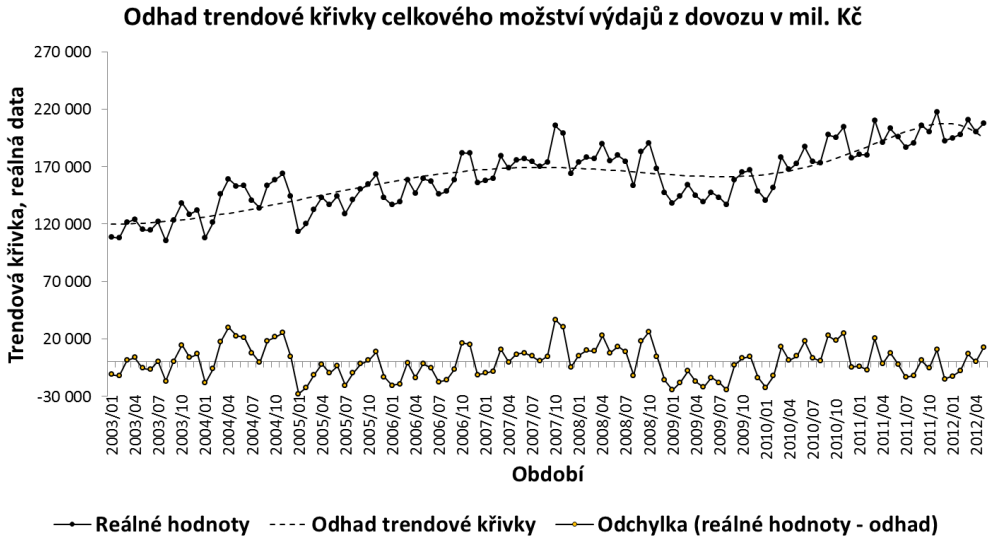
5.2.6 Odhad trendové (systematické) složky

Doposud prezentované hodnoty předpokládaly stacionární charakter získaných dat. V případě reálných dat jsou ale velmi časté potíže s jejich nestacionárním charakterem a tento předpoklad je hlavním důvodem pro analýzu odvozeného modelu k odhadu a predikci trendové křivky. Uvažovaná trendová křivka je sama o sobě aproximací získaného souboru dat pomocí spojité „čáry“ (tj. zatím bez bližší specifikace počtu zahrnutých složek ortonormálního systému), která shrnuje, kde se získané hodnoty nacházely v minulosti, kde se tyto hodnoty nalézají nyní ve vztahu k minulosti a kde by se tyto hodnoty mohly vyskytovat v krátkém budoucím časovém intervalu (po příslušném „prodloužení“ užitého ortonormálního systému). Samotná aplikace odvozeného modelu na soubory reálných dat je ukázána na níže uvedených obrázcích, kde jsou odhady

trendových křivek zobrazeny pomocí přerušované čáry. Hlavní myšlenka postupu spočívá v tom, že použití vygenerovaného ortonormálního systému nám umožní získat trendovou křivku (komponentu) uvažované časové řady. Poslední křivka na obou obrázcích, která je charakterizována přerušovanou čárou s body, ukazuje odchylky mezi pozorovanými hodnotami a odhadem trendové složky.



Obrázek 41: *Odhad trendové křivky celkového množství příjmů z vývozu ČR*

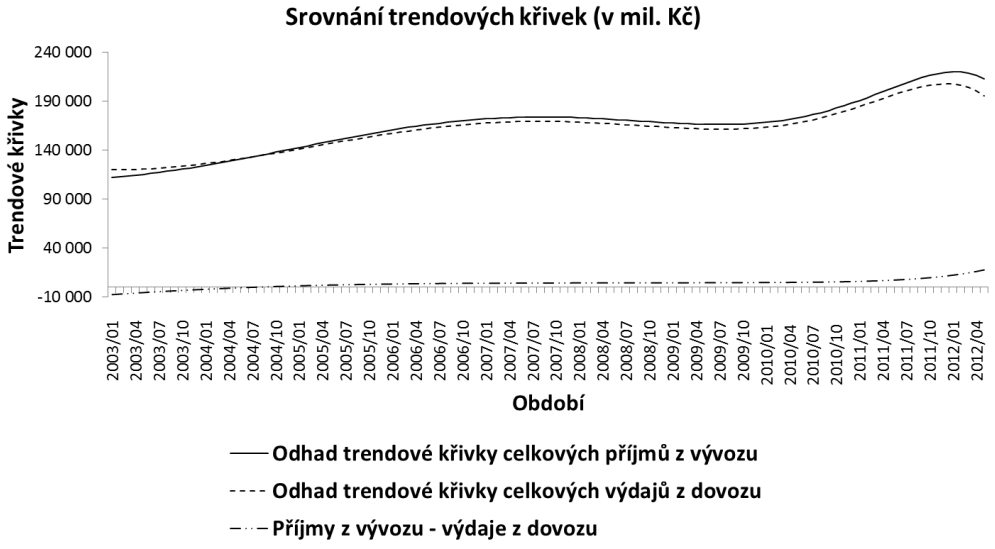


Obrázek 42: *Odhad trendové křivky celkového množství výdajů z dovozu do ČR*

Nyní lze stručně shrnout získané výsledky z ekonomického úhlu pohledu. O trendových křivkách na Obrázku 41 a 42 lze tvrdit, že demonstrují růstový potenciál, což může být způsobeno růstem ekonomické aktivity ve vybrané zemi. Tím je zde myšlen růst ekonomické aktivity v celé České republice, který je reprezentován celkovou výší výdajů z dovozu (i příjmů z vývozu). Dále lze detailněji popsat směr obou trendových křivek,

které mají od roku 2003 do roku 2008 růstový charakter, který lze vysvětlit pomocí všeobecné rostoucí ekonomické aktivity. Následná změna směru (recese) na konci tohoto časového úseku trvající do přelomu roku 2009/2010 je způsobena vzniklou celosvětovou „finanční krizí“. Tento směr je zachován až do přelomu let 2011/2012, kdy se na scénu vrátila původní celosvětová „krizová“ nervozita a obě trendové křivky až do současnosti vykazují klesající tendenci. Zde stojí za povšimnutí skutečnost, že od roku 2003 až do konce 2011 nedocházelo v dovozních ani vývozních objemech k výraznému poklesu, dokonce je patrná neklesající tendence a zároveň Česká republika navyšuje po odeznění první vlny „finanční krize“ své vývozy i dovozy. Uvedenou metodu lze dále využít pro krátkodobé prognózy (predikce).

Obě konkrétní aproximace neznámých trendových křivek (ve smyslu této práce) jsou dále užitečné i pro jejich vzájemné porovnání, zejména v dlouhodobém časovém horizontu. Vzniklá situace je znázorněna na níže uvedeném obrázku, kde přerušovaná křivka označuje trendovou křivku pro celkovou výši výdajů z dovozu a plná čára označuje trendovou křivku pro celkovou výši příjmů z vývozu. Rozdíl mezi vykreslenými trendovými křivkami představuje kladnou nebo zápornou hodnotu (aktuálních odečtů) zahraničního obchodu platební bilance, tedy obchodní bilanci. Obecně zde může být řečeno, že jestliže se nachází plná čára nad přerušovanou, potom hodnota vzniklé „trendové bilance“ je kladná a zahraniční obchod[62] má znaky dobré kondice.



Obrázek 43: *Srovnání odhadnutých trendových křivek z celkové výše příjmů a výdajů ČR*

Třetí zobrazená křivka je charakterizována pomocí přerušované čáry prokládané body a reprezentuje vzniklé rozdíly mezi oběma odhady trendových křivek. Z obrázku je evidentní neklesající, dokonce rostoucí tendence směru, který znamená „rychlejší“ růst celkové výše příjmů z vývozu než výdajů z dovozu. Stručně řečeno, ekonomická aktivita

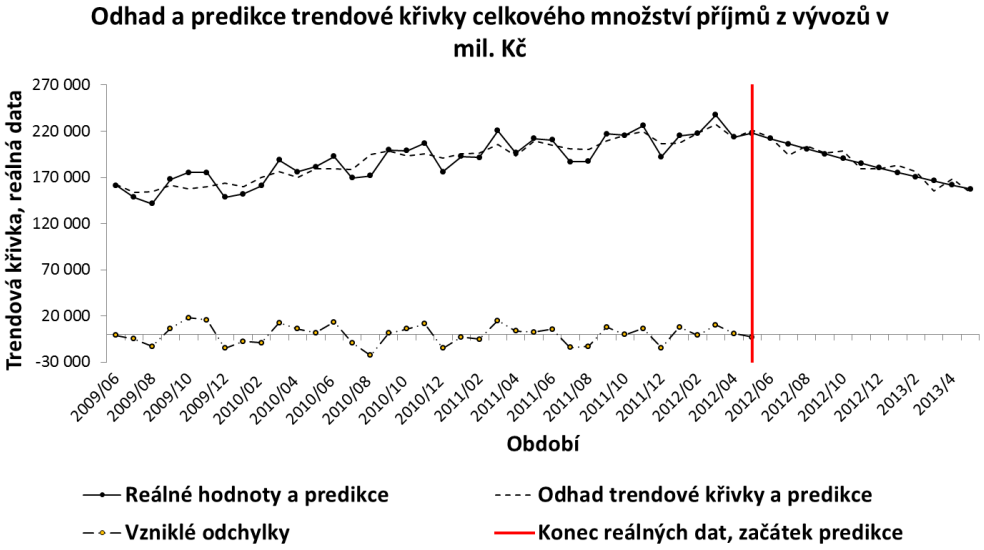
[62] V tomto speciálním případě.

v České republice zvyšuje celkové příjmy z vývozu rychleji, než celkovou výši výdajů z dovozu v milionech Kč již od roku 2003 (jistý vliv na to mohou mít i kurzové poměry, nejen).

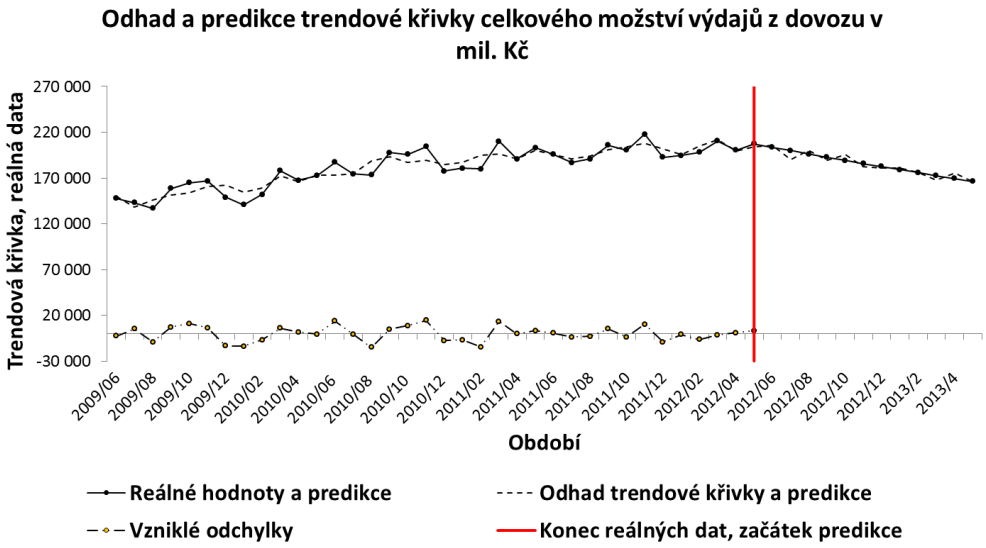
Poznámka 5.5 *Takto upravené časové řady mohou následně obsahovat další složky nestacionarity jako sezónní nebo cyklickou složku, ale ty budou předmětem dalšího výzkumu navazujícího na prezentovanou práci.*

5.2.7 Predikce trendové složky

V předchozí části byla neznámá trendová složka podrobena analýze, při které nebyl prozatím specifikován výběr a počet použitých koeficientů (též prvků ortonormálního systému). Koeficienty byly vybrány pragmatickým přístupem pro získání „vyhlazené“ trendové křivky za celé uvažované období, tj. bylo použito prvních 6 koeficientů (ortonormálních prvků). Dále je možné analyzovat použitý model s využitím heuristického přístupu, tj. vybrat „vhodné“ koeficienty a prvky ortonormálního systému. Tím je myšlen výběr těch koeficientů, které mají „největší“ (ve smyslu této práce) vliv na odhad (tvorbu) trendové křivky. V návaznosti na popsaný model s vybranými koeficienty je zároveň navržen model pro (možnou) predikci trendu v krátkodobém časovém horizontu. Ten je zvolen počtem 12 měsíců v závislosti na předchozích datech, které jsou vzhledem k věrohodnému popisu (váze) změn v poslední době zvoleny za dobu 3 let, tedy posledních 36 měsíců včetně posledně získaného měření (5/2012). Shrnuto, základem odhadu trendové křivky je posledních 36 pozorování (měsíčních) s predikcí na 12 pozorování (měsíců). Volba delšího predikčního intervalu by již vzhledem k dynamicky se měnícímu prostředí ztrácela smysl. Analýze jsou podrobeny oba získané reálné soubory dat (vývoz, dovoz) a získané výsledky jsou uvedeny na Obrázku 44 a 45 níže.

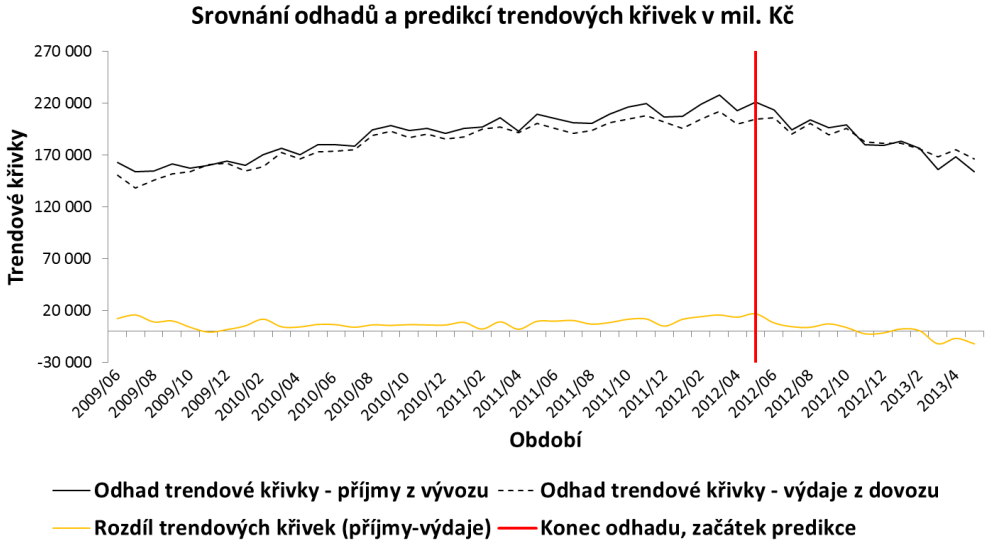


Obrázek 44: *Odhad trendové křivky (6/2009 – 5/2012) a 12 měsíční predikce celkového množství příjmů z vývozu ČR,* $n = 48$



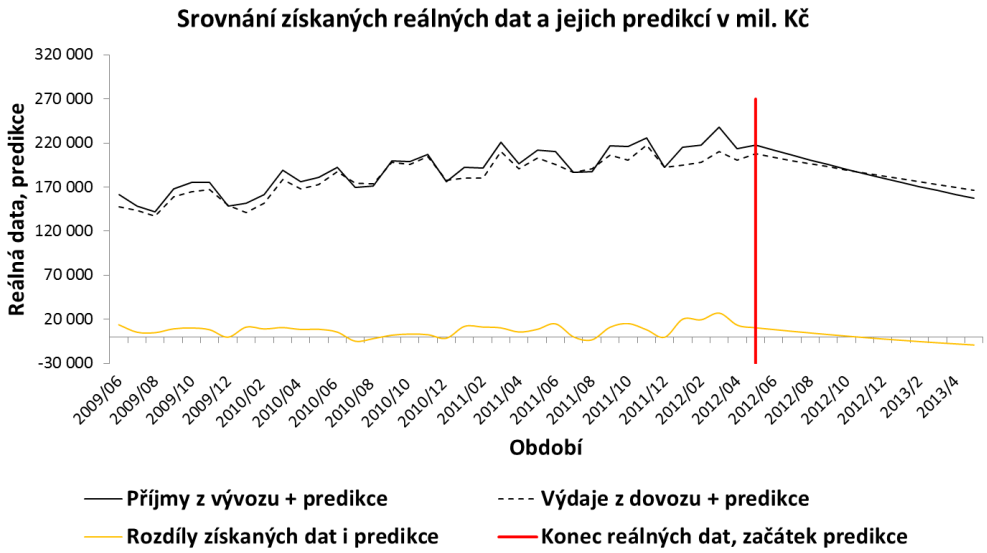
Obrázek 45: *Odhad trendové křivky (6/2009 – 5/2012) a 12 měsíční predikce celkového množství výdajů z dovozu ČR,* $n = 48$

Zde je patrná „růstová“ dynamika odhadovaných trendových složek s následným predikčním poklesem. Použitá heuristika vybrala pro soubor reálných dat vývozu 10 koeficientů (ortonormálních složek) a pro soubor reálných dat dovozu 11. Vykreslené odhady (čárkovanou čárou) trendové složky již nejsou vyhlazeny jako v předchozím případě a je patrný „vlnovitý“ tvar, který „upřesňuje“ získané odhady. Pro kontrolu lze použít vybrané testovací kritérium, např. obecný koeficient determinace popsaný výše. Srovnání vývojů (průběhů) trendových křivek je prezentováno na následujícím Obrázku 46, kde za povšimnutí stojí „vyšší“ dynamika v odhadu trendové křivky pro příjmy z vývozu, než výdajů z dovozu a jejich záměna v predikovaném období.



Obrázek 46: *Srovnání odhadnutých trendových křivek za 48 měsíců včetně 12 měsíční predikce z celkové výše příjmů a výdajů ČR*

Pro úplnost jsou zde zobrazeny získané soubory reálných dat (datových vektorů), respektive jejich vzájemné srovnání včetně predikčních odhadů na 12 měsíců.

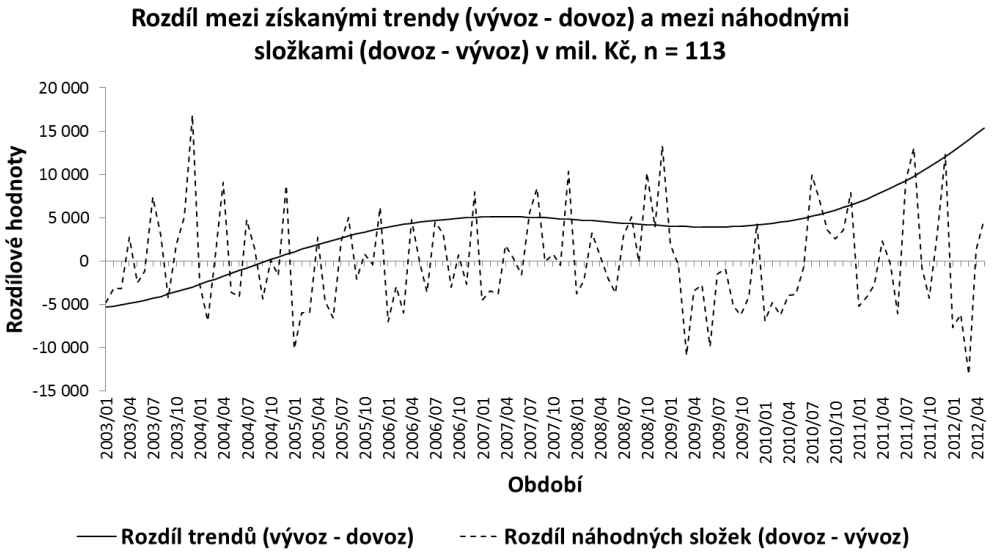


Obrázek 47: *Srovnání získaných souborů reálných dat z celkové výše příjmů a výdajů ČR za 36 měsíců a jejich 12 měsíční predikce*

V souvislosti s předchozími trendovými odhady zde nastala situace, kdy trendová křivka výdajů z dovozu převýší křivku příjmů z vývozu (v predikci), což může značit počátek (nejen) negativního prostředí pro ekonomiku dané země v uvažovaném časovém období, tj. zvyšující se negativní rozdíl mezi predikcemi. Dále jsou na Obrázku 47 patrné vyšší hodnoty získaných pozorování u příjmů z vývozu než výdajů z dovozu. Situace zároveň ukazuje na možnou „slabou stránku“ navržených modelů k odhadu trendových křivek, která již byla diskutována v předchozí kapitole. Ta se vyznačuje jak ve „vhodném“ určení prvotní hodnoty pro tvorbu trendové křivky (resp. datového vektoru reálných dat X), tak v určení délky časového úseku, pro který jsou trendové křivky odhadovány. Náznak řešení uvedené problematiky je uveden v Kapitole 6.

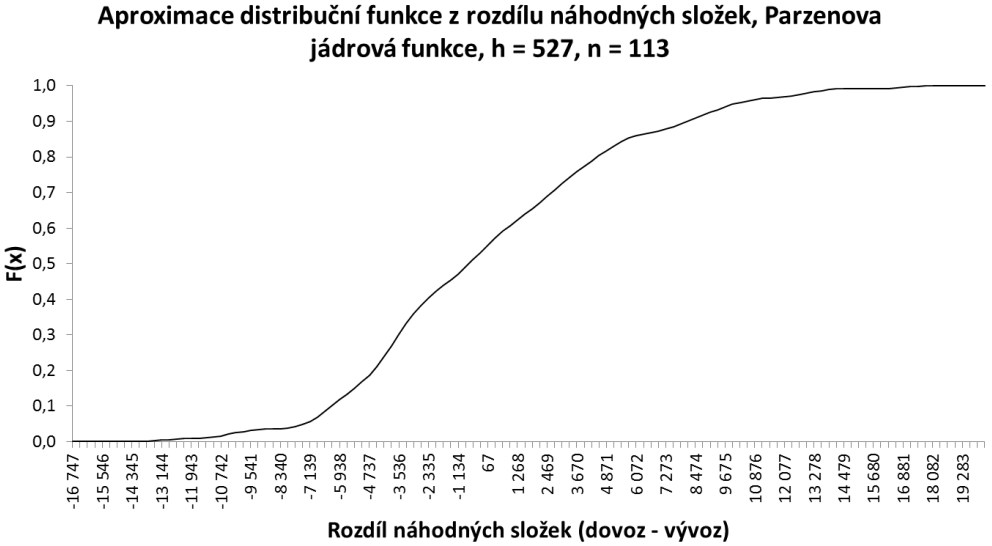
5.2.8 Výsledné hodnoty spolehlivosti (selhání) pro nestacionární data

Analýza spolehlivosti v případě nestacionárního charakteru dat je popsána v předchozí kapitole vztahem (4.100). V uvedené problematice je důležitá znalost celé distribuční funkce $F_{\varepsilon_x-\varepsilon_y}\big(Y_m(t)-X_m(t)\big),\ t=1,\dots,T$ pro získání spolehlivosti v jednotlivých komponentách (párovaných pozorování). Rozdíly mezi odhadovanými trendovými křivkami a náhodnými složkami jsou ukázány na Obrázku 48. Zde je patrný „vyhlazený“ tvar rozdílu mezi trendovými křivkami v závislosti na čase, který má potom vliv na průběh (vývoj v čase) získaných hodnot spolehlivosti. Volba trendových křivek pochází z Obrázku 43 a volby vyhlazené trendové křivky, nikoliv ze zvolené heuristiky, která (jen) vrací možný (nikoliv závazný) počet použitých koeficientů ovlivňující její tvar.



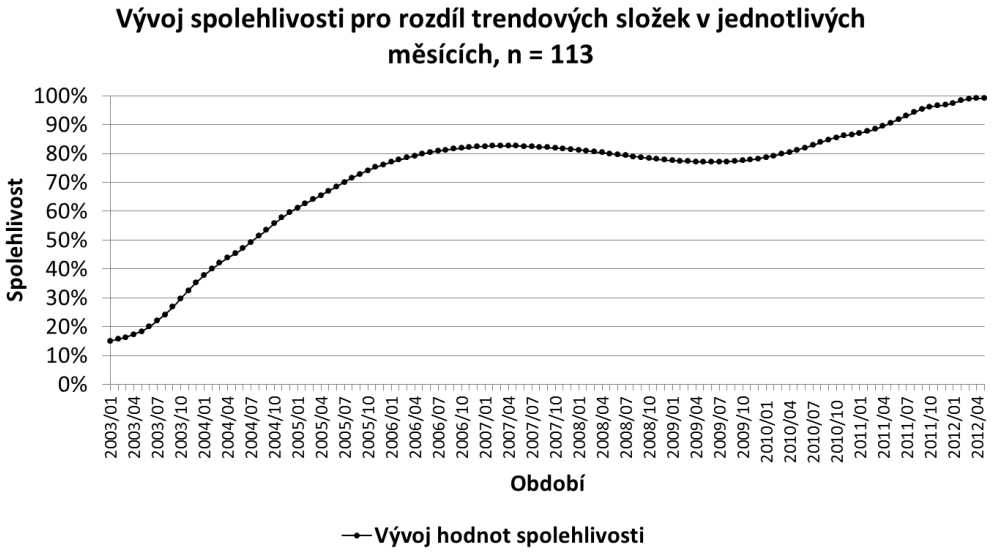
Obrázek 48: *Rozdíl mezi získanými trendy (vývoz – dovoz) a mezi náhodnými složkami (dovoz – vývoz) v mil. Kč pro n = 113*

Aproximace distribuční funkce z rozdílu náhodných složek $F_{\varepsilon_x-\varepsilon_y}\big(Y_m(t)-X_m(t)\big)$, $t=1,\dots,T$ pro výpočet spolehlivosti je vykreslena na Obrázku 49.



Obrázek 49: *Neparametrický odhad distribuční funkce z rozdílu náhodných složek, Parzenova jádrová funkce, h = 527, n = 113*

Potom vývoj spolehlivosti pro jednotlivé komponenty (párovaná pozorování v jednotlivých měsících) s využitím odhadované distribuční funkce pro rozdíl trendových složek je uveden na Obrázku 50.



Obrázek 50: *Vývoj hodnot spolehlivosti pro rozdíl trendových složek, n = 113*

Odhadem střední hodnoty z uvedených spolehlivostí s pomocí aritmetického průměru (tj. spolehlivost za celé vybrané období 113 pozorování) je získána „celková“ hodnota spolehlivosti $71{,}75\ \%$.

5.2.9 Finální výsledky

V případě stacionárního charakteru dat poskytuje model relativní frekvence dobrý odhad spolehlivosti vzhledem k jednoduchosti, náročnosti výpočtu na čas a možné chybovosti při výpočtech. Metoda je zde použita za účelem porovnání s „náročnějšími“ neparametrickými jádrovými modely, které obsahují odlišné typy jádrových funkcí a dosahují „téměř stejných“ výsledků. První neparametrický jádrový model používá Gaussovu jádrovou funkci a druhý prezentovaný model používá Parzenovu jádrovou funkci. Výběr Gaussovy jádrové funkce byl proveden na základě vlastností (vyhlazená funkce, „těžké“ konce, …) a obsahu v každém statistickém softwaru. Výběr Parzenovy jádrové funkce je založen na výpočetní (odvozovací) „složitosti“ (čas, hardware). Uvedené vlastnosti jsou demonstrovány na tvarech výsledných neparametrických aproximací hustot.

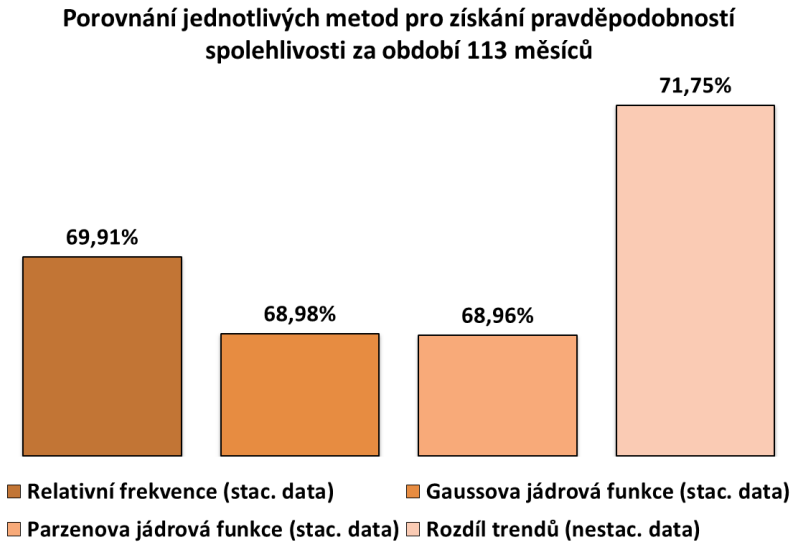
V případě nestacionárního charakteru dat je závěr odlišný. Zde je evidentní výhoda neparametrických jádrových odhadů, která představuje přístup k aproximaci distribuční funkce pro výpočet spolehlivosti (selhání).

Získané výsledné hodnoty spolehlivosti jsou prezentovány v následující Tabulce 5, kde si lze povšimnout „minimálních“ rozdílů.

Porovnání získaných pravděpodobností spolehlivosti za 113 *měsíců*	
Relativní frekvence (stac. data)	69,91 %
Gaussova jádrová funkce (stac. data)	68,98 %
Parzenova jádrová funkce (stac. data)	68,96 %
Rozdíl trendových křivek (nestac. data)	71,75 %

Tabulka 5: *Porovnání jednotlivých metod pro získání pravděpodobností spolehlivosti za období 113 pozorování*

Grafická prezentace konečných výsledků je uvedena na Obrázku 51. V případě předpokladu stacionárního charakteru dat je „největší“ rozdíl mezi metodami 0,95 procentních bodů a je mezi modelem používajícím Parzenovu jádrovou funkci a relativní frekvencí. Naopak „nejmenší“ rozdíl 0,02 procentních bodů je mezi oběma modely používajícími neparametrické jádrové odhady. Za předpokladu nestacionárního charakteru dat je výsledek spolehlivosti „nejvyšší“.



Obrázek 51: *Porovnání jednotlivých metod pro získání pravděpodobností spolehlivosti ze získaného souboru reálných dat čítajících 113 měsíců*

Výsledky lze formulovat tak, že v případě stacionárního charakteru dat jsou spolehlivostní rozdíly mezi uvedenými metodami v rozmezí jednoho procentního bodu. Nejvyšší spolehlivost je získána s použitím neparametrického modelu s Parzenovou jádrovou funkcí a naopak, nejnižší hodnota spolehlivosti je získána modelem relativní frekvence.

Získané výsledky potvrdily předchozí závěry o použitelnosti relativní frekvence v případě stacionárního charakteru dat a vhodnosti neparametrického přístupu v případě nestacionárního charakteru dat. Dále se získané výsledky jeví jako „téměř nezávislé“ na použité jádrové funkci, pokud zde nejsou uvažovány speciální jádrové funkce např. jádrová funkce ve tvaru „V“.

Poznámka 5.6 *Získané hodnoty pravděpodobnosti mohou být poté interpretovány i jako pravděpodobnost kladné bilance zahraničního obchodu (obchodní bilance) v náhodně vybraném měsíci bez respektování času.*

5.3 Výsledky v kumulacích

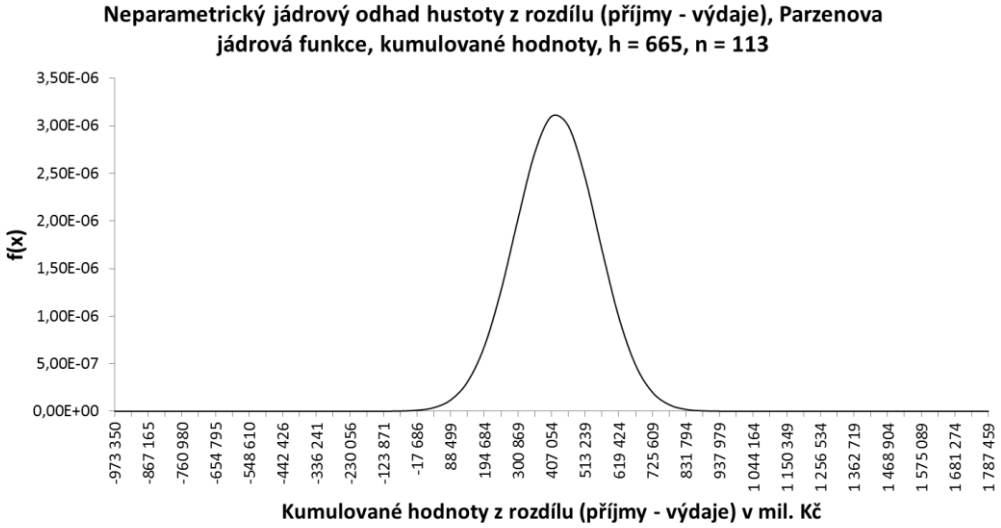
Pravděpodobnost spolehlivosti nebo selhání z načítaných (kumulovaných) hodnot pozorování v období pro $t = 1, \dots, T$ může být získána:

$$\begin{aligned} F &= P(Y < X) \\ &= P\left(\sum_{t=1}^{T} y_t < \sum_{t=1}^{T} x_t\right) \\ &= P\left(\sum_{t=1}^{T} (y_t - x_t) < 0\right). \end{aligned} \tag{5.7}$$

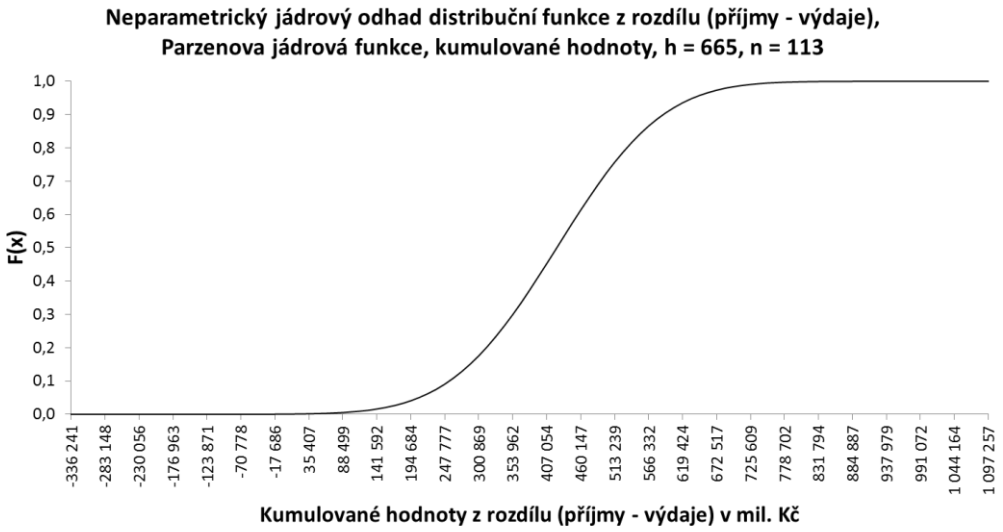
kde X jsou celkové výdaje z dovozu a Y celkové příjmy z vývozu. Tím je získána náhodná veličina z rozdílu obou hodnot v každém pozorování. Neparametrické jádrové odhady hustoty a distribuční funkce pro vzniklou náhodnou veličinu rozdílu jsou vykresleny na Obrázku 52 a 53. Výsledná hodnota (pravděpodobnost) selhání za celé uvažované období 113 měsíců (pro získané kumulace rozdílů) je $F = 0{,}076\ \%$. Naopak výsledná hodnota spolehlivosti za celé uvažované období je rovna:

$$R = 1 - F = 1 - 0{,}00076 = 0{,}99924 = 99{,}924\ \%. \tag{5.8}$$

Ekonomické opodstatnění pro kumulované hodnoty je založeno na faktu, že kumulací kladné platební bilance se vytváří rezervy na případné ztráty v příštím období. Z výsledné hodnoty je evidentní „nepatrná“ pravděpodobnost selhání vzhledem k dosud nashromážděným zdrojům.



Obrázek 52: *Neparametrický jádrový odhad hustoty z rozdílu (příjmy - výdaje), Parzenova jádrová funkce, kumulované hodnoty,* $h = 665$, $n = 113$



Obrázek 53: *Neparametrický jádrový odhad distribuční funkce z rozdílu (příjmy - výdaje), Parzenova jádrová funkce, kumulované hodnoty,* $h = 665, n = 113$

5.4 Popis ověřovacího systému

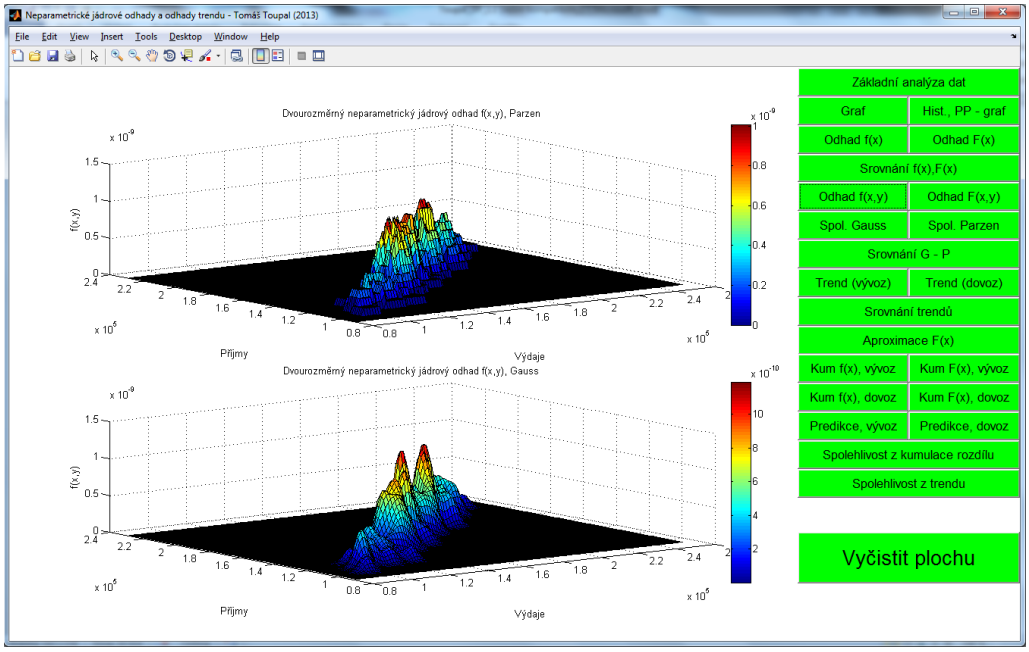
Společně s prací byl vytvořen program v softwaru Matlab verze 2010a. Cílem bylo v programu analyzovat navržené modely a z nich plynoucí prezentace získaných výsledků včetně grafických i textových výpisů v přehledné formě. Nevýhoda vytvořeného programu spočívá v nutnosti vlastnit software Matlab k zajištění plné funkčnosti.

Hlavní uživatelské rozhraní je zde prezentováno popisem hlavního okna obsahujícího prostor k interpretaci požadovaných výsledků a ovládací tlačítka pro co nejjednodušší možnou interakci mezi samotným uživatelem a naprogramovanými procedurami (funkcemi). Samozřejmostí je obsah celého zdrojového kódu, který se skládá se ze tří základních souborů:

- *"Uvod.m"* - soubor obsahuje grafické ztvárnění uživatelského rozhraní.
- *"uvod_seznam_funkci.m"* – soubor obsahuje databázi níže popsaných funkcí.
- *"vstupy.mat"* – soubor obsahuje soubory reálných pozorování obchodní bilance z České národní banky.

5.4.1 Hlavní uživatelské rozhraní

Hlavní uživatelské rozhraní představuje vstupní dveře do programu a je zde reprezentováno graficky přijatelným úvodním oknem obsahujícím základní tlačítka a prostor pro grafickou (textovou) prezentaci. Tlačítka s příslušným popisem funkce jsou umístěny na pravé straně od oblasti zobrazující požadované výsledky. Grafický vzhled je ukázán na následujícím obrázku, kde je oblast pro prezentaci výsledků rozdělena na jedno, dvě nebo čtyři podokna pro přehlednou vizualizaci.



Obrázek 54: *Hlavní uživatelské rozhraní vytvořeného programu*

5.4.2 Popis tlačítek

Jednotlivá tlačítka jsou příhodně pojmenována po výše prezentovaných modelech. Pro vyvolání jednotlivých funkcí je nezbytné používat „počítačovou myš“. Ta s využitím kurzoru a po následném kliknutí vyvolá (spustí) požadovanou funkci. Výsledky funkce jsou zobrazeny „uprostřed“ spuštěného okna.

Obsažené názvy všech použitých funkcí jsou:

- *“Základní analýza dat”* – výpočet základních statistických charakteristik.
- *“Graf”* – základní grafická interpretace časových řad ze získaných souborů reálných dat.
- *“Hist, PP-graf”* – grafická interpretace histogramů a $p-p$ grafů pro normální rozdělení.
- *“Odhad $f(x)$”* – neparametrické jádrové odhady hustot s použitím vybraných typů jádrových funkcí.
- *“Odhad $F(x)$”* – neparametrické jádrové odhady distribučních funkcí s použitím vybraných typů jádrových funkcí.
- *“Srovnání $f(x)$, $F(x)$ ”* – srovnání neparametrických jádrových odhadů hustot a distribučních funkcí pro získané soubory dat (vývoz, dovoz).
- *“Odhad $f(x,y)$”* – dvourozměrné neparametrické jádrové odhady hustot.
- *“Odhad $F(x,y)$”* – dvourozměrné neparametrické jádrové odhady distribučních funkcí.
- *“Spol. Gauss”* – výpočet spolehlivosti s použitím Gaussovy jádrové funkce.
- *“Spol. Parzen”* – výpočet spolehlivosti s použitím Parzenovy jádrové funkce.

- *"Srovnání G – P"* – srovnání spolehlivostí získaných z neparametrických jádrových modelů s použitím Gaussovy a Parzenovy jádrové funkce.
- *"Trend (vývoz)"* – odhad trendové složky z celkové výše příjmů z vývozu pro získaný soubor reálných dat (s použitím vygenerovaného ortonormálního systému, bez heuristiky určující počet zahrnutých prvků).
- *"Trend (dovoz)"* – odhad trendové složky z celkové výše výdajů z dovozu pro získaný soubor reálných dat (s použitím vygenerovaného ortonormálního systému, bez heuristiky určující počet zahrnutých prvků).
- *"Srovnání trendů"* – srovnání odhadovaných trendových křivek mezi celkovou výší příjmů z vývozu a výdajů z dovozu pro získaný soubor reálných dat.
- *„Aproximace $F(x)$"* – aproximace distribučních funkcí (vývoz, dovoz).
- *„Kum $f(x)$, vývoz"* – neparametrický jádrový odhad hustoty z kumulovaných dat příjmů z vývozu.
- *„Kum $F(x)$, vývoz"* – neparametrický jádrový odhad distribuční funkce z kumulovaných dat příjmů z vývozu.
- *„Kum $f(x)$, dovoz"* – neparametrický jádrový odhad hustoty z kumulovaných dat výdajů z dovozu.
- *„Kum $F(x)$, dovoz"* – neparametrický jádrový odhad distribuční funkce z kumulovaných dat výdajů z dovozu.
- „Predikce, vývoz" – odhad trendové křivky za posledních 36 měsíců a predikce na následujících 12 měsíců z příjmů za vývoz (použití uvedené heuristiky).
- „Predikce, dovoz" – odhad trendové křivky za posledních 36 měsíců a predikce na následujících 12 měsíců z výdajů za dovoz (použití uvedené heuristiky).
- „Spolehlivost z kumulace rozdílu" – výpočet spolehlivosti a odhad hustoty i distribuční funkce z kumulovaných dat rozdílu (příjmy – výdaje).
- „Spolehlivost z trendu" – výpočet spolehlivosti z odhadovaných trendových složek a náhodných složek.
- „Vyčistit plochu" – vyčištění plochy po předchozí interpretaci výsledků.

Poznámka 5.7 *U neparametrických jádrových funkcí, které odhadují pravděpodobnostní rozdělení náhodné veličiny z načítaných (kumulovaných) hodnot, jsou výpočty náročnější na čas tj. „Kum $f(x)$ a $F(x)$ pro dovoz i vývoz", "Spolehlivost z kumulace".*

Kapitola 6

Závěr

V kapitole jsou prezentovány možné (budoucí) rozšíření této práce, ať již jako pokračování nebo samostatné výzkumy a otevřené otázky vzniklé během výzkumu. Všechny zde uvedené oblasti jsou založeny na problematice vzniklé při odvození v předchozích kapitolách a jsou navrženy ke zlepšení „kvality“ prezentovaných modelů nebo vedou k následnému vytvoření nových modelů založených na informacích z přechozích závěrů.

6.1 Možné rozšíření práce a otevřené otázky

Jeden z důležitých problémů popsaných v této práci je pravděpodobnost rozdílu uvažovaných náhodných veličin, který může být vyjádřen jako

$$P(x(t) < y(t)) \leq \varphi, \tag{6.1}$$

kde $x(t), y(t)$ jsou získaná pozorování obou náhodných proměnných pro $t = 1, \dots, n$ a konstanta φ může být interpretována jako požadovaná bezpečnost nebo garance (pravděpodobnost). Podrobná analýza požadované garance při aplikaci na (jiné) soubory reálných dat (např. ve strojírenství, stavebnictví, …) a ověření modelů v rozsáhlejší aplikační sféře.

Zajímavé problémy, které mohou být analyzovány pro následný výzkum, byly doposud diskutovány v literatuře a opět je zde prostor pro zlepšení získaných modelů. Mezi uvedenou problematiku patří zejména analýza vyhlazovacího parametru h, jeho vlivu na „kvalitu“ neparametrických jádrových odhadů (zainteresovanost) a závislosti na apriorní informaci o pravděpodobnostním rozdělení.

Další rozšíření práce se týká navrhovaných modelů časových řad, resp. metody dekompozice, kterou je možné rozšířit o sezónní i cyklickou složku a odhady zahrnout do modelu. Zároveň analyzovat (matematicky, ekonomicky, …) první hodnotu intervalu a zvolenou délku pro odhady jednotlivých složek (např. volba první hodnoty průměrem roku, …). Upravené modely poté aplikovat pro získání spolehlivosti. Současně analyzovat predikci trendové křivky pro různé volby modelů (např. přímka, …).

Samostatnou kapitolou je analýza jádrových funkcí, přesněji kladných jádrových funkcí, které zde byly stručně interpretovány (tj. jádrové funkce pro kladné hodnoty získaných pozorování) a jejich vliv na tvorbu modelů.

6.2 Vlastní závěr

V první části práce jsou diskutovány vybrané pojmy z teorie zabývající se spolehlivostí, rozbory, odvozené důkazy i vztahy (vazby, modely). Pozornost je zaměřena na různá pojetí spolehlivosti v souvislostech, tj. různé typy spolehlivosti a statistické pojetí rizikové události. Následně jsou uvedeny některá klasická pravděpodobnostní pojetí v popisu tvorby existujících modelů a dvěma vybranými modely s rozdílnou „náročností“ k odhadu. Závěr kapitoly popisuje navržený neparametrický model spolehlivosti.

Druhá část práce popisuje jednorozměrné i dvourozměrné neparametrické jádrové odhady hustoty i distribuční funkce včetně vlastností. Dále jsou prezentovány vybrané typy jádrových funkcí a vzniklá problematika, zejména volba vyhlazovacího parametru u hustoty při znalosti i neznalosti „skutečné“ funkce hustoty. Volba vyhlazovacího parametru je detailně rozebrána i pro odhad distribuční funkce, včetně popisu vzniklého algoritmu. V případě náhodné veličiny vzniklé součtem náhodných veličin je navržen (neparametrický) model pro odhad hustoty a distribuční funkce. Model je uveden pro získání spolehlivosti za kumulované období, což vychází z ekonomické podstaty zkoumaných dat. Závěr kapitoly prezentuje využití neparametrických odhadů pro spolehlivost s použitím různých typů jádrových funkcí (Parzenova a Gaussova). Výběr „vhodné“ jádrové funkce není na první pohled tak důležitý, protože zde použité funkce modelují „téměř stejné“ distribuční funkce. Rozdíly mezi funkcemi jsou hlavně v aproximacích hustot, proto provedený výběr může být v souladu s výpočetní složitostí (náročností na čas, hardwaru, …) nebo s obtížností při odvození (při zanedbání jádrových funkcí, např. funkce ve tvaru “V”).

Třetí část práce je zaměřena na využití modelů v aplikační sféře. V úvodu je popsána platební bilance (stručný ekonomický význam). Následuje problematika časových řad, dekompozice na možné složky včetně popisu potíží vybrané trendové (systematické) složky. Získané poznatky jsou použity na vybraný aditivní model časové řady s předpokládanou (neznámou) trendovou složkou. Odhad trendové složky (retrospektivního trendu) je založen na soustavě bazických časových průběhů a z ní vytvořené ortonormální soustavy (báze). Vzniklá náhodná složka je podrobena statistické inferenci. V návaznosti na předchozí odhady trendové křivky je potom navržena její krátkodobá (roční) predikce včetně heuristického přístupu k výběru možné množiny vygenerovaných ortonormálních složek. Vzniklá systematická (trendová) a náhodná složka je použita k získání spolehlivosti. Závěr kapitoly se zabývá nezáporností platební bilance modelovanou spolehlivostí, obchodní bilancí v aktuálních hodnotách a kumulacích na získaných reálných datech včetně grafické prezentace dat.

Následuje konkrétní aplikace navržených modelů na souborech reálných dat z obchodní bilance České republiky a rozbory (ověření) získaných výsledků. Nejprve je zde prostor pro vlastní realizaci a využití základních vztahů (neparametrické jádrové odhady hustoty, distribuční funkce, spolehlivosti) na vygenerovaném souboru dat. Zde jsou analyzovány vlastnosti modelů a jejich citlivost na vstupní parametry. Z výsledků lze usuzovat na možné závěry, že „významný“ vliv na odhad tvaru hustoty má parametr měřítka i výběr jádrové funkce. Odhad distribuční funkce již není tak citlivý na volbě jádrové funkce a nejsou zde téměř „žádné“ rozdíly v odhadech. Získané spolehlivosti nevykazují „významné“ rozdíly pro vybrané jádrové funkce a parametr vyhlazování. Potom jsou navržené modely aplikovány na získané soubory reálných dat v aktuálních hodnotách, na kterých jsou přechozí závěry potvrzeny. Neparametrický jádrový odhad distribuční funkce podle navrženého přístupu nevykazuje rozdílné hodnoty oproti klasickému neparametrickému jádrovému odhadu, ale je dosaženo přiblížení k odhadu empirické distribuční funkci a „nižší“ náročnosti na výpočty. Detailněji jsou zde popsány analýzy spolehlivosti (selhání) pro jednotlivé měsíce a za celé období. Výsledné hodnoty analýz za celé období jsou poté porovnány s relativní frekvencí, kde není dosaženo významných změn (tj. více než 1 procentní bod), a proto se relativní frekvence jeví jako vhodnější metoda pro spolehlivost v případě stacionárního charakteru dat. V případě nestacionárních charakteru je odhadována trendová (systematická) složka, její predikce a výsledná hodnota spolehlivosti. Zde dosahuje spolehlivost vyšších hodnot, než v případě stacionárního charakteru dat. Následuje spolehlivost pro kumulovaná data, která vzhledem k dlouhodobému kladnému saldu dosahuje za celé období vysoké hodnoty. Kapitola je zakončena popisem uživatelského rozhraní vytvořeného programu.

V závěru jsou zde diskutovány nejen budoucí možná rozšíření této práce, ale i vzniklé otevřené otázky pro hledání různých přístupů v dalším osobním rozvoji.

Hlavním cílem práce je popis současného stavu spolehlivostní problematiky a rozšíření do oblasti neparametrických jádrových modelů jak pro stacionární tak nestacionární charakter dat. Získané poznatky (myšlenky) jsou poté aplikovány na soubory reálných dat z aplikační oblasti (platební resp. obchodní bilanci). Výhoda uvedených přístupů spočívá v získání spolehlivosti (selhání) bez apriorní znalosti pravděpodobnostního rozdělení pro získaný dvourozměrný soubor dat v případě stacionárních charakteru a v případě nestacionárního charakteru dat ve zlepšení „kvality“ vytvořených modelů.

Literatura

[1] L. Devroye and L. Gyorfi, *Nonparametric Density Estimation: The L1 View*, John Wiley, New York, 1985.

[2] A. Rényi, *Teorie pravděpodobnosti*, ACADEMIA, Praha, 1972.

[3] R. C. Gupta and S. Subramanian, *Estimation of reliability in a bivariate normal distribution with equal coefficients of variation,* Comm. Statist. Simulation Comput. 27, 1998.

[4] S. Nadarajah, *Reliability for some Bivariate Gamma Distributions*, Hindawi Publishing Corporation, 2005.

[5] D. D. Hangal, *Estimation of System Reliability under Bivariate Pareto Distribution*, University of Poona, India, 1996.

[6] S. Nadarajah and S. Kotz, *Reliability for some Bivariate Exponential Distributions*, Hindawi Publishing Corporation, 2005.

[7] Roger B. Nelsen, *An Introduction to Copulas*, Springer Science +Business Media, Inc., 2006.

[8] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, Springer Science + Business Media, LLC, 2006.

[9] M. Kvasnička and O. Vašíček, *Úvod do analýzy časových řad*, Skripta Masarykova Univerzita, 2001.

[10] T. Cipra, *Finanční ekonometrie*, EKOPRESS, s.r.o., 2008.

[11] G. Williams, *Linear Algebra with Applications*, Jones and Bartlett Publishers, Inc., 2008.

[12] M. Todinov, *Reliability and Risk Models*, John Wiley & Sons Ltd., 2005.

[13] P. C. B. Phillips, *The Mysteries of Trend*, Cowles Foundation for Research in Economics, Yale University, 2010.

[14] S. Kotz, N. Balakrishnan, N. L. Johnson, *Continuous Multivariate Distributions*, John Wiley & Sons, Inc., 2000.

[15] Ch. H. Skiadas, *Recent Advances in Stochastic Modeling and Data Analysis,* World Scientific Publishing Co. Pte. Ltd., 2007.

[16] Berwin A. Turlach, *Bandwidth Selection in Kernel Density Estimation: A Review,* C.O.R.E. and Institut de Statistique, Université Catholique de Louvain, Belium, 1993.

[17] W. Hardle, *Smoothing Techniques,* Springer – Verlag New York Inc., 1991.

[18] Murray R. Spiegel, Larry J. Stephens, *Statistics,* The McGraw-Hill Companies Inc., 2008.

[19] J. Hátle, J. Likeš, *Základy počtu pravděpodobnosti a matematické statistiky,* SNTL, Praha, 1974.

[20] W. Cheney, D. Kincaid, *Numerical Mathematics and Computing 6th Edition,* Thomson Learning, Inc., 2008.

[21] M. Merriman, *A text book on the Method of Least Squares,* Adamant Media Corporation, 2005.

[22] W. Hardle, M. Muller, S. Sperlich, A. Werwatz, *Nonparametric and Semiparametric Models,* Springer – Verlag, Berlin Heidelberg, 2004.

[23] E. Hewitt, K. R. Stromberg, *Real and Abstract Analysis,* Springer, 1975.

[24] M. Řezáč, *Jádrové odhady hustoty,* disertační práce, Masarykova univerzita, Brno, 2007.

[25] J. Orava, *Volba vyhlazovacího parametru při jádrových odhadech hustoty,* diplomová práce, Masarykova univerzita, Brno, 2008.

[26] J. Nováková, *Jádrové odhady distribuční funkce,* diplomová práce, Masarykova univerzita, Brno, 2009.

[27] Frank Deutsch, *Best Approximation in Inner Product Spaces,* Springer – Verlag New York, Inc., 2001.

[28] J. F. Cornwell, *Group Theory in Physics,* Academic press, 1997.

[29] N. Young, *An Introduction to Hilbert Space,* Cambridge University Press, 1988.

[30] L. N. Trefethen, D. Bau, *Numerical Linear Algebra,* Society for Industrial and Applied Mathematics, 1997.

[31] L. Collatz, *Funkcionální analýza a numerická matematika,* SNTL Praha, 1970.

[32] G. Bachman, L. Narici, *Functional Analysis,* George General publishing company, Ltd., 2000.

[33] F. Vávra, P. Nový a kol., *Nonparametric estimations of non-negative random variables distributions,* Kybernetika, vol. 39, issue 3, 2003.

[34] F. Zhang, *Linear Algebra: Challenging Problems for Students 2nd Edition,* The Johns Hopkins University Press, Baltimore, 2009.

[35] T. Amemiya, *Introduction to Statistics and Econometrics,* President and Fellows of Harvard College, 1994.

[36] J. Bečvář, *Vektorové prostory II,* Státní pedagogické nakladatelství, Praha, 1981.

[37] P. Neumann, P. Žamberský a M. Jiránková, *Mezinárodní ekonomie,* Státní pedagogické nakladatelství, Praha, 2010.

[38] Česká národní banka (ČNB), 2012. *Platební balance – měsíční*. [online] Zdroj: <http://www.cnb.cz/cs/statistika/platebni_bilance_stat/platebni_bilance_m/index.html> [Přístup 17. 8. 2012]

[39] A. Kufner, *Geometrie Hilbertova prostoru,* SNTL Praha, 1973.

[40] J. A. Rozanov, *Slučajnye processy,* Moskva: Nauka, 1979.

[41] Introduction to RiskMetrics™, November 21, 1995, *Fourth edition*. New York: Morgan Guaranty Trust Company, Risk Management Services, riskmetrics@jpmorgan.com.

[42] R. A. Johnson, J. W. Evans, D. W. Green, *Some Bivariate Distributions for Modeling the Strength Properties of Lumber*, Forest Products Laboratory, 1999.

[43] J. F. Kenney, E. S. Keeping, *Mathematics of statistics, Part two*, Princeton: D. Van Nostrand Company, 1951.

Příloha A

Příloha popisuje důkaz asymptotické nestrannosti a vydatnosti neparametrického jádrového odhadu. Nejprve je zde popsána „nestrannost“, která může být odvozena s pomocí následujícího postupu.

A.1 Asymptotická nestrannost

Nechť $\{x_1, \dots, x_n\}$ je i.i.d. výběr náhodné veličiny X rozsahu n, potom podmínky pro uvažovaný náhodný výběr dávají shodnou a nedostupnou funkci hustoty $f(x)$ pro všechny získaná pozorování v tomto souboru dat. Ta není samozřejmě k dispozici a jejich sdružená funkce hustoty může být vyjádřena jako

$$f(x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i)\,, \qquad n \in \mathbb{N}_+. \tag{A.1}$$

Střední hodnota z odhadované hustoty obsahující náhodný výběr, kde z podmínek pro náhodný výběr předpokládáme shodnou a nedostupnou hustotu $f(x)$ pro všechna pozorování, může být odvozena následujícím postupem, kde jsou ukázány postupné kroky pro lepší pochopení celé problematiky.

$$\begin{aligned} E\{\hat{f}(x;h)\} &= \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \hat{f}(x;h) \left[\prod_{i=1}^{n} f(x_i)\right] dx_1 \dots dx_n \\ &= \frac{1}{nh} \sum_{j=1}^{n} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} k\left(\frac{x - x_j}{h}\right) \left[\prod_{i=1}^{n} f(x_i)\right] dx_1 \dots dx_n \\ &=_a \frac{1}{h} \int_{-\infty}^{+\infty} k\left(\frac{x - z}{h}\right) f(z) dz \\ &= \int_{-\infty}^{+\infty} k(y)\, f(x - hy) dy, \qquad h > 0 \qquad n \in \mathbb{N}_+. \end{aligned} \tag{A.2}$$

Symbolem $=_a$ je v postupu použita substituce $y = \frac{x-z}{h} \Longrightarrow dz = -h dy$. Potom limita uvažované střední hodnoty pro $h \longrightarrow 0_+$ může být přepsána do tvaru

$$\lim_{h \to 0_+} E\{\hat{f}(x;h)\} = \lim_{h \to 0_+} \int_{-\infty}^{+\infty} k(y)\, f(x - hy) dy\,, \qquad h > 0 \qquad n \in \mathbb{N}_+. \tag{A.3}$$

Pokud jsou zároveň splněny předpoklady záměny limity a integrálu (zde v nevlastních mezích) a funkce hustoty je spojitá, potom platí následující vztah

$$\begin{aligned}\lim_{h\to 0_+} E\{\hat{f}(x;h)\} &= \lim_{h\to 0_+} \int_{-\infty}^{+\infty} k(y)\, f(x-hy)dy \\ &= \int_{-\infty}^{+\infty} k(y)\left(\lim_{h\to 0_+} f(x-hy)\right)dy \\ &= \int_{-\infty}^{+\infty} k(y)\left(f\left(x-\lim_{h\to 0_+}(hy)\right)\right)dy \\ &= f(x), \qquad h>0 \qquad n\in\mathbb{N}_+.\end{aligned} \tag{A.4}$$

Celý důkaz tohoto tvrzení za daleko obecnějších podmínek lze nalézt detailně zpracovaný v Devroy a Gyorfi (1985, s. 6 - 11). Dále pokud platí

$$h = h(n) \longrightarrow 0, \tag{A.5}$$

pro $n \longrightarrow \infty$ a $f(x)$ je spojitá, potom odhad hustoty $\hat{f}(x;h)$ je asymptoticky nestranný.

A.2 Asymptotická vydatnost

Zbývající důkaz asymptotické vydatnosti neparametrického jádrového odhadu může být opět odvozen níže popsaným postupem. Nejprve je potřeba odvodit střední hodnotu kvadrátu odhadu hustoty $\hat{f}^2(x;h)$ postupem

$$\begin{aligned}E\{\hat{f}^2(x;h)\} &= \int_{-\infty}^{+\infty}\dots\int_{-\infty}^{+\infty} \hat{f}^2(x;h)\left[\prod_{i=1}^{n} f(x_i)\right]dx_1\ \dots dx_n \\ &= \frac{1}{n^2h^2}\left[\sum_{i=1}^{n}\int_{-\infty}^{+\infty} k^2\left(\frac{x-x_k}{h}\right)f(x_k)dx_k\right. \\ &\quad \left. +2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} k\left(\frac{x-x_j}{h}\right)k\left(\frac{x-x_k}{h}\right)f(x_j)\,f(x_k)dx_j dx_k\right] \\ &=_a \frac{1}{nh^2}\int_{-\infty}^{+\infty} k^2\left(\frac{x-x_k}{h}\right)f(x_k)dx_k \\ &\quad +\frac{n-1}{nh^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} k\left(\frac{x-x_j}{h}\right)k\left(\frac{x-x_k}{h}\right)f(x_j)\,f(x_k)dx_j dx_k \\ &=_b \frac{1}{nh}\int_{-\infty}^{+\infty} k^2(y)\,f(x-hy)dy+\frac{n-1}{n}\left(\int_{-\infty}^{+\infty} k(y)\,f(x-hy)dy\right)^2,\end{aligned} \tag{A.6}$$

pro $h>0$ a $n\in\mathbb{N}_+$. Dále je pod použitým označením $=_a$ uvažován následující vztah pro $t=1,\dots,n$ ve tvaru

$$\frac{1}{n^2h^2}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}k^2\left(\frac{x-x_t}{h}\right)f(x_t)dx_t=\frac{1}{nh^2}\int_{-\infty}^{+\infty}k^2\left(\frac{x-x_t}{h}\right)f(x_t)dx_t \tag{A.7}$$

a

$$\frac{2}{n^2h^2}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\ldots=\frac{2}{n^2h^2}\cdot\frac{n(n-1)}{2}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\ldots, \tag{A.8}$$

pro $h>0$ a $n\in\mathbb{N}_+$, kde pod označením $=_b$ je charakterizována stejná substituce $y=\frac{x-z}{h}\Longrightarrow dz=-hdy$ jako v postupu asymptotické nestrannosti. Následně může být získán rozptyl odhadu již s použitím získaných výsledků

$$\begin{aligned}\sigma^2\{\hat{f}(x;h)\}&=E\{\hat{f}^2(x;h)\}-\left(E\{\hat{f}(x;h)\}\right)^2\\&=\frac{1}{nh}\int_{-\infty}^{+\infty}k^2(y)\,f(x-hy)dy-\frac{1}{n}\left(\int_{-\infty}^{+\infty}k(y)\,f(x-hy)dy\right)^2,\end{aligned} \tag{A.9}$$

pro $h>0$ a $n\in\mathbb{N}_+$.

Odtud okamžitě (a z asymptotické nestrannosti) plyne požadavek pro existenci rozptylu a asymptotickou vydatnost: $n\to+\infty\Rightarrow[(h\to0)\wedge(nh\to+\infty)]$. Samozřejmě je zde předpokládáno, že parametr měřítka je funkcí rozsahu pozorování $h=h(n)$.

Stručně řečeno, pro asymptoticky nestranný a vydatný odhad jsou postačující tři základní podmínky pro $n\to+\infty$:

$$\begin{aligned}&h\to0\\&nh\to+\infty\end{aligned}$$

a $f(x)$ je spojitá.

Seznam publikovaných prací

Články v časopisech a sbornících

ŤOUPAL T.: *Spatial Poisson Process Parameter Estimation using Information about Subset*. In 32nd Research Student's Conference in Probability and Statistics. Lancaster: Lancaster University, 2009. s. 35 - 36.

VÁVRA F., WAGNEROVÁ E., ŤOUPAL T., MAREK P.: *Economical and Financial View of Ageing*. Finanční řízení podniků a finančních institucí. Ostrava: Vysoká škola báňská – Technická univerzita, 2009. s. 453 - 460. ISBN: 978-80-248-2059-0.

ŤOUPAL T.: *Nonparametric Estimation of Reliability of Two Random Variables Using Kernel Estimation of Density*. In 33rd Research Student's Conference in Probability and Statistics. Warwick: University of Warwick, 2010. s. 85.

ŠEDIVÁ B., WAGNEROVÁ E., VÁVRA F., ŤOUPAL T., MAREK P.: *Statistical Monitoring of Failures – Methods and Use*. In Proceedings of the 11th International Scientific Conference Electric Power Engineering 2010. Brno: Brno University of Technology, 2010, s. 611 – 615. ISBN 978-80-214-4094-4.

ŤOUPAL T., MAREK P.: *Parameter Estimation of Spatial Poisson Process in Domain with Unknown Boundary*. Journal of Combinatorics, Information & System Sciences 35, New Delhi: MD Publications Pvt Ltd, 2010, s. 231-242. ISBN: 0250-9628.

VÁVRA F., WAGNEROVÁ E., ŤOUPAL T., MAREK P., HANZAL Z.: *Inflation rate prediction – a statistical approach*. Finanční řízení podniků a finančních institucí. Ostrava: Vysoká škola báňská – Technická univerzita, 2011, s. 552 - 559. ISBN: 978-80-248-2494-9.

ŤOUPAL T.: *Trend component Estimation*. In 35th Research Student's Conference in Probability and Statistics. Warwick: University of Southampton, 2012. s. 41.

ŤOUPAL T.: *Trend Component Estimation*. Řízení a modelování finančních rizik. Ostrava: Vysoká škola báňská – Technická univerzita, 2012, s. 618-627. ISBN: 978-80-248-2835-0.

ŤOUPAL T., VÁVRA F.: *Risk events, statistical point of view*. Řízení a modelování finančních rizik. Ostrava: Vysoká škola báňská – Technická univerzita, 2012, s. 628 - 633. ISBN: 978-80-248-2835-0.

Poloprovozy

VÁVRA F., ŠEDIVÁ B., NOVÝ P., ŤOUPAL T., MAREK P., WAGNEROVÁ E.: *Intervaly spolehlivosti intenzity poruch a spojená problematika 1 (Confidence Intervals for Failure Rate and Related Problems 1)*. ČEPS, a.s., 2009.

VÁVRA F., ŠEDIVÁ B., NOVÝ P., ŤOUPAL T., MAREK P., WAGNEROVÁ E.: *Aplikace vybraných obecných ekonomických a spolehlivostních modelů pro společnost ČEPS (Application of Selected Common Economic and Reliability Models for the Company ČEPS).* ČEPS, a.s., 2009.