



**MATEMATICKO-FYZIKÁLNÍ
FAKULTA**
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Patrik Nácovský

Modelování délky finančních transakčních dat

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Mgr. Hendrych Radek, Ph.D.

Studijní program: Matematika

Studijní obor: Finanční matematika

Praha 2019

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

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

V Praze dne 30. června 2019

Podpis autora

Rád bych tímto poděkoval svému vedoucímu bakalářské práce panu RNDr. Mgr. Radku Hendrychovi, Ph.D. nejen za, pro mě osobně, velmi zajímavé téma, ale především za trpělivost, cenné rady a veškerý věnovaný čas.

Název práce: Modelování durace finančních transakčních dat

Autor: Patrik Nácovský

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Mgr. Hendrych Radek, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato práce se zabývá modelem ACD (autoregressive conditional duration), který se používá k modelování durací časových řad finančních transakčních dat. Nejprve je představen pojem durace a časové řady, a to intuitivně i formálně. Následně je zaveden model ACD. Jsou představeny jeho základní typy, které jsou určeny různými pravděpodobnostními rozděleními reziduální složky. Rovněž je nastíněna možnost použití tohoto modelového rámce k predikcím. V další části jsou popsány jednotlivé kroky konstrukce modelu, tj. jak se ACD model identifikuje, odhaduje a diagnostikuje. Poslední část práce je věnována praktické ukázce konstrukce základních modelů EACD, WACD a GACD pro různé datové soubory s velkým množstvím pozorování. K dispozici jsou transakční data akcií Apple, kurzu EUR/USD a kurz zлата. Objem dat v každém ze tří souborů se pohybuje mezi 300 tisíc až 600 tisíc pozorování (jeden obchodní týden).

Klíčová slova: model ACD, durace, časové řady

Title: Modelling Duration of Financial Transaction Data

Author: Patrik Nácovský

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Mgr. Hendrych Radek, Ph.D., Department of Probability and Mathematical Statistics

Abstract: This bachelor thesis deals with ACD (autoregressive conditional duration) model, which is used to estimate durations of time series of financial transaction data. First, duration and time series are defined formally as well as with the intuitive way. Next, model ACD itself is defined and its basic types, which are determined with distribution of its residuals. Then way to use this model for predictions is introduced. In the second part, steps for model identification, construction and revision are described. In the last part models EACD, WACD and GACD are constructed for real data. There are three data sets of thick data, which are Apple stocks, EUR/USD and gold. Data sets contain from 300 thousands to 600 thousands elements (one trading week).

Keywords: ACD model, durations, time series

Obsah

Úvod	2
Seznam použitých zkratk	3
1 ACD model a jeho vlastnosti	4
1.1 Modelování časových řad	4
1.2 ADC model	5
1.3 Základní momenty modelu ACD	6
1.4 Metoda predikce	8
1.5 Typy modelu ACD	9
2 Konstrukce modelu ACD	12
2.1 Identifikace modelu	12
2.2 Odhad modelu	12
2.3 Diagnostika modelu	13
3 Aplikace na reálná data	15
3.1 Informace o použitých datech	15
3.2 Očištění dat	16
3.3 Odhad modelu	18
3.3.1 EACD a WACD	19
3.3.2 GACD	23
Závěr	26
Seznam použité literatury	27
Seznam obrázků	28
Seznam tabulek	29
A Přílohy	30
A.1 Rozdělení pravděpodobnosti	30
A.1.1 Exponenciální rozdělení	30
A.1.2 Weibullovo rozdělení	30
A.1.3 Zobecněné gama rozdělení	30

Úvod

Mechanismus trhu je starý jako lidstvo samo a lidé odjakživa využívali trh nejen k nakupování potřebných a prodeji přebytečných statků, ale také ke spekulacím a dosažení výdělku. S tím jde ruku v ruce také snaha modelovat ceny a odhadovat jejich chování v budoucnu.

Během posledních desetiletí jsme však svědky významného technologického pokroku, což způsobilo přesun velké části obchodování na finančních trzích na výpočetní techniku. Rychlost a frekvence obchodů roste, je možné ukládat velké množství dat. V takovém prostředí může být zajímavé, kromě ceny samotné, se zaměřit také na časové prodlevy mezi jednotlivými transakcemi, označované jako *durace*.

Úspěšné modelování a předpovídání *durací* může s předstihem upozornit na novou informaci na trhu, která může následně vést ke změně ceny. Protože intuitivně jsou prodlevy mezi obchody velké, pokud žádné nové informace na trhu nejsou, a naopak probíhá velké množství obchodů s malými prodlevami, pokud se na trh dostane nová informace.

V této práci si nejprve představíme pojem časové řady, formálně zavedeme vlastní pojem *durace* a model ACD se všemi jeho vlastnostmi, včetně různých typů podle pravděpodobnostního rozdělení reziduí. Pak probereme postup, jak takový model ACD vhodně zkonstruovat, ověřit splnění všech předpokladů a zhodnotit jeho kvalitu.

Nakonec si ukážeme konstrukci modelu na reálných datech. K dispozici máme tři balíčky transakčních dat, každý z jednoho obchodního týdne. Každý balíček obsahuje řádově okolo 300 tisíc až 600 tisíc pozorování. Zkonstruujeme několik základních typů modelu ACD, vyhodnotíme kvalitu a srovnáme vlastnosti.

Seznam použitých zkratek

- Ω – prostor elementárních jevů
- \mathcal{A} – σ -algebra na množině elementárních jevů Ω
- P – pravděpodobnostní funkce
- \mathbb{N} – obor přirozených čísel
- \mathbb{N}_0 – obor přirozených čísel včetně nuly
- \mathbb{Z} – obor celých čísel
- \mathbb{R} – obor reálných čísel
- $\text{Exp}(\lambda)$ – exponenciální rozdělení s parametrem $\lambda > 0$
- $\text{Weib}(\alpha, \beta)$ – Weibullovo rozdělení s parametry $\alpha, \beta > 0$
- $\Gamma(\alpha, \beta, \kappa)$ – Zobecněné gama rozdělení s parametry $\alpha, \beta, \kappa > 0$
- $\{X_t, t \in T\}$ – časová řada ve smyslu náhodných veličin
- $\{x_t, t \in T\}$ – konkrétní realizace časové řady
- $\{y_t, t \in T\}$ – časová řada očištěná o deterministickou složku (zda se jedná o náhodné veličiny nebo jejich realizace vyplývá z kontextu)

1. ACD model a jeho vlastnosti

1.1 Modelování časových řad

Všechny statistické modely uvedené v této práci se aplikují na časové řady. Časovou řadu lze chápat jako diskrétní množinu náhodných veličin, kterou je možné jednoznačně uspořádat z hlediska času.

Definice 1 (časová řada). *Mějme pravděpodobnostní prostor (Ω, \mathcal{A}, P) a množinu $T \subset \mathbb{N}_0$. Časovou řadou nazýváme množinu $\{X_t, t \in T\}$, kde X_t jsou reálné náhodné veličiny z (Ω, \mathcal{A}, P) .*

Definice 2 (slabá stacionarita). *Časovou řadu $\{X_t, t \in T\}$ nazýváme slabě stacionární, pokud má konečné druhé momenty, konstantní rozptyl a kovariance jejích členů je nezávislá na časovém posunutí. Tedy platí*

$$E [X_t^2] < \infty \quad \text{pro } \forall t \in T, \quad (1.1)$$

$$E X_t = \mu \quad \text{pro } \forall t \in T, \mu \in \mathbb{R}, \quad (1.2)$$

$$\text{Cov}(X_k, X_{k+d}) = \text{Cov}(X_l, X_{l+d}) \quad \text{pro } \forall k, l \in T, d \in \mathbb{N}_0. \quad (1.3)$$

Poznámka. V této práci budeme uvažovat časové řady a modely, které jsou slabě stacionární. V praxi je tato podmínka u časových řad splněna jen zřídka. Existují však způsoby, jak časové řady očistit o trendovou či periodickou složku, a transformovat je tak, aby byly slabě stacionární.

Vhodným příkladem časové řady může být cena aktiva v různých časových okamžicích. Avšak hodnota časové řady není jediná vlastnost, kterou je zajímavé a užitečné modelovat a predikovat. Další důležitou vlastností časové řady je rozptyl dat. Tato informace je zásadní pro určení rizika spojeného s nákupem a držetím aktiva. Pro investora je to ukazatel, jak moc se může cena vychýlit od předpovědi, ať už v jeho prospěch či neprospěch. Pro data s nekonstantním rozptylem je podstatná právě změna rozptylu v čase.

Až doposud jsme uvažovali časovou řadu jako hodnoty naměřené v ekvidistantních časových intervalech, což je pro základní modely žádoucí. Délka takového zvoleného časového intervalu se může velmi lišit v závislosti na volatilitě konkrétního aktiva či na potřebě konkrétního uživatele.

S rozvojem počítačových technologií došlo k přesunu obchodování mnoha aktiv na internetové platformy. To umožnilo vysokou frekvenci uskutečňování obchodů a pohybu cen společně s ukládáním velkého množství dat. Volba správného intervalu pro pozorování se tím komplikuje. Pokud zvolíme velký interval, jako je jeden obchodní den, pro modelování aktiva, jehož cena se mění v řádech sekund, ztrácíme tím velké množství informací. Když naopak zvolíme příliš malý interval, při většině měření zjistíme nulovou změnu od posledního stavu, což bude mít negativní dopad na zvolený model.

Cena aktiva je diskrétní veličina. Proto změna ceny nemůže být libovolně malá, ale je přesně definovaný minimální pohyb ceny a každý pohyb je jeho celočíselným násobkem. Změny ceny probíhají přijímáním takzvaných tickových dat. Každý jeden tick znamená novou příchozí informaci. Tato informace může být o změně ceny nebo také jen o nové uskutečněné transakci, která sama cenu

neovlivnila. Nedochozí tedy ke ztrátám dat v důsledku příliš velkých zvolených časových intervalů, ani se neobjevují prázdné intervaly bez nové informace. Ticky jsou diskrétní změny a lze je jednoznačně seřadit ve smyslu času. Přístupujeme k nim, jako by přicházely v ekvidistantních časových intervalech. To nám dává možnost nahlížet na skutečné časové intervaly mezi ticky jako na hodnoty časové řady ve smyslu definice 1, kde x_t , coby realizace časové řady X_t , je interval mezi příchodem t -ho a $(t + 1)$ -ho ticku.

Frekvence transakcí během dne má periodický charakter a také reaguje na příchod nových informací o obchodovaném aktivu. Periodickou část pozorování se budeme snažit odfiltrovat.

1.2 ADC model

Hlavní myšlenka modelování intervalů mezi transakcemi spočívá v jejich shlukování podle situace na trhu. Krátké intervaly napovídají, že se na trhu objevila nová informace a investoři na ni reagují zvýšenou aktivitou. Tato situace trvá, dokud se stav neuklidní. Pokud je naopak na trhu klid, transakce přicházejí s velkými prodlevami. Tedy časové intervaly následující těsně po sobě mají tendenci mít podobnou délku.

Budeme vycházet ze základního lineárního modelu ARMA (Tsay, 2002), který se používá na modelování (slabě stacionárních) časových řad. Jedná se o součet autoregresního modelu (AR) a klouzavého průměru (MA). Autoregresní část tvoří vážený průměr předchozích hodnot, zatímco klouzavý průměr je složený z chyb předchozích predikcí. Model ARMA však předpokládá shodný rozptyl dat, čehož je v praxi těžké docílit. Speciální modifikací modelu ARMA tak, aby modeloval právě rozptyl dat, vznikne model GARCH (generalized autoregressive conditional heteroskedasticity) (Tsay, 2002).

Statistický model ACD (autoregressive conditional duration) pro intervaly mezi transakcemi byl poprvé představen v článku Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data (Engle a Russell, 1998). Vznikl odvozením ekvivalentními úpravami z modelu ARMA, jako výše zmíněný model GARCH.

Nechť t_i je čas i -té transakce. Pak

$$\Delta t_i = t_i - t_{i-1}, i \in \mathbb{N}, \quad (1.4)$$

značí i -tý interval mezi dvěma transakcemi. Tento interval budeme označovat jako i -tou duraci. Durace Δt_i vyhovují definici 1 a tvoří časovou řadu.

Jak již bylo řečeno, je potřeba časovou řadu transformovat tak, aby byla slabě stacionární. Například je třeba eliminovat periodicitu a trend. Pro tento účel se volí funkce $f(t_i)$ tak, aby

$$y_i = \Delta t_i^* = \frac{\Delta t_i}{f(t_i)}, i \in \mathbb{N}, \quad (1.5)$$

označovaly durace očištěné o deterministickou složku. Existuje více postupů, jak funkci f vhodně volit a odhadnout. Nedá se však říct, že by některý výrazně

převládá. Obecně časté je použití splinů. Zde definovaný model ACD je popsán například v knize Analysis of Financial Time Series (Tsay, 2002).

Definice 3 (ACD(r, s) model). *ACD(r, s) model je definovaný předpisem*

$$\Delta t_i^* = y_i = \psi_i \varepsilon_i, \quad (1.6)$$

$$\psi_i = \alpha_0 + \sum_{j=1}^r \alpha_j \psi_{i-j} + \sum_{k=1}^s \beta_k y_{i-k}, \quad i \in \mathbb{N}_0, \quad (1.7)$$

$\alpha_0 > 0$, $\alpha_1, \dots, \alpha_r \geq 0$, $\alpha_j = 0$ pro $j > r$, $\beta_1, \dots, \beta_s \geq 0$, $\beta_k = 0$ pro $k > s$, $\sum_{j=1}^{\max(r,s)} (\alpha_j + \beta_j) < 1$, kde ε_i jsou nezávislé nezáporné stejně rozdělené náhodné veličiny, pro které platí $E(\varepsilon_i) = 1$ a $\psi_i = E(y_i | \mathcal{F}_{i-1})$ je očekávaná durace podmíněná \mathcal{F}_{i-1} , což představuje veškerou informaci o časové řadě dostupnou v čase $(i-1)$.

Poznámka. Omezení pro parametry modelu ACD uvedené v definici 3 jsou postačující podmínkou pro existenci konečné střední hodnoty ACD modelu (viz (1.11)).

Podle pravděpodobnostního rozdělení náhodných veličin ε_i se rozlišují různé druhy modelu ACD. Základním a nejjednodušším je model EACD, který předpokládá exponenciální rozdělení náhodných veličin ε_i . Budeme se zde zabývat také obecnějšími modely WACD a GACD, které předpokládají Weibullovo rozdělení, respektive gama rozdělení.

1.3 Základní momenty modelu ACD

Souvislost s modelem ARMA

Pro účely výpočtu momentů zavedeme chybu i -tého odhadu

$$\eta_i = y_i - \psi_i, \quad (1.8)$$

kterou následně aplikujeme na úpravu rovnice (1.7):

$$\begin{aligned} y_i - \eta_i &= \alpha_0 + \sum_{j=1}^r \alpha_j (y_{i-j} - \eta_{i-j}) + \sum_{k=1}^s \beta_k y_{i-k}, \\ y_i &= \eta_i + \alpha_0 + \sum_{j=1}^r \alpha_j y_{i-j} - \sum_{j=1}^r \alpha_j \eta_{i-j} + \sum_{k=1}^s \beta_k y_{i-k}, \\ y_i &= \alpha_0 + \sum_{j=1}^{\max(r,s)} y_{i-j} (\alpha_j + \beta_j) + \eta_i - \sum_{k=1}^r \alpha_k \eta_{i-k}. \end{aligned} \quad (1.9)$$

Takto lze model ACD(r, s) upravit zpět na model ARMA($\max(r, s), r$), ze kterého byl odvozen. Ve tvaru rovnice (1.9) je vidět součet autoregresní části tvořené předchozími hodnotami a části klouzavého průměru tvořené chybami předchozích odhadů.

Střední hodnota y_i

Střední hodnotu a rozptyl náhodné veličiny y_i řídící se ACD modelem spočteme aplikováním těchto funkcí na obě strany rovnice (1.9). Nejprve si odvodíme střední hodnotu náhodné veličiny η_i , pro kterou podle definice 3 a (1.8) platí

$$\begin{aligned}\mu_\eta = E \eta_i &= E (\psi_i(\varepsilon_i - 1)) = E (E (\psi_i(\varepsilon_i - 1)|\mathcal{F}_{i-1})) = \\ &= E (\psi_i E (\varepsilon_i - 1)) = E (0) = 0, \quad \forall i \in \mathbb{N}_0.\end{aligned}\quad (1.10)$$

O střední hodnotě modelu ACD se dozvíme více v následující větě. Rozptyl obecného modelu je příliš algebraicky náročný. Jeho výpočet si ukážeme dále a zvolíme si k tomuto účelu jednodušší model ACD(1, 1).

Věta 1. *Nechť je model ACD slabě stacionární, pak pro jeho střední hodnotu platí*

$$E y_i = E \psi_i = \frac{\alpha_0}{1 - \sum_{j=1}^{\max(r,s)} (\alpha_j + \beta_j)}.\quad (1.11)$$

Důkaz. Hledaný vztah se dostane aplikací střední hodnoty na obě strany rovnice (1.9) a ze vztahu (1.10)

$$E y_i = \alpha_0 + \sum_{j=1}^{\max(r,s)} E (y_{i-j})(\alpha_j + \beta_j) + E \eta_i - \sum_{k=1}^r \alpha_k E (\eta_{i-k}),\quad (1.12)$$

$$E y_i = \alpha_0 + \sum_{j=1}^{\max(r,s)} E (y_{i-j})(\alpha_j + \beta_j).\quad (1.13)$$

Dle předpokladu slabé stacionarity platí $E y_i = E y_{i-j} = \mu_y$ a tedy střední hodnota nezáleží na čase

$$\mu_y = \alpha_0 + \mu_y \sum_{j=1}^{\max(r,s)} (\alpha_j + \beta_j),\quad (1.14)$$

$$\mu_y = E y_i = \frac{\alpha_0}{1 - \sum_{j=1}^{\max(r,s)} (\alpha_j + \beta_j)}.\quad (1.15)$$

Pro střední hodnotu ψ_i navíc platí

$$E y_i = E (\psi_i \varepsilon_i) = E (E (\psi_i \varepsilon_i | \mathcal{F}_{i-1})) = E (\psi_i E \varepsilon_i) = E \psi_i E \varepsilon_i = E \psi_i.\quad (1.16)$$

□

Rozptyl y_i

Výpočet rozptylu si pro jednoduchost ukážeme na základním modelu ACD(1,1). Při dosazení do definice 3 a vztahu pro střední hodnotu (1.15) dostaneme

$$y_i = \psi_i \varepsilon_i,\quad (1.17)$$

$$\psi_i = \alpha_0 + \alpha_1 \psi_{i-1} + \beta_1 y_{i-1},\quad (1.18)$$

$$E y_i = \frac{\alpha_0}{1 - \alpha_1 - \beta_1}.\quad (1.19)$$

Rozptyl lze vyjádřit jako rozdíl středních hodnot, kde druhý člen již známe a první budeme dále upravovat:

$$\text{var } y_i = \mathbb{E} y_i^2 - (\mathbb{E} y_i)^2, \quad (1.20)$$

$$\mathbb{E} y_i^2 = \mathbb{E} (\psi_i^2 \varepsilon_i^2) = \mathbb{E} (\mathbb{E} (\psi_i^2 \varepsilon_i^2 | \mathcal{F}_{i-1})) = \mathbb{E} (\psi_i^2 \mathbb{E} \varepsilon_i^2) = \mathbb{E} \psi_i^2 \mathbb{E} \varepsilon_i^2. \quad (1.21)$$

Za předpokladu slabé stacionarity je zajištěna rovnost středních hodnot a také rovnost rozptylů pro různé časové okamžiky, tj.

$$\mathbb{E} \psi_i^2 = \mathbb{E} \psi_{i-1}^2. \quad (1.22)$$

Zbývá vyjádřit tuto střední hodnotu ψ_i^2 . Dosazením z (1.18) a roznásobením obdržíme

$$\begin{aligned} \mathbb{E} \psi_i^2 &= \mathbb{E} (\alpha_0 + \alpha_1 \psi_{i-1} + \beta_1 y_{i-1})^2 = \alpha_0^2 + \alpha_1^2 \mathbb{E} \psi_{i-1}^2 + \beta_1^2 \mathbb{E} y_{i-1}^2 + \\ &\quad + 2\alpha_0 \alpha_1 \mathbb{E} \psi_{i-1} + 2\alpha_0 \beta_1 \mathbb{E} y_{i-1} + 2\alpha_1 \beta_1 \mathbb{E} (\psi_{i-1} y_{i-1}). \end{aligned} \quad (1.23)$$

Pouze poslední vzniklá střední hodnota nebyla dosud zmíněna, proto si ji také upravíme s využitím analogického principu, jako u (1.21):

$$\mathbb{E} (\psi_{i-1} y_{i-1}) = \mathbb{E} (\psi_{i-1}^2 \varepsilon_{i-1}) = \mathbb{E} \psi_{i-1}^2. \quad (1.24)$$

Z definice 3, slabé stacionarity a vztahů (1.11), (1.21), (1.22) a (1.24) po dosazení dostaneme

$$\mathbb{E} \psi_i^2 = \alpha_0^2 + \alpha_1^2 \mathbb{E} \psi_i^2 + \beta_1^2 \mathbb{E} \psi_i^2 \mathbb{E} \varepsilon_i^2 + 2\alpha_0 \alpha_1 \mu_y + 2\alpha_0 \beta_1 \mu_y + 2\alpha_1 \beta_1 \mathbb{E} \psi_i^2, \quad (1.25)$$

$$\mathbb{E} \psi_i^2 = \frac{\alpha_0^2 + 2\alpha_0 \mu_y (\alpha_1 + \beta_1)}{1 - (\alpha_1^2 + 2\alpha_1 \beta_1 + \beta_1^2 \mathbb{E} \varepsilon_i^2)}, \quad (1.26)$$

kde $(\alpha_1^2 + 2\alpha_1 \beta_1 + \beta_1^2 \mathbb{E} \varepsilon_i^2) < 1$ je podmínka existence konečného kladného rozptylu.

Použitím spočtených členů z výrazů (1.19) a (1.26) k výpočtu rozptylu podle vztahu (1.20) získáme explicitní vyjádření

$$\text{var } y_i = \frac{\alpha_0^2 + 2\alpha_0 \mu_y (\alpha_1 + \beta_1)}{1 - (\alpha_1^2 + 2\alpha_1 \beta_1 + \beta_1^2 \mathbb{E} \varepsilon_i^2)} - \left(\frac{\alpha_0}{1 - \alpha_1 - \beta_1} \right)^2. \quad (1.27)$$

Z explicitního vyjádření rozptylu y_i je patrné, že určení postačujících podmínek pro jeho existenci a konečnost by bylo velmi komplikované.

1.4 Metoda predikce

K předpovědi ACD modelu o l kroků přistupujeme jako k očekávané hodnotě náhodné veličiny y_{h+l} , kde h předchozích kroků známe. Tento postup je popsán například v knize Analysis of Financial Time Series (Tsay, 2008). Předpověď budeme značit $y_h(l)$ a dle (1.16) platí

$$y_h(l) = \mathbb{E} (y_{h+l} | \mathcal{F}_h) = \mathbb{E} (\psi_{h+l} \varepsilon_{h+l} | \mathcal{F}_h) = \mathbb{E} (\psi_{h+l} | \mathcal{F}_h). \quad (1.28)$$

Princip výpočtu předvedeme nejprve na předpovědi o jeden a dva kroky. Chybu předpovědi o l kroků označíme $e_h(l)$. Při predikci o jeden krok je ψ_{h+1} známé, proto v čase h platí

$$y_h(1) = \mathbf{E}(\psi_{h+1}|\mathcal{F}_h) = \psi_{h+1}, \quad (1.29)$$

$$e_h(1) = y_{h+1} - y_h(1) = \psi_{h+1}(\varepsilon_{h+1} - 1). \quad (1.30)$$

Střední hodnota a rozptyl chyby $e_h(1)$ mají tvar

$$\mathbf{E}(e_h(1)|\mathcal{F}_h) = 0 \text{ analogicky jako u (1.16),} \quad (1.31)$$

$$\text{var}(e_h(1)|\mathcal{F}_h) = \mathbf{E}(\psi_{h+1}^2(\varepsilon_{h+1} - 1)^2|\mathcal{F}_h) = \psi_{h+1}^2 \mathbf{E}(\varepsilon_{h+1}^2 - 1). \quad (1.32)$$

Předpověď o dva a více kroků ukážeme pro jednoduchost opět na modelu ACD(1, 1). Po úpravě a dosazení podle (1.28) a (1.29) dostaneme

$$\begin{aligned} y_h(2) &= \mathbf{E}(\psi_{h+2}|\mathcal{F}_h) = \alpha_0 + \alpha_1 \mathbf{E}(y_{h+1}|\mathcal{F}_h) + \beta_1 \psi_{h+1} = \\ &= \alpha_0 + y_h(1)(\alpha_1 + \beta_1). \end{aligned} \quad (1.33)$$

Předpověď pro l kroků má tedy rekurentní tvar. Pomocí částečného roznásobení a součtu geometrické řady ji vyjádříme explicitně v závislosti pouze na parametrech a $y_h(1)$.

$$y_h(l) = \alpha_0 + (\alpha_1 + \beta_1)y_h(l-1), \quad (1.34)$$

$$\begin{aligned} y_h(l) &= \alpha_0 + \alpha_0(\alpha_1 + \beta_1) + \alpha_0(\alpha_1 + \beta_1)^2 + \dots \\ &\quad \dots + \alpha_0(\alpha_1 + \beta_1)^{l-1} + (\alpha_1 + \beta_1)^{l-1}y_h(1) = \\ &= \sum_{k=0}^{l-1} \alpha_0(\alpha_1 + \beta_1)^k + (\alpha_1 + \beta_1)^{l-1}y_h(1) = \\ &= \alpha_0 \frac{1 - (\alpha_1 + \beta_1)^l}{1 - (\alpha_1 + \beta_1)} + (\alpha_1 + \beta_1)^{l-1}y_h(1). \end{aligned} \quad (1.35)$$

Z definice 3 víme, že $(\alpha_1 + \beta_1) < 1$. Tato podmínka aplikovaná na poslední tvar (1.35) ukazuje, že predikce konverguje k průměru μ_y pro velký počet předpovídaných kroků ($l \rightarrow \infty$).

1.5 Typy modelu ACD

Jednotlivé typy modelu ACD se rozlišují podle rozdělení reziduální složky $\{\varepsilon_i\}$. Zde si představíme tři nejpoužívanější zástupce, kterými jsou EACD (exponenciální rozdělení), WACD (Weibullovo rozdělení) a GACD (obecné gama rozdělení).

EACD

Nejjednodušším typem je model EACD. Reziduální složka $\{\varepsilon_i\}$ má exponenciální rozdělení, viz příloha A.1.1. Podle definice 3 musí platit $\mathbf{E} \varepsilon_i = 1$, a proto

pro parametr exponenciálního rozdělení λ a rozdělení reziduí platí:

$$\lambda = 1, \quad (1.36)$$

$$f_\varepsilon(x) = \begin{cases} e^{-x}, & x > 0, \\ 0 & \text{jinak,} \end{cases} \quad (1.37)$$

$$\mathbf{E} \varepsilon_i = 1, \quad (1.38)$$

$$\text{var } \varepsilon_i = 1, \text{ pro } i \in \mathbb{N}. \quad (1.39)$$

Výhody i nevýhody tohoto modelu spočívají v jeho jednoduchosti. Nemá žádný parametr, který by bylo potřeba odhadovat, proto je vhodný pro ilustraci mnoha vlastností modelu ACD a nenáročný na výpočet. Zároveň však má omezené možnosti vystihnout realitu s takovou přesností, jaké lze dosáhnout u modelů s parametry.

Podmíněnou hustotu modelu lze určit pomocí věty o transformaci náhodné veličiny, která je uvedena například v knize Matematická statistika (Anděl, 1978). Podle (1.6) lze vyjádřit rezidua ε_i a zapsat hustotu modelu, tedy náhodné veličiny y_i , pomocí hustoty exponenciálního rozdělení ε_i

$$\varepsilon_i = \frac{y_i}{\psi_i}, \quad (1.40)$$

$$f(y_i | \mathcal{F}_{i-1}) = \begin{cases} f_\varepsilon\left(\frac{y_i}{\psi_i}\right) \left|\frac{1}{\psi_i}\right| = \frac{1}{\psi_i} e^{-\frac{y_i}{\psi_i}}, & y_i > 0, \\ 0 & \text{jinak,} \end{cases} \quad (1.41)$$

$$y_i | \mathcal{F}_{i-1} \sim \text{Exp}\left(\frac{1}{\psi_i}\right), \text{ pro } i \in \mathbb{N}. \quad (1.42)$$

WACD

Dalším používaným typem je model WACD. Reziduální složka $\{\varepsilon_i\}$ má Weibullovo rozdělení, viz příloha A.1.2. Parametr α budeme vybírat při volbě modelu. Pak pro parametr β a rozdělení reziduí podle definice 3 platí:

$$\beta = \Gamma\left(1 + \frac{1}{\alpha}\right), \quad (1.43)$$

$$f_\varepsilon(x) = \begin{cases} \alpha \left[\Gamma\left(1 + \frac{1}{\alpha}\right)\right]^\alpha x^{\alpha-1} \exp\left\{-\left[x\Gamma\left(1 + \frac{1}{\alpha}\right)\right]^\alpha\right\}, & x \geq 0, \\ 0 & \text{jinak,} \end{cases} \quad (1.44)$$

$$\mathbf{E} \varepsilon_i = 1, \quad (1.45)$$

$$\text{var } \varepsilon_i = \frac{\Gamma\left(1 + \frac{2}{\alpha}\right)}{\left[\Gamma\left(1 + \frac{1}{\alpha}\right)\right]^2} - 1, \text{ pro } i \in \mathbb{N}, \alpha > 0, \quad (1.46)$$

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt, x > 0 \text{ (gamma funkce)}. \quad (1.47)$$

Model WACD je zobecněním exponenciálního EACD, neboť při volbě parametru $\alpha = 1$ dostaneme právě exponenciální rozdělení. Výhodou modelu WACD je přítomnost parametru α . Díky tomu je možné rozdělení lépe přizpůsobit datům, se kterými pracujeme. Parametr je možné zvolit předem, pokud máme o datech

dostatek informací a očekáváme určité chování. Standardně však budeme parametr odhadovat spolu s ostatními parametry modelu odhadnout pomocí metody maximální věrohodnosti.

Podmíněné rozdělení samotného modelu určíme analogicky, jako v případě EACD, pomocí věty o transformaci náhodné veličiny (Anděl, 1978)

$$\varepsilon_i = \frac{y_i}{\psi_i}, \quad (1.48)$$

$$\begin{aligned} f(y_i|\mathcal{F}_{i-1}) &= f_\varepsilon\left(\frac{y_i}{\psi_i}\right) \left|\frac{1}{\psi_i}\right| = \\ &= \alpha \left[\Gamma\left(1 + \frac{1}{\alpha}\right)\right]^\alpha \left(\frac{y_i}{\psi_i}\right)^{\alpha-1} \exp\left\{-\left[\frac{y_i}{\psi_i}\Gamma\left(1 + \frac{1}{\alpha}\right)\right]^\alpha\right\} \frac{1}{\psi_i} = \\ &= \begin{cases} \alpha \left[\frac{\Gamma(1+\frac{1}{\alpha})}{\psi_i}\right]^\alpha y_i^{\alpha-1} \exp\left\{-\left[y_i\frac{\Gamma(1+\frac{1}{\alpha})}{\psi_i}\right]^\alpha\right\}, & y_i \geq 0, \\ 0 & \text{jinak,} \end{cases} \end{aligned} \quad (1.49)$$

$$y_i|\mathcal{F}_{i-1} \sim \text{Weib}\left(\alpha, \frac{\Gamma\left(1 + \frac{1}{\alpha}\right)}{\psi_i}\right), \text{ pro } i \in \mathbb{N}, \alpha > 0. \quad (1.50)$$

GACD

Posledním zde představeným typem je model GACD. Reziduální složka $\{\varepsilon_i\}$ má obecné gama rozdělení se třemi parametry, což dělá z tohoto typu nejobecnější model uvedený v této práci. Parametrizace je opět volena s ohledem na snadnost výpočtů, viz příloha A.1.3. Parametry α a κ volíme při výběru modelu. Vhodnou volbou parametru β splníme podmínku na střední hodnotu v definici 3:

$$\beta = \frac{\Gamma\left(\kappa + \frac{1}{\alpha}\right)}{\Gamma(\kappa)}, \quad (1.51)$$

$$f_\varepsilon(x) = \begin{cases} \frac{\alpha[\Gamma(\kappa+\frac{1}{\alpha})]^\kappa x^{\kappa\alpha-1}}{[\Gamma(\kappa)]^{\kappa\alpha+1}} \exp\left\{-\left[\frac{\Gamma(\kappa+\frac{1}{\alpha})}{\Gamma(\kappa)}x\right]^\alpha\right\}, & x \geq 0, \\ 0 & \text{jinak,} \end{cases} \quad (1.52)$$

$$\mathbb{E} \varepsilon_i = 1, \quad (1.53)$$

$$\text{var } \varepsilon_i = \frac{\Gamma(\kappa)\Gamma\left(\kappa + \frac{2}{\alpha}\right)}{[\Gamma(\kappa + \frac{1}{\alpha})]^2} - 1, \quad (1.54)$$

$$y_i|\mathcal{F}_{i-1} \sim \Gamma\left(\alpha, \frac{\Gamma\left(\kappa + \frac{1}{\alpha}\right)}{\Gamma(\kappa)}, \kappa\right), \text{ pro } i \in \mathbb{N}, \alpha > 0, \kappa > 0. \quad (1.55)$$

Podmíněná hustota tohoto modelu se určí analogicky jako v případě EACD a WACD, viz (1.41) a (1.49). Pro volbu parametru $\kappa = 1$ dostáváme model WACD, jde tedy o další zobecnění. Z toho také plyne výhoda tohoto modelu. Pokud se pro popis dat hodí WACD nebo EACD dostatečně dobře, lze tyto modely dostat pouze volbou parametrů. Parametry rozdělení lze i zde volit deterministicky předem nebo spolu s ostatními parametry modelu pomocí metody maximální věrohodnosti.

2. Konstrukce modelu ACD

Při konstrukci modelu $ACD(r, s)$ se budeme držet postupu popsaného v knize (Cipra, 2008). Je zde předveden na modelu ARMA, ze kterého je model ACD odvozen. Samotná konstrukce probíhá ve třech krocích:

1. Identifikace modelu: Stanovení řádu r , s a rozdělení náhodné složky ε .
2. Odhad modelu: Odhad parametrů $\boldsymbol{\alpha} = \{\alpha_0, \dots, \alpha_r\}$, $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_s\}$ zvoleného modelu a vektor parametrů $\boldsymbol{\lambda}$ zvoleného pravděpodobnostního rozdělení.
3. Diagnostika modelu: Ověření vlastností zkonstruovaného modelu.

Pokud výsledky diagnostiky nedopadnou dle našeho očekávání, je potřeba tento postup iterovat. Výsledky diagnostiky mohou napovědět, jaké úpravy model vyžaduje.

2.1 Identifikace modelu

Identifikací modelu určíme jeho řád, tedy hodnotu parametrů r a s v případě modelu $ACD(r, s)$, a zvolíme rozdělení reziduí $\{\varepsilon_i\}$. Odhad parametrů provedeme pomocí informačního kritéria AIC (Akaike information criterion) popsaného například v (Cipra, 2008)

$$AIC(k, l) = -2\ell_{k,l}(\mathbf{y}_n|\boldsymbol{\theta}) + 2(k + l + 1), \quad (2.1)$$

kde $\ell_{k,l}(\mathbf{y}_n|\boldsymbol{\theta})$ je hodnota logaritmické věrohodnostní funkce modelu $ACD(k, l)$, n je počet pozorování a $\boldsymbol{\theta}$ je vektor všech parametrů odhadovaných metodou maximální věrohodnosti. Minimalizací kritéria přes různá k , l získáme odhad pro parametry r , s respektive. Informační kritérium penalizuje velké množství parametrů, v rámci tohoto testu se tak lze omezit například na $k, l \in \{1, \dots, 5\}$.

V různých zdrojích se tvar tohoto kritéria může drobně lišit. Například ve zmíněném (Cipra, 2008) je navíc vydělený počtem pozorování n , což je konstanta a na výsledek testu nemá vliv.

Pro identifikaci modelu existují také další postupy. Kromě volby jiného informačního kritéria to může být například identifikace pomocí tvaru autokorelační a parciální autokorelační funkce. Tato metoda však neumožňuje postup automatizovat a nevyhnutelně dochází k subjektivním zásahům analytika.

2.2 Odhad modelu

Správné fungování modelu závisí na vhodné volbě parametrů. Je potřeba určit $\boldsymbol{\alpha} = \{\alpha_0, \dots, \alpha_r\}$, $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_s\}$ a také vektor $\boldsymbol{\lambda}$ parametrů charakterizující rozdělení náhodných veličin ε_i (viz definice 3). Použijeme upravenou metodu maximální věrohodnosti uvedenou v (Engle a Russell, 1998). Tato metoda se nazývá QMLE (quasi maximum likelihood estimation method). Liší se v tom, že nám stačí aproximace věrohodnostní funkce, neboť její přesný tvar nemusí být možné v reálném případě získat.

Mějme $\mathbf{y}_n = \{y_1, y_2, \dots, y_n\}$ vektor dat, $\boldsymbol{\theta}$ vektor neznámých parametrů a $g = \max(r, s)$, pak věrohodnostní funkce modelu ACD(r, s) má tvar

$$\mathcal{L}(\mathbf{y}_n|\boldsymbol{\theta}) = f(y_g|\boldsymbol{\theta}) \prod_{i=g+1}^n f(y_i|\mathcal{F}_{i-1}, \boldsymbol{\theta}). \quad (2.2)$$

Hustotu y_g může být obtížné získat a také s rostoucím počtem pozorování klesá význam tohoto členu. Metoda QMLE nám dovoluje tento člen vynechat a tím získat aproximaci věrohodnostní funkce. Podle (Engle a Russell, 1998) je tento odhad za určitých podmínek regularity konzistentní a asymptoticky normální, avšak není eficientní, pokud ε_i nemá exponenciální rozdělení.

Logaritmická kvazi-věrohodnostní funkce má tvar

$$\ell(\mathbf{y}_n|\boldsymbol{\theta}) = \sum_{i=g+1}^n f(y_i|\mathcal{F}_{i-1}, \boldsymbol{\theta}). \quad (2.3)$$

Připomeňme si, že $y_i = \psi_i \varepsilon_i$. To nám dává možnost transformovat náhodnou veličinu y_i a vyjádřit její hustotu pomocí hustoty ε_i , pokud ji známe, viz kapitola 1.5. Na základě (2.3) je pak již možné provést odhad jednotlivých parametrů.

2.3 Diagnostika modelu

Model je odhadnut a nyní je potřeba ověřit, zda vyhovuje předpokladům. Z metody maximální věrohodnosti jsme získali odhad $\hat{\psi}_i$. Vypočteme odhady reziduí $\{\varepsilon_i\}$ pomocí definice 3:

$$\hat{\varepsilon}_i = \frac{y_i}{\hat{\psi}_i}. \quad (2.4)$$

U takto získaných odhadů reziduí se kontroluje jejich nekorelovanost, a také zda mají požadované rozdělení, které jsme pro rezidua zvolili. Shodnost rozdělení lze nahlédnout z grafu QQ-plot, kde jsou vykresleny teoretické kvantily předpokládaného rozdělení a empirické kvantily pozorovaných reziduí. Podle míry překrytí lze odhadnout, zda se teoretické a empirické rozdělení shodují.

Nekorelovanost budeme posuzovat prohlédnutím grafu autokorelační funkce. Díky stacionaritě procesů závisí korelace mezi y_k a $y_{k+\tau}$ pouze na prodlevě τ .

Definice 4 (Autokorelační funkce). *Mějme stacionární proces $\{y_k, k \in \mathbb{N}_0\}$ se střední hodnotou $E(y_k) = \mu$ a konečným rozptylem $\text{var}(y_k) = \sigma^2$. Pak funkce definovaná předpisem*

$$ACF(\tau) = \frac{\text{Cov}(y_k, y_{k+\tau})}{\sigma^2} = \frac{1}{\sigma^2} E[(y_k - \mu)(y_{k+\tau} - \mu)], \tau \in \mathbb{N}_0 \quad (2.5)$$

se nazývá autokorelační funkce procesu $\{y_k\}$.

Uvedli jsme si teoretický tvar autokorelační funkce. Při diagnostice však potřebujeme odhadnout její hodnoty ze získaných dat. Využijeme k tomu zápis autokorelační funkce ve tvaru střední hodnoty. Výběrový průměr je nestranný a konzistentní odhad střední hodnoty. Pokud známe střední hodnotu a rozptyl

náhodného procesu, pak nestranný a konzistentní odhad autokorelační funkce má tvar

$$\widehat{ACF}(\tau) = \frac{1}{\sigma^2(n-\tau)} \sum_{k=1}^{n-\tau} (Y_k - \mu)(Y_{k+\tau} - \mu), \tau = 1, 2, \dots, n-1. \quad (2.6)$$

Pokud neznáme střední hodnotu a rozptyl, je možné je nahradit výběrovým průměrem a výběrovým rozptylem. Odhad se pak stává pouze asymptoticky nestranným, což v praxi při dostatečném množství pozorování nevádí. Pro praktický význam je vhodné, aby $n > 50$ a $\tau < n/4$.

V praxi se často ve jmenovateli namísto $(n-\tau)$ nachází pouze n . Takto upravený odhad má sice větší vychýlení, tedy zhoršenou nestrannost, ale dosahuje menší čtvercové chyby odhadu. Navíc je pro velká n také asymptoticky nestranný.

Autokorelační funkci lze chápat jako testovou statistiku. Testujeme hypotézu, že data jsou nekorelovaná. Pro hladinu spolehlivosti α se používá interval spolehlivosti

$$|\widehat{ACF}(\tau)| < \frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{n}}, \tau = 1, 2, \dots, n-1, \quad (2.7)$$

kde Φ^{-1} značí kvantilovou funkci normovaného normálního rozdělení, viz (Cipra, 2008).

Je třeba vhodně zvolit rozsah τ , pro které graf autokorelační funkce spočítat. Software nám automaticky přidá do grafu také vyznačené hranice pro 95% interval spolehlivosti. Podle dodržení této hranice pak posuzujeme nekorelovanost časové řady. Při tom je třeba mít na paměti, že 95% interval spolehlivosti dává prostor pro 5% hodnot mimo interval spolehlivosti i za platnosti hypotézy.

3. Aplikace na reálná data

V této části předvedeme aplikování výše zmíněných modelů na reálná data. Pro tento účel jsme zvolili ticková data tří různých podkladových aktiv s vysokou frekvencí obchodování¹. Jsou jimi akcie společnosti Apple Inc., eurodolar (EUR/USD) a zlato. Všechna data pocházejí z CFD (contract for difference) trhu a zobrazují každou jednotlivou změnu ceny, ne však úplně každou jednotlivou transakci podkladového aktiva. Úplná transakční ticková data bohužel nejsou volně dostupná. Díky vysoké frekvenci obchodů však tato aproximace nemá vliv na fungování modelů.

3.1 Informace o použitých datech

V následující tabulce 3.1 jsou shrnuty základní informace o již zmíněných třech datových sadách, se kterými budeme v celé této kapitole pracovat. Každý datový balíček obsahuje vektor durací za jeden týden a časů, kdy k nim došlo. Všechna data jsou z týdne 8. – 12. dubna 2019. Obchoduje se pouze v pracovní dny. Všechny uvedené časy jsou vzhledem k našemu času v České republice.

Aktivum	Obchodní hodiny	Počet pozorování	Min	Průměr	Max
Apple Inc.	15:30 – 22:00	318 383	0.050	0.3675	11.786
EUR/USD	00:00 – 23:59	659 130	0.014	0.6499	96.451
Zlato	00:00 – 23:59	459 532	0.050	0.9000	139.122

Tabulka 3.1: Informace o použitých datech.

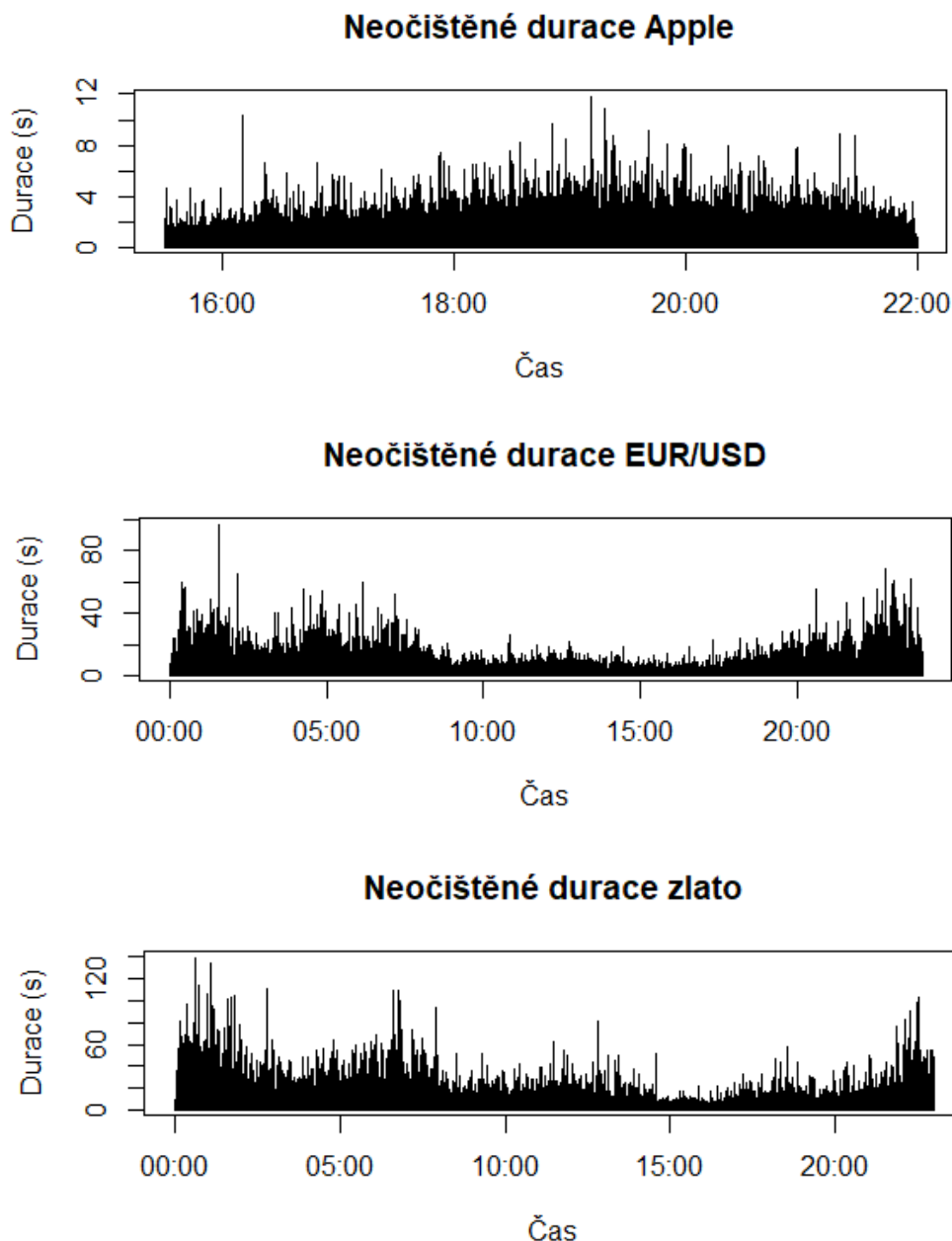
Durace získané z tickových dat pravděpodobně nebudou splňovat podmínky pro použití v modelu ACD kvůli denním periodicitám a trendům. Je nutné je alespoň graficky zkontrolovat a o tyto vlivy očistit, viz kapitola 1.2.

Na obrázku 3.1 jsou znázorněny neočištěné durace všech tří podkladových aktiv. U akcií Applu je patrná větší volatilita po otevření trhu a ke konci obchodních hodin. Také zde nejsou tak velké extrémy a obvyklé hodnoty durace jsou několikanásobně menší, než v případě zbylých dvou aktiv.

Pár eura s dolarem je nejvíce obchodovaný během dopoledních a odpoledních hodin našeho časového pásma. Zejména pak aktivita vzroste po deváté hodině ránní, kdy otevírá burza v Londýně, a po druhé hodině odpolední, kdy otevírá burza v New Yorku. Naopak v době, kdy jsou otevřeny burzy v Tokiu a Sydney, jsou prodlevy mezi transakcemi o poznání větší.

Zlato má podobné tendence jako eurodolar, avšak při přesunu obchodování na burzy do Tokia a Sydney během našich nočních hodin neklesá volatilita tak výrazně.

¹Zdroj dat viz Ticková data



Obrázek 3.1: Neočištěné durace.

3.2 Očištění dat

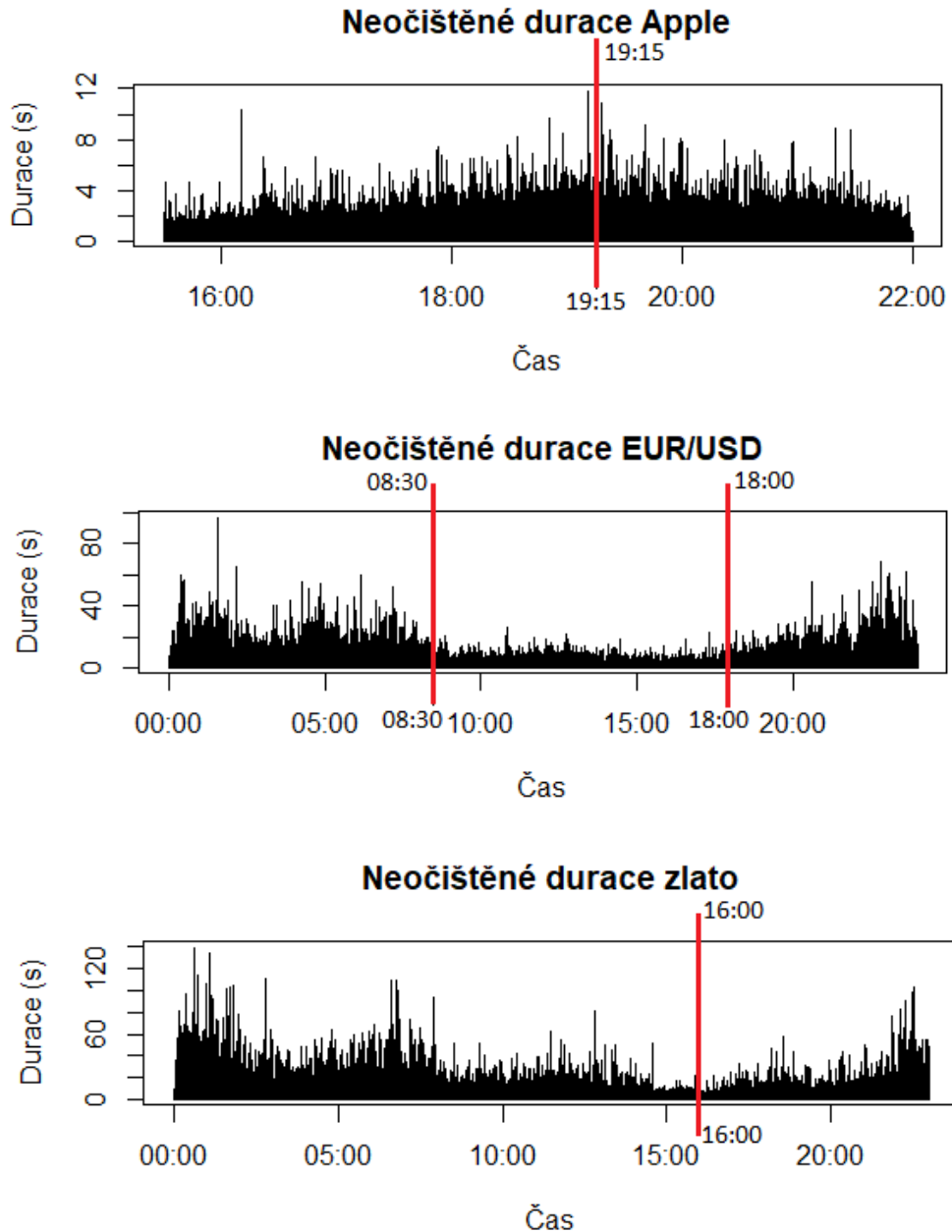
Očištěná data dostaneme ze vztahu (1.5). Funkce $f(t_i)$ představuje trendovou složku, které se chceme zbavit. Pro jednoduchost zvolíme lineární semiparametrický přístup k odhadnutí $f(t_i)$. Každý graf rozdělíme podle potřeby na dva až tři podintervaly, na kterých je pozorovatelný jednoduchý trend, který následně proložíme přímkou. Složitější a také efektivnější metody jsou uvedeny například v knize *Analysis of Financial Time Series* (Tsay, 2002).

Pro jednotlivé grafy subjektivně zvolíme rozdělení na podintervaly uvedené v následující tabulce 3.2.

Aktivum	1. interval	2. interval	3. interval
Apple	15:30 – 19:15	19:15 – 22:00	
EUR/USD	00:00 – 08:30	08:30 – 18:00	18:00 – 23:59
Zlato	00:00 – 16:00	16:00 – 23:59	

Tabulka 3.2: Intervaly k prokládání přímkou pro jednotlivá aktiva.

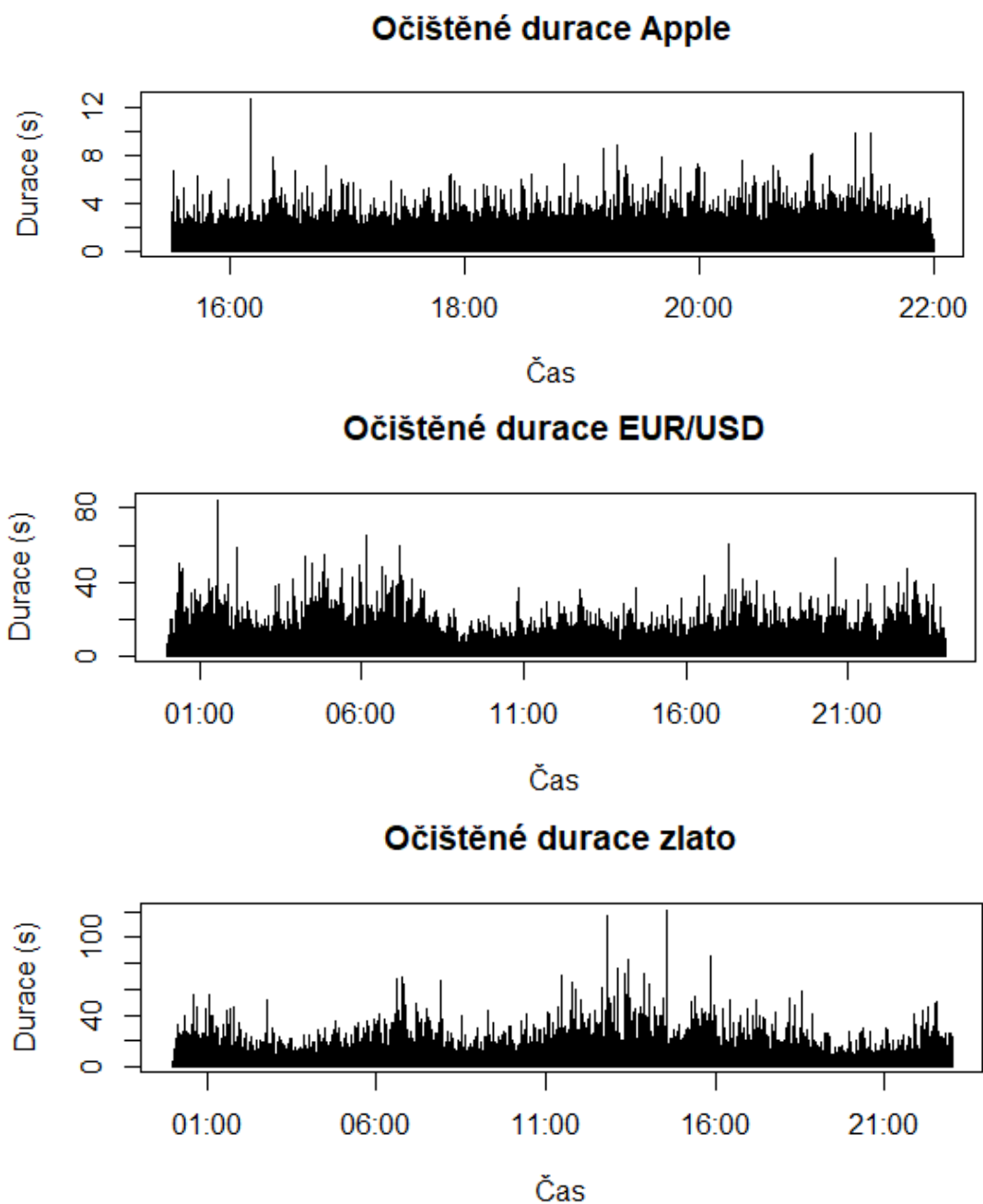
Na obrázku 3.2 je rozdělení jednotlivých grafů podle změny trendu vyznačeno červenými svislými čarami.



Obrázek 3.2: Vyznačené intervaly pro proložení přímkou.

Na obrázku 3.3 jsou zobrazeny již očištěné grafy. Nejlepší výsledek je pozorovatelný u durací Applu, které se držely hlavního trendu. Naopak u durací zlata

je vidět, že hlavní trend je složen z menších trendů na menších časových úsecích, proto zde jednoduché proložení přímkou odstraňující pouze hlavní trend není tolik efektivní. Také graf EUR/USD vykazuje složitější trendovost při bližším pohledu.



Obrázek 3.3: Očištěné durace.

3.3 Odhad modelu

Na takto očištěných datech si ukážeme odhad základních modelů EACD(1,1), WACD(1,1) a GACD(1,1). Postupně pro jednotlivé sady dat odhadneme každý z modelů a srovnáme jejich vlastnosti.

Pro názorné předvedení zde volíme $r = 1$ a $s = 1$ hlavně z toho důvodu, že modely v tomto řádu jsou výpočetně stabilní. Pracujeme s velkým objemem dat a očištění o deterministickou složku proběhlo základní metodou odstraňující pouze hlavní trend. Je velmi pravděpodobné, že data i nadále deterministickou složku obsahují. S rostoucími řády modelu roste i náchylnost na nesplněné podmínky stacionarity. Tyto modely pak vyžadují komplikovanější a efektivnější metody očištění. V opačném případě dochází k odhadnutí parametrů, které nevyhovují definici stacionárního modelu ACD. Aplikace pokročilejších metod očištění časových řad durací o deterministické složky by překračovala rámec stanovený pro tuto práci.

3.3.1 EACD a WACD

V tabulkách 3.3 a 3.4 jsou parametry odhadnuté metodou maximální věrohodnosti pro každý z datových balíčků. Sloupečky S.E. označují standardní chybu odhadu.

	Apple		EUR/USD		Zlato	
	Hodnota	S.E.	Hodnota	S.E.	Hodnota	S.E.
α_0	0.00595	0.000426	0.00252	0.000087	0.00396	0.000109
α_1	0.03259	0.001207	0.02588	0.000355	0.04277	0.000472
β_1	0.95128	0.002269	0.97152	0.000413	0.95418	0.000511

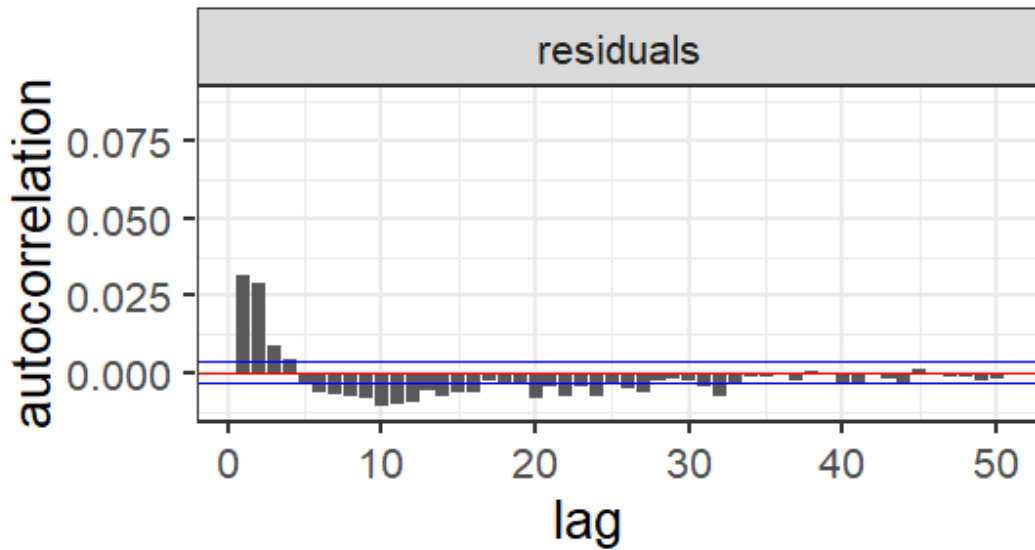
Tabulka 3.3: Odhadnuté parametry modelu EACD.

	Apple		EUR/USD		Zlato	
	Hodnota	S.E.	Hodnota	S.E.	Hodnota	S.E.
α_0	0.00613	0.000414	0.00355	0.000123	0.00514	0.000170
α_1	0.03427	0.001169	0.02502	0.000415	0.03723	0.000617
β_1	0.94935	0.002196	0.97022	0.000527	0.95565	0.000742
γ	1.07626	0.001387	0.83746	0.000725	0.72296	0.000784

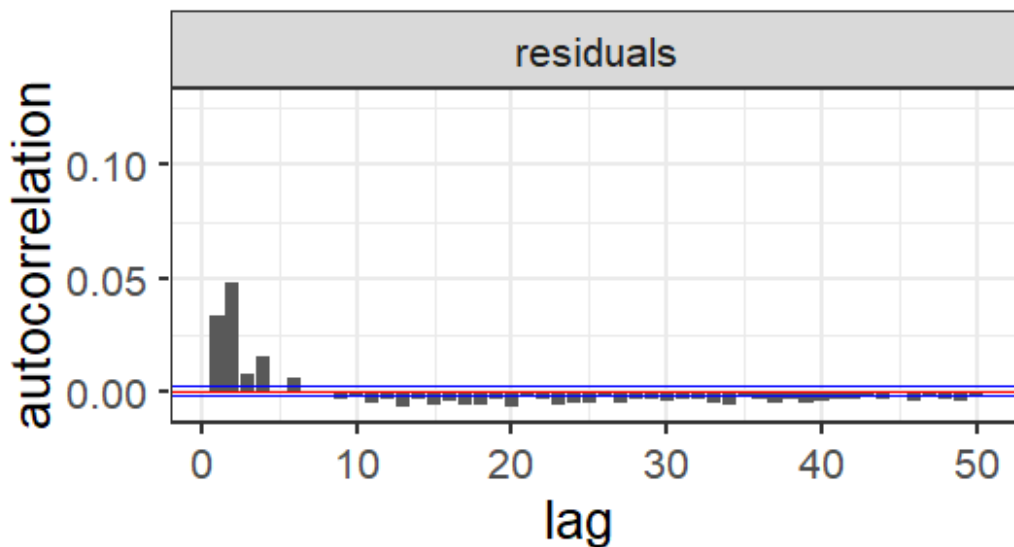
Tabulka 3.4: Odhadnuté parametry modelu WACD.

Parametry vyhovují definici ACD modelu, můžeme tedy přejít k vlastnostem odhadu. Hodnota parametru Weibullova rozdělení γ je u všech tří souborů dat blízká jedničce. Z toho plyne, že přechod od EACD k obecnějšímu WACD pravděpodobně nenabízí významné zlepšení. Nejméně se EACD a WACD liší pro data Apple. Nejvýznamnější rozdíl je pak u zlata.

Následující grafy ukazují hodnotu autokorelační funkce (ACF) reziduí se zpožděním 1 až 50. Pro velkou podobnost modelů EACD a WACD u aplikace na Apple a EUR/USD byla vykreslena funkce pouze pro model WACD, neboť mezi nimi není pozorovatelný rozdíl. U zlata jsou pro názornost vykresleny obě funkce, protože právě pro zlato se tyto modely liší nejvíce. I přes to je rozdíl těžko pozorovatelný (Obrázek 3.6).



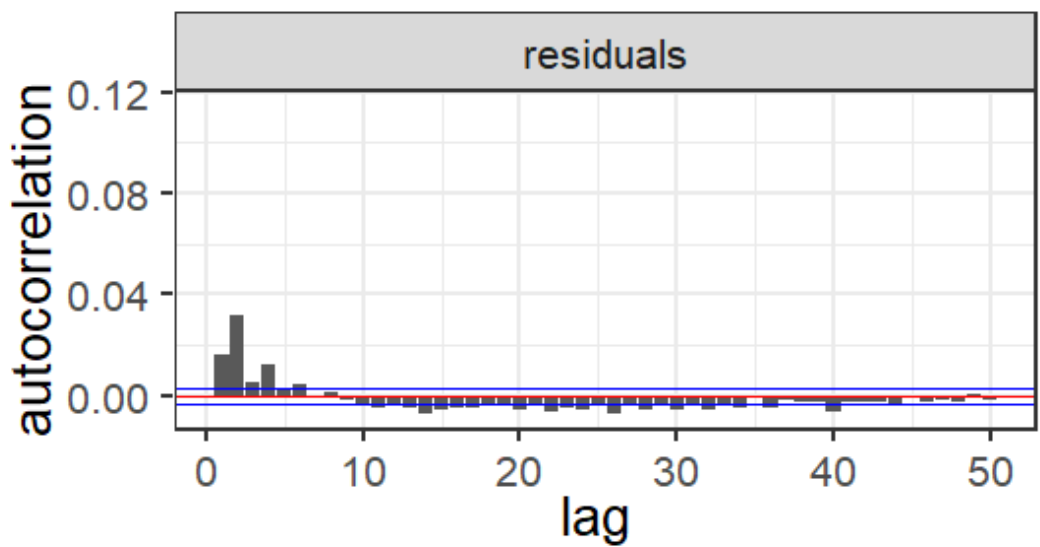
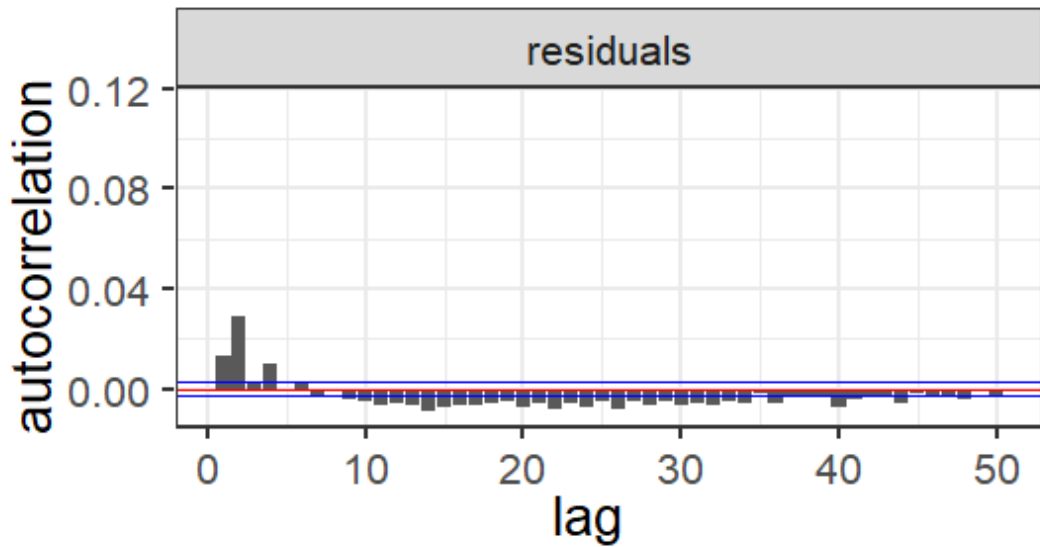
Obrázek 3.4: ACF modelu EACD a WACD aplikovaných na Apple.



Obrázek 3.5: ACF modelu EACD a WACD aplikovaných na EUR/USD.

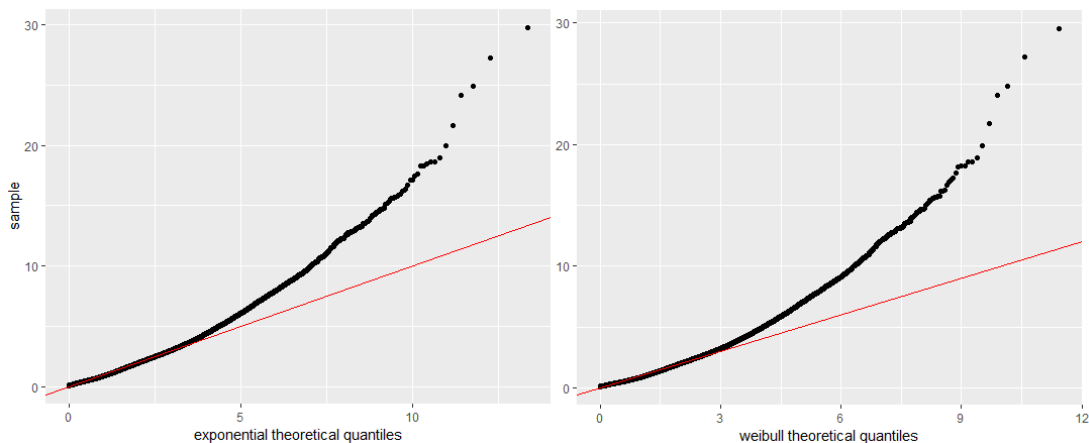
Autokorelační funkce dosahuje relativně nízkých hodnot. Pro lepší zobrazení detailu je vynechána hodnota v nule, která je vždy rovna jedné. Šířka 95 procentního intervalu spolehlivosti je nepřímě závislá na počtu pozorování, tedy čím více pozorování máme k dispozici, tím je odhad přesnější a interval spolehlivosti užší. Proto i takto relativně nízké hodnoty přesahují interval spolehlivosti a ukazují velmi pravděpodobně na nedokonalosti v odstraňování deterministické složky dat.

Rezidua všech zkoumaných modelů vykazují pozitivní korelaci při malém zpoždění o jeden až dva kroky. Rezidua modelů aplikovaných na data Apple navíc vykazují také negativní korelaci při zpoždění okolo deseti kroků. Nejméně korelovaná jsou rezidua modelu WACD aplikovaného na zlato.

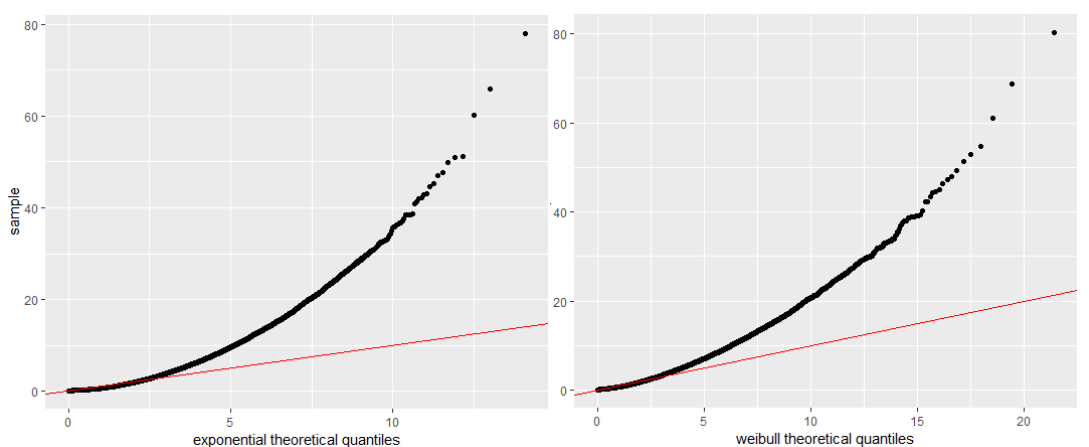


Obrázek 3.6: ACF srovnání modelu EACD a WACD aplikovaných na zlato.

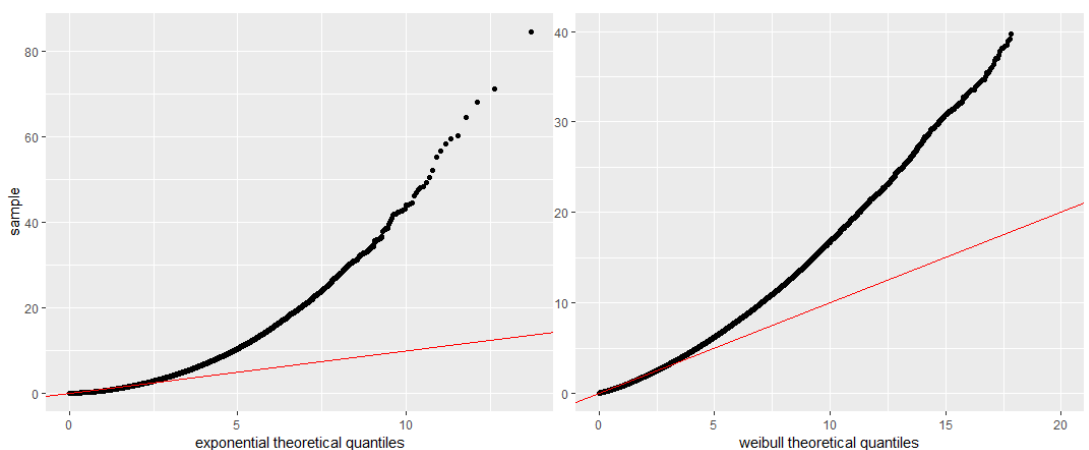
Dalším důležitou vlastností dobrého modelu je správné rozdělení reziduí. Pozorované hodnoty reziduí by se měli co nejlépe shodovat s teoretickým rozdělení daného modelu. To ověříme pomocí kvantilového diagramu. Přímka představuje teoretické hodnoty pro dané rozdělení a body představují pozorované hodnoty. Pokud body sledují přímku, rozdělení se shodují. Jestliže body tvoří přímku, ale ta se nepřekrývá s teoretickou přímku, pak stačí lineární transformace k tomu, aby se teoretické a pozorované rozdělení shodovaly. Pokud body tvoří jinou křivku, rozdělení se neshodují.



Obrázek 3.7: Q-Q diagram modelu EACD a WACD aplikovaných na Apple.



Obrázek 3.8: Q-Q diagram modelu EACD a WACD aplikovaných na EUR/USD.



Obrázek 3.9: Q-Q diagram modelu EACD a WACD aplikovaných na zlato.

Body na kvantilových diagramech se od přímky výrazně odchyľují u všech modelů. Lze tedy říct, že pozorované rozdělení reziduí neodpovídá teoretickému předpokladu. Použití modelu EACD tedy není pro tato data vhodné a rozšíření na model WACD není dostatečné a nepřináší významné zlepšení. Opět je největší rozdíl pozorovatelný u zlata, pro které se parametr γ rozdělení modelu WACD

lišil od jedničky nejvíce.

Ukázalo se, že pro takováto reálná data v takovémto objemu je model EACD příliš jednoduchý. Ani jeho zobecnění na model WACD a přidání parametru ovlivňujícího vlastnosti rozdělení reziduí nepřidává dostatečnou flexibilitu.

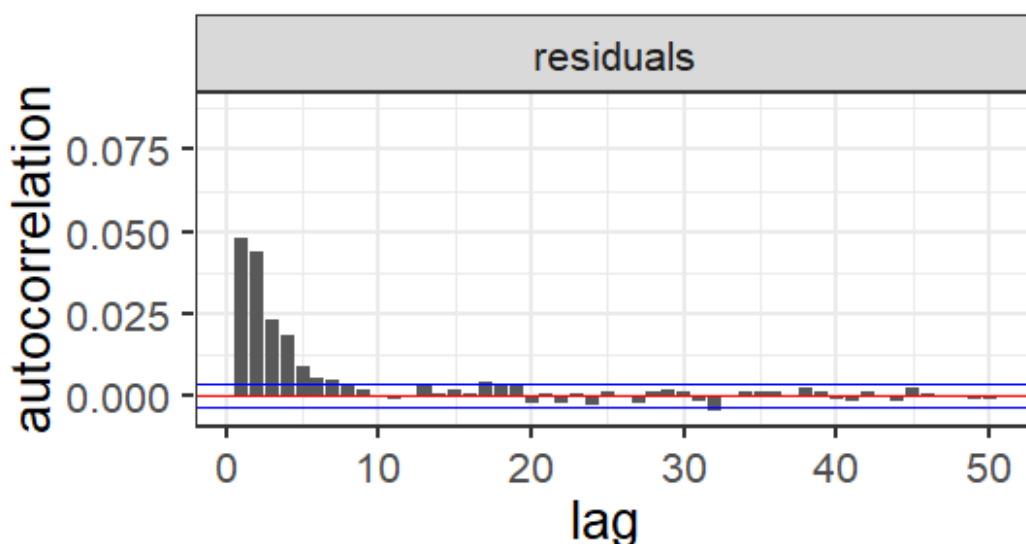
3.3.2 GACD

Další zobecnění na model GACD s sebou přináší navíc parametr κ . Jak můžete vidět v tabulce 3.5, odhad tohoto parametru se od jedničky liší velmi výrazně. Lze tedy předpokládat, že dojde k výraznější změně oproti EACD, než při zobecnění na WACD.

	Apple		EUR/USD		Zlato	
	Hodnota	S.E.	Hodnota	S.E.	Hodnota	S.E.
α_0	0.00459	0.000063	0.0108	0.0005745	0.00759	0.000220
α_1	0.01617	0.000126	0.0170	0.0007041	0.01943	0.001545
β_1	0.97092	0.000221	0.9662	0.0015324	0.97019	0.001762
γ	0.03344	0.000371	0.0705	0.0000865	0.06757	0.000108
κ	1027.9	16.2	162.8	0.01275	113.6	0.314

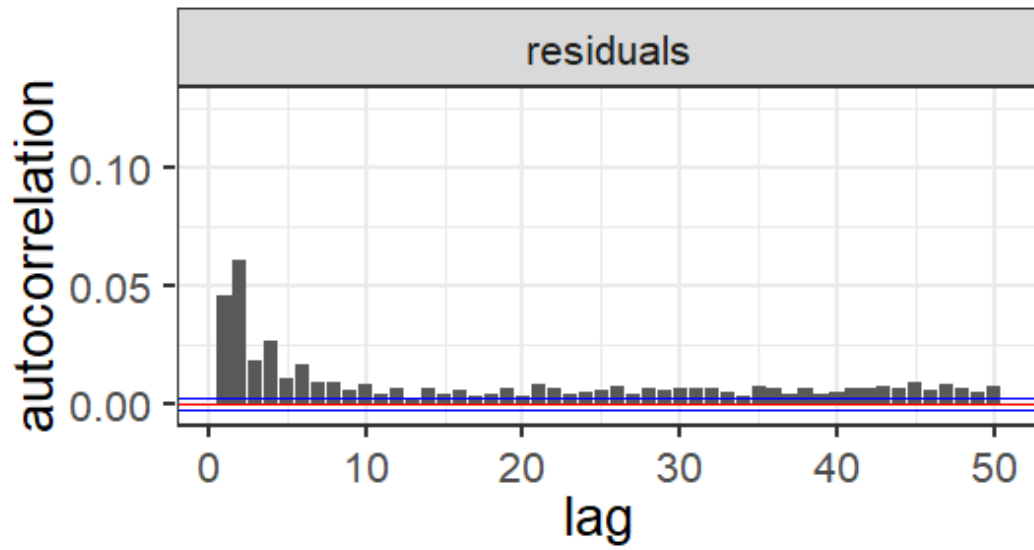
Tabulka 3.5: Odhadnuté parametry modelu GACD.

Odhadnuté parametry vyhovují definici stacionárního ACD modelu. Můžeme tedy přistoupit k vlastnostem modelu, které budeme opět zkoumat pomocí autokorelační funkce a kvantilového diagramu, to vše aplikované na získaná rezidua.

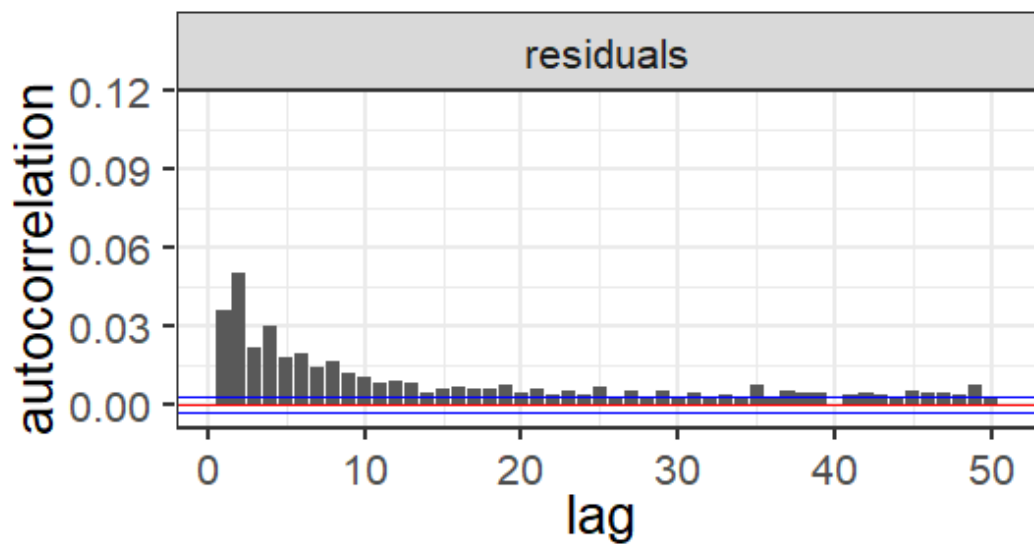


Obrázek 3.10: ACF modelu GACD aplikovaného na Apple.

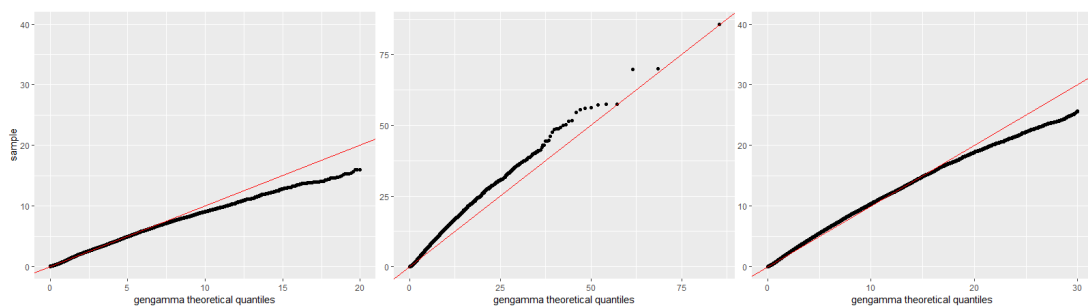
Rezidua opět vykazují kladnou korelaci pro malé hodnoty zpoždění autokorelační funkce. Tato korelace je výraznější, než v případě modelů EACD a WACD. Rezidua modelu aplikovaného na Apple se chovají pro větší hodnoty zpoždění nekorelovaně. Pro zbylé dva soubory dat je patrná pozitivní korelace i pro větší zpoždění autokorelační funkce.



Obrázek 3.11: ACF modelu GACD aplikovaného na EUR/USD.



Obrázek 3.12: ACF modelu GACD aplikovaného na zlato.



Obrázek 3.13: Q-Q diagram modelu GACD aplikovaných na Apple, EUR/USD a zlato (zleva).

Na kvantilových diagramech je vidět výrazné zlepšení oproti modelům EACD a WACD. Přidaný parametr zajistil, že se body výrazně přiblížily tvaru teoretické přímky. Pro vyšší hodnoty dochází k mírnému odchýlení, avšak při srovnání s kvantilovými diagramech modelů EACD a WACD je vylepšení jasně patrné. Rozdělení reziduí modelu aplikovaného na EUR/USD se odchyluje od přímky nejvýrazněji, avšak dává nejlepší shodu pro extrémní hodnoty. Je tedy možné, že by se dalo dosáhnout ještě trochu lepšího výsledku pouhou lineární transformací.

Na základě grafické analýzy reziduí lze rozhodnout, že model GACD je z testovaných modelů pro tyto sady dat nejvhodnější. Zároveň je však nutné připomenout, že korelovanost reziduí je výrazná a narušuje stacionaritu.

Závěr

V této práci jsme se zabývali modelem ACD (autoregressive conditional duration), který se používá k modelování durací u časových řad. Nejprve jsme představili pojmy durace a časová řada intuitivně. Následně jsme vše zavedli formálně, včetně modelu ACD a jeho základních typů podle pravděpodobnostního rozdělení reziduální složky. Také byly diskutovány základní vlastnosti uvedených modelů a byla nastíněna možnost použití těchto modelů k predikcím.

V další části byly popsány jednotlivé kroky, jak se ACD model identifikuje, odhaduje a diagnostikuje.

Poslední část je věnována praktické ukázce odhadu základních modelů EACD, WACD a GACD na různé datové soubory s velkým množstvím pozorování. K dispozici jsou transakční data akcií Apple, kurzu EUR/USD a kurz zlata. Objem dat v každém ze tří souborů se pohybuje mezi 300 tisíc až 600 tisíc pozorování (jeden obchodní týden).

Použili jsme pouze základní metodu očištění dat od hlavní trendové složky a odhadli jsme parametry uvedených modelů pomocí metody maximální věrohodnosti. Ukázalo se, že data i nadále vykazovala mírnou trendovost, což narušovalo podmínky stacionarity. Zřejmě z tohoto důvodu autokorelační funkce reziduí ukázaly na mírnou korelovanost.

Kvantilový diagram ukázal, že pro tyto sady dat nebylo vhodné volit model EACD, neboť rozdělení reziduí odhadnutého modelu se od exponenciálního rozdělení výrazně lišilo. Také se ukázalo, že rozšíření na model WACD a přidání parametru nepřineslo významné zlepšení. Autokorelační funkce a kvantilové diagramy vypadaly téměř identicky a přidání odhadnutý parametr Weibullova rozdělení se velmi blížil hodnotě jedna, což odpovídá exponenciálnímu rozdělení, a tedy modelu EACD.

Model GACD díky dalšímu parametru dosáhl výrazně lepších výsledků. Autokorelační funkce sice vyšla lépe pouze u durací Apple a v ostatních případech vykazoval model o něco silnější korelovanost reziduí, než předchozí EACD a WACD, ale kvantilový diagram ukázal, že rozdělení reziduí se téměř shoduje s předpokládaným teoretickým rozdělením, a tedy model splňuje předpoklady. Odhady parametrů vybraných modelů vyhovovaly podmínkám uvedeným v definici 3, které zaručují existenci a konečnost střední hodnoty.

Seznam použité literatury

ANDĚL, J. (1978). *Matematická statistika*. SNTL, Praha.

CIPRA, T. (2008). *Finanční ekonometrie*. Ekopress, Praha.

ENGLE, R. a RUSSELL, J. R. (1998). Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica*, **66**(5), 1127–1162. URL <https://EconPapers.repec.org/RePEc:ecm:emetrp:v:66:y:1998:i:5:p:1127-1162>.

TICKOVÁ DATA. <https://www.dukascopy.com/swiss/english/marketwatch/historical/>. Staženo: 5.5.2019.

TSAY, R. (2002). *Analysis of Financial Time Series*. Wiley, New York. URL <https://books.google.cz/books?id=kE-RjgEACAAJ>.

TSAY, R. (2008). Lecture 15: Autoregressive Conditional Duration Models. URL <http://faculty.chicagobooth.edu/ruey.tsay/teaching/uts/lec15-08.pdf>. Staženo: 20.2.2019.

Seznam obrázků

3.1	Neočištěné durace.	16
3.2	Vyznačené intervaly pro proložení přímkou.	17
3.3	Očištěné durace.	18
3.4	ACF modelu EACD a WACD aplikovaných na Apple.	20
3.5	ACF modelu EACD a WACD aplikovaných na EUR/USD.	20
3.6	ACF srovnání modelu EACD a WACD aplikovaných na zlato.	21
3.7	Q-Q diagram modelu EACD a WACD aplikovaných na Apple.	22
3.8	Q-Q diagram modelu EACD a WACD aplikovaných na EUR/USD.	22
3.9	Q-Q diagram modelu EACD a WACD aplikovaných na zlato.	22
3.10	ACF modelu GACD aplikovaného na Apple.	23
3.11	ACF modelu GACD aplikovaného na EUR/USD.	24
3.12	ACF modelu GACD aplikovaného na zlato.	24
3.13	Q-Q diagram modelu GACD aplikovaných na Apple, EUR/USD a zlato (zleva).	24

Seznam tabulek

3.1	Informace o použitých datech.	15
3.2	Intervaly k prokládání přímkou pro jednotlivá aktiva.	17
3.3	Odhadnuté parametry modelu EACD.	19
3.4	Odhadnuté parametry modelu WACD.	19
3.5	Odhadnuté parametry modelu GACD.	23

A. Přílohy

A.1 Rozdělení pravděpodobnosti

A.1.1 Exponenciální rozdělení

Značení:	$X \sim \text{Exp}(\lambda)$
Parametry:	$\lambda > 0$
Nosič:	$X \in (0, \infty)$
Hustota:	$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$
Distribuční funkce:	$F(x) = \begin{cases} 1 - e^{-\lambda x} & x > 0 \\ 0 & x \leq 0 \end{cases}$
Střední hodnota:	$E X = \frac{1}{\lambda}$
Rozptyl:	$\text{var } X = \frac{1}{\lambda^2}$

A.1.2 Weibullovo rozdělení

Značení:	$X \sim \text{Weib}(\alpha, \beta)$
Parametry:	$\alpha, \beta > 0$
Nosič:	$X \in (0, \infty)$
Hustota:	$f(x) = \begin{cases} \alpha \beta^\alpha x^{\alpha-1} e^{-(\beta x)^\alpha} & x \geq 0 \\ 0 & x < 0 \end{cases}$
Distribuční funkce:	$F(x) = \begin{cases} 1 - e^{-(\beta x)^\alpha} & x \geq 0 \\ 0 & x < 0 \end{cases}$
Střední hodnota:	$E X = \frac{1}{\beta} \Gamma\left(1 + \frac{1}{\alpha}\right)$
Rozptyl:	$\text{var } X = \frac{1}{\beta^2} \left\{ \Gamma\left(1 + \frac{2}{\alpha}\right) - \left[\Gamma\left(1 + \frac{1}{\alpha}\right) \right]^2 \right\}$

A.1.3 Zobecněné gama rozdělení

Značení:	$X \sim \Gamma(\alpha, \beta, \kappa)$
Parametry:	$\alpha, \beta, \kappa > 0$
Nosič:	$X \in (0, \infty)$
Hustota:	$f(x) = \begin{cases} \frac{\alpha \beta^\kappa x^{\kappa-1}}{\Gamma(\kappa)} \exp[-(\beta x)^\alpha] & x > 0, \\ 0 & x \leq 0 \end{cases}$
Střední hodnota:	$E X = \frac{1}{\beta} \frac{\Gamma(\kappa + \frac{1}{\alpha})}{\Gamma(\kappa)}$
Rozptyl:	$\text{var } X = \frac{1}{\beta^2} \left\{ \frac{\Gamma(\kappa + \frac{2}{\alpha})}{\Gamma(\kappa)} - \left[\frac{\Gamma(\kappa + \frac{1}{\alpha})}{\Gamma(\kappa)} \right]^2 \right\}$