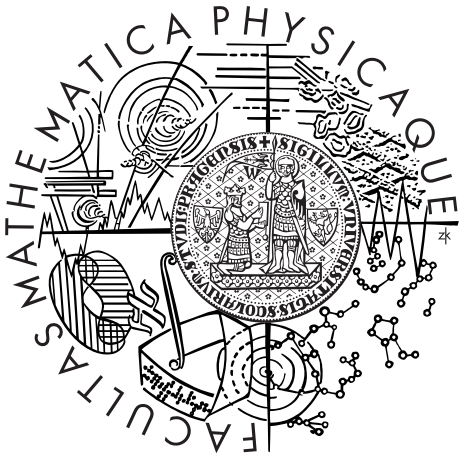
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Bohuš Nemčovič

Odhady v Markovských řetězcích se spojitým časem

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	RNDr. Michaela Prokešová, Ph.D.
Studijní program:	Matematika
Studijní odbor:	Finanční matematika

Praha 2014

V prvom rade by som chcel poďakovať vedúcej mojej bakalárskej práce
RNDr. Michaele Prokešovej, Ph.D. za všetkú jej pomoc, cenné rady a poskytnuté materiály.
Ďalej by som chcel poďakovať RNDr. Jakubovi Staněkovi, Ph.D., ktorý mi veľmi ochotne pomáhal s bakalárskou prácou.
Taktiež ďakujem doc. RNDr. Janovi Hurtovi, CSc. za pomoc s programovaním v Mathematice.
V neposlednom rade ďakujem svojim rodičom za možnosť študovať a za podporu počas celého štúdia.

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 v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis autora

Název práce: Odhady v Markovských řetězcích se spojitým časem

Autor: Bohuš Nemčovič

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: RNDr. Michaela Prokešová, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V této práci se zabýváme odhadováním matic intenzit spojitých Markovských řetězců, v případě, že máme k dispozici úplné pozorování jeho trajektorie a v případě, že pozorujeme řetězec pouze ve vybraných diskrétních časech. Pro získání odhadu používáme metodu maximální věrohodnosti. Ve druhé kapitole nejprve představíme obecný EM algoritmus a následně ho upravíme na hledání odhadu matice intenzity na základě pozorování řetězce v jednotlivých diskrétních časech. V poslední kapitole ukážeme EM algoritmus na numerických příkladech a budeme ilustrovat vliv velikosti diskretizačného kroku na kvalitu odhadu matice intezity.

Klíčová slova: Markovské řetězce, matice intenzity, metoda maximální věrohodnosti, EM algoritmus

Title: Estimation in continuous time Markov chains

Author: Bohuš Nemčovič

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Michaela Prokešová, Ph.D., Department of Probability and Mathematical Statistics

Abstract: In this work we deal with estimating the intensity matrices of continuous Markov chains in the case of complete observation and observation at selected discrete time points. To obtain an estimate we use the maximum likelihood method. In the second chapter we first introduce the general EM algorithm and then adjust it for finding the intensity matrix estimate based on observations at discrete time points. In the last chapter we will illustrate the impact of the discrete step size on the quality of intensity matrix estimate.

Keywords: Markov chains, intensity matrix, maximum likelihood estimation, EM algorithm

Názov práce: Odhady v Markovských reťazcoch so spojitým časom

Autor: Bohuš Nemčovič

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedúci bakalárskej práce: RNDr. Michaela Prokešová, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: V tejto práci sa zaoberáme odhadovaním matíc intenzít spojitých Markovských reťazcov v prípade, že máme k dispozícii úplné pozorovanie jeho trajektórie a v prípade, že pozorujeme reťazec iba vo vybraných diskrétnych časoch. Na získanie odhadu používame metódu maximálnej vierohodnosti. V druhej kapitole najprv predstavíme všeobecný EM algoritmus a následne ho upravíme na hľadanie odhadu matice intenzity na základe pozorovania reťazca v jednotlivých diskrétnych časoch. V poslednej kapitole ukážeme EM algoritmus na numerických príkladoch a budeme ilustrovať vplyv veľkosti diskretizačného kroku na kvalitu odhadu matice intenzity.

Kľúčové slová: Markovské reťazce, matice intenzity, metóda maximálnej vierohodnosti, EM algoritmus

Obsah

Použité značenie **2**

Úvod **3**

1 Odhady matice intenzity **5**
1.1 Maximálne vierohodný odhad matice intenzity pri úplnom pozorovaní 5
1.2 Maximálne vierohodný odhad matice intenzity pri pozorovaní v diskrétnych časových bodoch 7

2 EM algoritmus **10**
2.1 O algoritme 10
2.2 Aplikácia algoritmu na hľadanie odhadu matice intenzity 11
2.3 Implementácia 13

3 Numerické príklady použitia EM algoritmu **15**
3.1 Príklad 1 15
3.2 Príklad 2 19
3.3 Príklad 3 22
3.4 Zhrnutie a záver 25

Literatúra **26**

Zoznam obrázkov **27**

Zoznam tabuliek **28**

Použité značenie

S	Množina všetkých stavov Markovského reťazca
$\mathcal{Q}$	Množina všetkých matíc intenzít
$\mathcal{P}$	Množina všetkých matíc prechodu
$\{X_t, t \geq 0\}$	Spojitý Markovský reťazec
$\{X_{t_k}\}_{k=1}^{n}$	Markovský reťazec s diskrétnym časom
p_{ij}	Pravdepodobnosť prechodu zo stavu i do stavu j po jednom kroku
q_{ij}	Intenzita prechodu zo stavu i do stavu j
$\widehat{\theta}_{MLE}$	Maximálne vierohodný odhad parametra θ
$N_{ij}(t)$	Počet prechodov zo stavu i do stavu j v časovom intervale $[0, t]$
$R_i(t)$	Čas strávený v stave i v časovom intervale $[0,t]$
L_τ^c	Vierohodnostná funkcia pre spojitý Markovský reťazec do času τ
L_n	Vierohodnostná funkcia pre diskrétny Markovský reťazec do času n
$K_{ij}(n)$	Pozorovaný počet prechodov zo stavu i do stavu j v diskrétnom Markovskom reťazci $\{X_{t_1}, \ldots, X_{t_n}\}$ do času n
Θ	Parametrický priestor

Úvod

Markovské reťazce majú využitie ako nástroj na modelovanie rôznych procesov, používajú sa na opis fyzikálnych udalostí, modelovanie vo finančnej sfére alebo napríklad aj v medicíne. V našej práci sa zaoberáme prevažne homogénnymi spojitými Markovskými reťazcami. Spojitý Markovský reťazec definujeme podľa skrípt [1].

Definícia 1. *Systém celočíselných náhodných veličín* $\{X_t, t \geq 0\}$ *definovaných na pravdepodobnostnom priestore* $(\Omega, \mathcal{A}, P)$ *sa nazýva Markovský reťazec so spojitým časom a spočítateľnou množinou stavov* S*, keď*

$$P(X_t = j | X_s = i, X_{t_n} = i_n, \ldots, X_{t_1} = i_1) = P(X_t = j | X_s = i) \tag{1}$$

pre všetky $i,j,i_1,\ldots,i_n \in S$ *a pre všetky* $0 \leq t_1 < t_2 < \ldots < t_n < s < t$*, pre ktoré* $P(X_s = i, X_{t_n} = i_n, \ldots, X_{t_1} = i_1) > 0$.

Spojité markovské reťazce podobne ako u diskrétnych reťazcoch splňujú markovskú vlastnosť, čo znamená, že to ako sa reťazec chová v čase $t+1$ záleží čisto na stave, v ktorom sa reťazec nachádza v čase t a nie na slede udalostí, ktoré ho predchádzali. Spojité Markovské reťazce sú často definované maticou intenzity. Definujme celkovú intenzitu q_i a intenzity prechodu q_{ij} zo stavu i do stavu j podľa Vety 3.3 v [1].

Definícia 2. *Nech* $i,j \in S$ *potom*

$$q_i := \lim_{h \to 0_+} \frac{1 - p_{ii}(h)}{h}$$

pre každé $i,j \in S$*,* $i \neq j$ *potom*

$$q_{ij} := \lim_{h \to 0_+} \frac{p_{ij}(h)}{h}$$

a pre každé $i \in S$ *platí* $\sum_{j \neq i} q_{ij} \leq q_i$.

Kde $p_{ij}(h)$ sa z homogenity reťazca rovná $p_{ij}(s,s+h) = P(X_{s+h} = j | X_s = i)$, čo je pravdepodobnosť prechodu zo stavu i v čase s do stavu j v čase $s+h$. Matica $\boldsymbol{Q} = \{q_{ij}\}_{i,j \in S}$, kde $q_{ii} = -q_i$, sa nazýva matica intenzity. Intenzitám q_{ij} môžeme rozumieť aj ako mieru, ako rýchlo nastane skok zo stavu i do stavu j.

V praxi často chceme zistiť s akou intenzitou nastanú skoky z jedného stavu do druhého. Ale väčšinou nemáme k dispozícii úplné pozorovanie a musíme si poradiť s pozorovaniami v diskrétnom čase. V tom prípade je jednou z možností použit EM algoritmus na zistenie intenzít. V prvej kapitole podrobne odvodíme

odhad matice intenzity pri úplnom pozorovaní a pri pozorovaní v diskrétnych časoch, kde výpočet preberáme z knihy [4]. Použijeme na to metódu maximálnej vierohodnosti. V druhej kapitole stručne opíšeme ako funguje EM algoritmus podľa [5] a ukážeme aplikáciu tohto algoritmu na odhadnutie matice intenzity, ktorú preberáme z [3]. V tretej kapitole sa zameriame na numerické príklady, ktoré budú ilustrovať vplyv veľkosti diskretizačného kroku na kvalitu odhadu matice intenzity. Uvedieme celkovo tri príklady, na ktorých prebehne výpočet. Okrem toho porovnáme aj vplyv veľkosti kroku na konvergenciu algoritmu.

Kapitola 1

Odhady matice intenzity

1.1 Maximálne vierohodný odhad matice intenzity pri úplnom pozorovaní

Majme spojitý homogénny Markovský reťazec $\{X_t, t \geq 0\}$. Pomocou metódy maximálnej vierohodnosti (MLE - anglicky maximum likelihood estimation) odvodíme odhad matice intenzity pri úplnom pozorovaní Markovského reťazca. Predpokladáme, že $S = \{0,1,...,m\}$ je množina stavov Markovského reťazca. Najprv potrebujeme odvodiť vierohodnostnú funkciu. K tomu nám pomôžu tvrdenia z [1] a to konkrétne:

Veta 1. *Ak $q_i = 0$, potom $p_{ii}(t) = 1$ pre všetky $t \geq 0$. Ak je $0 < q_i < \infty$, má doba, po ktorú reťazec zotrvá v stave i, exponenciálne rozdelenie so strednou hodnotou $\frac{1}{q_i}$, kde $i \in S$.*

Veta 2. *Nech $0 < q_i < \infty$, nech τ_i je čas prvého výstupu zo stavu i. Potom*

$$P(X_{\tau_i} = j | X_0 = i) = \frac{q_{ij}}{q_i} \quad i,j \in S, i \neq j, \tag{1.1}$$

tj. pravdepodobnosť, že reťazec z počiatočného stavu i prejde najprv do stavu j je rovné $\frac{q_{ij}}{q_i}$.

Z Vety 1 vieme, že reťazec, ktorý vstúpi do stavu i v určitom čase $t_0 > 0$, zotrvá v i po dobu T, ktorá má exponenciálne rozdelenie s parametrom q_i a hustotou $q_i \exp\{-q_i(t_k - t_{k-1})\}$, kde $(t_k - t_{k-1})$ je doba zotrvania v stave i. Z homogenity reťazca vieme, že rozdelenie doby zotrvania v stave i nezávisí na okamihu, v ktorom reťazec vstúpi do stavu i, ale len na q_i. Doby zotrvania $(t_k - t_{k-1})$ v danom stave sú nezávislé náhodné veličiny. Na odvodenie vierohodnostnej funkcie použijeme aj Vetu 2, podľa ktorej vieme pravdepodobnosť, s ktorou reťazec prejde najprv zo stavu i do stavu j, a tá je rovná $\frac{q_{ij}}{q_i}$.

Vytvorme funkciu:

$$G(t_k) = \begin{cases} 1, & \text{ak } X_t \neq i, \\ q_i \exp\left\{-q_i(t_k - t_{k-1})\right\}, & \text{ak } X_t = i, \end{cases}$$

kde t_k sú označené časy jednotlivých prechodov v Markovskom reťazci, X_t označuje stav Markovského reťazca v čase $t \in (t_k - t_{k-1})$ a $q_i = \sum_{j \neq i} q_{ij}$. Hodnota t_K

označuje čas posledného prechodu Markovského reťazca pred časom τ. Nesmieme zabudnúť ani na hustotu v čase $(\tau - t_K)$, tá je rovná $\exp\{-q_i(\tau - t_K)\}$, pretože v čase τ nastane prechod do iného stavu s pravdepodobnosťou 0. Pomocou funkcie $G(t_k)$ odvodíme vierohodnostnú funkciu s parametrom $\boldsymbol{Q} = \{q_{ij}\}$:

$$L_\tau^{(c)}(\boldsymbol{Q}) = \prod_{i=1}^{m}\prod_{k=1}^{K}\left(G(t_k)\exp\{-q_i(\tau - t_K)\}\right)\prod_{i=1}^{m}\prod_{j\neq i}\left(\frac{q_{ij}}{q_i}\right)^{N_{ij}(\tau)}, \tag{1.2}$$

kde $N_{ij}(t)$ je počet prechodov zo stavu i do stavu j v čase [0,t]. Exponent (c) označuje, že sa jedná o úplné pozorovanie spojitého Markovského reťazca. Označme $R_i(t)$ ako čas strávený v stave i v časovom intervale [0,t]:

$$R_i(t) = \int_0^t I\left\{X_s = i\right\} ds. \tag{1.3}$$

Použijeme fakt, že

$$\exp\left\{-\sum_{j\neq i} q_{ij}R_i(\tau)\right\} = \prod_{j\neq i}\exp\left\{-q_{ij}R_i(\tau)\right\}.$$

Vierohodnostná funkcia po úprave:

$$\begin{aligned}
L_\tau^{(c)}(\boldsymbol{Q}) &= \prod_{i=1}^{m}\prod_{j\neq i} q_i^{N_{ij}(\tau)}\exp\left\{-q_iR_i(\tau)\right\}\left(\frac{q_{ij}}{q_i}\right)^{N_{ij}(\tau)} \\
&= \prod_{i=1}^{m}\exp\left\{-\sum_{j\neq i} q_{ij}R_i(\tau)\right\}\prod_{j\neq i} q_{ij}^{N_{ij}(\tau)} \\
&= \prod_{i=1}^{m}\prod_{j\neq i} q_{ij}^{N_{ij}(\tau)}\exp\left\{-q_{ij}R_i(\tau)\right\}.
\end{aligned}$$

Metódou maximálnej vierohodnosti odhadneme maticu $\boldsymbol{Q}$. Najprv zlogaritmujeme vierohodnostnú funkciu:

$$\begin{aligned}
\log L_\tau^{(c)}(\boldsymbol{Q}) &= \log\left(\prod_{i=1}^{m}\prod_{j\neq i} q_{ij}^{N_{ij}(\tau)}\exp\left\{-q_{ij}R_i(\tau)\right\}\right) \\
&= \sum_{i=1}^{m}\sum_{j\neq i}\log\left(q_{ij}^{N_{ij}(\tau)}\exp\left\{-q_{ij}R_i(\tau)\right\}\right) \\
&= \sum_{i=1}^{m}\sum_{j\neq i}\left(N_{ij}(\tau)\log\left(q_{ij}\right) - q_{ij}R_i(\tau)\right)
\end{aligned}$$

Položíme deriváciu $\log L_\tau^{(c)}(\boldsymbol{Q})$ rovnú 0, aby sme zistili $\tilde{\boldsymbol{Q}}$, teda maximum odhadu.

$$\frac{\partial \log L_\tau^{(c)}(\tilde{\boldsymbol{Q}})}{\partial \tilde{\boldsymbol{Q}}} = 0$$
$$\sum_{i=1}^{m} \sum_{j \neq i} \left(\frac{N_{ij}(\tau)}{\tilde{q}_{ij}} - R_i(\tau) \right) = 0$$

Po úpravách dostávame maximálne vierohodný odhad matice $\boldsymbol{Q}$:

$$\tilde{q}_{ij} = \frac{N_{ij}(\tau)}{R_i(\tau)}, \qquad i \neq j$$
$$\tilde{q}_{ii} = -\sum_{j \neq i} \tilde{q}_{ij}$$

1.2 Maximálne vierohodný odhad matice intenzity pri pozorovaní v dis-krétnych časových bodoch

Bežne avšak nemáme kompletnú informáciu o Markovskom reťazci. Nevieme ako sa proces chová medzi dvoma diskrétnymi pozorovaniami a tým pádom nepoznáme ani hodnoty $R_i(t)$ a $N_{ij}(t)$. Pokiaľ máme k dispozícii diskrétne pozorovania stavov Markovského reťazca v jednotlivých časových bodoch $\{t_1, \ldots, t_n\}$, musíme upraviť vierohodnostnú funkciu. Proces $Y_i = X_{t_i}$ je Markovský reťazec s diskrétnym časom. Budeme predpokladať, že časy, v ktorých pozorujeme stav Markovského reťazca sú ekvidistantné, to znamená $t_k - t_{k-1} = \Delta$ pre všetky $k = 1,2,\ldots,n$. Pretože sa jedná o homogénny Markovský reťazec vieme, že:

$$P(X_{t_k} = j | X_{t_{k-1}} = i) = p_{ij}(\Delta) \quad \text{pre} \quad i,j \in S \quad \text{a} \quad k = 1,2,\ldots,n. \tag{1.4}$$

Ďalej budeme pre jednoduchosť používať značenie $p_{ij} = p_{ij}(\Delta)$. Prenásobením všetkých pravdepodobností tých prechodov, ktoré nastali, získame vierohodnostnú funkciu pre diskrétne pozorovania:

$$L_n(\boldsymbol{P}) = \prod_{i=1}^{m} \prod_{j=1}^{m} p_{ij}^{K_{ij}(n)}, \qquad \boldsymbol{P} \in \mathcal{P} \tag{1.5}$$

kde $\boldsymbol{P} = \{p_{ij}\}_{i,j \in S}$ je matica prechodu Markovského reťazca s rozmermi $m \times m$, $\mathcal{P}$ označuje množinu všetkých matíc prechodu s rozmermi $m \times m$ a $K_{ij}(n)$ je pozorovaný počet prechodov zo stavu i do stavu j do času n. Podobne ako u spojitého Markovského reťazca odvodíme maximálne vierohodný odhad p_{ij}. Uvedomíme si, že naša vierohodnostná funkcia je podľa [3] rovnaká ako vierohodnostná funkcia pre m nezávislých multinomických rozdelení, a keďže $\boldsymbol{P} = \{p_{ij}\}_{i,j \in S}$ je stochastická matica, tak platí $\sum_{j=1}^{m} p_{ij} = 1$ pre $j = 1,2,\ldots,m$. Podľa [4] vypočítame maximálne vierohodný odhad p_{ij}.

Máme

$$\log L_n(\boldsymbol{P}) = \sum_{i=1}^{m}\sum_{j=1}^{m} K_{ij}(n)\log p_{ij}.$$

Keďže nám ide o maximalizáciu funkcie $\sum_{i=1}^{m}\sum_{j=1}^{m} K_{ij}(n)\log p_{ij}$ pri vedľajšej podmienke $\sum_{j=1}^{m} p_{ij} = 1$, použijeme metódu Lagrangeových multiplikátorov. Položíme

$$f(p_{ij},\lambda) = \sum_{i=1}^{m}\sum_{j=1}^{m} K_{ij}(n)\log p_{ij} - \lambda\left(\sum_{j=1}^{m} p_{ij} - 1\right). \tag{1.6}$$

Deriváciou funkcie (1.6) podľa λ a jednotlivých p_{ij} pre $i,j = 1,2,...,m$ a položením derivácie rovnej nule získame sústavu rovníc:

$$\frac{\partial f}{\partial \lambda} = 0, \qquad \frac{\partial f}{\partial p_{ij}} = 0 \quad \text{pre} \quad i,j = 1,2,\ldots,m.$$

Máme

$$\sum_{j=1}^{m} \tilde{p}_{ij} = 1, \qquad \frac{K_{ij}(n)}{\tilde{p}_{ij}} - \lambda = 0 \quad \text{pre} \quad i,j = 1,2,\ldots,m,$$

z čoho výpočtom dostaneme maximálne vierohodný odhad:

$$\tilde{p}_{ij} = \frac{K_{ij}(n)}{K_{i.}(n)} \quad \text{pre} \quad i,j = 1,2,\ldots,m, \tag{1.7}$$

kde

$$K_{i.}(n) = \sum_{j=1}^{m} K_{ij}(n).$$

Nech $\tilde{\boldsymbol{P}} = \{\tilde{p}_{ij}\}$ pre $i,j = 1,2,...,m$. V prípade, že máme odhad $\tilde{p}_{ij}$ môžeme použiť Vety 3.10 a 3.11 z [1], ktoré nám dávajú do súvislosti pravdepodobnosti prechodu a intenzity prechodu. Vieme tým pádom z matice prechodu $\tilde{\boldsymbol{P}}$ zinvertovaním vypočítať odhad matice intenzity $\tilde{\boldsymbol{Q}}$ pomocou nasledujúceho vzorca:

$$\boldsymbol{P} = e^{\boldsymbol{Q}\Delta} \tag{1.8}$$

Definujme množinu matíc prechodu, ktoré sa zhodujú s diskrétnymi pozorovaniami spojitého Markovského procesu:

$$\mathcal{P}_0 = \{\exp(\boldsymbol{Q}) | \boldsymbol{Q} \in \mathcal{Q}\}, \tag{1.9}$$

kde $\mathcal{Q}$ je množina všetkých matíc intenzít. Predpokladajme, že vypočítame $\tilde{\boldsymbol{P}}$ podľa rovnice (1.7) založené na diskrétnych pozorovaniach spojitého Markovského reťazca. Ak $\tilde{\boldsymbol{P}} \in \mathcal{P}_0$ potom existuje $\tilde{\boldsymbol{Q}} \in \mathcal{Q}$ také, že spĺňuje $\exp(\tilde{\boldsymbol{Q}}\Delta) = \tilde{\boldsymbol{P}}$ a vierohodnostná funkcia nadobúda svoju maximálnu hodnotu v $\tilde{\boldsymbol{Q}}$, čo je tým pádom maximálne vierohodný odhad. Avšak situácia je zložitejšia. Jednak množina $\mathcal{P}_0$ je veľmi komplikovaná a navyše matica exponenciály nie je prostá funkcia na celom svojom definičnom obore. Výpočet $\tilde{\boldsymbol{Q}}$ pomocou (1.8) preto nemusí byť jednoznačný a teda nemusí sa jednať o maximálne vierohodný odhad. Ak $\tilde{\boldsymbol{P}} \notin \mathcal{P}_0$, situácia je nejasná kvôli komplikovanej štruktúre $\mathcal{P}_0$. Bližšia diskusia ohľadom existencie a jednoznačnosti maximálneho vierohodného odhadu sa nachádza v článku [3]. My sa touto cestou nevydáme, namiesto toho použijeme EM algoritmus v ďalšej kapitole na výpočet odhadu matice intenzity $\tilde{\boldsymbol{Q}}$.

Kapitola 2

EM algoritmus

2.1 O algoritme

V nasledujúcej kapitole stručne opíšeme všeobecný EM algoritmus a potom ho aplikujeme na náš konkrétny problém. EM (anglicky Expectation-Maximization) algoritmus je iteračná metóda na hľadanie odhadov parametrov maximálnej vierohodnosti z množiny dát, ktorá je neúplná alebo má chýbajúce hodnoty. Majme k dispozícii pozorované (neúplné) data y, s odpovedajúcou hustotou $p(y|\theta)$, ktoré čiastočne opisujú úplné data x s hustotou $p(x|\theta)$ (viď [5]).

Našim cieľom bude nájsť maximálne vierohodný odhad θ.

$$\widehat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} p(y|\theta) \tag{2.1}$$

Θ označuje parametrický priestor. Často je jednoduchšie vypočítať θ, ktoré maximalizuje log-vierohodnostnú funkciu.

$$\widehat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} \log p(y|\theta) \tag{2.2}$$

Keďže logaritmus je monotónna rastúca funkcia, (2.1) má rovnaké riešenie ako (2.2). Avšak niekedy je ťažké vypočítať (2.2) i (2.1). Vtedy môžeme skúsiť EM algoritmus. Na to, aby sme ho mohli použiť potrebujeme mať nejaké napozorované dáta y s hustotou $p(y|\theta)$, určitý opis kompletných dát x, ktoré chceme mať s ich hustotou $p(x|\theta)$. Predpokladáme, že kompletné dáta môžu byť modelované ako spojitý náhodný vektor X s hustotou $p(x|\theta)$, kde $\theta \in \Theta$. Najprv spravíme odhad kompletných dát X, následne zmaximalizujeme očakávanú log-vierohodnostnú funkciu X podľa θ a získame nový odhad, vďaka ktorému môžeme lepšie odhadnúť dáta X. Ďalej pokračujeme iterovaním.

Podľa [5] môžeme EM algoritmus zhrnúť v piatich krokoch:

Krok 1: Nech $m = 0$ a zvoľme prvotný odhad $\theta^{(m)}$ pre θ.
Krok 2: Máme napozorované dáta y a predpokladajme, že odhad $\theta^{(m)}$ je správny. Zformulujme podmienenú pravdepodobnostnú hustotu $p(x|y,\theta^{(m)})$ pre úplné data x.
Krok 3: Pomocou podmienenej pravdepodobnostnej hustoty $p(x|y,\theta^{(m)})$ z kroku 2 zformulujeme podmienenú očakávanú log-vierohodnostnú funkciu nazývanú Q-funkciou:

$$Q(\theta|\theta^{(m)}) = \int_{\chi(y)} \log p(x|y,\theta^{(m)})\,\mathrm{d}x \tag{2.3}$$

$$= E_{X|y,\theta^{(m)}} \log p(X \mid \theta), \tag{2.4}$$

kde $\chi(y)$ je uzáver množiny $\{x \mid p(x|y,\theta) > 0\}$ a predpokladáme, že $\chi(y)$ nezávisí na θ.
Krok 4: Nájdeme θ, ktoré maximalizuje Q-funkciu z (2.3). Výsledkom je nový odhad $\theta^{(m+1)}$.
Krok 5: Nech $m := m + 1$ a pokračujme znova od kroku 2. EM algoritmus pokračuje iterovaním, až kým pre nejaké stanovené $\epsilon > 0$ nastane
$\| \theta^{(m+1)} - \theta^{(m)} \| < \epsilon$, viď [5]. Zvolením ϵ určíme kedy sa algoritmus zastaví.

Podľa [5], EM algoritmus zaručuje, že odhad $\theta^{(m+1)}$ nikdy nebude menej pravdepodobný ako odhad $\theta^{(m)}$. Algoritmus zvyčajne nájde maximum, ale nezaručuje, že sa bude jednať o globálne maximum. Jednoducho sa dá EM algoritmus rozložiť na dve časti **E-krok**(anglicky expectation), patria sem kroky 2 a 3 a **M-krok**(anglicky maximization, krok 4).

E-krok: Pomocou odhadu $\theta^{(m)}$ z predošlej iterácie algoritmu spočítame $Q(\theta|\theta^{(m)})$ z (2.3).
M-krok: Odhad $(m + 1)$ získame maximalizáciou Q-funkcie podľa parametru θ:

$$\theta^{(m+1)} = \arg\max_{\theta\in\Theta} Q(\theta|\theta^{(m)}). \tag{2.5}$$

Keďže E-krok je počítanie Q-funkcie, ktorá sa použije v M-kroku, EM algoritmus môžeme charakterizovať ako iteračný proces počítajúci M-krok. Pri implementovaní EM algoritmu je dobré vziať na vedomie, že nie je treba zakaždým počítať časti Q-funkcie, ktoré nezávisia na θ. Urýchli sa tým výpočtový proces.

2.2 Aplikácia algoritmu na hľadanie odhadu matice intenzity

Pre spojitý Markovský reťazec $\{X_t, 0 \le t \le \tau\}$, ktorý pozorujeme v diskrétnych časoch $t_i, i = 1,...,n$, je vhodné použiť EM algoritmus na hľadanie maximálne vierohodného odhadu matice intenzity. Nech $t_1 = 0$, $t_n = \tau$, kde τ je čas, v ktorom poslednýkrát pozorujeme v akom stave sa reťazec nachádza a $\boldsymbol{Q}_0$ je daná matica intenzít. Podľa **E-kroku** treba nájsť

$$\begin{aligned}\mathbb{E}_{\boldsymbol{Q}_0}[\log L_\tau^{(c)}(\boldsymbol{Q})|\boldsymbol{Y} = y] = \sum_{i=1}^{m}\sum_{j\neq i} \log(q_{ij})\mathbb{E}_{\boldsymbol{Q}_0}[N_{ij}(\tau)|\boldsymbol{Y} = y] - \\ - \sum_{i=1}^{m}\sum_{j\neq i} q_{ij}\mathbb{E}_{\boldsymbol{Q}_0}[R_i(\tau)|\boldsymbol{Y} = y],\end{aligned} \tag{2.6}$$

kde $\boldsymbol{Y} = \{Y_i|i = 1,...,n\}$ a zložky vektora $\boldsymbol{Y}$ sú pozorované stavy Markovského reťazca v jednotlivých časoch $Y_i = X_{t_i}$. Jedná sa o očakávanú hodnotu log-

vierohodnostnej funkcie, ktorú maximalizujeme (**M-krok**) ako funkciu s parametrom $\boldsymbol{Q}$. Z markovskej vlastnosti a homogenity reťazca nám stačí podľa [3] nájsť očakávanú hodnotu $R_k(t)$ a očakávanú hodnotu $N_{kl}(t)$ za podmienky, že v čase 0 proces začína v stave i a v čase t je v stave j. Označíme si ich

$$\tilde{M}_{ij}^k(t) = \mathbb{E}_{\boldsymbol{Q}_0}[R_k(t)|X_t = j, X_0 = i], \tag{2.7}$$

$$\tilde{f}_{ij}^{kl}(t) = \mathbb{E}_{\boldsymbol{Q}_0}[N_{kl}(t)|X_t = j, X_0 = i], \tag{2.8}$$

Stačí vypočítať $\tilde{M}_{ij}^k(t)$ a $\tilde{f}_{ij}^{kl}(t)$, pretože platí

$$\mathbb{E}_{\boldsymbol{Q}_0}[N_{ij}(t)|\boldsymbol{Y} = y] = \sum_{k=1}^{n-1} \tilde{f}_{y_k,y_{k+1}}^{ij}(t_{k+1} - t_k), \tag{2.9}$$

$$\mathbb{E}_{\boldsymbol{Q}_0}[R_l(\tau)|\boldsymbol{Y} = y] = \sum_{k=1}^{n-1} \tilde{M}_{y_k,y_{k+1}}^{l}(t_{k+1} - t_k). \tag{2.10}$$

Uvedieme zjednodušený výpočet (2.7) a (2.8). Detailnejší postup je popísaný v [3]. Zvolíme $\lambda \geq \max_{i=1,\dots,m}(-q_{ii})$ a definujme maticu $\boldsymbol{B}$, ktorú použijeme pri výpočte $M_{ij}^k(t)$.

$$\boldsymbol{B} = \boldsymbol{I} + \frac{1}{\lambda}\boldsymbol{Q} = \frac{1}{\lambda}(\lambda\boldsymbol{I} + \boldsymbol{Q}),$$

kde $\boldsymbol{I}$ je jednotková matica. Takto definovaná matica $\boldsymbol{B}$ je stochastická matica. Označme $\boldsymbol{M}^k = \{M_{ij}^k(t)\}_{i,j\in S}$, kde

$$M_{ij}^k(t) = \mathbb{E}_{\boldsymbol{Q}_0}[R_k(t)I\{X_t = j\}|X_0 = i], \tag{2.11}$$

potom

$$\boldsymbol{M}^k(t) = \exp(-\lambda t)\lambda^{-1} \sum_{n=0}^{\infty} \frac{(\lambda t)^{n+1}}{(n+1)!} \sum_{l=0}^{n} \boldsymbol{B}^l(\boldsymbol{e}_k\boldsymbol{e}_k^T)\boldsymbol{B}^{n-l}, \tag{2.12}$$

kde $\boldsymbol{e}_k$ je jednotkový vektor s k-tou súradnicou rovnou 1 a $\boldsymbol{e}_k^T$ je jeho transpozícia. $\tilde{M}_{ij}^k(t)$ vypočítame podľa vzorca z [3]:

$$\tilde{M}_{ij}^k(t) = \frac{M_{ij}^k(t)}{\boldsymbol{e}_i{}^T \exp(\boldsymbol{Q}t)\boldsymbol{e}_j}. \tag{2.13}$$

Pomocou uniformizačnej metódy (viď [6]) môžeme vypočítať maticu exponenciály:

$$\exp(\boldsymbol{Q}t) = \exp(-\lambda t\boldsymbol{I} + \lambda t\boldsymbol{B}) = \sum_{n=0}^{\infty} e^{-\lambda t}\frac{(\lambda t)^n}{n!}\boldsymbol{B}^n.$$

Podobne postupujme aj pri výpočte (2.8). Najprv uvažujme:

$$f_{ij}^{kl}(t) = \mathbb{E}_{\boldsymbol{Q}_0}[N_{kl}(t)I\{X_t = j\}|X_0 = i]. \tag{2.14}$$

Označme $\boldsymbol{f}^{kl} = \{f_{ij}^{kl}(t)\}_{ij\in S}$, potom

$$\boldsymbol{f}^{kl}(t) = q_{kl}\exp(-\lambda t)\lambda^{-1}\sum_{n=0}^{\infty}\frac{(\lambda t)^{n+1}}{(n+1)!}\sum_{j=0}^{n}\boldsymbol{B}^j(\boldsymbol{e}_k\boldsymbol{e}_l^T)\boldsymbol{B}^{n-j}. \tag{2.15}$$

Hodnotu $\tilde{f}_{ij}^{kl}(t)$ dostaneme pomocou vzorca:

$$\tilde{f}_{ij}^{kl}(t) = \frac{f_{ij}^{kl}(t)}{\boldsymbol{e}_i{}^T\exp(\boldsymbol{Q}t)\,\boldsymbol{e}_j}. \tag{2.16}$$

EM algoritmus na hľadanie maximálne vierohodného odhadu matice intenzity $\tilde{\boldsymbol{Q}}$ môžeme zhrnúť nasledovne. Nech $\boldsymbol{Q}_0$ je ktorákoľvek matica intenzity pre Markovský reťazec s množinou stavou S. Na začiatku zvoľme $\boldsymbol{Q} := \boldsymbol{Q}_0$ ako náš prvotný odhad.

Krok 1: Spočítame $\tilde{M}_{y_i,y_{i+1}}^{k}(t_{i+1} - t_i)$ a $\tilde{f}_{y_i,y_{i+1}}^{kl}(t_{i+1} - t_i)$ pre všetky k,$l \in S$ príslušné modelu s maticou intenzity $\boldsymbol{Q}$ podľa (2.13) a (2.16).
Krok 2: Spočítame $\mathbb{E}_{\boldsymbol{Q}}[N_{ij}(t)|\boldsymbol{Y} = y]$ a $\mathbb{E}_{\boldsymbol{Q}}[R_i(\tau)|\boldsymbol{Y} = y]$ (**E-krok**) pomocou (2.9) a (2.10).
Krok 3: Určíme nový odhad $\tilde{\boldsymbol{Q}}_{ij} = \mathbb{E}_{\boldsymbol{Q}}[N_{ij}(t)|\boldsymbol{Y} = y]/\mathbb{E}_{\boldsymbol{Q}}[R_i(\tau)|\boldsymbol{Y} = y]$ pre všetky $i \neq j$, čo je vlastne maximalizácia (**M-krok**) matice $\boldsymbol{Q}$.
Krok 4: S novým odhadom $\boldsymbol{Q} := \tilde{\boldsymbol{Q}}$ pokračujeme iteračne znova od kroku 1, až kým $|\boldsymbol{Q}_{k+1} - \boldsymbol{Q}_k| \leq \epsilon$, kde ϵ je nami dopredu stanovená tolerancia a $\boldsymbol{Q}_0$,$\boldsymbol{Q}_1$, $\boldsymbol{Q}_2$,... je postupnosť matíc intenzit získaných EM algoritmom.

Platí, že $L_n(\boldsymbol{Q}_{k+1}) \geq L_n(\boldsymbol{Q}_k)$ pre k = 0,1,2,.... Podľa [3] je vhodné zvoliť prvotný odhad matice $\boldsymbol{Q}_0$ tak, aby $\det\{\exp(\boldsymbol{Q}_k)\}$ bol ďaleko od 0. Ak zvolíme prvotný odhad tak, že pre nejaké i,$j \in S$ je $(\boldsymbol{Q}_0)_{ij} = 0$, potom očakávaný počet preskokov z i do j zostane rovný 0 cez všetky iterácie algoritmu.

2.3 Implementácia

Na naprogramovanie simulácii Markovských reťazcov a výpočtov odhadu matíc intenzít sme použili program Wolfram Mathematica 9.0. Počas výpočtu je dôležité nepočítať zbytočne výpočtovo náročné hodnoty $\tilde{M}_{y_i,y_{i+1}}^{k}(t_{i+1}-t_i)$ a $\tilde{f}_{y_i,y_{i+1}}^{kl}(t_{i+1}-t_i)$

viackrát. Vhodné je si výpočty $\boldsymbol{M}^k(t)$, $\boldsymbol{f}^{kl}(t)$ a $\exp(\boldsymbol{Q}t)$ uložiť ako maticu a k jednotlivým prvkom následne opakovane pristupovať. V opačnom prípade budeme mať algoritmus s neprijateľnou zložitosťou. Pre urýchlenie výpočtu sme sumy (2.12) a (2.15) počítali iba do určitého n, pri ktorom sa nám hodnota sumy už prakticky nemenila. Pri výpočte je takisto vhodné zvoliť prvotný odhad $\boldsymbol{Q}_0$ tak, aby matica exponenciály $\boldsymbol{Q}_0$ obsahovala iba reálne čísla, inak pravdepodobne náš algoritmus nezvládne vypočítať odhad matice intenzity. Ako prvotný odhad sa nám osvedčilo použiť maticu intenzity, ktorá má mimo diagonály samé jednotky. Výpočtová zložitosť nášho algoritmu je pre 1 iteráciu podľa [7] rovná $O(r \cdot d^5)$, kde r je počet vybraných diskrétnych časov t_k a d je počet stavov Markovského reťazca.

Kapitola 3

Numerické príklady použitia EM algoritmu

V tejto kapitole budeme na troch numerických príkladoch ilustrovať vplyv dĺžky času simulácie spojitého Markovského reťazca a zároveň vplyv veľkosti diskretizačného kroku na kvalitu odhadu matice intenzity. Pozrieme sa tiež na rýchlosť konvergencie EM algoritmu. Naše dáta budú pochádzať zo simulácii spojitých Markovských reťazcov $\{X_t,0 \leq t \leq T\}$ s piatimi, štyrmi a šiestimi stavmi, ktoré obsahujú kompletnú informáciu o chovaní Markovského reťazca. Z týchto reťazcov získame nekompletné pozorovania $\boldsymbol{Y} = \{y_0 = X_{t_0}, \ldots, y_N = X_{t_N}\}$ tým, že na základe daných diskretizačných krokoch $t_{k+1} - t_k = \Delta_j$ vyberieme, v ktorom stave sa v danom čase nachádza Markovský reťazec. Δ_j sú rôzne diskretizačné kroky pre $j = 1,2,\ldots$. Postupne aplikujeme EM algoritmus na simulované Markovské reťazce s danou maticou intenzity $\boldsymbol{Q}$ pre časy $T = 100, 250, 500, 10^3, 10^4, 10^5$ a u jednotlivých časoch použijeme rôzne diskretizačné kroky, konkrétne v prvom príklade $t_{k+1} - t_k = \{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}\}$, v druhom $t_{k+1} - t_k = \{1, \frac{1}{10}, \frac{1}{100}\}$ a posledného $t_{k+1} - t_k = \{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}\}$. Zdôvodníme aj prečo sme ich takto vybrali. Pre jednotlivé príklady zvolíme na začiatku prvotný odhad $\boldsymbol{Q}_0$ pre všetky diskretizačné kroky. EM algoritmus necháme bežať dovtedy, kým maximálna zmena prvkov matice $\boldsymbol{Q}_{k+1}$ oproti matici predchádzajúcej iterácie $\boldsymbol{Q}_k$ bude väčšia ako $\epsilon = 10^{-6}$ alebo počet iterácii algoritmu dosiahne hodnotu 500. Zvolili sme toto obmedzenie, pretože sa môže stať, že EM algoritmus nebude konvergovať alebo počet iterácii nutný na konvergenciu bude neakceptovateľne veľký. Obzvlášť oceníme túto voľbu pri použití diskretizačného kroku $t_{k+1} - t_k = 1$, kde nám algoritmus často nekonverguje ani po 1000 iteráciách.

3.1 Príklad 1

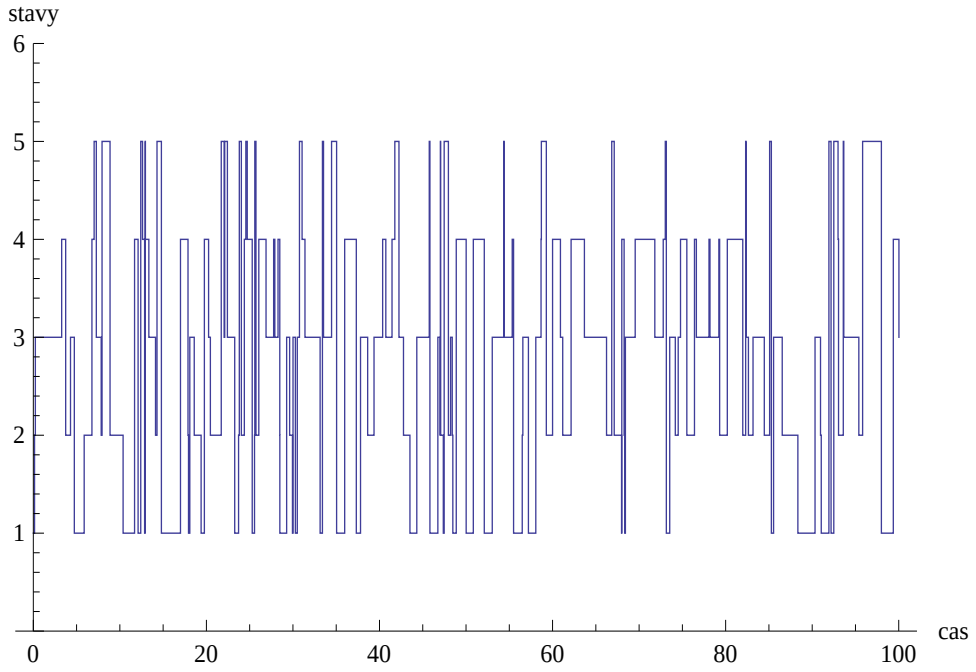
Uvažujme maticu intenzity $L1$:

$$L_1 = \begin{pmatrix} -1.5 & 0.2 & 0.8 & 0.25 & 0.25 \\ 0.5 & -2 & 0.7 & 0.3 & 0.5 \\ 0.1 & 0.6 & -1 & 0.2 & 0.1 \\ 0.4 & 0.2 & 0.7 & -1.5 & 0.2 \\ 0.5 & 0.9 & 0.8 & 0.5 & -2.7 \end{pmatrix}. \tag{3.1}$$

Nasimulujeme priebeh spojitého Markovského reťazca pre interval [0,100] s ma-

ticou intenzity L_1, viď Obr.3.1. Pre väčšie časové realizácie sú obrázky neprehľadné a preto ich nebudeme zobrazovať.

Obr. 3.1: Simulácia spojitého Markovského reťazca s maticou intenzity L_1 v čase $T = 100$.



Naše dáta sú stavy tohto procesu v ekvidistančných časoch s diskretizačným krokom $t_{k+1} - t_k = 1$. Označme $\hat{\boldsymbol{Q}}_{EM}(T,t_{k+1} - t_k)$ maticu, ktorá vznikla po aplikácii EM algoritmu na dáta s časovou realizáciou T a diskretizačným krokom $t_{k+1} - t_k$. EM algoritmus nám nekonvergoval ani po 500 iteráciách, preto sme výpočet ukončili a výsledok po 500 iteráciách, ktoré trvali 498 sekúnd, bol nasledovný:

$$\hat{\boldsymbol{Q}}_{EM}(100,1) = \begin{pmatrix} -0.82289 & 0.00977 & 0.30954 & 3.768 \times 10^{-18} & 0.50358 \\ 0.58449 & -1.40824 & 0.54508 & 0.27868 & 1.627 \times 10^{-39} \\ 0.08306 & 0.55255 & -0.983 & 0.24247 & 0.10492 \\ 4.399 \times 10^{-6} & 1.15394 & 0.96177 & -2.46576 & 0.35004 \\ 0.60824 & 5.446 \times 10^{-10} & 0.73691 & 0.86952 & -2.21466 \end{pmatrix}.$$

Keď porovnáme výslednú maticu s maticou (3.1) vidíme, že EM algoritmus neaproximuje maticu intenzity ani zďaleka. Chyba počítaná cez euklidovu normu $\parallel L_1 - \hat{Q}_{EM}(100,1) \parallel$ je 1.8064. Hodnoty chyby odhadu, počítanú cez euklidovu normu pre časy $T = 100, 250, 500, 10^3, 10^4, 10^5$ a diskretizačné kroky 1, $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}$ nájdeme v Tabuľke 3.1.

Z hodnôt v tabuľke vyplýva, že pre čas $T = 100$ sa so zjemňujúcim sa diskretizačným krokom zmenšuje chyba odhadu, avšak trend nie je spočiatku úplne jednoznačný. Konkrétne pre diskretizačný krok $\frac{1}{8}$ a čas $T = 100$ nám algoritmus konverguje po 29 iteráciách, ale chyba oproti pôvodnej matici je väčšia ako pri diskretizačnom kroku $\frac{1}{4}$, a to 1.04 oproti 0.907. Avšak v dostatočnom počte dát, napríklad pre časy $T = 10^4, 10^5$, je trend jasne klesajúci. Matica pre $T = 100$ a $t_{k+1} - t_k = \frac{1}{8}$ vyšla viď (3.2).

Tabuľka 3.1: Závislosť dĺžky časovej realizácie T a veľkosti diskretizačných krokov na veľkosť chyby počítanej cez euklidovu normu $\| L_1 - \hat{Q}_{EM}(T,t_{k+1} - t_k) \|$.

	Diskretizačný krok			
T	1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$
100	1.806	1.250	0.907	1.040
250	1.040	0.764	0.706	0.575
500	2.073	0.733	0.562	0.526
1 000	2.153	0.523	0.594	0.405
10 000	0.337	0.219	0.162	0.108
100 000	0.124	0.055	0.050	0.043

$$\hat{\boldsymbol{Q}}_{EM}(100,\frac{1}{8}) = \begin{pmatrix} -0.91045 & 0.03815 & 0.51095 & 0.05769 & 0.30367 \\ 0.80226 & -2.20197 & 0.62996 & 0.48848 & 0.28127 \\ 0.09248 & 0.63594 & -1.32178 & 0.33769 & 0.25567 \\ 0.44212 & 0.45466 & 1.22981 & -2.19496 & 0.06837 \\ 0.22628 & 1.34485 & 0.53097 & 0.81462 & -2.91672 \end{pmatrix}. \tag{3.2}$$

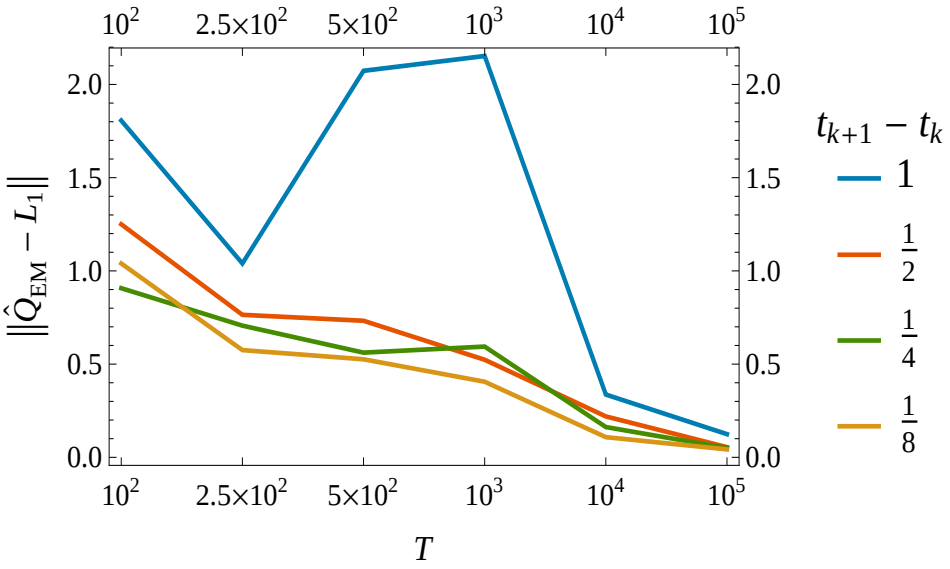
Od pôvodnej matice intenzity má však takisto ďaleko. Môže to byť spôsobené tým, že počet dát pre $T = 100$ nie je dostačujúci ani pre približnú aproximáciu, pretože daná realizácia Markovského reťazca neobsahuje dostatok informácii v dátach pre presnú aproximáciu pomocou EM algoritmu. Naopak pre $T = 10^5$ a diskretizačný krok $\frac{1}{8}$ nám vyšla chyba odhadu iba 0.043, čo naznačuje, že náš odhad bude veľmi presný. EM algoritmus konvergoval po 16 iteráciách, ktoré trvali približne 110 minút. Matica tohto odhadu vyzerá následovne:

$$\hat{\boldsymbol{Q}}_{EM}(10^5,\frac{1}{8}) = \begin{pmatrix} -1.50676 & 0.19823 & 0.81109 & 0.24913 & 0.2483 \\ 0.50884 & -2.02733 & 0.71623 & 0.30367 & 0.4986 \\ 0.097155 & 0.60977 & -1.00714 & 0.20313 & 0.09708 \\ 0.3923 & 0.20153 & 0.69534 & -1.48374 & 0.19458 \\ 0.48135 & 0.89586 & 0.79247 & 0.49463 & -2.66431 \end{pmatrix}. \tag{3.3}$$

Keď porovnáme maticu (3.3) s maticou (3.1), vidíme, že naša matica aproximuje pôvodnú maticu intenzity celkom presne. Grafické znázornenie závislosti kvality odhadu, dĺžky času a veľkosti diskretizačného kroku, v ktorom sme simulovali Markovský reťazec môžeme vidieť na Obr. 3.2.

Z obrázka vidíme, že na kvalitu odhadu matice intenzity má najväčší vplyv počet empirických pozorovaní. Po pár spustených EM algoritmoch sme zistili, že počet empirických pozorovaní a takisto počet stavov Markovského reťazca ovplyvňujú rýchlosť konvergencie. Na porovnanie rýchlosti konvergencie sa nám bude hodiť nasledujúca tabuľka (Tabuľka 3.2), kde pre jednotlivé časy a diskretizačné kroky je počet iterácií, pri ktorom EM algoritmus konverguje a vedľa toho čas, za ktorý sa tak stalo.

Obr. 3.2: Veľkosť chyby odhadu matice intenzity v závislosti na diskretizačnom kroku a dĺžke časovej realizácie Markovského reťazca.

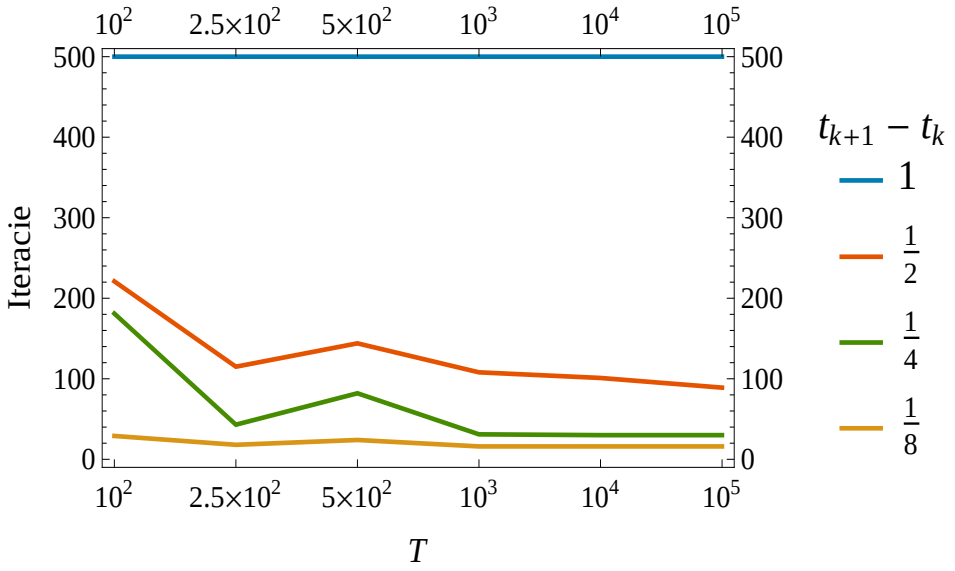


Tabuľka 3.2: Závislosť dĺžky časovej realizácie T a veľkosti diskretizačných krokov na rýchlosť konvergercie z hľadiska počtu iterácií a času trvania algoritmu v sekundách.

	Diskretizačný krok							
	1		$\frac{1}{2}$		$\frac{1}{4}$		$\frac{1}{8}$	
T	Iterácie	Čas	Iterácie	Čas	Iterácie	Čas	Iterácie	Čas
100	500	498s	221	229s	181	216s	29	44s
250	500	541s	115	142s	43	65s	18	40s
500	500	573s	144	206s	82	155s	24	72s
1 000	500	653s	108	208s	31	90s	16	80s
10 000	500	2 092s	101	1 090s	30	651s	16	718s
100 000	500	16 255s	89	8 814s	30	5 781s	16	6 631s

Z Tabuľky 3.2 vyplýva, že rýchlosť konvergencie silno závisí na počte dát a diskretizačnom kroku. Obzvlášť je zaujímavé pozorovanie, ako so zjemňujúcim sa diskretizačným krokom sa predlžuje dĺžka výpočtu jednej iterácie, avšak počet iterácií nutný na konvergenciu EM algoritmu sa naopak znižuje. Takisto môžeme z Tabuľky 3.2 usudzovať, že so zvyšujúcim sa počtom dát klesá počet iterácií nutný na konvergenciu EM algortimu. Celkový čas výpočtu sa so zjemňujúcim sa diskretizačným krokom zvyčajne zmenšuje, ale podľa dát v tabuľke sa tak deje iba do určitého kroku. Pre časy $T = 10^4, 10^5$ sa od kroku $\frac{1}{4}$ opäť zvyšuje. Pri výpočtoch s $t_{k+1} - t_k = 1$ nám algoritmus nekonvergoval ani po 500 iteráciach. Na čo najrýchlejšiu konvergenciu algoritmu je z údajov v tabuľke vhodné použiť menší diskretizačný krok. Ďalej si môžeme všimnúť, ako sa až na čas $T = 500$, so zvyšujúcim sa časom zmenšuje počet iterácií nutný na konvergenciu algoritmu. Môžeme to vidieť aj na Obr. 3.3.

Obr. 3.3: Počet iterácií nutný na konvergenciu algoritmu v závislosti na diskretizačnom kroku a dĺžke časovej realizácie Markovského reťazca.



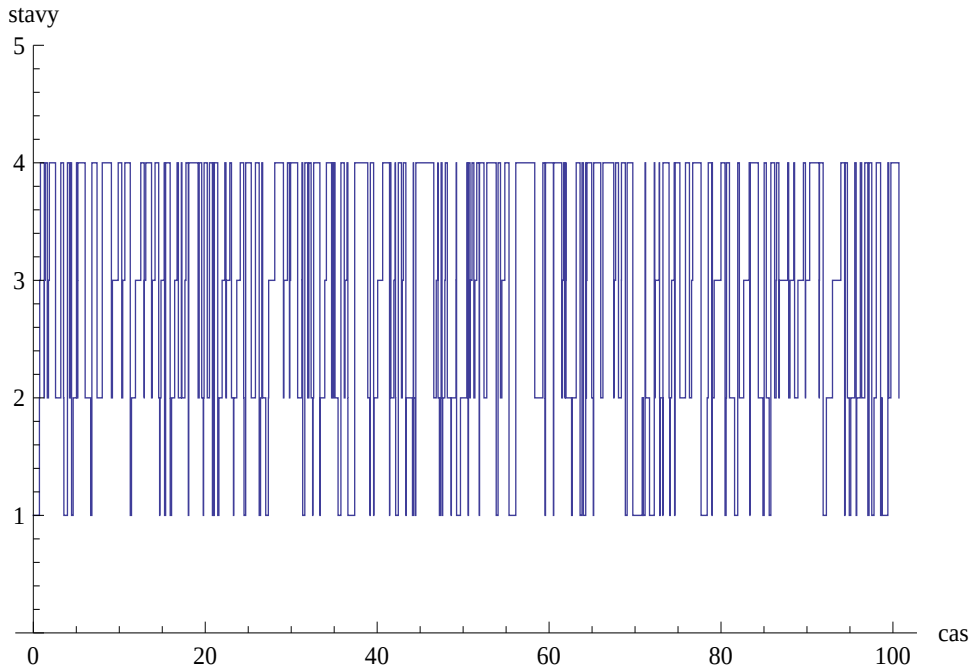
3.2 Príklad 2

V tejto sekcii ukážeme, že EM algoritmus funguje aj v iných príkladov, ktoré majú intenzity rovné 0 alebo sú rovné celým číslam. Majme maticu intenzity L_2:

$$L_2 = \begin{pmatrix} -5 & 2 & 0 & 3 \\ 1 & -4 & 2 & 1 \\ 0 & 0 & -3 & 3 \\ 1 & 2 & 0 & -3 \end{pmatrix}. \tag{3.4}$$

Podobne ako v Príklade 3.1 nasimulujeme kompletný priebeh spojitého Markovského reťazca pre interval [0,100], viď 3.4. Obr. 3.4 má oproti Obr. 3.1 väčší počet skokov z jedného stavu do druhého. Je to spôsobené väčšími intenzitami. V tomto príklade nám pre diskretizačné kroky $t_{k+1} - t_k = \{1, \frac{1}{2}, \frac{1}{4}\}$ často nekonvergoval algoritmus ani po 500 iteráciach. Preto sme zvolili nové $t_{k+1} - t_k = \{1, \frac{1}{10}, \frac{1}{100}\}$ na porovnanie kvality odhadu matice intenzity. V krokoch $\frac{1}{10}$ a $\frac{1}{100}$ už algoritmus zvyčajne konvergoval. Pre vyššie intenzity odporúčame voliť jemnejšie diskretizačné kroky, pretože so zvyšujúcimi sa intenzitami sa zvýši aj frekvencia preskokov z jedného stavu do druhého. Pri zvolení príliš veľkého kroku je možné, že naše dáta nebudú obsahovať určité preskoky, ktoré nastali medzi časmi t_k a t_{k+1}, čo samozrejme ovplyvní kvalitu odhadu. EM algortimus sme spustili pre časy $T = 100, 250, 500, 10^3, 10^4$. Výpočet pre čas $T = 10^5$ sme vynechali, pretože dĺžka výpočtu pre krok $\frac{1}{100}$ by bola neprimerane dlhá.

Obr. 3.4: Simulácia spojitého Markovského reťazca s maticou intenzity L_2 v čase $T = 100$.



Pre čas $T = 100$ a krok 1 nám EM algoritmus nekonvergoval ani po 500 iteráciach, výpočet sme ukončili a vyšla nám matica (3.5).

$$\hat{\boldsymbol{Q}}_{EM}(100{,}1) = \begin{pmatrix} -1.4799 & 0.66269 & 0.00615 & 0.81105 \\ 1.01454 & -2.51115 & 1.25851 & 0.2381 \\ 3.414 \times 10^{-10} & 0.03081 & -2.73825 & 2.70744 \\ 5.487 \times 10^{-7} & 1.88576 & 0.07915 & -1.96491 \end{pmatrix}. \tag{3.5}$$

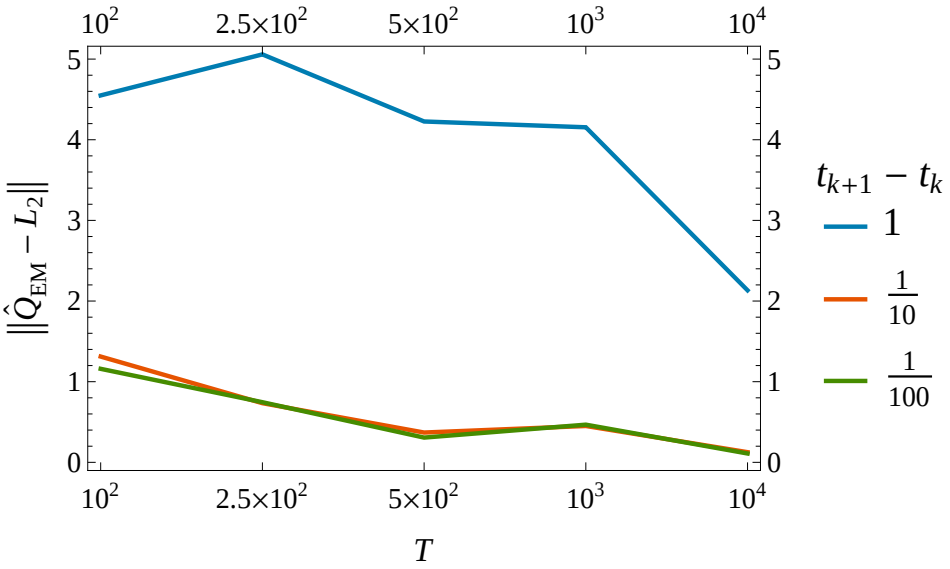
Táto matica má od pôvodnej matice intenzity ďaleko. Ukazuje to aj hodnota chyby počítanej cez euklidovu normu, ktorá sa rovná 4.54962. Uvedieme Tabuľku 3.3, ktorá obsahuje chyby odhadu matice, podobne ako v Príklade 3.1.

Tabuľka 3.3: Závislosť dĺžky časovej realizácie T a veľkosti diskretizačných krokov na veľkosť chyby počítanej cez euklidovu normu $\| L_2 - \hat{Q}_{EM}(T,t_{k+1} - t_k) \|$.

	Diskretizačný krok		
T	1	$\frac{1}{10}$	$\frac{1}{100}$
100	4.549	1.312	1.159
250	5.058	0.734	0.746
500	4.226	0.369	0.307
1 000	4.154	0.451	0.467
10 000	2.133	0.126	0.110

Znova z Tabuľky 3.3 a Obr. 3.5 vidíme, že kvalita odhadu matice intenzity je priamo úmerná počtu dát. Z minulého príkladu sme zistili, že s jemnejším diskretizačným krokom je náš odhad matice kvalitnejší. Pre kroky 1 a $\frac{1}{10}$ to rozhodne platí aj v tomto prípade. Pre kroky $\frac{1}{10}$ a $\frac{1}{100}$ to neplatí vždy, i keď to môže byť

Obr. 3.5: Veľkosť chyby odhadu matice intenzity v závislosti na diskretizačnom kroku a dĺžke časovej realizácie Markovského reťazca.



spôsobené nedostatočným počtom dát. Pre $T = 10^5$, by sme rozhodne očakávali, že chyba odhadu bude menšia v kroku $\frac{1}{100}$ ako u kroku $\frac{1}{10}$. Vo výpočte pre čas $T = 10^4$ a krok $\frac{1}{100}$ nám vyšla matica (3.6).

$$\hat{\boldsymbol{Q}}_{EM}(10^4, \frac{1}{100}) = \begin{pmatrix} -5.03957 & 1.98173 & 1.006 \times 10^{-21} & 3.05784 \\ 1.01006 & -4.05722 & 2.04218 & 1.00498 \\ 8.672 \times 10^{-8} & 5.427 \times 10^{-7} & -2.93541 & 2.93541 \\ 0.98328 & 1.98586 & 0.00002 & -2.96917 \end{pmatrix}. \tag{3.6}$$

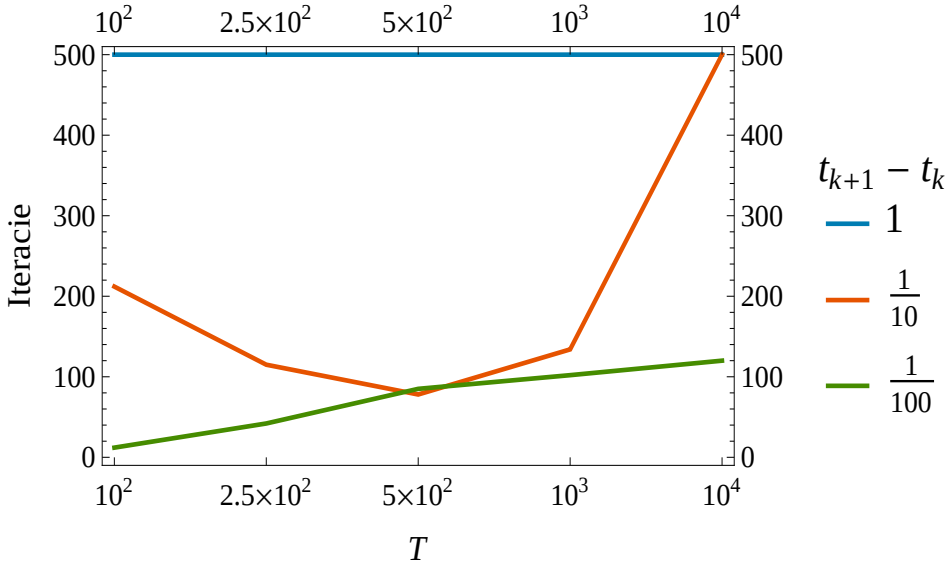
Matica (3.6) sa už veľmi podobá na našu pôvodnú maticu intenzity o čom svedčí aj veľkosť chyby, ktorá je rovná 0.11. Porovnáme rýchlosť konvergencie algoritmu.

Tabuľka 3.4: Závislosť dĺžky časovej realizácie T a veľkosti diskretizačných krokov na rýchlosť konvergercie z hľadiska počtu iterácií a času trvania algoritmu v sekundách.

	Diskretizačný krok					
	1		$\frac{1}{10}$		$\frac{1}{100}$	
T	Iterácie	Čas	Iterácie	Čas	Iterácie	Čas
100	500	187s	212	136s	12	42s
250	500	212s	115	121s	42	335s
500	500	239s	78	138s	85	1 362s
1 000	500	284s	134	436s	102	3 200s
10 000	500	1 134s	500	15 921s	120	42 554s

Z Tabuľky 3.4 a Obr. 3.6 môžeme pozorovať, že znova ako v Príklade 3.1 nám algoritmus pri kroku 1 nekonverguje ani po 500 iteráciach. So zvyšujúcim sa počtom dát sa logicky zvyšuje aj čas potrebný na výpočet. Ďalej môžeme

Obr. 3.6: Počet iterácií nutný na konvergenciu algoritmu v závislosti na diskretizačnom kroku a dĺžke časovej realizácie Markovského reťazca.



pozorovať ako sa pre krok $\frac{1}{100}$ s rastúcim časom zvyšuje počet iterácií nutný na konvergenciu algoritmu a tým pádom aj čas výpočtu. Oproti Príkladu 3.1 je to opačný trend.

3.3 Príklad 3

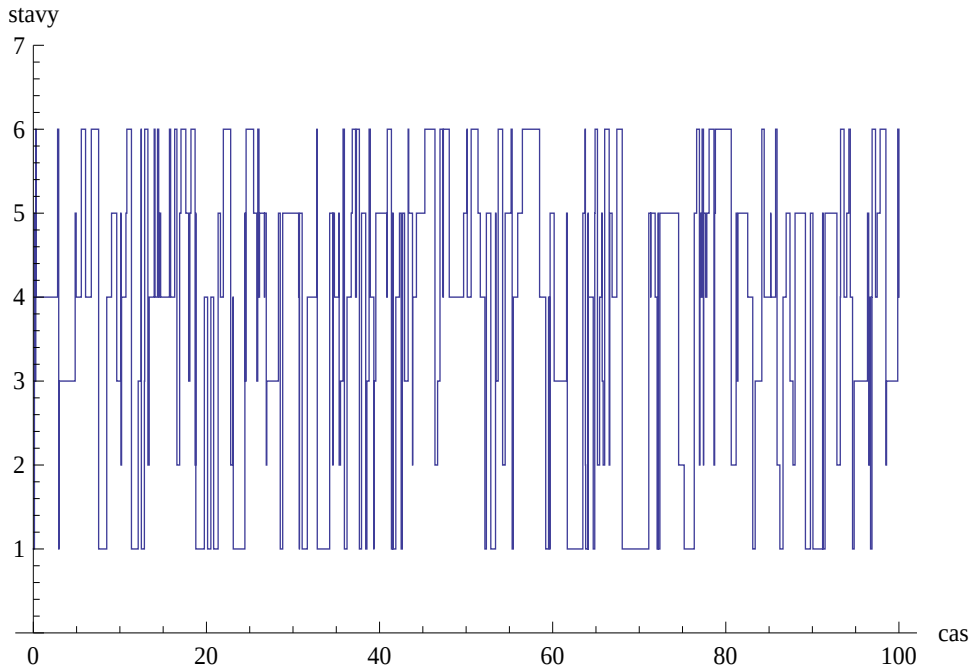
Majme maticu intenzity L_3:

$$L_3 = \begin{pmatrix} -1.8 & 0 & 0.4 & 0.6 & 0.7 & 0.1 \\ 0.5 & -3.65 & 0.15 & 0.9 & 2 & 0.1 \\ 0 & 0.2 & -2 & 0 & 0.7 & 1.1 \\ 0.7 & 0.1 & 0 & -2.7 & 0.9 & 1 \\ 0.5 & 0.5 & 0.5 & 0.5 & -2.5 & 0.5 \\ 0.4 & 0.6 & 0 & 0.6 & 0.1 & -1.7 \end{pmatrix}. \tag{3.7}$$

Zvolili sme menšie intenzity, aby nám výpočet netrval príliš dlho. Menšie intenzity vidieť aj na Obr. 3.7, ktorý má oproti Obr. 3.4 z minulého príkladu menšiu hustotu skokov. V predchádzajúcom príklade sme pre vyššie intenzity museli použiť menšie diskretizačné kroky, aby nám algoritmus konvergoval v rozumnom čase, čo zvyšovalo aj zložitosť algoritmu. V Sekcií 2.3 sme spomenuli, že zložitosť algoritmu polynomiálne závisí na počte stavov Markovského reťazca. V tomto príklade má náš reťazec šesť stavov, čím sa výrazne zvýši zložitosť. Aj preto sme počítali odhad matice intenzity maximálne do času $T = 10^4$. Po prvom spustený výpočtu pre čas $T = 100$ a krok $\frac{1}{2}$ nám vyšla matica:

$$\hat{\boldsymbol{Q}}_{EM}(100, \tfrac{1}{2}) = \begin{pmatrix} -1.8193 & 2.225 \times 10^{-46} & 0.0472 & 0.68667 & 1.08544 & 2.902 \times 10^{-75} \\ 2.45354 & -7.80672 & 0.05409 & 4.05 \times 10^{-19} & 3.72478 & 1.57431 \\ 0.16484 & 0.20909 & -1.68295 & 0.39374 & 0.55476 & 0.36053 \\ 0.28527 & 0.00011 & 0.53867 & -2.31472 & 0.50805 & 0.98261 \\ 0.46618 & 0.00336 & 0.33287 & 0.59456 & -2.34947 & 0.95251 \\ 0.00469 & 2.02485 & 6.566 \times 10^{-24} & 0.78495 & 0.00026 & -2.81472 \end{pmatrix}.$$

Obr. 3.7: Simulácia spojitého markovského reťazca s maticou intenzity L_3 v čase $T = 100$.



Chyba odhadu bola vysoká, a to 5.514. Znova je naša matica výrazne odlišná od pôvodnej matice intenzity (3.7). Naopak pri $T = 10^4$ a kroku $\frac{1}{16}$ je chyba odhadu 0.119 a výsledná matica sa už celkom podobá na (3.7). Ak by sme chceli väčšiu presnosť odhadu, tak musíme zvýšiť počet dát, teda predĺžiť časovú realizáciu Markovského reťazca.

$$\hat{\boldsymbol{Q}}_{EM}(10^4,\frac{1}{16}) = \begin{pmatrix} -1.76833 & 0.00061 & 0.39416 & 0.56368 & 0.70035 & 0.10953 \\ 0.54262 & -3.69797 & 0.17434 & 0.93032 & 1.96009 & 0.09059 \\ 0.00637 & 0.20659 & -2.06774 & 1.203 \times 10^{-6} & 0.72433 & 1.13045 \\ 0.69157 & 0.082 & 3.975 \times 10^{-8} & -2.72876 & 0.89603 & 1.05915 \\ 0.53044 & 0.51509 & 0.48226 & 0.52921 & -2.58897 & 0.53197 \\ 0.39898 & 0.59705 & 3.675 \times 10^{-6} & 0.58046 & 0.09912 & -1.67562 \end{pmatrix}.$$

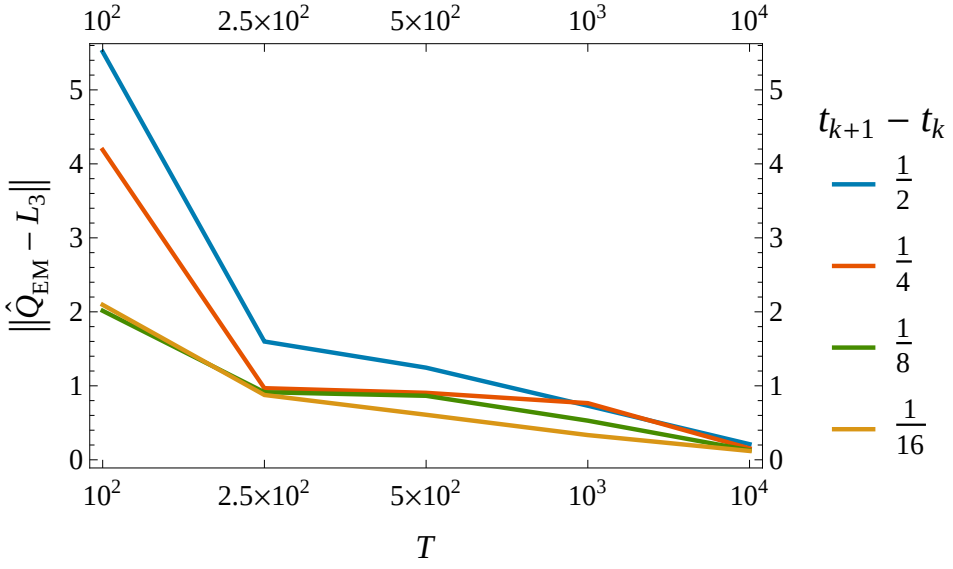
Na porovnanie závislosti počtu dát a diskretizačných krokov máme Tabuľku 3.5.

Tabuľka 3.5: Závislosť dĺžky časovej realizácie T a veľkosti diskretizačných krokov na veľkosť chyby počítanej cez euklidovu normu $\| L_3 - \hat{Q}_{EM}(T,t_{k+1} - t_k) \|$.

	Diskretizačný krok			
T	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$
100	5.514	4.188	2.010	2.090
250	1.599	0.968	0.913	0.876
500	1.245	0.905	0.865	0.609
1 000	0.731	0.766	0.529	0.334
10 000	0.211	0.154	0.128	0.119

Pre rovnaké časové realizácie Markovského reťazca a rovnaké kroky sú chyby v odhade v porovnaní s Tabuľkou 3.1 v prvom príklade menšie. Pre väčšie dáta je však rozdiel chýb menší. Na Obr. 3.8 môžeme vidieť, ako sa so zjemňujúcim diskretizačným krokom znižuje aj chyba odhadu matice intenzity. Avšak na kvalitu odhadu je dôležitejšia dĺžka časovej realizácie Markovského reťazca.

Obr. 3.8: Veľkosť chyby odhadu matice intenzity v závislosti na diskretizačnom kroku a dĺžke časovej realizácie Markovského reťazca.

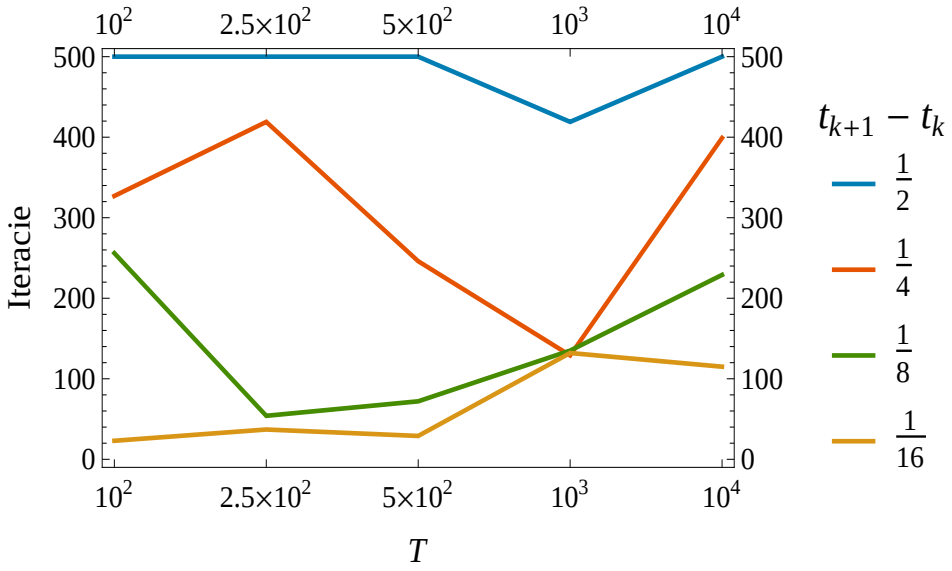


V Tabuľke 3.6 vidíme, že oproti Tabuľke 3.2 v prvom príklade potrebuje algoritmus väčší počet iterácií na konvergenciu pri tých istých diskretizačných krokoch. So zjemňujúcim krokom znova klesá počet iterácií nutný na konvergenciu. Toto pozorovanie je vidieť i na Obr. 3.9.

Tabuľka 3.6: Závislosť dĺžky časovej realizácie T a veľkosti diskretizačných krokov na rýchlosť konvergercie z hľadiska počtu iterácií a času trvania algoritmu v sekundách.

	Diskretizačný krok							
	$\frac{1}{2}$		$\frac{1}{4}$		$\frac{1}{8}$		$\frac{1}{16}$	
T	Iterácie	Čas	Iterácie	Čas	Iterácie	Čas	Iterácie	Čas
100	500	955s	327	706s	256	597s	23	81s
250	500	1 071s	419	1 028s	54	172s	37	173s
500	500	1 187s	246	731s	72	320s	29	210s
1 000	419	1 124s	129	576s	135	953s	132	1 726s
10 000	500	7 901s	399	12 825s	229	13 898s	115	14 306s

Obr. 3.9: Počet iterácií nutný na konvergenciu algoritmu v závislosti na diskretizačnom kroku a dĺžke časovej realizácie Markovského reťazca.



3.4 Zhrnutie a záver

Zhrnieme naše výsledky a pozorovania z príkladov. V priebehu výpočtov EM algoritmu sme zistili, že je dôležité vhodne zvoliť veľkosť diskretizačného kroku. Pre príliš veľké kroky algoritmus nemusí konvergovať, respektíve bude konvergovať až po extrémne veľa iteráciách. Pre príliš malé kroky zas výrazne narastá zložitosť algoritmu. U Markovských reťazcoch s vyššími intenzitami prechodu nastáva skok do iného stavu v menších časových intervaloch, preto je dobré prispôsobiť aj veľkosť krokov. Zvyčajne pre reťazce s vysokými intenzitami, viď Príklad 3.2, je vhodné zvoliť kroky veľkosti $\{\frac{1}{10}, \frac{1}{100}\}$. Pre reťazce s menšími intenzitami, viď Príklad 3.1, môžeme zvoliť kroky veľkosti $\{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}\}$. Diskretizačný krok $t_{k+1} - t_k = 1$ neodporúčame voliť, pretože nám ani pri jednom výpočte nekonvergoval algoritmus do hranice 500 iterácií.

Vo všetkých troch príkladoch sme pozorovali nárast kvality odhadu matice intenzity pri zjemňovaní diskretizačných krokoch. Obzvlášť sa zlepšuje kvalita odhadu pri časovej realizácii $T < 10^4$. V druhom príklade, viď Príklad 3.2, bol rozdiel medzi krokom $\frac{1}{10}$ a $\frac{1}{100}$ veľmi malý. U časoch $T = 250, 1000$ bola chyba kvality odhadu matice väčsia u $\frac{1}{100}$ v porovnaní s $\frac{1}{10}$. Môže to byť spôsobené nedostatočnou časovou realizáciou Markovského reťazca.

Najväčší vplyv na kvalitu odhadu mala dĺžka časovej realizácie Markovského procesu. Pre čas $T = 10^4$ sa chyba odhadu pohybovala okolo 10^{-1}, pre $T = 10^5$ okolo 10^{-2}.

S jemnejším diskretizačným krokom sme pozorovali, ako sa zmenšuje počet iterácií nutných na konvergenciu EM algoritmu, ale na druhej strane sa zvyšuje čas výpočtu jednej iterácie, pretože sme zvýšili počet vybraných diskrétnych pozorovaní stavov z realizácie Markovského procesu.

Literatúra

[1] Z. Prášková and P. Lachout. *Základy náhodných procesů I.* Prvé vydanie. Matfyzpress, Praha, 2012.

[2] Michal Kulich. *Přehledový větník, Statistika pro finanční matematiky.* 2014.

[3] M. Bladt and Sorensen. Statistical inference for discretely observed markov jump processes. *Journal of the Royal Statistical Society, Series B*, 67(3):395–410, 2005.

[4] J. Anděl. *Základy matematické statistiky.* Tretie vydanie. Matfyzpress, Praha, 2011.

[5] Maya R. Gupta and Yihua Chen. Theory and use of the EM algorithm. *Foundations and Trends in Signal Processing*, 4:224–233, 2011.

[6] Marcel F. Neuts. Algorithmic probability: a collection of problems. *Journal of Applied Mathematics and Stochastic Analysis*, 3:472, 1995.

[7] P. Metzner, I. Horenko, and Ch. Schütte. Generator estimation of markov jump processes based on incomplete observations nonequidistant in time. *Journal of the American Statistical Association*, 2007.

Zoznam obrázkov

3.1 Simulácia spojitého Markovského reťazca s maticou intenzity L_1 v čase $T = 100$. 16
3.2 Veľkosť chyby odhadu matice intenzity v závislosti na diskretizačnom kroku a dĺžke časovej realizácie Markovského reťazca. 18
3.3 Počet iterácií nutný na konvergenciu algoritmu v závislosti na diskretizačnom kroku a dĺžke časovej realizácie Markovského reťazca. 19
3.4 Simulácia spojitého Markovského reťazca s maticou intenzity L_2 v čase $T = 100$. 20
3.5 Veľkosť chyby odhadu matice intenzity v závislosti na diskretizačnom kroku a dĺžke časovej realizácie Markovského reťazca. 21
3.6 Počet iterácií nutný na konvergenciu algoritmu v závislosti na diskretizačnom kroku a dĺžke časovej realizácie Markovského reťazca. 22
3.7 Simulácia spojitého markovského reťazca s maticou intenzity L_3 v čase $T = 100$. 23
3.8 Veľkosť chyby odhadu matice intenzity v závislosti na diskretizačnom kroku a dĺžke časovej realizácie Markovského reťazca. 24
3.9 Počet iterácií nutný na konvergenciu algoritmu v závislosti na diskretizačnom kroku a dĺžke časovej realizácie Markovského reťazca. 25

Zoznam tabuliek

3.1 Závislosť dĺžky časovej realizácie T a veľkosti diskretizačných krokov na veľkosť chyby počítanej cez euklidovu normu $\| L_1 - \hat{Q}_{EM}(T, t_{k+1} - t_k) \|$. 17
3.2 Závislosť dĺžky časovej realizácie T a veľkosti diskretizačných krokov na rýchlosť konvergercie z hľadiska počtu iterácií a času trvania algoritmu v sekundách. 18
3.3 Závislosť dĺžky časovej realizácie T a veľkosti diskretizačných krokov na veľkosť chyby počítanej cez euklidovu normu $\| L_2 - \hat{Q}_{EM}(T, t_{k+1} - t_k) \|$. 20
3.4 Závislosť dĺžky časovej realizácie T a veľkosti diskretizačných krokov na rýchlosť konvergercie z hľadiska počtu iterácií a času trvania algoritmu v sekundách. 21
3.5 Závislosť dĺžky časovej realizácie T a veľkosti diskretizačných krokov na veľkosť chyby počítanej cez euklidovu normu $\| L_3 - \hat{Q}_{EM}(T, t_{k+1} - t_k) \|$. 23
3.6 Závislosť dĺžky časovej realizácie T a veľkosti diskretizačných krokov na rýchlosť konvergercie z hľadiska počtu iterácií a času trvania algoritmu v sekundách. 24