Univerzita Pardubice
Fakulta ekonomicko-správní

Modely zpracování signálů v MATLAB/Simulink-u

Ing. Ondřej Rozinek

Bakalářská práce

2009

Univerzita Pardubice
Fakulta ekonomicko-správní
Ústav systémového inženýrství a informatiky
Akademický rok: **2008/2009**

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení: **Ondřej ROZINEK**

Studijní program: **B6209 Systémové inženýrství a informatika**

Studijní obor: **Informatika ve veřejné správě**

Název tématu: **Modely zpracování signálů v MATLAB/Simulink-u**

Zásady pro vypracování:

Předpokládá se, že bakalářská práce bude zaměřena na:
- možnost zpracování signálů,
- analýza možnosti využití daného zpracování minimálně ve dvou oblastech,
- návrh algoritmů dvou příkladů z předmětné oblasti MATLAB-u.

Rozsah grafických prací:

Rozsah pracovní zprávy:

Forma zpracování bakalářské práce: **tištěná/elektronická**

Seznam odborné literatury:

HLAVÁČ, M., SEDLÁČEK, M. Zpracování signálů a obrazů, 2. vydání, Praha, Vydavatelství ČVUT, 2005, 255 s., ISBN 80-01-03110-1.
UHLÍŘ, J., SOVKA, P. Číslicové zpracování signálů, 2. vydání, Praha, Vydavatelství ČVUT, 2002, 328 s., ISBN 80-01-02613-2.
ZAPLATÍLEK, K., DOŇAR, B. Matlab - pro začátečníky, Praha, BEN – technická literatura, 2005, 2. vyd., 152 s., ISBN 80-7300-175-6.
ZAPLATÍLEK, K., DOŇAR, B. Matlab – tvorba uživatelských aplikací, Praha, BEN – technická literatura, 2004, 1. vyd., 216 s., ISBN 80-7300-133-0.

Vedoucí bakalářské práce: **doc. Ing. Jiří Křupka, Ph.D.**
Ústav systémového inženýrství a informatiky

Datum zadání bakalářské práce: **6. října 2008**

Termín odevzdání bakalářské práce: **1. května 2009**

L.S.

doc. Ing. Renáta Myšková, Ph.D.
děkanka

doc. Ing. Jiří Křupka, Ph.D.
vedoucí ústavu

V Pardubicích dne 6. října 2008

Prohlašuji:

Tuto práci jsem vypracoval samostatně. Veškeré literární prameny a informace, které jsem v práci využil, jsou uvedeny v seznamu použité literatury.

Byl jsem seznámen s tím, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorský zákon, zejména se skutečností, že Univerzita Pardubice má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona, a s tím, že pokud dojde k užití této práce mnou nebo bude poskytnuta licence o užití jinému subjektu, je Univerzita Pardubice oprávněna ode mne požadovat přiměřený příspěvek na úhradu nákladů, které na vytvoření díla vynaložila, a to podle okolností až do jejich skutečné výše.

Souhlasím s prezenčním zpřístupněním své práce v Univerzitní knihovně.

V Pardubicích dne 24. 8. 2009

Ondřej Rozinek

Na tomto místě bych rád poděkoval doc. Ing. Jiřímu Křupkovi, Ph.D. za věnovaný čas při konzultacích a vedení při plánování, tvorbě a kompletování finální podoby této bakalářské práce. Dále bych rád poděkoval svým rodičům a přátelům za podporu při studiu.

Ondřej Rozinek

ANOTACE

Práce se zabývá třemi modelovými příklady zpracování signálů v prostředí MATLAB/Simulink-u. První příklad řeší analýzu ekonomické časové řady – identifikaci, odhadování parametrů a ověřování modelu. Na druhém příkladu je ukázáno použití Expectation-Maximization algoritmu pro shlukování v obraze a vztah ke K-means. Třetí příklad demonstruje shlukování mean-shift algoritmem v obraze.

KLÍČOVÁ SLOVA

časové řady, stochastický proces, Box–Jenkinsova metodologie, AR, MA, ARMA, ARIMA, EM algoritmus, směs Gaussových rozdělení, shluková analýza, K-means, mean-shift algoritmus, zpracování obrazu

TITLE

Models in Signal Processing using MATLAB/Simulink

ANNOTATION

The work deals with three model examples of signal processing using MATLAB/Simulink environment. The first example solves an analysis of economic time series – identification, parameter estimate and model verification. The second instance shows the application of the Expectation-Maximization algorithm for the cluster analysis in the image processing and its relation to K-means. The third example demonstrates the clustering in the image with mean-shift algorithm.

KEYWORDS

time series, stochastic process, Box–Jenkins methodology, AR, MA, ARMA, ARIMA, EM algorithm, Gaussian mixture model, K-means, cluster analysis, mean-shift algorithm, image processing

Obsah

Seznam obrázků

1 Úvod

Práce se zabývá třemi modelovými příklady zpracování signálů. V kapitole 2 je řešen modelový příklad analýzy ekonomické časové řady z [3]. Nejdříve je popsána základní teorie vztahující se k tomuto příkladu. V druhé polovině této kapitoly je uvedena analýza tohoto konkrétního příkladu a popis výsledků ze zpracování tohoto příkladu.

V kapitole 3 je uveden v první části teoretický aspekt Expectation-Maximization (EM) algoritmu. V druhé části kapitoly je sestaven vývojový diagram implementovaného algoritmu. Použití EM algoritmu je ukázáno na shlukování dat. K tomuto účelu jsou použity modelové obrázky, na kterých je patrný vizuální efekt shlukováním EM algoritmem.

Kapitola 4 uvádí vynikající metodu pro shlukování dat nazývanou mean-shift algoritmus. Tento algoritmus je používán např. v počítačovém vidění a hraje významnou roli ve shlukové analýze.

V kapitole 5 jsou shrnuty dosažené výsledky a řešena problematika jednotlivých příkladů. Všechny příklady jsou implementovány v programovém prostředí MATLAB/Simulink-u ve verzi 2009a [47], [48].

V tomto úvodu je v další podkapitole charakterizován signál a jsou ukázány příklady možných podob signálů.

1.1 Charakteristika signálu

Signál můžeme obecněji definovat jako jev fyzikální, chemické, biologické, ekonomické či jiné materiální povahy, nesoucí informaci o stavu systému, který jej generuje. Signály dále můžeme klasifikovat:

- spojité a diskrétní signály (analogové a digitální),
- reálné a komplexní,
- deterministické a náhodné,
- sudé a liché,
- periodické a neperiodické,
- jednorozměrné a vícerozměrné signály.

Jednou z podmnožin signálů jsou časové řady. Časová řada je posloupnost věcně a prostorově srovnatelných dat, která jsou jednoznačně uspořádána z hlediska času ve směru

minulost – přítomnost. Časová řada svým charakterem se řadí do diskrétních signálů. S časovou řadou se setkáváme v různých oblastech života. V medicíně je to např. EKG (elektrokardiografie), EEG (elektroencefalografie), EMG (elektromyografie), ve fyzice např. seismický záznam, řada nejvyšších denních teplot v metrologii, průběh výstupního signálu určitého elektrického přístroje, v ekologii např. sledování různých parametrů znečištění ovzduší. V ekonomii je teorie časových řad jedna z nejdůležitějších kvantitativních metod pro analýzu ekonomických dat, např. vývoj kurzu akcií na burze, poptávka po určitém výrobku, vývoj kurzů cizích měn, inflace, nezaměstnanosti atd. [24], [14].

Z výše uvedeného plyne, že časová řada je signál. Ve třech příkladech se zabývám zpracováním jednorozměrného signálu, resp. ekonomickou časovou řadou a zpracováním dvourozměrného signálu a to obrazem, nebo-li také dvourozměrnou časovou řadou.[25]

2 Box-Jenkinsova metodologie

Nejprve se budeme zabývat možnými přístupy k modelování časových řad. Definujme si výchozí jednorozměrný model časové řady [24]

$$X_t = f(t, \varepsilon_t), \tag{2.1}$$

kde X_t je hodnota modelované časové řady v čase t ($t = 1, 2, ..., n$) a ε_t je hodnota náhodné složky v čase t. K tomuto modelu je možno přistoupit třemi způsoby:

- klasickým modelem,
- Boxovou-Jenkinsovou metodologií,
- spektrální analýzou.

Nyní bude uvedena stručná charakteristika klasického modelu. Je založen na tom, že časovou řadu lze rozložit do čtyř složek:

1. trendovou složku T_t
2. sezonní složku S_t
3. cyklickou složku C_t
4. náhodnou složku ε_t

Přitom jsou uvažovány dva možné způsoby dekompozice časové řady:

- aditivní, pro který platí

  $$X_t = T_t + C_t + S_t + \varepsilon_t = Y_t + \varepsilon_t, \qquad t = 1, 2, ..., n, \tag{2.2}$$

  kde Y_t se označuje jako deterministická složka.

- multiplikativní, definovaný jako

  $$X_t = T_t C_t S_t \varepsilon_t, \qquad t = 1, 2, ..., n, \tag{2.3}$$

  který lze logaritmickou transformací převést na aditivní model.

Tento přístup se zabývá především identifikací i modelováním zejména systematických složek, především trendové a sezonní složky.

S dalším přístupem k modelování časových řad je spektrální analýza [2]. Ta je založena na tom, že periodický signál by se dal rozložit pomocí Fourierovy řady na součet sinů a cosinů s rozdílnou amplitudou, frekvencí a fází. Pro neperiodický signál se dá odvodit při použití určitých předpokladů Fourierova transformace, která převádí signál z časové oblasti do frekvenční. Při výpočtech na počítači se především používá rychlý algoritmus FFT (Fast Fourier Transform) [25], [42].

Posledním přístupem k modelování časových řad je Box-Jenkinsova metodologie, která bude více popsána, protože se k ní vztahuje modelový příklad.

Tato metodologie bere v úvahu při konstrukci modelu časové řady reziduální složku, která může obsahovat korelované náhodné veličiny a dokáže tak zpracovávat časové řady s navzájem závislými pozorováními. Dokonce je zaměření těchto postupů založeno právě na vyšetřování těchto závislostí. Kombinují se často např. autoregresivní modely AR(p) s modely klouzavých průměrů reziduální složky MA(q) [3], [4]. Box-Jenkinsova metodologie umožňuje modelovat trend, sezonnost i cyklický průběh. Je určena především pro analýzu náhodných procesů s diskrétním časem. Nelze ji použít pro analýzu procesu se spojitým časem. Vzhledem k digitalizaci dat (na počítači) se setkáváme většinou jen s diskrétním časem.

Analýza podle Boxovy-Jenkinsovy metodologie se provádí podle postupu daného z teorie modelování. Je ji možno rozdělit na tři obecné kroky [31]:

1. identifikaci modelu,
2. odhad parametrů modelu,
3. ověřování modelu.

V následujících podkapitolách je podrobněji specifikována analýza pomocí Boxovy-Jenkinsovy metodologie po jednotlivých výše uvedených krocích, dále je stručně popsána a vysvětlena základní teorie stacionárního stochastického procesu, autokorelační funkce (ACF) a parciální korelační funkce (PACF) a modely AR, MA, ARMA a ARIMA, které souvisí s konkrétním příkladem analýzy ekonomické časové řady. Vybraným příkladem ekonomické časové řady a jeho analýzou a zpracováním v MATLAB/Simulink-u ve verzi 2009a se zabývá poslední podkapitola. Pro analýzu časových řad se také využívají jiné počítačové programy a knihovny, např.: BMDP, SAS, SPSS, NAG, IMSL, [31] GiveWin2, TSM [3], STATISTICA atd.

2.1 Analýza časových řad

Identifikace modelu je počátečním krokem Boxovy-Jenkinsovy metodologie. Její úlohou je vybrat typ modelu. V Boxově-Jenkinsově metodologii je řada modelů, z kterých se dá vybrat určitý model pro konkrétní případ. V literaturách jsou často zmíněny některé z níže uvedených modelů, které bychom mohli rozdělit podle jejich vlastností následovně [3], [4]:

- modely stacionárních časových řad
  - autoregresní procesy (AR)
  - procesy klouzavých průměrů (MA)
  - smíšené procesy (ARMA)
- modely nestacionárních časových řad
  - procesy náhodné procházky (Random Walk Process)
  - procesy ARIMA,
- modely sezonních časových řad
  - sezonní autoregresní procesy (SAR)
  - sezonní procesy klouzavých průměrů (SMA)
  - smíšené sezonní a nesezonní procesy (SARMA)
  - modely sezonních integrovaných časových řad (SARIMA)
- modely časových řad s dlouhou pamětí
  - frakcionálně integrované procesy (FI)
  - procesy ARFIMA
- modely s režimy určenými pozorovatelnými veličinami
  - modely SETAR
  - modely STAR
- modely s režimy určenými nepozorovatelnými veličinami
  - model MSW (Markov-Switching)
- lineární modely volatility
  - modely ARCH (Autoregressive Conditional Heteroscedasticity)
  - modely GARCH (Generalized ARCH)
  - modely IGARCH (Integrated GARCH)
  - modely FIGARCH (Fractionaly IGARCH)
- nelineární modely volatility
  - modely EGARCH (Exponential GARCH)
  - modely IEGARCH (Integrated EGARCH)

Výběr vhodného modelu se určuje na základě zkoumání odhadu $\hat{\rho}_k$ korelační funkce ρ_k a odhadu $\hat{\beta}_{kk}$ parciální korelační β_{kk}. Většinou stačí použít prvních 20 hodnot těchto funkcí. V těchto funkcích se snažíme najít identifikační bod k_0 [31]. V dalších podkapitolách jsou stručně vysvětleny tyto funkce pro MA, AR a ARMA model spolu s jejich grafy a popisem.

2.2 Lineární proces, autokorelační funkce, parciální autokorelační funkce a bílý šum

V této podkapitole jsou vysvětleny některé ze základních vlastností lineárního stochastického procesu [3], [4], [31], [10].

Nechť funkcí středních hodnot je

$$\mu_t = E(X_t), \tag{2.4}$$

variační funkcí

$$\sigma_t^2 = D(X_t) = E(X_t - \mu_t)^2, \tag{2.5}$$

kovarianční funkcí mezi X_t a X_{t-k} pro $k = \ldots -2, -1, 0, 1, 2, \ldots$

$$\gamma(t, t-k) = E(X_t - \mu_t)(X_{t-k} - \mu_{t-k}), \tag{2.6}$$

a korelační funkci

$$\rho(t, t-k) = \frac{\gamma(t, t-k)}{\sigma_t \sigma_{t-k}}. \tag{2.7}$$

Platí-li pro všechna t, že $\mu_t = \mu$, $\sigma_t^2 = \sigma^2$ a kovarianční a korelační funkce závisí pouze na časové vzdálenosti náhodných veličin, tj.

$$\gamma(t, t-k) = \gamma(t, t+k) = \gamma_k, \tag{2.8}$$

$$\rho(t, t-k) = \rho(t, t+k) = \rho_k, \tag{2.9}$$

potom se takový proces nazývá stacionární nebo také kovariančně stacionární. Jinými slovy stochastický proces je stacionární, pokud charakteristiky jeho náhodných veličin jsou v čase neměnné.

Jak již bylo v předešlé podkapitole zmíněno, autokorelační a parciální autokorelační funkce je důležitou charakteristikou pro identifikaci modelu.

Autokorelační funkci (ACF) lze formulovat jako

$$\rho_k = \frac{C(X_t, X_{t-k})}{\sqrt{D(X_t)}\sqrt{D(X_{t-k})}} = \frac{\gamma_k}{\gamma_0}, \tag{2.10}$$

kdy vzhledem ke stacionaritě procesu můžeme uvažovat $D(X_t) = D(X_{t-k}) = \gamma_0$. Obecně jsou parametry μ, γ_0, ρ_k neznámé. Při předpokladu stacionarity můžeme tyto parametry odhadovat. Střední hodnota procesu μ se odhaduje jako výběrový průměr

$$\overline{X} = \frac{\sum_{t=1}^{T} X_t}{T}, \tag{2.11}$$

kde T je počet hodnot časové řady, rozptyl procesu γ_0 může být odhadován pomocí výběrového rozptylu

$$s^2 = \frac{\sum_{t=1}^{T}(X_t - \overline{X})^2}{T}, \tag{2.12}$$

Odhad ρ_k se vypočítá pomocí výběrové korelace se zpožděním k

$$\hat{\rho}_k = \frac{\sum_{t=k+1}^{T}(X_t - \overline{X})(X_{t-k} - \overline{X})}{\sum_{t=1}^{T}(X_t - \overline{X})^2}. \tag{2.13}$$

Parciálni autokorelační funkce dává informaci o korelaci veličin X_t a X_{t-k}, která je očištěná o vliv veličin ležících mezi nimi. Parciální autokorelaci se zpožděním k vyjadřuje parciální regresní koeficient β_{kk} v autoregresi k-tého řádu

$$X_t = \beta_{k1}X_{t-1} + \beta_{k2}X_{t-2} + ... + \beta_{kk}X_{t-k} + \varepsilon_t \tag{2.14}$$

β_{kk} jako funkce zpoždění k se nazývá parciální autokorelační funkcí (PACF). Postupnou úpravou v maticové formě vyjádříme PACF řešením pomocí Cramerova pravidla jako podíl dvou determinantů.

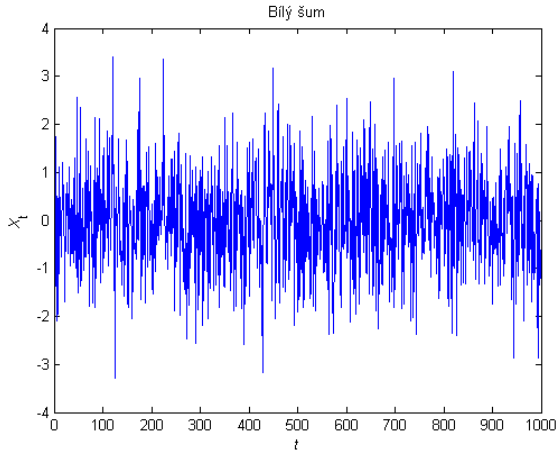
$$\beta_{kk} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & \rho_k \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_{k-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & 1 \end{vmatrix}} \tag{2.15}$$

Parciální autokorelační funkce se odhaduje výběrovou parciální autokorelační funkcí $\hat{\beta}_{kk}$, ve které se nahrazuje ρ jejím odhadem $\hat{\rho}$. Místo počítání složitějších determinantů se používá rekurzivní vztah

$$\hat{\beta}_{kk} = \frac{\hat{\rho} - \sum_{j=1}^{k-1}\hat{\beta}_{k-j}\hat{\rho}_{k-j}}{1 - \sum_{j=1}^{k-1}\hat{\beta}_{k-j}\hat{\rho}_j} \tag{2.16}$$

Proces bílého šumu je základním prvkem časových řad v praxi. Často je nazýván čistý náhodný proces. Je definován jako posloupnost nekorelovaných náhodných veličin. Může, ale také nemusí být stacionární, nezávislý či mající normální rozdělení. Sledovaná proměnná X_t je modelována jako posloupnost nekorelovaných náhodných hodnot ε_t s průměrem a rozptylem jako funkcí t, když je proces nestacionární:

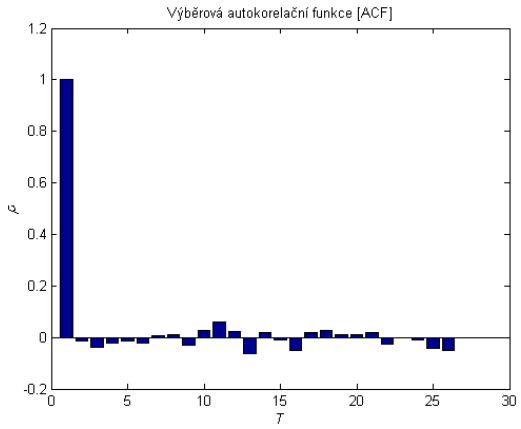
$$X_t = \varepsilon_t \tag{2.17}$$



Obr. 1: Bílý šum generovaný pro 1000 hodnot s normálním rozdělením, střední hodnotou 0 a rozptylem 1

Pro stacionární procesy střední hodnota μ_ε a rozptyl σ_ε^2 bílého šumu ε_t a tak časové řady X_t jsou konstantní. Autokorelační funkce ρ_k je dána

$$\begin{aligned} \rho_k &= 1 \; pro \; k = 0, \\ \rho_k &= 0 \; pro \; k \neq 0. \end{aligned} \tag{2.18}$$



Obr. 2: Výběrová autokorelační funkce vypočtena z 1000 hodnot z obr. 1. Odhad autokorelační funkce se přibližuje k výše uvedeným teoretickým hodnotám s rostoucí délkou časové řady podle zákonu velkých čísel.

2.3 Proces klouzavých průměrů řádu *q*

Proces klouzavých průměrů řádu q, MA(q) proces (z angl. *moving average*), je váženou lineární kombinací kombinací q+1 posunutých nebo zpožděných dat bílého šumu. Tento model časové řady lze formulovat [11], [14]

$$X_t = \varepsilon_t + \alpha_1 \varepsilon_{t-1} + ... + \alpha_q \varepsilon_{t-q}. \tag{2.19}$$

První parametr α_0 je obecně brán jako jedna a ε_t je stacionární, čistě náhodný proces se střední hodnotou μ_ε a rozptylem σ_ε^2. Pomocí operátoru zpoždění, MA rovnice může být zapsána jako

$$X_t = B(z)\varepsilon_t \tag{2.20}$$

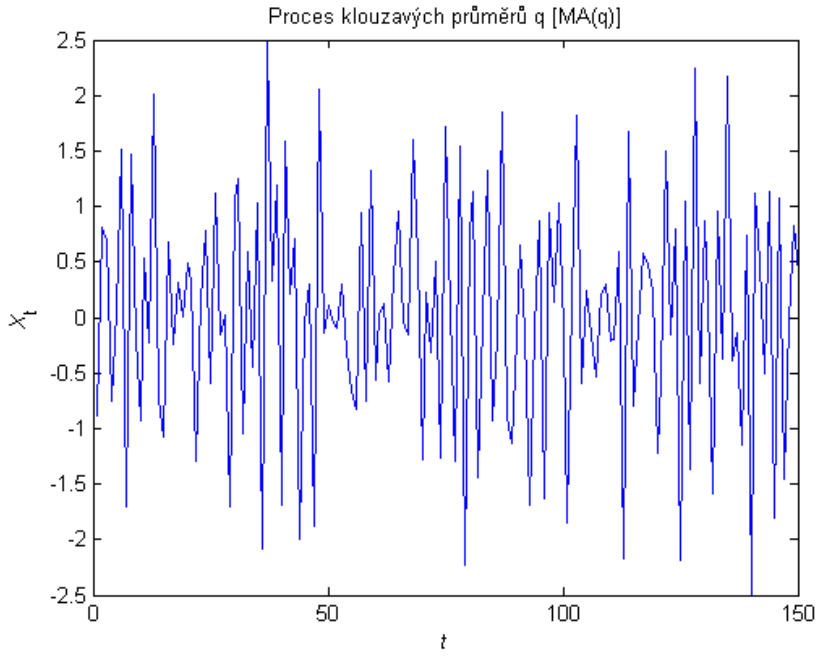
s

$$B(z) = 1 + \alpha_1 z^{-1} + ... + \alpha_q z^{-q} \tag{2.21}$$

a z^{-1} je diferenční operátor zpoždění definován:

$$\begin{aligned} z^{-1}\varepsilon_t &= \varepsilon_t \\ z^{-k}\varepsilon_t &= \varepsilon_{t-k} \\ z\varepsilon_t &= \varepsilon_{t+1} \end{aligned} \tag{2.22}$$

Kořeny $B(z)$ jsou nazývány nuly MA modelu. MA proces je nazýván invertibilní, pokud nuly jsou uvnitř jednotkového kruhu. Výraz $B(z)$ je transformací v časové oblasti s q zpožděním.



Obr. 3: Proces klouzavých průměrů

Střední hodnota procesu se vypočítá

$$E(X_t) = E(\varepsilon_t + \alpha_1\varepsilon_{t-1} + ... + \alpha_q\varepsilon_{t-q}) = E(\varepsilon_t) + \alpha_1 E(\varepsilon_{t-1}) + ... + \alpha_q E(\varepsilon_{t-q}) =$$
$$= \mu_\varepsilon + \alpha_1\mu_\varepsilon + ... + \alpha_q\mu_\varepsilon = \mu_\varepsilon \sum_{i=0}^{q} \alpha_i \tag{2.23}$$

při předpokladu, že $\mu_t = 0$ se odvodí autokovarianční funkce procesu jako

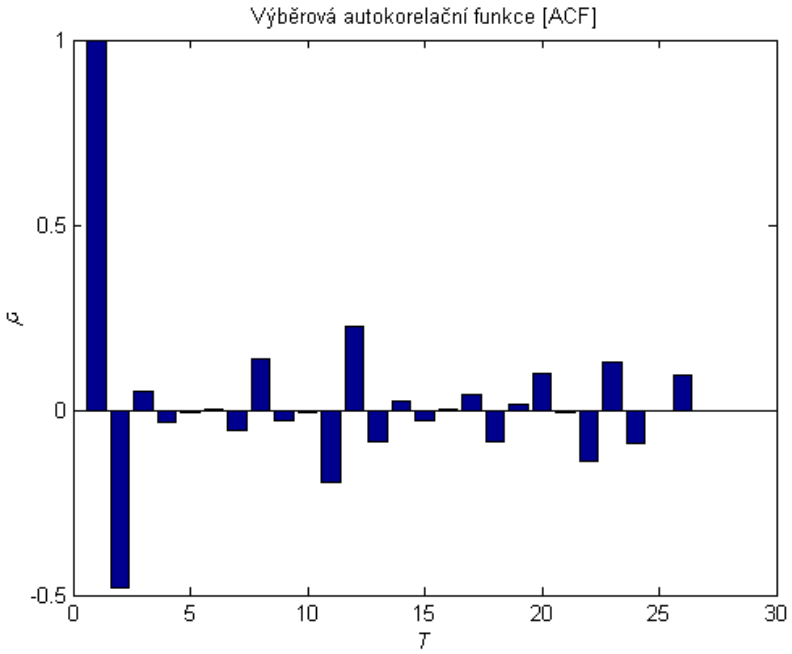
$$r(k) = \sigma_t^2 \rho(k) = E(X_t X_{t+k}) = E[(\varepsilon_t + \alpha_1\varepsilon_{t-1} + ... + \alpha_q\varepsilon_{t-q})(\varepsilon_{t+k} + \alpha_1\varepsilon_{t+k-1} + ... + \alpha_q\varepsilon_{t+k-q})] =$$
$$= \sigma_\varepsilon^2 \sum_{i=0}^{q-k} \alpha_i\alpha_{i+k} \quad pro\, 0 \le k \le q \tag{2.24}$$
$$= 0 \qquad pro\, k > q$$

z autokovarianční funkce vyjádříme rozptyl:

$$\sigma_t^2 = r(0) = E(X_t X_t) = E[(\varepsilon_t + \alpha_1\varepsilon_{t-1} + ... + \alpha_q\varepsilon_{t-q})(\varepsilon_t + \alpha_1\varepsilon_{t-1} + ... + \alpha_q\varepsilon_{t-q})] =$$
$$= \sigma_\varepsilon^2 \sum_{i=0}^{q} \alpha_i^2 \tag{2.25}$$

Nyní můžeme formulovat autokorelační funkci procesu ve tvaru

$$\rho(k) = \frac{E(X_t X_{t+k})}{\sigma_t^2} = \frac{\sum_{i=0}^{q-k} \alpha_i\alpha_{i+k}}{\sum_{i=0}^{q} \alpha_i^2} = \frac{-\alpha_k + \alpha_1\alpha_{k+1} + ... + \alpha_{q-k}\alpha_q}{1 + \alpha_1^2 + ... + \alpha_q^2} \qquad pro\, 0 \le k \le q \tag{2.26}$$
$$= 0 \qquad pro\, k > q$$



Obr. 4: Výběrová autokorelační funkce MA procesu

2.4 Autoregresní proces řádu *p*

Autoregresní proces řádu p, AR(*p*) proces (z angl. *autoregressive*), je definován jako vážená lineární kombinací $p+1$ posunutých nebo zpožděných dat časové řady. Tento model časové řady lze formulovat [11]

$$X_t + \beta_1 X_{t-1} + ... + \beta_p X_{t-p} = \varepsilon_t. \tag{2.27}$$

První parametr β_0 je obecně brán jako nula. Pomocí operátoru zpoždění můžeme tuto rovnici zapsat stručněji ve tvaru

$$A(z) X_t = \varepsilon_t \tag{2.28}$$

s

$$A(z) = 1 + \beta_1 z^{-1} + ... + \beta_p z^{-p} \tag{2.29}$$

Kořeny $A(z)$ jsou nazývány póly AR(*p*) procesu. Procesy jsou nazývány stacionární, pokud jsou všechny póly uvnitř jednotkového kruhu.

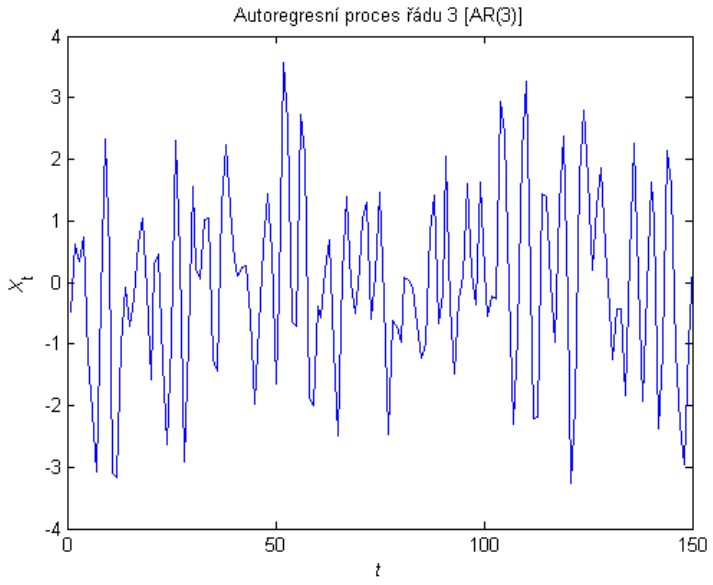
Hodnoty autokorelační funkce se získají na základě diferenční rovnice

$$\rho(k) + \beta_1 \rho(k-1) + ... + \beta_p \rho(k-p) = 0 \tag{2.30}$$

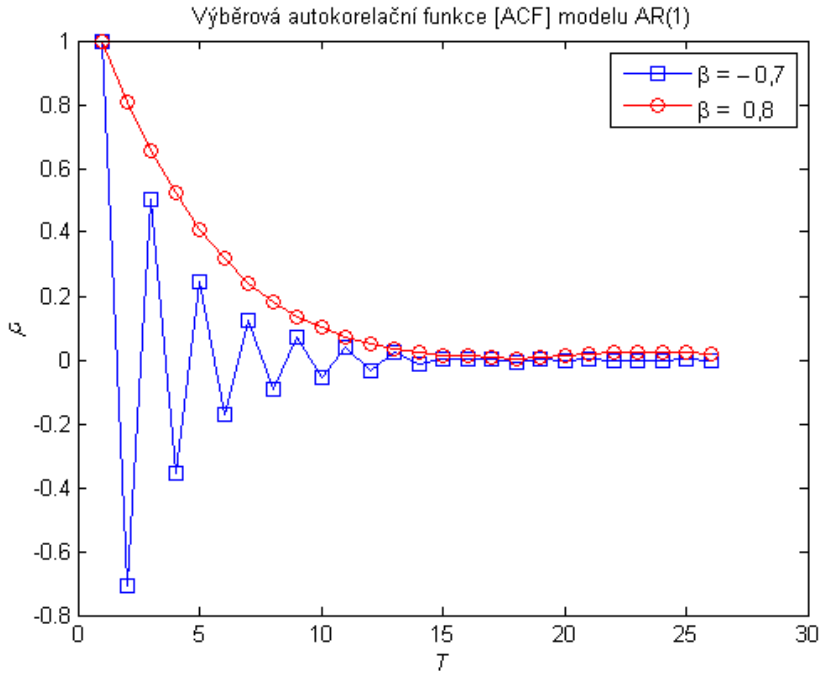
řešením této rovnice je výraz

$$\rho(k) = C_1 p_1^k + C_2 p_2^k + ... + C_p p_p^k,\ k \geq 0 \tag{2.31}$$

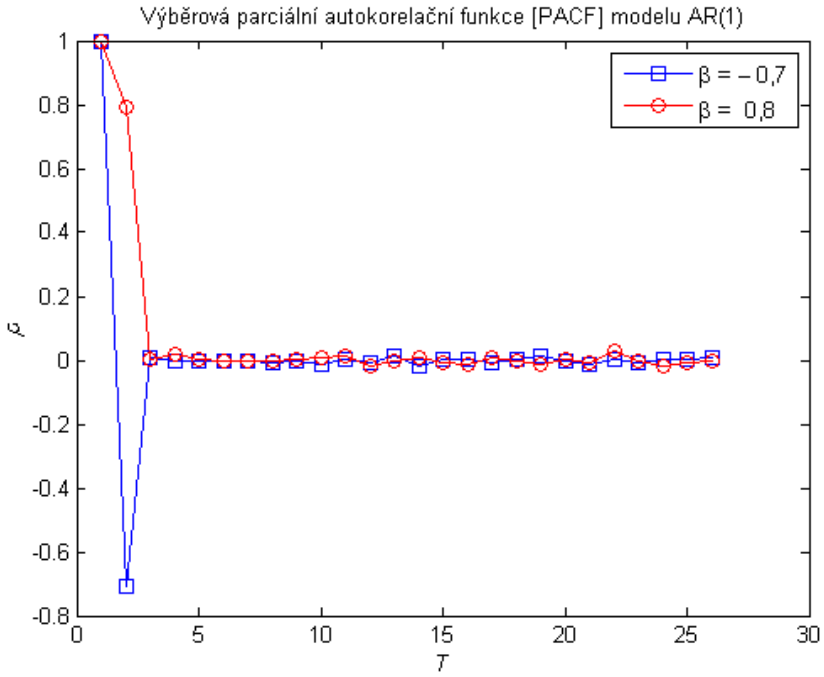
kde p_i jsou póly a C_i jsou konstanty určené okrajovými podmínkami pro $\rho(0), \rho(1), ..., \rho(p-1)$. Autokorelační funkce se skládá z kombinace exponenciálně klesajících pohybů v případě reálných kořenů rovnice $A(z) = 0$ a exponenciálně klesajících sinusoidních pohybů v případě komplexních kořenů stejné rovnice.



Obr. 5: Model časové řady – autoregresní proces řádu 3



Obr. 6: Výběrová ACF modelu AR(1) pro reálné kořeny rovnice A(z) = 0



Obr. 7: Výběrová PACF modelu AR(1) pro reálné a komplexní kořeny rovnice A(z) = 0

2.5 Autoregresní proces klouzavých průměrů řádu *p, q*

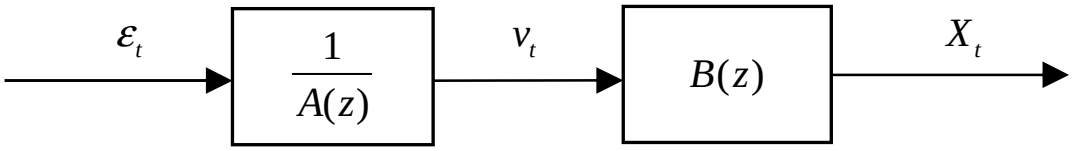
Rovnice generující *autoregresní proces klouzavých průměrů řádu p, q* (také nazýván *smíšený proces*) ARMA(*p*,*q*) (z angl. *autoregressive moving average*) je dána tvarem [11]

$$X_t + \beta_1 X_{t-1} + \ldots + \beta_p X_{t-p} = \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \ldots + \alpha_q \varepsilon_{t-q} \tag{2.32}$$

pomocí operátorů lze rovnici zapsat ve zkrácené formě

$$A(z) X_t = B(z) \varepsilon_t \tag{2.33}$$

Tento proces má *p* pólů a *q* nul. ARMA(*p*,*q*) proces je ekvivalentní sériovému zapojení procesů AR(*p*) a MA(*q*).



Obr. 8: Reprezentace ARMA procesu jako sériového zapojení AR a MA procesu.

Rovnice pro reprezentaci uvedenou na obr. 8 jsou ve tvarech

$$X_t = \frac{B(z)}{A(z)} \varepsilon_t \tag{2.34}$$

$$v_t = \frac{1}{A(z)} \varepsilon_t \tag{2.35}$$

$$X_t = B(z) v_t \tag{2.36}$$

Podmínka stacionarity modelu je totožná s podmínkou stacionarity modelu AR(*p*) a podmínka invertibility s podmínkou invertibility modelu MA (*q*).

2.6 Autoregresní integrovaný proces klouzavých průměrů řádu *p, d, q*

Model ARIMA (*p*,*d*,*q*) se nazývá *autoregresním integrovaným procesem klouzavých průměrů řádu p, d, q* (z angl. *autoregressive integrated moving average*). Dokáže analyzovat nestacionární stochastický proces. Transformací tohoto integrovaného procesu pomocí diference *d*-tého řádu jej lze poté vyjádřit ve formě stacionárního a invertibilního modelu ARMA(*p*,*q*) [3], [4]

$$A(z)(1-z)^d X_t = B(z)\varepsilon_t \tag{2.37}$$

Z modelu ARIMA(p,d,q) se stává po diferenci d-tého řádu model ARMA(p,q), to můžeme uvést symbolicky jako

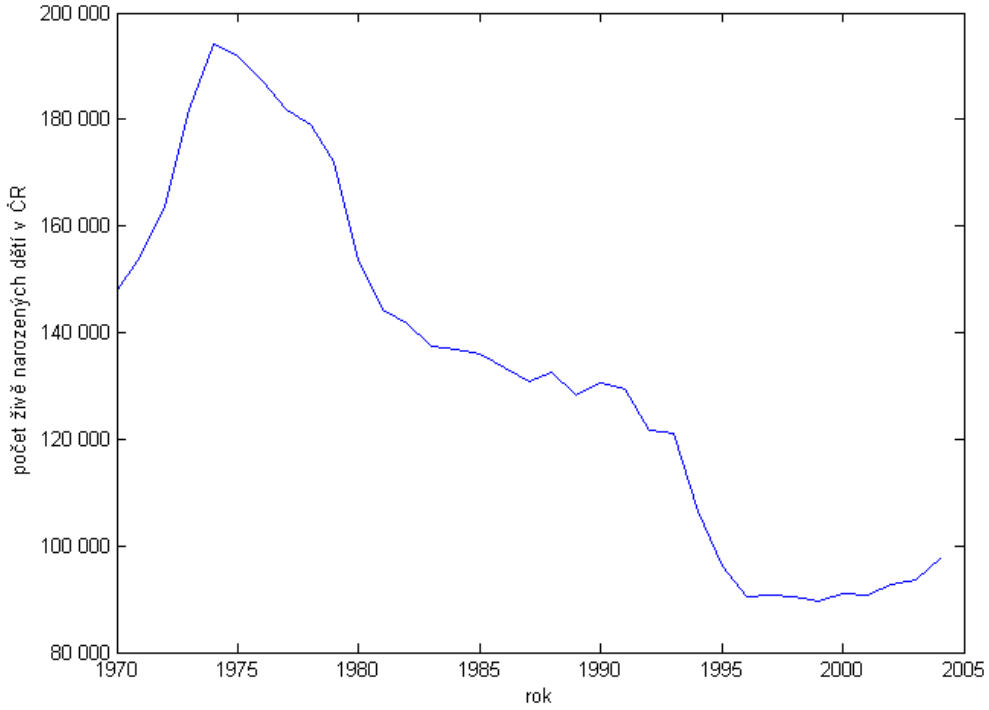
$$X_t \sim ARIMA(p,d,q) \Leftrightarrow \Delta^d X_t \sim ARMA(p,q), \tag{2.38}$$

kde Δ je operátor diference.

2.7 Příklad zpracování ekonomické časové řady

Na příkladu ekonomické časové řady v této kapitole se použije analýza časové řady uváděná v předchozích podkapitolách.

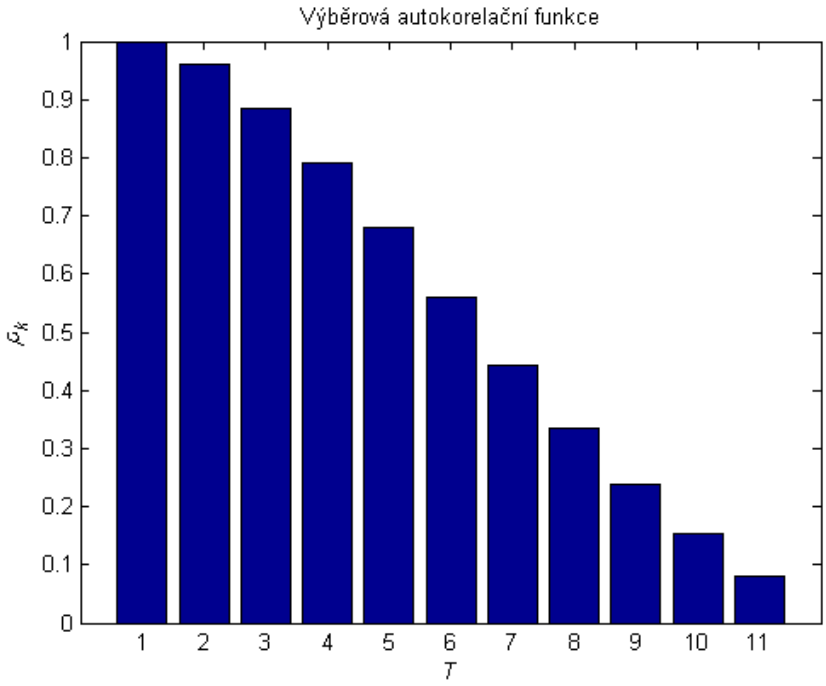
Je dána ekonomická časová řada počtu živě narozených dětí v České republice v letech 1970 až 2004 [3] na obr. 9. Cílem je identifikovat model, odhadnout jeho parametry a ověřit jeho kvalitu diagnostickou kontrolou.



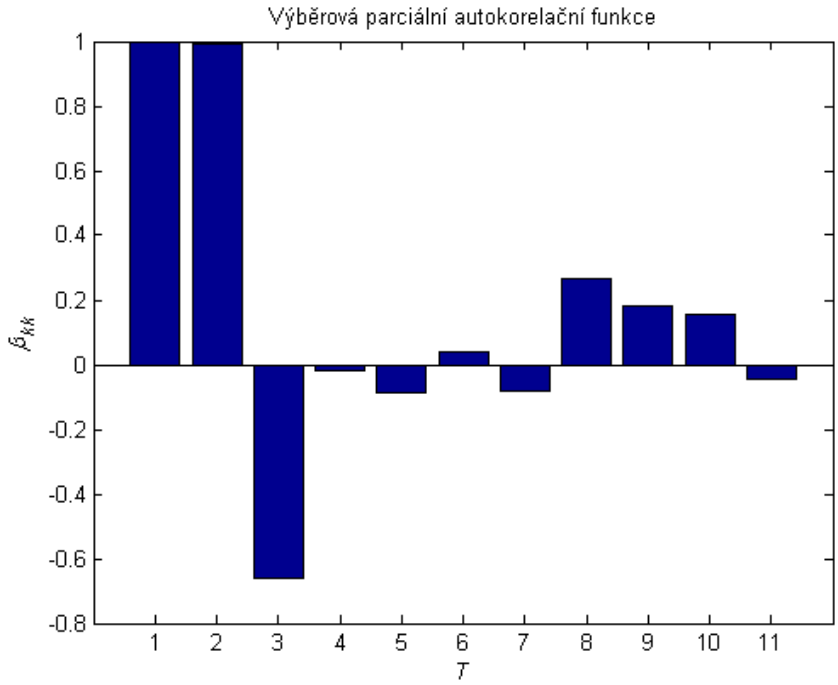
Obr. 9: Počet živě narozených dětí v ČR v letech 1970 – 2004

Analýza časové řady začíná identifikací modelu. Z klesajícího tvaru časové řady a z tvaru výběrové autokorelační funkce obr. 10 a parciální autokorelační funkce obr. 11 je možno se domnívat, jak je dále vysvětleno, že se jedná o nestacionární integrovanou časovou řadu a je vhodné zvolit pro tuto časovou řadu model ARIMA.

Určení a ověření řádu diferencování d při výstavbě modelu ARIMA je velmi důležité. Při analýze ekonomických a časových řad jsou většinou integrované časové řady maximálně řádu dva, $d = 2$. Nejčastěji se však vyskytují řady s $d = 1$. Z tohoto důvodu se časové řady nejčastěji stacionarizují většinou prostřednictvím první či druhé diference. Existují subjektivní metody, jak určit a ověřit řád diferencování. Jednou z nich je posouzení grafu časové řady. Porovnává se původní nediferencovaná časová řada s časovou řadou prvních či druhých diferencí. Tato subjektivní metoda je závislá na empirické zkušenosti analytika, který časovou řadu posuzuje. Další metodou je posouzení tvaru autokorelační funkce. Klesá-li tato funkce pomalu, přibližně lineárním tempem obr. 10, jde o situaci, kdy alespoň jeden kořen rovnice $A(z) = 0$ je blízký jedné a je vhodné provést diferencování. Je-li pozorováno, že původní časová řada je nestacionární, analyzují se první diference. Pokud je opět řada prvních diferencí nestacionární, analyzuje se řada druhých diferencí atd.



Obr. 10: Autokorelační funkce časové řady počtu živě narozených dětí v ČR



Obr. 11: Parciální autokorelační funkce časové řady počtu živě narozených dětí v ČR

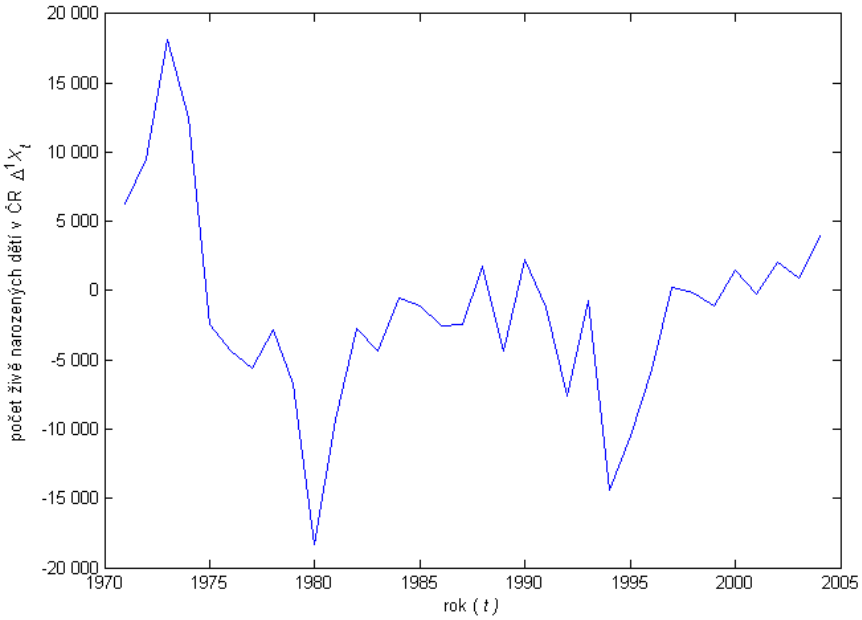
Skutečnost, že se nejedná o stacionární časovou řadu, potvrzuje rozšířený Dickeyův – Fullerův test (ADF; z angl. augmented Dickey – Fuller test). Nulová hypotéza je, že skutečný základ procesu tvoří nulový drift modelu ARIMA(p,1,0) ve tvaru

$$X_t = X_{t-1} + \sum_{i=1}^{p} \xi_i \Delta X_{t-i} + \varepsilon_t . \tag{2.39}$$

Tato rovnice je ekvivalentní integrovanému modelu AR($p+1$). Alternativní hypotézou je odhadovaný regresní model metodou nejmenších čtverců (the estimated OLS regression model)

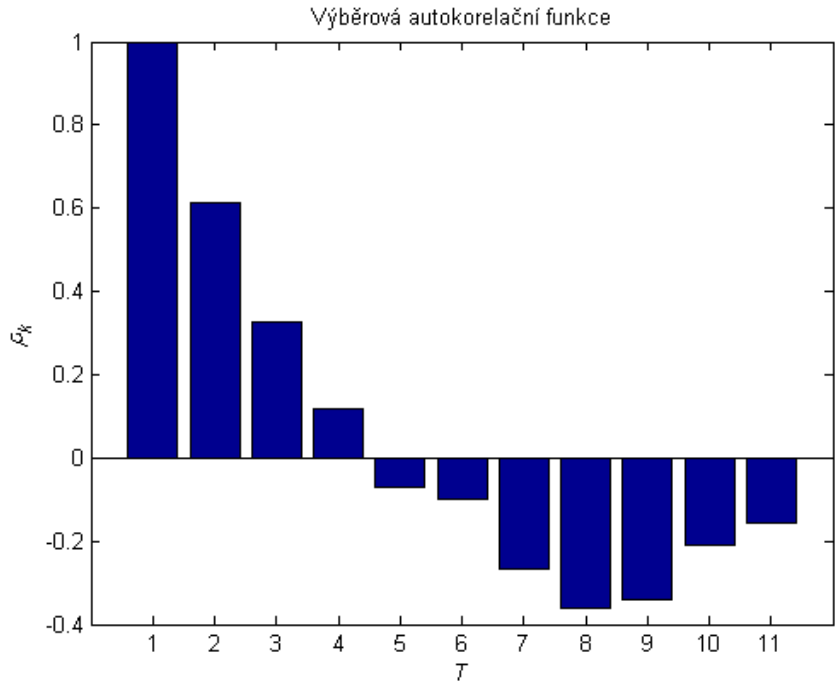
$$X_t = C + \beta_1 X_{t-1} + \sum_{i=1}^{p} \xi_i \Delta X_{t-i} + \varepsilon_t \tag{2.40}$$

pro nějakou konstantu C a model AR(1) s koeficientem $\beta_1 < 1$. V MATLABu je tento test implementován. Např. dosazením zpoždění 2 se dostane hodnota testovacího kritéria $-1,342$. Protože hodnota testovacího kritéria není menší než kritická hodnota $-2,958$ na hladině významnosti 5%, lze konstatovat tak, že nebyla na 5% hladině významnosti prokázána hypotéza, že analyzovaná časová řada je stacionární. Stacionarizace časové řady se provede pomocí první diference. To je znázorněno na obr. 12.

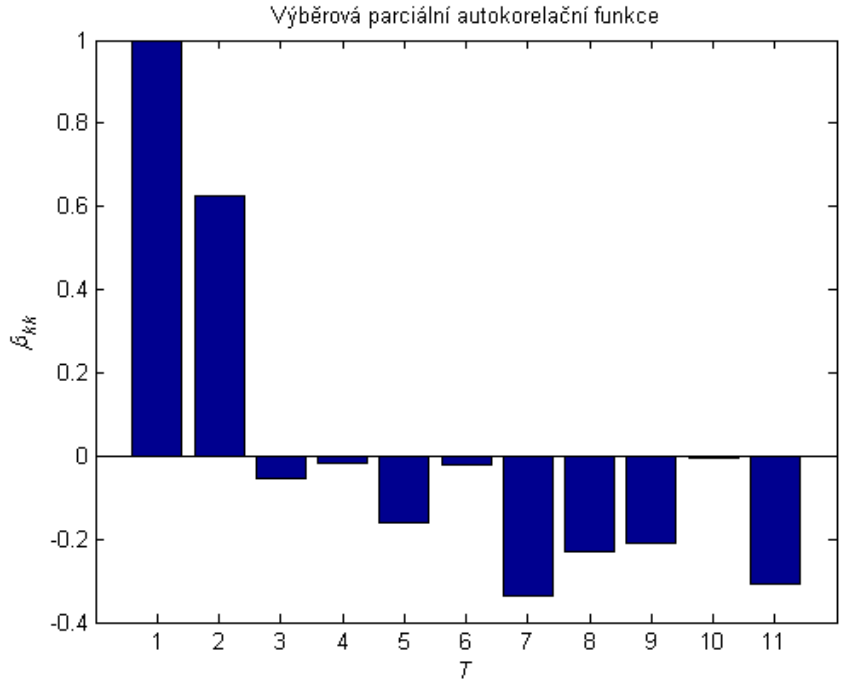


Obr. 12: První diference počtu živě narozených dětí

I když u výběrové autokorelační a parciálně autokorelační funkce je druhý odhad statisticky nevýznamný, lze pozorovat, že hodnoty autokorelační funkce klesají pomaleji. Autokorelační funkce na obr. 13 má exponenciálně klesající sinusoidní pohyb. Je zřejmé, že identifikační bod se může zvolit $k_0 = 1$(tab. 1). Pro diferencovanou řadu se tak dá vybrat model AR(1) a model původní nediferencované časové řady lze označit jako ARIMA(1,1,0).



Obr. 13: ACF diferencované časové řady počtu živě narozených dětí v ČR



Obr. 14: PACF diferencované časové řady počtu živě narozených dětí v ČR

Tab. 1. Tvar autokorelační a parciální autokorelační funkce pro jednotlivé modely [31], [3].

	AR(p)	MA(q)	ARMA(p,q)
ρ_k	Neexistuje identifikační bod k_0; ρ_k ve tvaru kombinace exponenciálně klesajících pohybů nebo exponenciálně klesajících sinusoidních pohybů.	Identifikační bod $k_0 = q$ $\rho_k \neq 0 \quad pro \quad k = 1,2,...,q$ $\rho_k = 0 \quad pro \quad k > q$	Neexistuje identifikační bod k_0; ρ_k ve tvaru kombinace exponenciálně klesajících pohybů nebo exponenciálně klesajících sinusoidních pohybů po prvních $p-q$ hodnotách.
β_{kk}	Identifikační bod $k_0 = p$ $\beta_{kk} \neq 0 \quad pro \quad k = 1,2,...,p$ $\beta_{kk} = 0 \quad pro \quad k > p$	Neexistuje identifikační bod k_0; β_{kk} ve tvaru kombinace exponenciálně klesajících pohybů nebo exponenciálně klesajících sinusoidních pohybů.	Neexistuje identifikační bod k_0; β_{kk} ve tvaru kombinace exponenciálně klesajících pohybů nebo exponenciálně klesajících sinusoidních pohybů po prvních $p-q$ hodnotách.

Nejprve použijeme t-test na 5% hladině významnosti, který indikuje použití odhadu konstanty. Nulová hypotéza je, že diferencovaná časová řada má normální rozdělení se střední hodnotou 0. Vůči tomu alternativní hypotéza je, že střední hodnota je různá od nuly. Nulovou hypotézu nezamítáme, protože hodnota testovacího kritéria padla do oblasti přípustných hodnot.

Nyní se přistoupí k odhadu modelu ARIMA(1,1,0) bez konstanty. Odhadovaná rovnice je zapsána pomocí operátoru zpoždění ve tvaru

$$A(z)X_t = \varepsilon_t, \tag{2.41}$$

kde

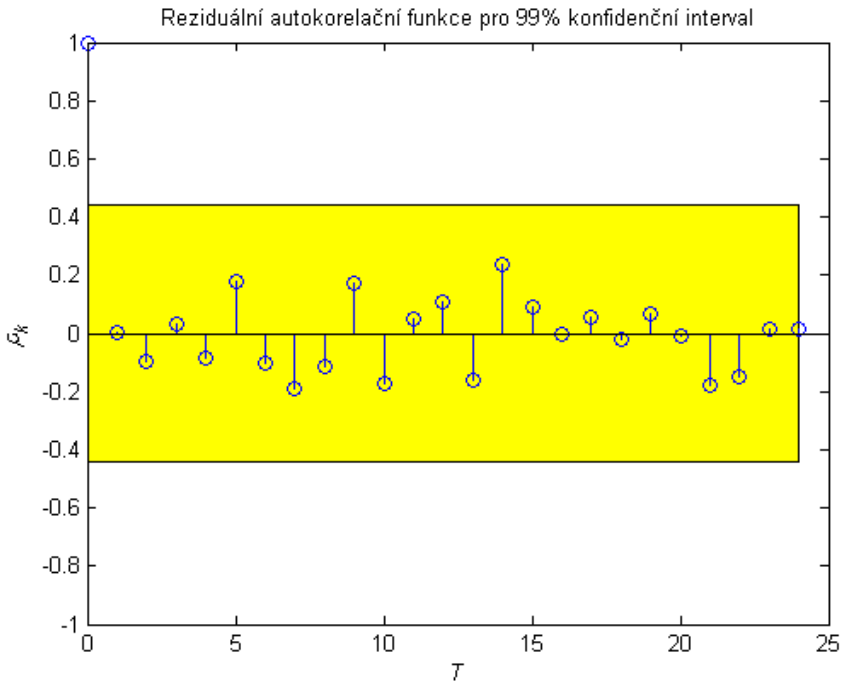
$$A(z) = 1 + \beta_1 z^{-1} \tag{2.42}$$

Odhadnutý parametr je $\beta_1 = -0,6463$, potom lze zapsat výslednou rovnici pomocí operátoru zpoždění ve formě

$$(1 - 0,6463z^{-1})X_t = \varepsilon_t \tag{2.43}$$

a v klasickém tvaru

$$X_t - 0,6463X_{t-1} = \varepsilon_t \tag{2.44}$$



Obr. 15: Reziduální ACF, žlutým pásmem je označen 99% konfidenční interval

V této fázi se přistoupí k diagnostické kontrole modelu ARIMA(1,1,0). Výpočtem reziduální autokorelační funkce se zjistí, že její hodnoty leží uvnitř tolerančních mezí daných 99% konfidenčním intervalem obr. 15. Lze tak konstatovat, že rezidua nevykazují autokorelaci. Tato hypotéza se může otestovat Jarque-Bera testem, který indikuje normalitu nesystematické složky. Testuje se nulová hypotéza, že časová řada má normální rozdělení nesystematické složky s neznámou střední hodnotou a rozptylem vůči alternativní hypotéze, že časová řada nemá normální rozdělení. Výsledkem je, že nulovou hypotézu na 5% hladině významnosti nezamítáme, protože hodnota testovacího kritéria je v oblasti přípustných hodnot. Další test, který může být použit na otestování normality nesystematické šložky je Lilliersfor test, u kterého také nezamítáme nulovou hypotézu o normálním rozdělení. Možnost, jak zjistit zda nesystematická složka není autokorelována, je použití tzv. portmanteau testu, nebo-li Ljung-Box Q-test. Tímto testem nebyla prokázána autokorelace nesystematické složky. Engle's ARCH test neprokazuje přítomnost podmíněné heteroskedasticity.

Protože z tvaru autokorelační a parciální autokorelační funkce na obr. 13 a obr. 14 nelze určit jednoznačně model, bude taky uvažován model MA(1), resp. model ARIMA(0,1,1) pro nediferencovanou časovou řadu. Nyní se přistoupí k odhadu modelu ARIMA (0,1,1) bez konstanty. Odhadovaná rovnice je zapsána pomocí operátoru zpoždění ve tvaru

$$X_t = B(z)\varepsilon_t, \tag{2.45}$$

kde

$$B(z) = 1 + \alpha_1 z^{-1} \tag{2.46}$$

Odhadnutý parametr je $\alpha_1 = 0,6887$ potom lze zapsat výslednou rovnici pomocí operátoru zpoždění ve formě

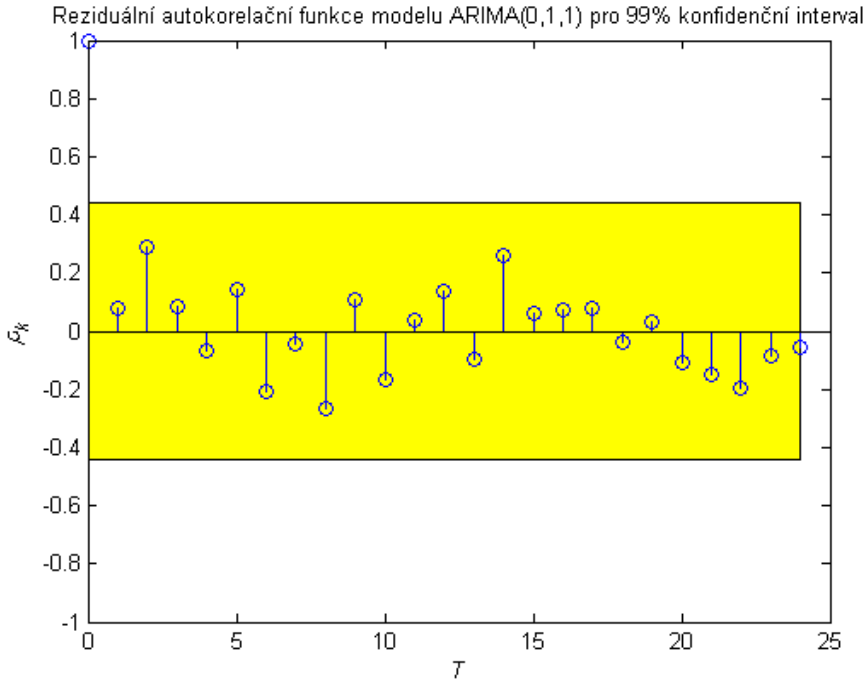
$$X_t = (1 + 0,6887 z^{-1})\varepsilon_t \tag{2.47}$$

a v klasickém tvaru

$$X_t = \varepsilon_t + 0,6887\varepsilon_{t-1} \tag{2.48}$$

Obdobně jako na předešlý model se použijí stejné testy. Opět u tohoto modelu rezidua nevykazují autokorelaci, podmíněnou heteroskedasticitu a ani nenormalitu. Testy bylo dosaženo u modelu ARIMA(0,1,1) stejných závěrů jako u modelu ARIMA(1,1,0). Může se tak konstatovat, že oba modely jsou pro časovou řadu akceptovatelné. V tomto případě se použije kritérium pro volbu modelu, např. AIC (Akaikeho kritérium). Což je kritérium, které

hodnotí správnost volby modelu. Čím je nižší hodnota kritéria, tím je vhodnější model. Avšak vzhledem k tomu, že model ARIMA(0,1,1) má vyšší hodnotu AIC (Akaikeho kritérium), je zřejmé, že pro časovou řadu je vhodnější použít model ARIMA(1,1,0).



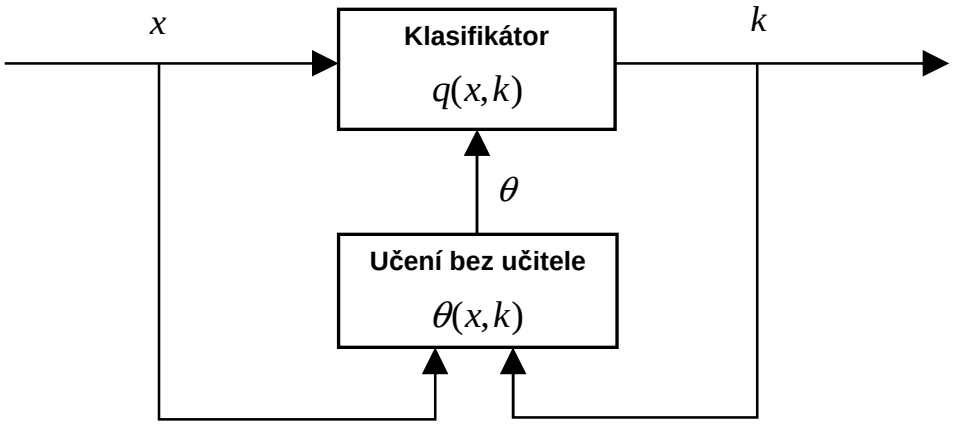
Obr. 16: Reziduální ACF, žlutým pásmem je označen 99% konfidenční interval

3 Expectation-Maximization algoritmus

Tato kapitola je věnována Expectation-Maximization (EM) algoritmu. Postupně je v podkapitolách vysvětlen teoretický úvod, princip a řešení EM algoritmu.

3.1 Teoretický aspekt Expectation-Maximization algoritmu

Expectation-Maximization (EM) byl poprvé vysvětlen a pojmenován v roce 1977 [19]. Čtenáři najdou obsáhlý popis této problematiky v [28], [44]. Zde jsou uvedeny také modifikované rychlejší algoritmy. EM algoritmus je používán především pro shlukování dat ve strojovém učení bez učitele (obr. 17) a počítačovém vidění. EM algoritmus je obecná metoda pro nalezení maximálně-věrohodného odhadu parametrů výchozího rozdělení pravděpodobnosti z daného souboru dat, když jsou data nekompletní nebo mají chybějící hodnoty. EM algoritmus je tedy iterativní optimalizační metoda k odhadnutí nějakých neznámých parametrů θ rozdělení pravděpodobnosti.



Obr. 17: Blokové schéma učení bez učitele, x – pozorovaná vstupní data, k – skrytý stav (nebo také výsledek rozpoznání), θ – parametr, na kterém závisí rozhodovací strategie, q – rozhodovací strategie [20], [13]

Předpokládáme vícerozměrné normální rozdělení (nazývané též vícerozměrné Gaussovo rozdělení) v souladu s centrální limitní větou datového vektoru $x = [x_1, x_1, ..., x_n]^{\mathrm{T}}$ ve formě [23]

$$N(x|\theta) = N(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(\frac{1}{2}(x-\mu)^{\mathrm{T}} \Sigma^{-1} (x-\mu)\right) \tag{3.1}$$

kde $\mu = [\mu_1, \mu_1, ..., \mu_n]^T$ je vektor středních hodnot, $|\Sigma|$ je determinant kovarianční matice Σ. Kovarianční matice Σ je symetrická pozitivně semidefinitní matice ve tvaru

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1n} \\ \Sigma_{21} & \Sigma_{22} & & \\ \vdots & & \ddots & \vdots \\ \Sigma_{n1} & & \cdots & \Sigma_{nn} \end{bmatrix} = \begin{bmatrix} \sigma_{11}^2 & \Sigma_{12} & \cdots & \Sigma_{1n} \\ \Sigma_{21} & \sigma_{22}^2 & & \\ \vdots & & \ddots & \vdots \\ \Sigma_{n1} & & \cdots & \sigma_{nn}^2 \end{bmatrix} \tag{3.2}$$

kde diagonální členy σ^2 jsou rozptyly. Je důležité si uvědomit, že kovarianční matice může být singulární (potom není definována inverzní kovarianční matice Σ^{-1}). Tento případ vede k tomu, že řádky této matice nejsou lineárně nezávislé.

Předpokládejme datové vektory generované nezávisle a identicky s pravděpodobností p. Potom výsledné rozdělení pravděpodobnosti vzorku je

$$p(x|\theta) = \prod_{i=1}^{N} p(x_i|\theta) = L(\theta|x) \tag{3.3}$$

Funkce $L(\theta|x)$ je nazývána věrohodnostní funkcí. Naším cílem je najít parametry θ, které maximalizují L. Chceme najít odhad $\hat{\theta}$ parametrů danou θ rovnicí

$$\hat{\theta} = \arg\max_{\theta} L(\theta|x) \tag{3.4}$$

Často raději maximalizujeme tvar $\ln(L(\theta|x))$, protože je analyticky snadněji řešitelný. Pro nalezení maxima položíme derivaci věrohodnostní funkce nule následovně

$$\frac{\partial \ln(L(\theta|x))}{\partial \theta} = 0 \tag{3.5}$$

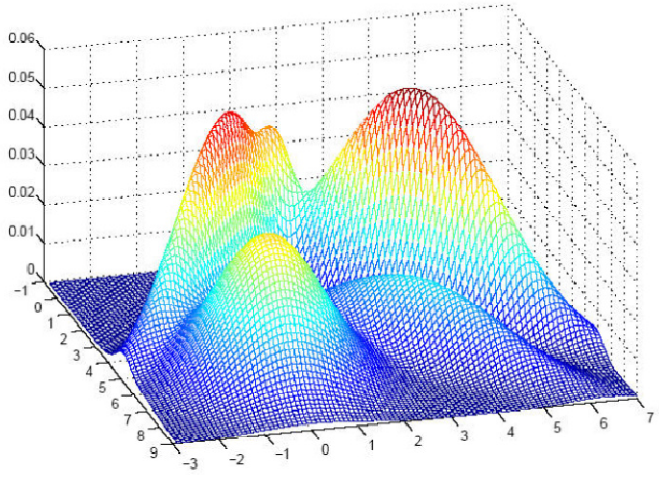
Nicméně pro mnoho problémů není možné najít takové analytické řešení. Musíme se uchýlit tak k více sofistikovanějším numerickým iterativním metodám, jednou z nich je právě EM algoritmus.

Zkoušíme aproximovat data modelem, v našem případě směsí vícerozměrných normálních rozdělení. Definujeme obecně smíšený pravděpodobnostní model (směs Gaussových rozdělení) ve formě

$$p(x|\theta) = \sum_{j=1}^{K} \alpha_j p(x|\theta_j) = \sum_{j=1}^{K} \alpha_j N(x|\mu_j, \Sigma_j) \tag{3.6}$$

kde K je počet komponent pravděpodobnostního rozdělení a platí $\sum_{j=1}^{K} \alpha_j = 1$ váhových koeficientů α_j. Pro nekompletní data je definována logaritmická věrohodnostní funkce jako

$$\ln\left(L\left(\theta|x\right)\right)=\ln\prod_{i=1}^{N}p\left(x_i|\theta\right)=\sum_{i=1}^{N}\ln\left(\sum_{j=1}^{K}\alpha_j p_j\left(x_i|\theta_j\right)\right). \tag{3.7}$$



Obr. 18: Příklad směsi Gaussových rozdělení ve dvourozměrném prostoru ukazující čtyři komponenty [36].

EM algoritmus očekává, že data jsou nekompletní (nebo mají chybějící hodnoty), ale my máme pozorovaná kompletní data. Hlavní trik spočívá v předpokladu existence hodnot pro přidané, ale chybějící (skryté) parametry [5]. Tímto způsobem může být věrohodnostní funkce zjednodušena a analyticky tvárná. Jako předtím předpokládejme, že datový vektor x je pozorovaný a generovaný nějakým pravděpodobnostním rozdělením (v tomto případě vícerozměrným normálním rozdělením). Budeme nazývat vektor x nekompletními daty. Nechť existuje kompletní soubor dat $z=\left(x,y\right)$ a můžeme tak zavést podmíněné pravděpodobnostní rozdělení vyjádřeno rovnicí

$$p\left(z|\theta\right)=p\left(x,y|\theta\right)=p\left(y|x,\theta\right)p\left(x|\theta\right). \tag{3.8}$$

Můžeme také definovat změněnou věrohodností funkci následovně

$$L\left(\theta|z\right)=L\left(\theta|x,y\right). \tag{3.9}$$

Proměnná y (neznámá data) je chybějící informací danou pozorovanými daty x a aktuálními odhady parametrů θ. EM algoritmus nejprve najde očekávané hodnoty kompletních dat logaritmickou věrohodnostní funkcí $\log p\left(x,y|\theta\right)$. Uvažujme x jako nekompletní data a předpokládejme existenci nepozorovaných dat $y=\left\{y_i\right\}_{i=1}^{N}$, jejichž hodnoty nás informují, která pravděpodobnostní komponenta modelu „generuje“ datový bod, věrohodnostní funkce je tak signifikantně zjednodušena. Když známe hodnoty y pro každý datový bod i je generována komponenta $y_i=j$, poté můžeme věrohodnostní funkci zapsat jako

$$\ln\left(L\left(\theta|x,y\right)\right)=\ln\left(p\left(x,y|\theta\right)\right)=\sum_{i=1}^{N}\ln\left(p\left(x_i|y_i\right)p\left(y\right)\right)=\sum_{i=1}^{N}\ln\left(\alpha_j p_j\left(x_i|\theta_j\right)\right) \quad (3.10)$$

Nejprve hádáme parametry pro směs rozdělení (inicializační nastavení parametrů). $\theta^t=\left[\alpha_1^t,...,\alpha_{k,}^t\theta_1^t...\theta_k^t\right]$ jsou náhodné parametry (pro neznámé podmínky) pro věrohodnostní funkci $L\left(\theta^t\middle|x,y\right)$. Pro každý datový bod i komponenty j počítáme podmíněnou pravděpodobnost $p_j\left(x_i|\theta_j^{\ t}\right)$. α_j^t jsou váhové koeficienty. Použitím Bayesovy věty odvodíme

$$p_j\left(j|x_i,\theta^t\right)=\frac{\alpha_j^t p_j\left(x_i|\theta_j^t\right)}{p_j\left(x_i|\theta^t\right)}=\frac{\alpha_j^t p_j\left(x_i|\theta_j^t\right)}{\sum_{j=1}^{K}\alpha_j^t p_j\left(x_i|\theta_j^t\right)} \quad (3.11)$$

kde $p_j\left(j|x_i,\theta^t\right)$ je aposteriorní pravděpodobnost. Vyhodnocení očekávání je nazýváno E-krok algoritmu.

Můžeme zkonstruovat nyní věrohodnostní funkci ve formě

$$p\left(j|x,\theta^t\right)=\prod_{i=1}^{N}p\left(j_i|x_i,\theta^t\right) \quad (3.12)$$

Vypočítáme odhady parametrů, které chceme maximalizovat očekávanou logaritmickou věrohodnostní funkcí podmíněného jevu. Tato část algoritmu se nazývá M-krok. Formulujeme rovnici:

$$\begin{aligned}Q\left(\theta^t\right)&=E\left[\ln p\left(x,j|\theta^t\right)|x,\theta^t\right]\\&=\sum_{j=1}^{K}\sum_{i=1}^{N}\ln\left(\alpha_j p_j\left(x_i|\theta_j^t\right)\right)p\left(j|x_i,\theta^t\right)\\&=\sum_{j=1}^{K}\sum_{i=1}^{N}\ln\left(\alpha_j p_j\left(x_i|\theta_j^t\right)\right)p\left(j|x_i,\theta^t\right)\\&=\sum_{j=1}^{K}\sum_{i=1}^{N}\ln\left(\alpha_j\right)p\left(j|x_i,\theta^t\right)+\sum_{j=1}^{K}\sum_{i=1}^{N}\ln\left(p_j\left(x_i|\theta_j^t\right)\right)p\left(j|x_i,\theta^t\right)\end{aligned} \quad (3.13)$$

Detailní odvození je ukázáno v [5]. K maximalizaci této rovnice můžeme maximalizovat člen obsahující α_j a člen obsahující θ_j^t nezávisle, protože spolu nesouvisí. Pro odvození EM algoritmu je použita Jensenova nerovnost [40].

Najdeme výraz pro α_j použitím Lagrangova multiplikátoru λ. Lagrangův multiplikátor poskytuje strategii pro nalezení maxima funkce s podmínkou, že omezení je $\sum_{j=1}^{K}\alpha_j=1$. Najdeme extrém funkce parciální derivací rovnou nule:

$$\frac{\partial \sum_{j=1}^{K}\sum_{i=1}^{N} \ln\left(\alpha_j\right) p\left(j \middle| x_i, \theta^t\right) + \lambda\left(\sum_j \alpha_j - 1\right)}{\partial \alpha_j} = 0$$
$$\sum_{i=1}^{N} \frac{p\left(j \middle| x_i, \theta^t\right)}{\alpha_j} + \lambda = 0 \tag{3.14}$$

Stanovíme $\lambda = -N$, kde N je počet datových bodů. Vyjádříme výsledné α_j^{t+1} v čase $t+1$

$$\alpha_j^{t+1} = \frac{1}{N}\sum_{i=1}^{N} p\left(j \middle| x_i, \theta^t\right) = 0 \tag{3.15}$$

Analogicky odhadujeme parametr vektoru středních hodnot μ a kovarianční matici Σ. Vezmeme druhý člen $Q\left(\theta^t\right)$

$$\begin{aligned} &\sum_{j=1}^{K}\sum_{i=1}^{N} \ln\left(p_j\left(x_i \middle| \mu_j, \Sigma_j\right)\right) p\left(j \middle| x_i, \theta^t\right) \\ &= \sum_{j=1}^{K}\sum_{i=1}^{N}\left(-\frac{1}{2}\ln\left(\left|\Sigma_j\right|\right) - \frac{1}{2}\left(x_i - \mu_j\right)^{\mathrm{T}} \Sigma_j^{-1}\left(x_i - \mu_j\right)\right) p\left(j \middle| x_i, \theta^t\right) \end{aligned} \tag{3.16}$$

Pro odhad μ_j je parciální derivace definovaná ve formě

$$\frac{\partial \sum_{j=1}^{K}\sum_{i=1}^{N}\left(-\frac{1}{2}\ln\left(\left|\Sigma_j\right|\right) - \frac{1}{2}\left(x_i - \mu_j\right)^{\mathrm{T}} \Sigma_j^{-1}\left(x_i - \mu_j\right)\right) p\left(j \middle| x_i, \theta^t\right)}{\partial \mu_j} = 0$$
$$\sum_{i=1}^{N} \Sigma_j^{-1}\left(x_i - \mu_j\right) p\left(j \middle| x_i, \theta^t\right) = 0 \tag{3.17}$$

Řešíme tuto rovnici a vyjádříme odhad μ_j^{t+1} v čase $t+1$

$$\mu_j^{t+1} = \frac{\sum_{i=1}^{N} x_i\, p\left(j \middle| x_i, \theta^t\right)}{\sum_{i=1}^{N} p\left(j \middle| x_i, \theta^t\right)} \tag{3.18}$$

Podobně můžeme zapsat nalezení odhadu kovarianční matice Σ_j

$$\frac{\partial \sum_{j=1}^{K}\sum_{i=1}^{N}\left(-\frac{1}{2}\ln\left(\left|\Sigma_j\right|\right) - \frac{1}{2}\left(x_i - \mu_j\right)^{\mathrm{T}} \Sigma_j^{-1}\left(x_i - \mu_j\right)\right) p\left(j \middle| x_i, \theta^t\right)}{\partial \Sigma_j} = 0$$
$$\sum_{i=1}^{N} p\left(j \middle| x_i, \theta^t\right)\left(\Sigma_j - \left(x_i - \mu_j\right)\left(x_i - \mu_j\right)^{\mathrm{T}}\right) = 0 \tag{3.19}$$

Řešíme odhad kovarianční matice Σ_j^{t+1} v čase $t+1$

$$\Sigma_j^{t+1} = \frac{\sum_{i=1}^{N} p\left(j \middle| x_i, \theta^t\right)\left(x_i - \mu_j\right)\left(x_i - \mu_j\right)^{\mathrm{T}}}{\sum_{i=1}^{N} p\left(j \middle| x_i, \theta^t\right)} \tag{3.20}$$

Takto jsme odvodili nové odhady parametrů shrnuté v E-kroku a M-kroku následovně (tyto výrazy použijeme pro implementaci) [49], [21], [6], [7], [26], [20], [1], [39], [45]

- E-krok

$$p_j^t\left(j \middle| x_i, \theta^t\right) = \frac{\alpha_j^t p_j\left(x_i \middle| \mu_j^t, \Sigma_j^t\right)}{\sum_{j=1}^{K} \alpha_j^t p_j\left(x_i \middle| \mu_j^t, \Sigma_j^t\right)} = \frac{\alpha_j^t N_j\left(x_i \middle| \mu_j^t, \Sigma_j^t\right)}{\sum_{j=1}^{K} \alpha_j^t N_j\left(x_i \middle| \mu_j^t, \Sigma_j^t\right)} \tag{3.21}$$

- M-krok

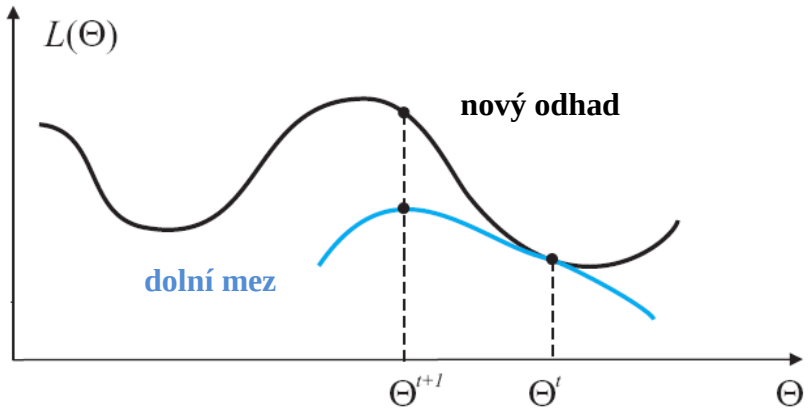
$$\alpha_j^{t+1} = \frac{1}{N}\sum_{i=1}^{N} p_j^t\left(j \middle| x_i, \theta^t\right) \tag{3.22}$$

$$\mu_j^{t+1} = \frac{\sum_{i=1}^{N} x_i p_j^t\left(j \middle| x_i, \theta^t\right)}{\sum_{i=1}^{N} p_j^t\left(j \middle| x_i, \theta^t\right)} \tag{3.23}$$

$$\Sigma_j^{t+1} = \frac{\sum_{i=1}^{N} p_j^t\left(j \middle| x_i, \theta^t\right)\left(x_i - \mu_j^{t+1}\right)\left(x_i - \mu_j^{t+1}\right)^{\mathrm{T}}}{\sum_{i=1}^{N} p_j^t\left(j \middle| x_i, \theta^t\right)} \tag{3.24}$$

a tyto iterace se opakují do konvergence. M-krok najde odhady parametrů θ^{t+1}, které maximalizují odhadovanou dolní mez [18]. Obecně je splněno $L\left(\theta^0\right) \leq L\left(\theta^1\right) \leq \ldots \leq L\left(\theta^t\right)$ [46]. Věrohodnostní funkce konverguje v $t \to \infty$ buď

- k lokálnímu maximu,
- k sedlovému bodu,
- nebo ke globálnímu maximu.



Obr. 19: Optimalizace dolní meze graficky. [13]

3.2 Vztah Expectation-Maximization algoritmu ke K-means

EM algoritmus má přímý vztah ke K-means. K-means je speciální případ EM algoritmu [7], [8]. EM algoritmus vytváří „měkké“ přiřazení datových bodů do shluků založené na pravděpodobnostech. Na druhé straně K-means algoritmus vykoná „tvrdé“ přiřazení datových bodů do shluků.

K-means algoritmus může být odvozen jako partikulární limita EM algoritmu pro pravděpodobnostní model směsi vícerozměrného normálního rozdělení, ve kterém jsou kovarianční matice komponent směsi dány $\sigma^2\mathbf{I}$, kde jsou σ^2 rozptyly, které náleží jednotlivým komponentám a $\mathbf{I}$ je identická matice. Pravděpodobnostní model vícerozměrného normálního rozdělení z rovnice (3.1) může být přepsán do tvaru:

$$N\left(x\middle|\theta\right)=N\left(x\middle|\mu,\Sigma\right)=\frac{1}{\left(2\pi\sigma^2\right)^{\frac{1}{2}}}\exp\left(\frac{1}{2\sigma^2}\left\|x-\mu\right\|^2\right) \tag{3.25}$$

EM algoritmus se předpokládá pro směs K vícerozměrných normálních rozdělení ve formě, ve které je σ^2 jako pevná konstanta na místo odhadovaného parametru. Z rovnice (3.21) je aposteriorní pravděpodobnost pro jednotlivé datové body x_i

$$p_j^t\left(j\middle|x_i,\theta^t\right)=\frac{\alpha_j^t N_j\left(x_i\middle|\mu_j^t,\Sigma_j^t\right)}{\sum_{j=1}^{K}\alpha_j^t N_j\left(x_i\middle|\mu_j^t,\Sigma_j^t\right)}=\frac{\alpha_j^t\exp\left(\frac{1}{2\sigma^2}\left\|x_i-\mu\right\|^2\right)}{\sum_{j=1}^{K}\alpha_j^t\exp\left(\frac{1}{2\sigma^2}\left\|x_i-\mu\right\|^2\right)} \tag{3.26}$$

Je předpokládána limita $\sigma^2 \to 0$ a touto limitou je dosaženo tvrdé přiřazení datových bodů do shluků stejně jako v K-means algoritmu, tak že $p_j^t\left(j\middle|x_i,\theta^t\right) \to r_{ij}$. Jinými slovy *i*-tý datový bod je jednoduše přiřazen do shluku, který má nejbližší střední hodnotu. To se může více formálně vyjádřit

$$r_{ij} = \begin{cases} 1 & pro\ j = \arg\min \left\| x_i - \mu_j \right\|^2 \\ 0 & jinak. \end{cases} \tag{3.27}$$

EM nový odhad pro μ_j daný rovnicí (3.23) se potom redukuje na K-means odhad vyjádřený

$$\mu_j = \frac{\sum_{i=1}^{N} r_{ij} x_i}{\sum_{i=1}^{N} r_{ij}} \tag{3.28}$$

Konečně v limitě pro logaritmickou věrohodnostní funkci pro očekávaná kompletní data se z rovnice (3.13) stává tvar

$$E\left[\ln p\left(x, j\middle|\theta^t\right)\middle|x,\theta^t\right] \to -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{K} r_{ij}\left\|x_i - \mu_j\right\|^2 + konst. \tag{3.29}$$

V této rovnici maximalizovaná logaritmická věrohodnostní funkce pro očekávaná kompletní data je ekvivalentní minimalizaci funkcionálu kvality rozkladu J pro K-means algoritmus daný rovnicí

$$J = \sum_{i=1}^{N}\sum_{j=1}^{K} r_{ij}\left\|x_i - \mu_j\right\|^2, \tag{3.30}$$

který reprezentuje součet čtverců vzdáleností každého datového bodu přiřazeného k vektoru μ_j. Cílem je najít hodnoty pro $\{r_{ij}\}$ a $\{\mu_j\}$, a tak minimalizovat J. Tento cíl může být udělán iterativní procedurou, ve které každá iterace zahrnuje dva po sobě jdoucí kroky korespondující k postupné optimalizaci parametrů r_{ij} (3.27) a μ_j (3.28). Je předpokládaná tzv. batch verze K-means, ve které jsou všechna data updatována. Z toho se může odvodit on-line stochastický algoritmus (MacQueen, 1967). Ten vede k sekvenčnímu updatu postupně pro každý datový bod x_i. Vektor středních hodnot μ_j je updatován použitím rovnice

$$\mu_j^{new} = \mu_j^{old} + \eta_i\left(x_i - \mu_j^{old}\right) \tag{3.31}$$

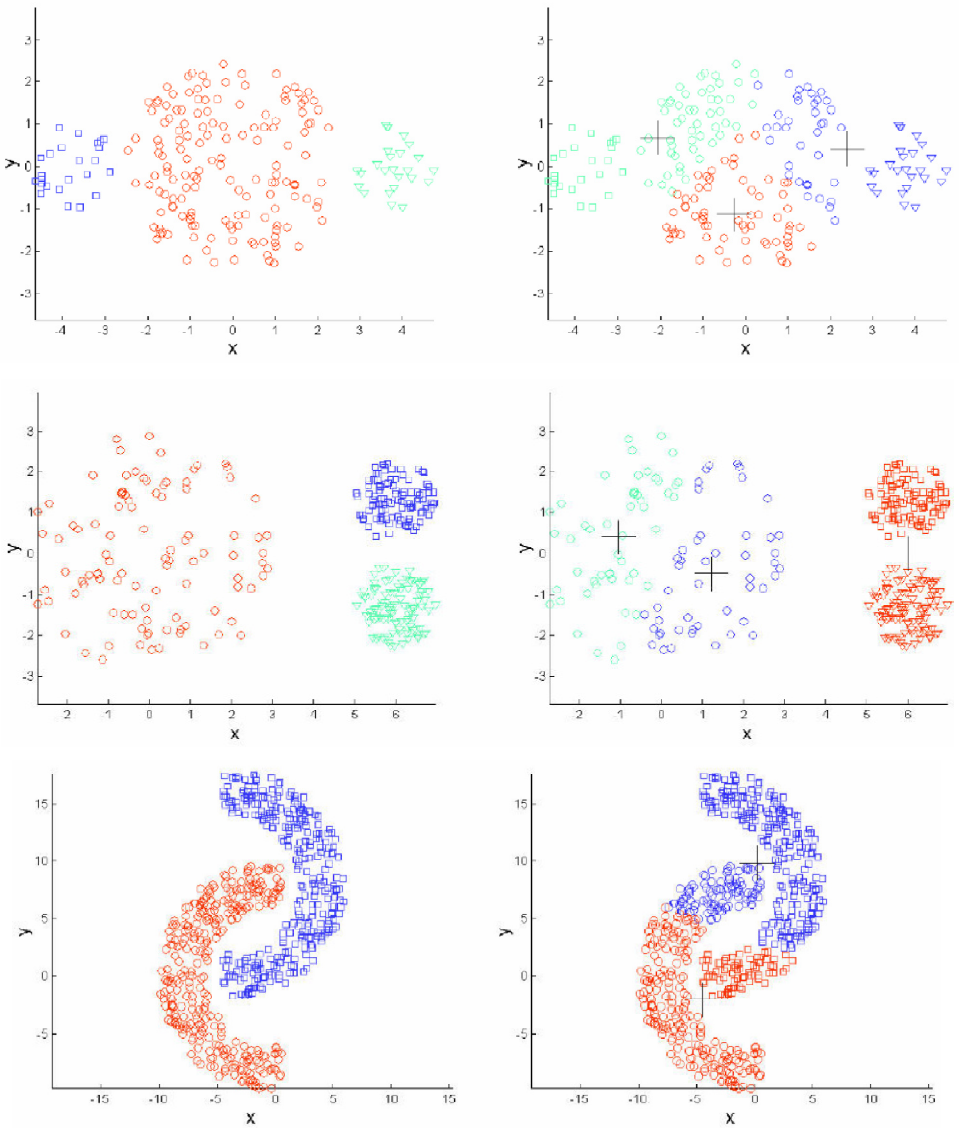
kde η_i je parametr učení.

Je důležité poznamenat, že K-means algoritmus je sám o sobě často používán k inicializaci parametrů pravděpodobnostního modelu směsí vícerozměrných normálních

rozdělení před použitím EM algoritmu. V praxi se také používají podobné algoritmy založené na shlukové analýze jako je ISODATA algoritmus nebo mean-shift algoritmus, kterému je věnována následující kapitola [16], [22], [50], [35].

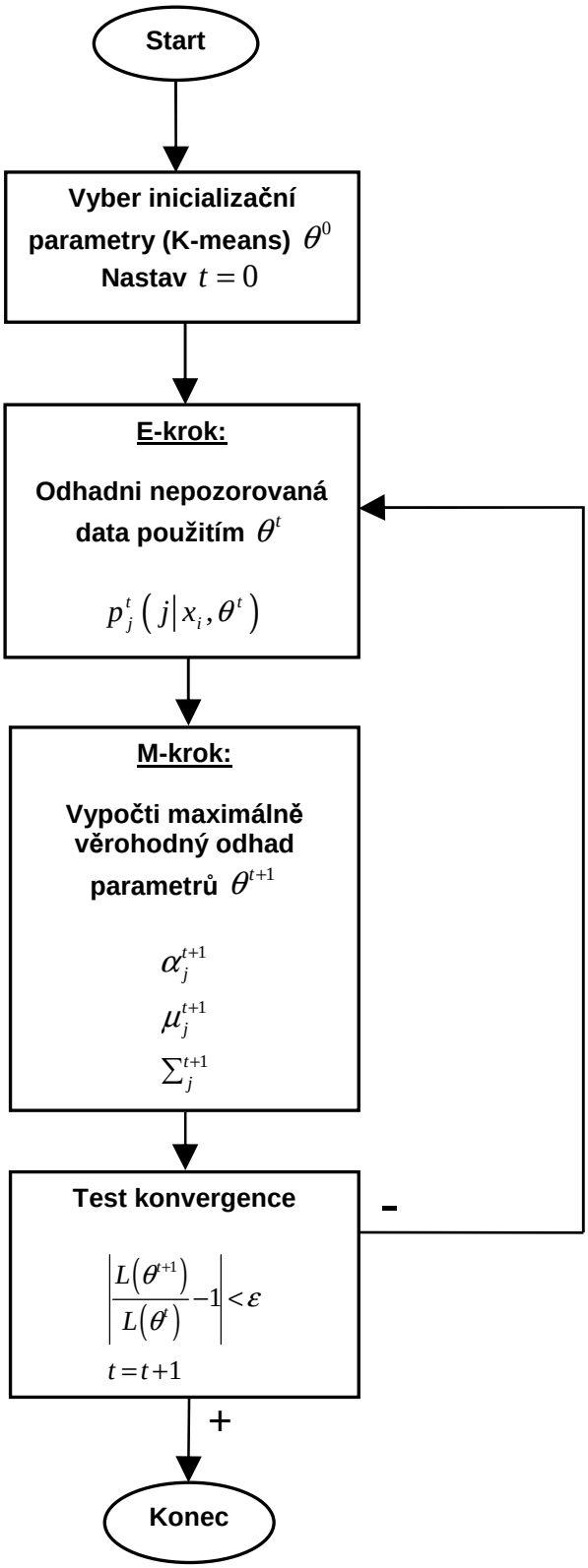
Výhodou K-means algoritmu je, že je výpočetně méně náročný a stabilnější než EM algoritmus. Pro větší počet proměnných může být K-means výpočetně rychlejší než hierarchické shlukování. K-means může vytvářet těsnější shluky než hierarchické shlukování, speciálně v případě kulových shluků.

Nevýhodou K-means je obtížné porovnání kvality vytvořených shluků (např. rozdílné inicializační hodnoty způsobí jiný výsledek). Pevné určení počtu shluků vytváří problém při predikci vhodného počtu shluků. K-means nepracuje dobře s nekulovými shluky. Rozdílné inicializační úseky mohou ve výsledku dávat odlišné konečné shluky.



Obr. 20: Nevýhody K-means; vlevo originalní body, vpravo nashlukované body; odshora – rozdílná velikost shluků, rozdílná hustota, nekulaté shluky [36]

3.3 Řešení Expectation-Maximization algoritmu



Obr. 21: Vývojový diagram EM algoritmu pro odhad parametrů směsi vícerozměrných Gaussových rozdělení [29]

Test konvergence EM algoritmu je definován výrazem

$$\left|L\left(\theta^{t+1}\right)-L\left(\theta^{t}\right)\right|<\varepsilon, \tag{3.32}$$

nebo může být také ve tvaru

$$\left|\frac{L\left(\theta^{t+1}\right)}{L\left(\theta^{t}\right)}-1\right|<\varepsilon, \tag{3.33}$$

kde ε je kladné reálné číslo, empiricky určené pro ukončení iterací a vyplývající z logaritmické věrohodnostní funkce.

Výběr vhodného modelu je důležitý při použití EM algoritmu. K tomuto účelu slouží informační kritéria. Mezi nejčastěji používaná informační kritéria patří Schwarzovo kritérium (BIC – Bayesian Information Criterion), Akaikeho informační kritérium (AIC), křížová validace (Cross validation), délka minimálního popisu (MDL – Minimum Description Length) [12]. Model je lepší než jiný v případě, že má nižší hodnotu AIC (nebo BIC). Obě AIC a BIC informační kritéria mají pevné teoretické základy [27]: Kullback-Leibler vzdálenost v informační teorii (pro AIC) a integrovaná věrohodnostní funkce v Bayesově teorii (pro BIC). Když se složitost správného modelu nezvyšuje s velikostí dat, je v tomto případě preferováno kritérium BIC, opačně AIC. Nicméně lepších výsledků pro smíšené modely je dosahováno s BIC ve srovnání s jinými kritérii [17]. BIC (na rozdíl od AIC) je konzistentní odhad. [9]

Schwarzovo kritérium bylo navrženo Gideonem Schwarzem [33] k vyhodnocení kvality odhadu. BIC je dáno rovnicí

$$BIC=-2L\left(\theta\right)+k\log\left(N\right), \tag{3.34}$$

kde $L\left(\theta\right)$ je maximálně věrohodný odhad, k je počet odhadovaných parametrů modelu a N je počet vzorků (pixelů).

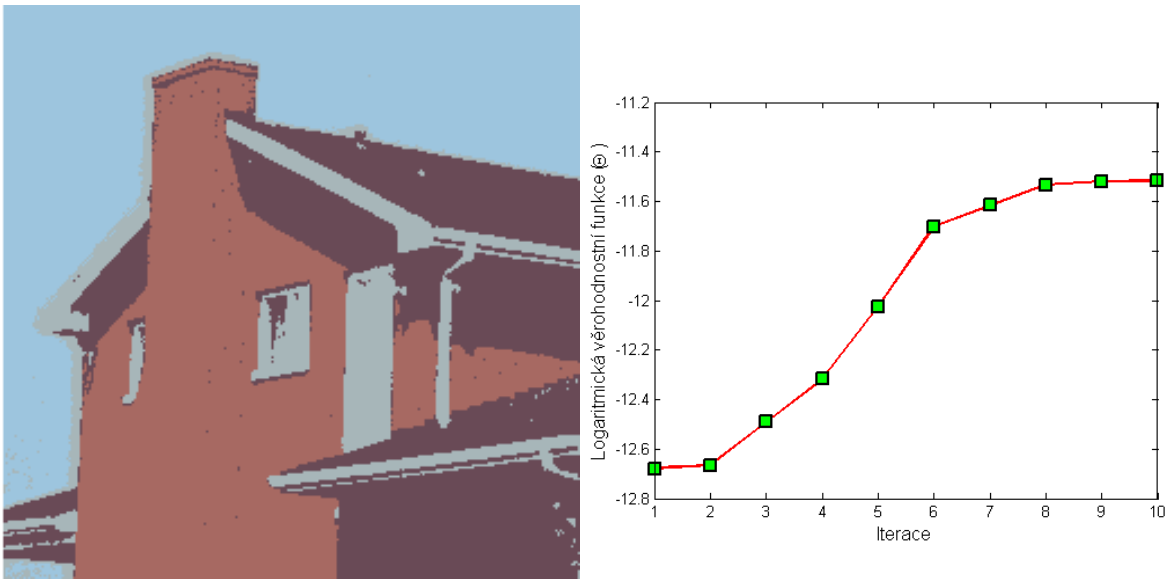
Některé příklady použití EM algoritmu můžou být dobře ukázáné na segmentaci obrazu. K tomuto záměru je definováno 5 příznaků pro každý datový bod:

1. hodnota jasu červené složky obrazu (R),
2. hodnota jasu zelené složky obrazu (G),
3. hodnota jasu modré složky obrazu (B),
4. x – souřadnice obrazu,
5. y – souřadnice obrazu.

Je vytvořena vlastní implementace v MATLAB použitím for-cyklů pro lepší názornost, jak algoritmus pracuje. Pro zpracování výsledků je ale použit modifikovaný

rychlejší algoritmus (až 20×) v důsledku použití funkcí reshape a repmat podle [12], [37]. K inicializaci parametrů je používán K-means algoritmus nebo náhodná inicializace parametrů.

Obr. 22: Vlevo původní vstupní obraz, napravo segmentovaný (shlukovaný) obraz K-means algoritmem do 4 shluků



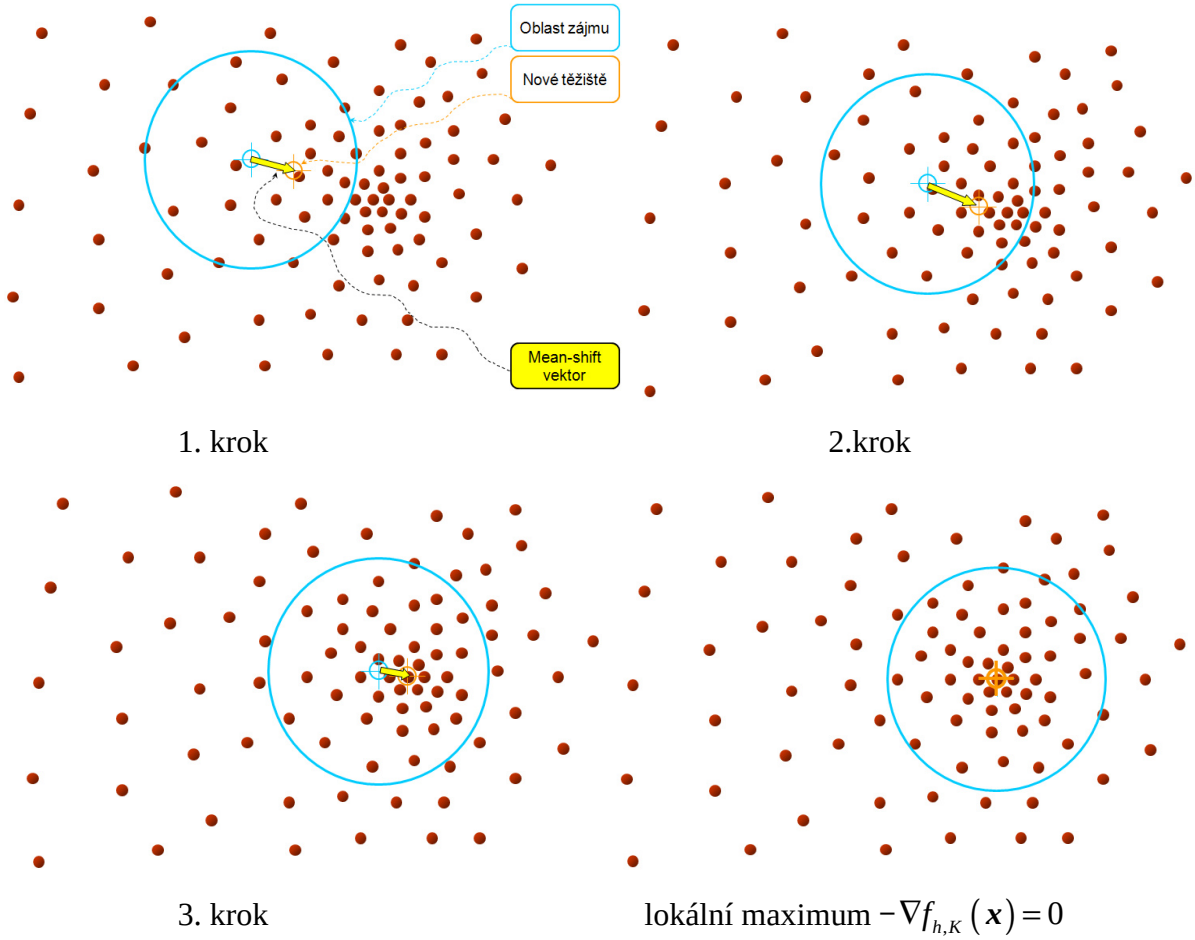
Obr. 23: Vlevo segmentovaný obraz EM algoritmem do 4 shluků, napravo nárůst logaritmické věrohodnostní funkce

4 Mean-shift algoritmus

Mean-shift algoritmus je vynikající metoda pro shlukování dat popsaná v následujících podkapitolách s vlastními výsledky. Odstraňuje většinu nevýhod K-means algoritmu.

4.1 Teoretický úvod do mean-shift algoritmu

Mean-shift algoritmus je shlukovací metoda vytvořena v roce 1975 K. Fukunagou a L. D. Hostetlerem. Mean-shift algoritmus na rozdíl např. od EM algoritmu se vyhýbá odhadování pravděpodobnostní funkce hustoty rozdělení. Mean-shift algoritmus je obecná neparametrická metoda pro analýzu komplexního multimodálního příznakového prostoru a identifikaci příznaků shluků. Mean-shift algoritmus je často používán v úlohách počítačového vidění, např. pro real-time tracking objektu, segmentaci, adaptivní vyhlazování a filtrování (discontinuity preserving filtering) [34].



Obr. 24: Princip mean-shift algoritmu – nejhustší oblast dat je identifikována iterativním procesem [43].

Jedinými volnými parametry této procedury jsou velikost a tvar oblasti zájmu – nebo přesněji identifikace estimátoru vícerozměrného jádra hustoty [16]. Pro všechny praktické účely jsou používána radiálně symetrická jádra

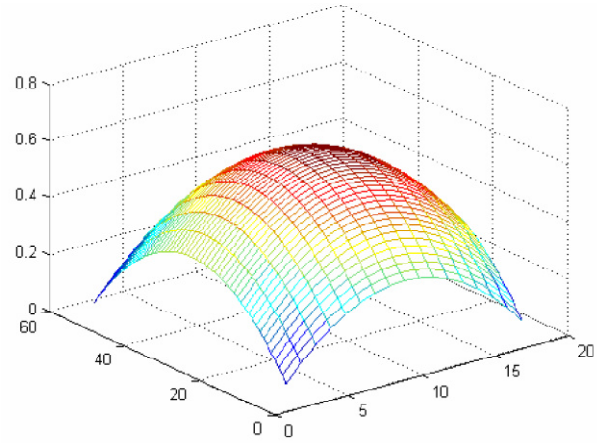
$$K(\boldsymbol{x}) = ck\left(\|\boldsymbol{x}\|^2\right), \tag{4.1}$$

kde c je striktně pozitivní konstanta, která činí $K(\boldsymbol{x})$ integrovatelné jedné. Typická jádra používáná pro mean-shift algoritmus jsou normální jádra $K_N(\boldsymbol{x})$ a Epanechnikova jádra $K_E(\boldsymbol{x})$. Normální jádro je definováno

$$K_N(\boldsymbol{x}) = c\exp\left(-\frac{1}{2}\|\boldsymbol{x}\|^2\right). \tag{4.2}$$

Normální jádro je často symetricky zkrácené k získání konečné podpory. Epanechnikovo jádro je definováno jako

$$K_E(\boldsymbol{x}) = \begin{cases} c\left(1-\|\boldsymbol{x}\|^2\right) & pro \|\boldsymbol{x}\| \le 1 \\ 0 & jinak \end{cases} \tag{4.3}$$



Obr. 25: Ukázka Epanechnikova jádra [30]

a není diferencovatelné na okrajích.

Je dáno n datových bodů (např. pixelů v obraze) $\boldsymbol{x}_i$ v d-rozměrném prostoru R^d, potom estimátor vícerozměrného jádra hustoty $\tilde{f}_{h,K}(\boldsymbol{x})$ je vypočítán v bodě $\boldsymbol{x}$

$$\tilde{f}_{h,K}(\boldsymbol{x}) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right), \tag{4.4}$$

kde h reprezentuje velikost jádra.

Mean-shift algoritmus se snaží identifikovat $\boldsymbol{x}$, ve kterém ideálně platí nulový gradient:

$$\nabla f_{h,K}(\boldsymbol{x}) = 0 \tag{4.5}$$

Mean-shift procedura je elegantní způsob identifikování těchto lokalit bez odhadování výchozí pravděpodobnostní funkce hustoty rozdělení. Jinými slovy z odhadování hustoty problém přejde na odhadování hustotního gradientu

$$\nabla \tilde{f}_{h,K}(\boldsymbol{x}) = \frac{1}{nh^d}\sum_{i=1}^{n} \nabla K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right). \tag{4.6}$$

Použitím tvaru jádra, ve kterém $k(x)$ je profil jádra a předpokladem, že existuje derivace taková, že $-k'(x) = g(x)$ pro všechny $x \in \langle 0, \infty)$, opomeneme-li konečnou množinu bodů

$$K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right) = c_k k\left(\left\|\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right\|^2\right), \tag{4.7}$$

kde c_k je normalizující konstanta a h reprezentuje velikost jádra.

Je důležité poznamenat, že profil funkce $g_E(x)$ je uniformní, když $K(\boldsymbol{x}) = K_E(\boldsymbol{x})$ a pro $g_N(x)$ je definovaný stejný exponenciální výraz jako pro profil jádra $K_N(\boldsymbol{x})$. Při použití funkce $g(x)$ definující profil jádra $G(\boldsymbol{x}) = c_g g\left(\|\boldsymbol{x}\|^2\right)$ se rovnice (4.6) změní na [16]

$$\begin{aligned}
\nabla \tilde{f}_{h,K}(\boldsymbol{x}) &= \frac{2c_k}{nh^{d+2}}\sum_{i=1}^{n}(\boldsymbol{x}-\boldsymbol{x}_i)\,k'\left(\left\|\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right\|^2\right) \\
&= \frac{2c_k}{nh^{d+2}}\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{x})\,g\left(\left\|\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right\|^2\right) \\
&= \frac{2c_k}{nh^{d+2}}\,g\left[\sum_{i=1}^{n}\left(\left\|\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right\|^2\right)\right]\left[\frac{\sum_{i=1}^{n}\boldsymbol{x}_i\,g\left(\left\|\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n}g\left(\left\|\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right\|^2\right)} - \boldsymbol{x}\right]
\end{aligned} \tag{4.8}$$

První člen rovnice (4.8) je úměrný estimátoru hustoty $\tilde{f}_{h,G}$ vypočítaný s jádrem G

$$\tilde{f}_{h,G}(\boldsymbol{x}) = \frac{c_g}{nh^d}\sum_{i=1}^{n} g\left(\left\|\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right\|^2\right). \tag{4.9}$$

Druhý člen reprezentuje mean-shift vektor $m_{h,G}(\boldsymbol{x})$

$$m_{h,G}(\boldsymbol{x}) = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i g\left(\left\|\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right\|^2\right)} - \boldsymbol{x}. \tag{4.10}$$

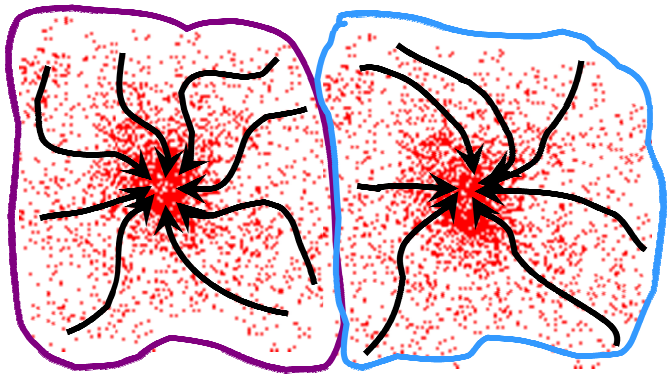
Následné lokality $\{\boldsymbol{y}_j\}_{j=1,2\ldots}$ jádra G jsou vypočítány rovnicí

$$\boldsymbol{y}_{j+1} = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i g\left(\left\|\frac{\boldsymbol{y}_j-\boldsymbol{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\boldsymbol{y}_j-\boldsymbol{x}_i}{h}\right\|^2\right)}, \tag{4.11}$$

kde je $\boldsymbol{y}_1$ inicializovaná pozice jádra G.

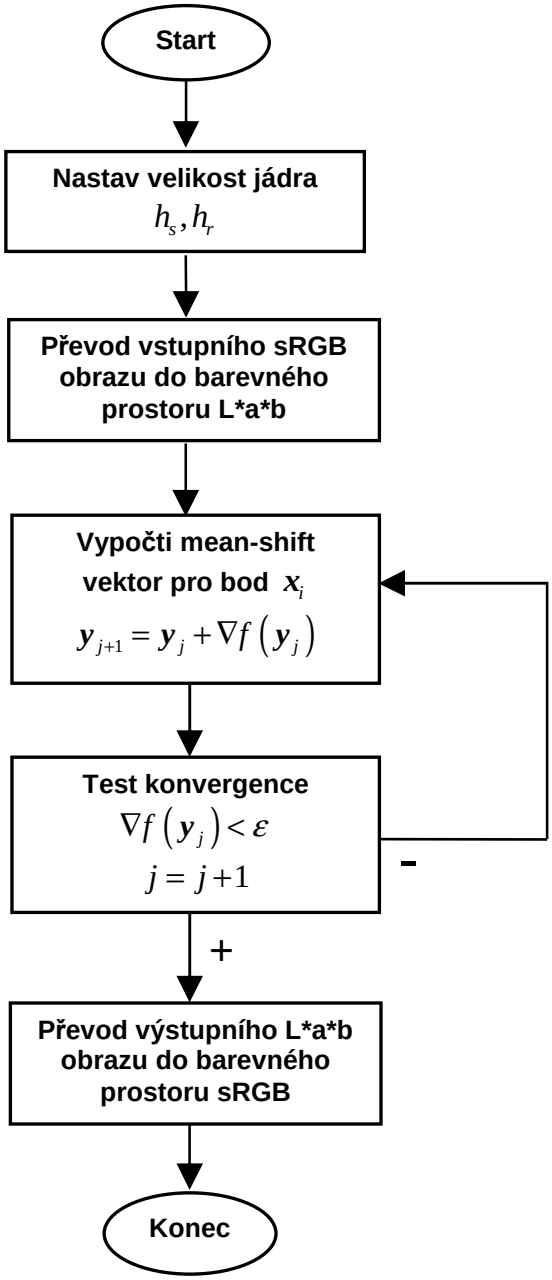
Když má jádro K konvexní a monotónnní klesající profil, sekvence $\{\boldsymbol{y}_j\}_{j=1,2\ldots}$ a $\tilde{f}_{h,K}(j)$ konverguje a monotónně roste. Důkaz je uveden v [16]. Tím je zaručena konvergence do lokálních maxim pravděpodobnostní funkce hustoty nebo-li do hustotních módů adaptováním velikosti mean-shift vektoru (mean-shift vektor konverguje k nule). Rychlost konvergence je dána vybraným jádrem. Když je použito Epanechnikovo jádro (uniformní profil jádra) na diskrétní data, konvergence je dosaženo konečným počtem kroků. Pokud jsou datové body váhovány např. použitím normálního jádra, mean-shift konverguje s nekonečným počtem kroků. Z tohoto důvodu se používá pro ukončení iterací hranice velikosti mean-shift vektoru ε, kde ε je kladné reálné číslo.

Množině všech lokalit, která konverguje do stejného módu (maxima) s určitou tolerancí, se říká „basin of attraction“. V principu je tím vyjádřena příslušnost k danému shluku, jak je ukázáno na obr. 26.



Obr. 26: Dva barevně označené okraje shluků a trajektorie ukazující konvergenci bodů do lokálních maxim [43].

4.2 Řešení mean-shift algoritmu



Obr. 27: Vývojový diagram mean-shift algoritmu

Mean-shift procedura má množství výhod a některé nevýhody. Mezi nejsilnější výhody patří obecná platnost tohoto nástroje. Kvůli značné šumové robustnosti je tento přístup vhodný pro skutečné praktické aplikace. V metodě se nastavuje pouze jeden parametr, a to velikost jádra h, který má fyzikální a srozumitelný význam. Nicméně reakce na výběr h je důležitým omezením, protože vhodné nastavení této hodnoty není triviální. Metody pro adaptivní identifikaci h existují, ale jsou výpočetně velice náročné. [16], [34].

Pro mean-shift algoritmus lze použít jeho implementaci v C++ v softwarovém balíku EDISON [15], [16]. Vlastní implementace je vytvořena v MATLAB/Simulink-u modifikací [35]. Vlastní *d*-rozměrný obraz je reprezentován *d*-rozměrnou mřížkou (prostorová doména) a *p*-rozměrnými pixely (voxely), kde *p* reprezentuje počet spektrálních pásem asociované s obrazem (rozsahovou doménou); $p=1$ pro šedotónový obraz, $p=3$ pro barevný obraz (R, G, B barevné složky). Za předpokladu Euklidovské metriky pro obě domény mají prostorové a rozsahové vektory kompletní informaci o lokalitě pixelů a vlastnosti obrazu mohou být sloučeny do formy vzájemné prostorově-rozsahové domény. Výsledné vzájemné doménové jádro $K_{h_s,h_r}(\boldsymbol{x})$ se skládá ze dvou radiálně symetrických jader s parametry h_s a h_r reprezentující velikost jádra prostorově-rozsahové domény, respektive *p* a *d* určují počet rozměrů prostoru.

$$K_{h_s,h_r}(\boldsymbol{x}) = \frac{c}{h_s^d h_r^d} \sum_{i=1}^{n} k\left(\left\|\frac{\boldsymbol{x}^s}{h_s}\right\|^2\right) k\left(\left\|\frac{\boldsymbol{x}^r}{h_r}\right\|^2\right), \tag{4.12}$$

kde $\boldsymbol{x}^s$ a $\boldsymbol{x}^r$ jsou prostorové a rozsahové části příznakového vektoru, $k(x)$ je běžný profil použitý v obou doménách a *c* je normalizační konstanta. Epanechnikova a normální jádra poskytují dobrý výkon. Z tohoto důvodu se nastavení úrovně rozlišení detekce módů získá dvěma parametry jediného vektoru $\boldsymbol{h}=(h_s,h_r)$. Příklady výsledků jsou ukázány v příloze. Převzaté testovací obrázky jsou z [38], [41].

Obr. 28: Filtrování mean-shift algoritmem (discontinuity preserving filtering)

5 Závěr

Na prvním příkladu byla ukázána analýza ekonomické časové řady skládající se z postupných kroků: identifikace, odhad parametrů a ověření modelu. Při identifikaci modelu se zkoumali především autokorelační a parciální autokorelační funkce. Dále bylo nutné zjistit, zda daná ekonomická časová řada je stacionární. Testováním bylo zjištěno, že se jedná o nestacionární časovou řadu a bylo přistoupeno postupně k odhadům parametrů modelů ARIMA(0,1,1) a ARIMA(1,1,0). Při diagnostické kontrole modelu byla použita řada statistických testů pro otestování normality reziduí, heteroskedasticity a autokorelace reziduí. Oba modely byly vyhovující, avšak pro konečný model ARIMA(1,1,0) rozhodla nižší hodnota AIC (Akaikeho informační kritérium). Analýza této ekonomické časové řady byla implementována v prostředí MATLAB/Simulink 2009a. Dále byly implementovány generátory stochastických procesů jako AR(p), MA(q), ARMA(p,q) a proces náhodné procházky (Random Walk).

V druhém příkladu byl uveden stručně teoretický aspekt Expectation-Maximization (EM) algoritmu a naznačeno jeho odvození. Příklad použití EM algoritmu je ukázán na shlukování dat v obrázcích. K tomuto účelu posloužili modelové obrázky běžně používané ve vědní disciplíně zpracování obrazu. Inicializovat parametry se mohou buď K-means algoritmem, anebo randomizovaně. V této práci je také popsán úzký přímý vztah mezi EM algoritmem a K-means, kde K-means je limitní případ EM algoritmu. Některé výsledky jsou ukázány v příloze. Vlastní implementace je vytvořena v MATLAB/Simulink 2009a názorně pomocí for cyklů, ale pro experimenty bylo použito části zdrojových kódů, které používají funkcí MATLABu a tj. funkce reshape a repmat, které namísto použití for cyklů zrychlují výpočet v MATLABu až 20×. EM algoritmus byl také porovnáván s jinými segmentačními metodami ve vlastní práci [32], kde byl příspěvek oceněn 3. místem v sekci Informatics and Cybernetics.

Třetí příklad se zabývá mean-shift algoritmem. Použití mean-shift algoritmu je ukázáno na segmentaci obrazu (nebo také filtraci – discontinuity preserving filtering). Na příkladech je aplikováno Epanechnikovo nebo normální jádro a je ukázán vliv převodu do jiného barevného prostoru. Implementace je provedena v MATLABu ve verzi 2009a modifikací [35]. Doba výpočtu je závislá na velikosti jádra a velikosti obrázku. Pro větší obrázky s větší velikostí jádra může výpočet trvat až několik hodin.

6 Použitá literatura

[1] ALDER, Michael. *An Introduction to Pattern Recognition*, 2001, 561 p.

[2] ANDĚL, Jiří. *Statistická analýza časových řad.* SNTL, Praha, 1976.

[3] ARLT, Josef – ARLTOVÁ, Markéta. *Ekonomické časové řady: vlastnosti, metody modelování, příklady a aplikace.* Grada, 2007.

[4] ARLT, Josef – ARLTOVÁ, Markéta. *Finanční časové řady: vlastnosti, metody modelování, příklady a aplikace.* Grada, 2003.

[5] BILMES, A. Jeff. *A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models.* International Computer Science Institute, Berkley, USA, 1998.

[6] BISHOP, M. Christopher. *Latent Variables, Mixture Models and EM.* Tutorial, Microsoft Research, Cambridge, 2004.

[7] BISHOP, M. Christopher. *Machine Learning Techniques for Computer Vision*. Tutorial, Microsoft Research, Cambridge, 2004.

[8] BISHOP, M. Christopher. *Pattern Recognition and Machine Learning*. Springer, 2006, 748 p.

[9] BROADWATER, Joshua. Bayesian Information Criterion. *Elements of Statistical Learning*. 2003.

[10] BROCKWELL, Peter J. – DAVIS, A. Richard. *Time Series: Theory and Methods.* Springer-Verlag. 1987.

[11] BROERSEN, Piet M.T. Automatic Autocorrelation and Spectral Analysis. Springer-Verlag, 2006.

[12] CALINON, Sylvain – FLORENT, Guenter – BILLARD, Aude. On Learning, Representing and Generalizing a Task in a Humanoid Robot. *IEEE Transactions on Systems, Man and Cybernetics, Part B.* Special issue on robot learning by observation, demonstration and imitation. Vol. 36, No. 5, 2006.

[13] *Center for Machine Perception* [online]. 2009 [cit. 2009-3-9]. Dostupné z URL: <http://cmp.felk.cvut.cz/cmp/courses/recognition/>.

[14] CIPRA, Tomáš. *Analýza časových řad s aplikacemi v ekonomii.* SNTL, Praha, 1986.

[15] Code for the Edge Detection and Image SegmentatiON system (EDISON). [on-line] Dostupné z URL: <http://www.caip.rutgers.edu/riul/research/code.html> [cit. 2009-8-19]

[16] COMANICIU, Dorin – MEER, Peter. Mean Shift: A Robust Approach Toward Feature Space Analysis. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 5, 2002.

[17] DE MELO, C. O. A. – DE MORAES, M. R. – MACHADO, L. dos S. *Gaussian Mixture Models for Supervised Classification of Remote Sensing Multispectral Images.*

[18] DELLAERT, Frank. *The Expectation Maximization Algorithm.* Technical Report number GIT-GVU-02-20. College of Computing, Georgia Institute of Technology, February 2002.

[19] DEMPSTER, P. A. – LAIRD, M. N. – RUBIN, B. D. *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society. Vol. 39, No. 1. 1977, pp. 1-38.

[20] FRANC, Vojtěch. *EM algoritmus*. Lectures. Center for Machine Perception, FEL ČVUT, Praha.

[21] GRIM, Jiří. *Součinové distribuční směsi.* Ústav teorie informace a automatizace AV ČR. Oddělení rozpoznávání obrazců. WWW: <http://www.utia.cas.cz/RO/>

[22] GU, Y. H. Irene – GUI, Vasile – XU, Zhifei. Video Segmentation Using Joint Space-Time Range Adaptive Mean Shift. *Advances in Multimedia Information Processing - PCM 2006* (2006), pp. 740-748.

[23] HEBÁK, Petr – HUSTOPECKÝ, Jiří – JAROŠOVÁ, Eva – PECÁKOVÁ, Iva. *Vícerozměrné statistické metody [1]*. 2. vydání, Informatorium, Praha 2007, 253 s.

[24] HINDLS, Richard; HRONOVÁ, Stanislava; SEGER, Jan; FISHER, Jakub. *Statistika pro ekonomy.* Osmé vydání, Professional Publishing, 2007.

[25] HLAVÁČ, M. – SEDLÁČEK, M. Zpracování signálů a obrazů, 2. vydání, Praha, Vydavatelství ČVUT, 2005, 255 s., ISBN 80-01-03110-1

[26] HOFMAN, Thomas. *Machine Learning*. Lecture 11. June 2, 2005.

[27] LI, Wentian – NYHOLT, R. Dale. *Marker Selection by AIC and BIC.* Laboratory of Statistical Genetics, The Rockefeller University, New York.

[28] MCLACHLAN, J. Geoffrey – KRISHNAN, Thriyambakam. *The EM Algorithm and Extensions.* A Wiley-Intersience Publication, United States of America, 1997, 274 s.

[29] MOON, T.K. The Expectation-Maximization Algorithm. *Signal Processing Magazine, IEEE*. November 1996, Vol. 13, ISSN: 1053-5888.

[30] NEDOVIĆ, Vladimir. *Tracking moving video objels using mean-shift algorithm*. Project report.

[31] ROGALEWICZ, Vladimír. *Stochastické procesy*. Vydavatelství ČVUT, 1993, 106 s.

[32] ROZINEK, O. *Comparison of Segmentation Methods for Marker Tracking in Video*. POSTER 2009, 13th International Student Conference on Electrical Engineering, Czech Technical University in Prague, 2009. Dostupné z URL: <http://radio.feld.cvut.cz/conf/poster2009/>

[33] SHWARZ, Gideon. Estimating the Dimension of a Model. *The Annals of Statistics*. 1978, Vol. 6, No. 2, 461-464 s.

[34] SONKA, M. – HLAVAC, V. – BOYLE, R. *Image Processing, Analysis, and Machine Vision*. 3rd ed.: Cengage-Engineering, 2007, 864 s.

[35] SVOBODA, T. – KYBIC, J. – HLAVÁČ, V. *Image processing, analysis and machine vision: The MATLAB Companion*. Thomson Learning, Toronto, Canada, 2007, 255 s.

[36] ŠPANĚL, M. *Klasifikace a rozpoznávání*. [on-line] Fakulta informačních technologií, VUT Brno. [cit. 2009-8-20]. Dostupné z URL: <http://www.fit.vutbr.cz/study/courses/IKR/public/prednasky/>

[37] *The Mathworks – MATLAB Central – File Exchange* [online]. 2009 [cit. 2009-3-9]. Dostupné z URL: <http://www.mathworks.com/matlabcentral/fileexchange/>.

[38] *The USC-SIPI Image Database*. [databáze on-line]. Signal and Image Processing Institute, University of Southern California [cit. 2009-8-21]. Dostupné z URL <http://sipi.usc.edu/database>

[39] THEODORIDIS, Sergios – KOUTROUMBAS, Konstantinos. *Pattern Recognition*. Third Edition. Elsevier, 2006, 840 s.

[40] TOMASI, Carlo. *Estimating Gaussian Mixture Densities wit EM – A Tutorial*. Duke University.

[41] *True-color Kodak test images*. [databáze on-line]. [cit. 2009-8-21]. Dostupné z URL: <http://r0k.us/graphics/kodak/>

[42] UHLÍŘ, J.; SOVKA, P.: Číslicové zpracování signálů, 2. vydání, Praha, Vydavatelství ČVUT, 2002, 328 s., ISBN 80-01-02613-2.

[43] UKRAINITZ, Yarin – SAREL, Bernard. *Mean Shift: Theory and Application.*

[44] WATANABE, Michiko – YAMAGUCHI, Kazunori. *The EM Algorithm and Related Statistical Models.* Marcel Dekker, Inc. 2004.

[45] WEBB R. Andrew. *Statistical Pattern Recognition*. Second Edition, 2002, 515 s.

[46] WU, C. F. Jeff. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 1983, Vol. 11, No. 1, s. 95-103.

[47] ZAPLATÍLEK, K. – DOŇAR, B. Matlab – pro začátečníky, Praha, BEN – technická literatura, 2005, 2. vyd., 152 s., ISBN 80-7300-175-6

[48] ZAPLATÍLEK, K. – DOŇAR, B. Matlab – tvorba uživatelských aplikací, Praha, BEN – technická literatura, 2004, 1. vyd., 216 s., ISBN 80-7300-133-0

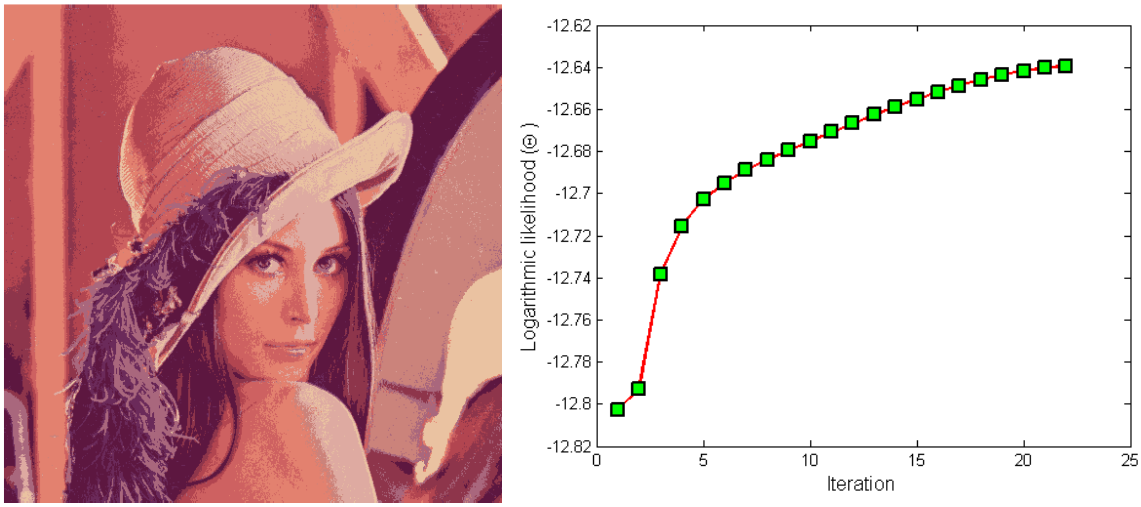
[49] ZHAI, ChengXiang. *A Note on the Expectation-Maximization (EM) Algorithm.* University of Illinois at Urbana-Champaign, November 2, 2004.

[50] ZHANG, Yu-Jin. *Advances in Image and Video Segmentation*. IRM Press, 2006. 473 s.

7 Přílohy

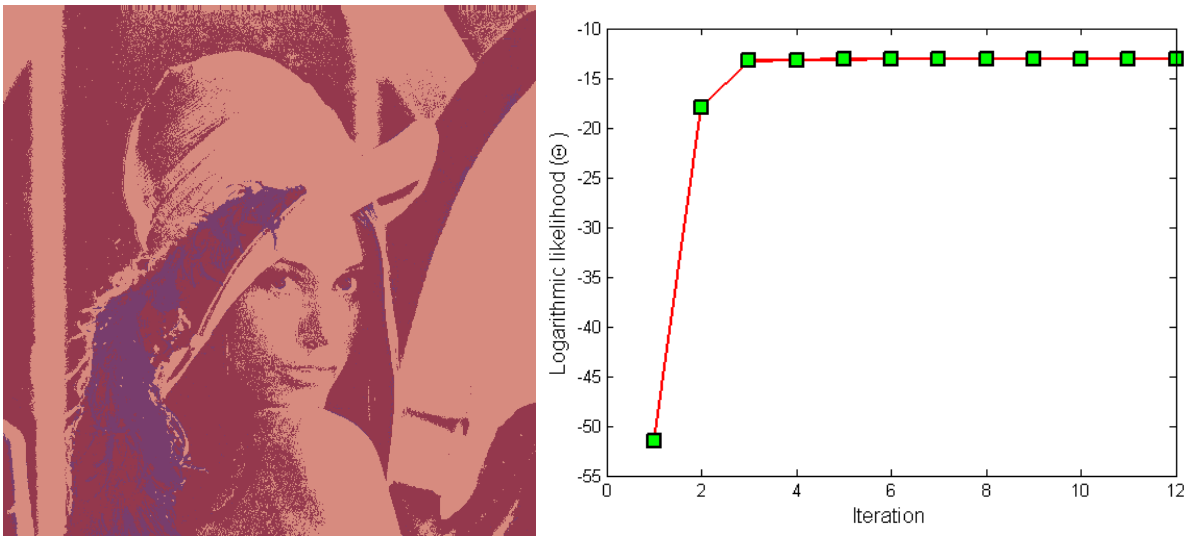
7.1 Výsledky dosažené s EM algoritmem a K-means

a) b)



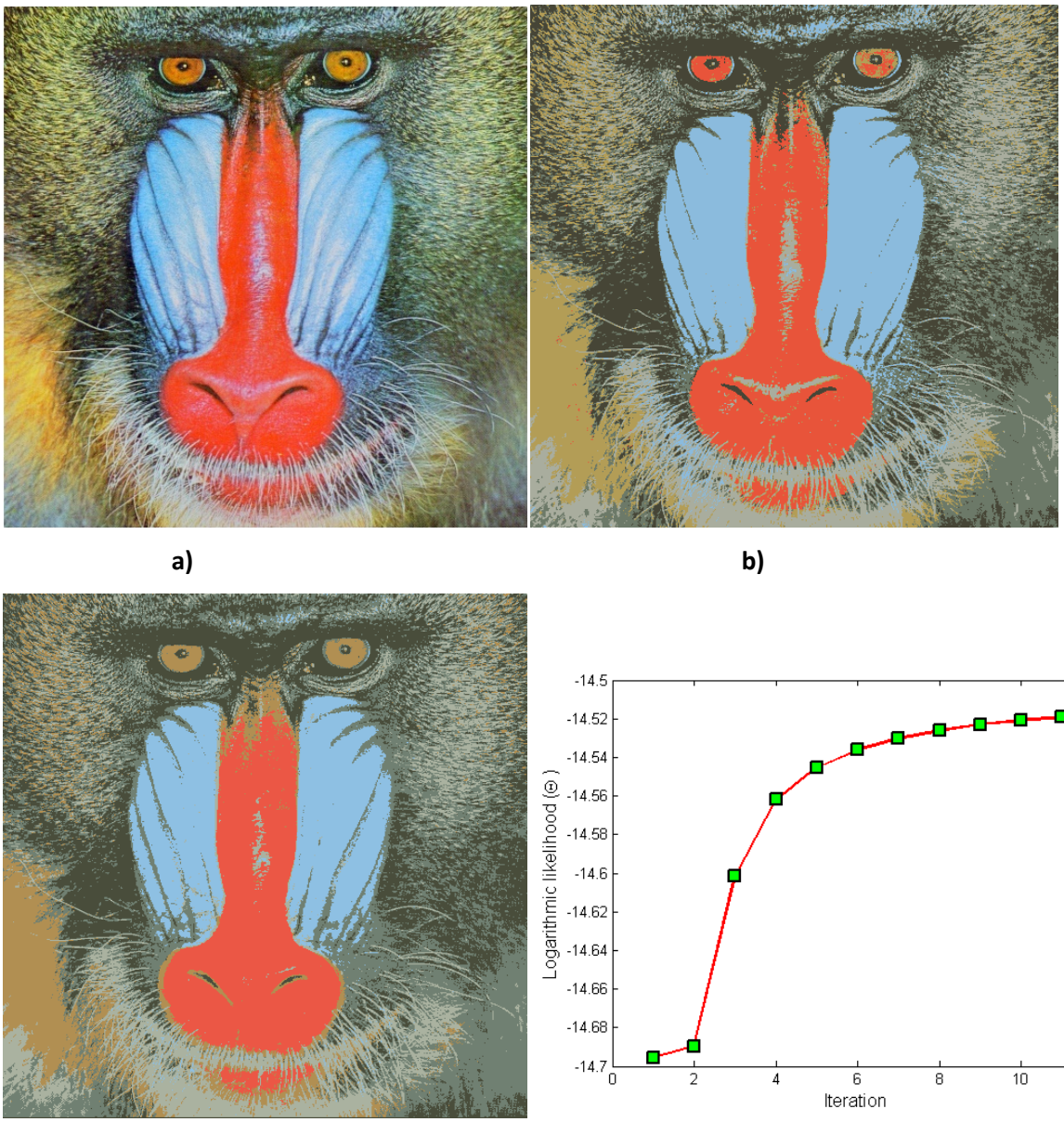
c) d)

Obr. 29: **Shlukování do 10 shluků – pravděpodobnostní model směsi vícerozměrných normálních rozdělení s 10 komponentami a 3 příznaky (R, G, B): a) vstupní původní modelový obraz (Lena Söderberg), b) inicializace parametrů algoritmem K-means, c) EM algoritmus, d) konvergence věrohodnostní funkce**



a) **b)**

Obr. 30: **a) Shlukování EM algoritmem do 3 shluků – pravděpodobnostní model směsi vícerozměrných normálních rozdělení s 3 komponentami a 3 příznaky (R, G, B) s náhodnou inicializací parametrů, b) konvergence věrohodnostní funkce**



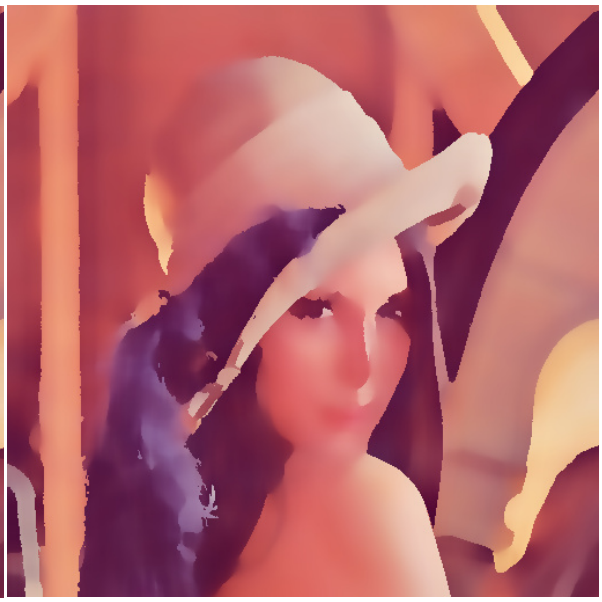
Obr. 31: **Shlukování do 6 shluků – pravděpodobnostní model směsi vícerozměrných normálních rozdělení s 3 komponentami a 3 příznaky (R, G, B) s inicializací parametrů použitím K-means: a) vstupní původní obraz (baboon.tif), b) K-means, c) EM algoritmus, d) konvergence věrohodnostní funkce**

7.2 Výsledky dosažené s mean-shift algoritmem

$h_s = 10, h_r = 40,$ L * a * b, Epanechnikov jádro

$h_s = 8, h_r = 16,$ RGB, normální jádro

$h_s = 10, h_r = 40,$ L * a * b, Epanechnikovo jádro



$h_s = 8, h_r = 16,$ RGB, normální jádro

Původní obraz

$h_s = 15, h_r = 35,$ L * a * b, Epanechnikovo jádro

Původní obraz

$h_s = 20, h_r = 35,$ L * a * b, Epanechnikovo jádro

Původní obraz

$h_s = 20, h_r = 35,$ L * a * b, Epanechnikovo jádro

Původní obraz

$h_s = 15, h_r = 35,$ L*a*b, Epanechnikovo jádro

Obr. 32: Příklady zpracování obrazu mean-shift algoritmem pro Epanechnikovo a normální jádro a pro odlišné barevné prostory a parametry velikosti jádra