Akademie věd České republiky
Ústav teorie informace a automatizace

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

Pavel Boček, Karel Vrbenský:

Implementace algoritmu MIDIA v prostředí Google Spreadsheets

No. 2325 Listopad 2012

ÚTIA AV ČR, P. O. Box 18, 182 08 Prague, Czech Republic
E-mail: utia@utia.cas.cz

Implementace algoritmu MIDIA v prostředí Google Spreadsheets

Pavel Boček, Karel Vrbenský

November 21, 2012

1 Úvod

Aplikační prostředí *Google Docs* je novou platformou pro sdílení dat a algoritmů prostřednictvím sítě Internet. Nabízí široké možnosti pro spolupráci na výzkumných projektech, včetně statistických aplikací. Zároveň poskytuje uživatelům intuitivní a známé prostředí, velmi podobné aplikacím pro PC z balíku Microsoft Office. Přepis metody MIDIA, popsané například v [1], přináší, prostřednictvím jejího zveřejnění v knihovně *Google script Gallery*, možnost jejího zařazení do libovolné aplikace založené na platformě *Google docs*.

Pro účely použití algoritmu v novém prostředí, bylo nutné stávající implementaci v jazyce Visual Basic převést do jazyku Java Script. Předchozí verze programu, naprogramovaná jako soustava maker pro Microsoft Excel, měla spíše podobu samostatného programu, zatímco pro novou implementaci byla zvolena podoba klasické funkce tabulkového kalkulátoru, která umožňuje uživateli zcela volné použití pro zkoumání a prezentaci vlastních dat.

Tento výzkum je podporován grantem GAP202/10/0618

2 Popis úlohy

Nechť jsou dána celá čísla $I \geq 2$ a $J \geq 2$, označme $\mathbf{p} = (p_{ij})$ matici řádu $I \times J$ s nezápornými prvky, jejichž součet je roven jedné. Takovou matici nazveme pravděpodobnostní matice budeme jí reprezentovat nějaké diskrétní dvojrozměrné rozdělení pravděpodobnosti s konečným počtem možných dvojic hodnot, jejichž pravděpodobnosti odpovídají prvkům $\mathbf{p}$. Dále zavedeme

hvězdičkové operátory řádkových a sloupcových marginálních součtů

$$\mathbf{p}^* = (p_j^*)_{j=1}^J, \text{ kde } p_j^* = \sum_{i=1}^I p_{ij}, \text{ pro } 1 \le j \le J,$$

$$\mathbf{p}_* = (p_{*i})_{i=1}^I, \text{ kde } p_{*i} = \sum_{j=1}^J p_{ij}, \text{ pro } 1 \le i \le I.$$

Řádkové i sloupcové marginální součty budeme chápat jako sloupcové vektory. Šipka bude označovat vektorizaci matice po sloupcích, tj.

$$\vec{\mathbf{p}} = \mathsf{Vec}\,\mathbf{p} = (p_{11}, \ldots, p_{I1}, p_{12}, \ldots, p_{I2}, \ldots, p_{1J}, \ldots, p_{IJ})^T,$$

kde horní index ^{T} označuje transpozici.

Množinu všech pravděpodobnostních matic řádu $I \times J$ označíme T, symbol $\vec{T}$ budeme užívat pro množinu vektorů, které vzniknou vektorizací matic množiny T. Je zřejmé, že množina $\vec{T}$ je konvexní omezená část nadroviny v nezáporné části prostoru $\mathbb{R}^{IJ}$. Extremální body množiny $\vec{T}$ jsou vektory vzniklé vektorizací matic $\mathbf{e}^{ij} = (e_{k\ell}^{ij})$, kde

$$e_{k\ell}^{ij} = \begin{cases} 1, & \text{když } k = i \text{ a } \ell = j; \\ 0, & \text{jinak.} \end{cases}$$

Označme

$$E = \left\{ \mathbf{e}^{ij} \,:\, 1 \le i \le I,\, 1 \le j \le J \right\}.$$

Pro každou pravděpodobnostní matici $\mathbf{p} \in T$ platí

$$\mathbf{p} = \sum_{i=1}^I \sum_{j=1}^J \alpha_{ij} \mathbf{e}^{ij},$$

přičemž koeficienty α_{ij} jsou nezáporné a $\sum_{ij} \alpha_{ij} = 1$. Zároveň je množina E nejmenší množina matic, jejichž konvexní lineární kombinace generují celou množinu T. Body $\vec{\mathbf{p}}$ množiny $\vec{T}$ je také možné vyjádřit pomocí prvních $IJ - 1$ souřadnic, neboť poslední souřadnici lze z ostatních dopočítat. Platí

$$\vec{T} = \left\{ \vec{\mathbf{p}} \,:\, 0 \le p_{ij} \le 1, \text{ pro } (i,j) \ne (I,J) \text{ a zároveň } 0 \le \sum_{(i,j)\ne(I,J)} p_{ij} \le 1 \right\}.$$

Jsou-li dány vektory $\mathbf{a} = (a_i) \in \mathbb{R}^I$, $\mathbf{b} = (b_j) \in \mathbb{R}^J$ s nezápornými prvky splňující podmínku $\sum a_i = \sum b_j = 1$, označíme $T_{\mathbf{ab}}$ množinu všech pravděpodobnostních matic řádu $I \times J$, jejichž marginály jsou $\mathbf{a}$ a $\mathbf{b}$, tj.

$$T_{\mathbf{ab}} = \left\{ \mathbf{p} \in T \,:\, \mathbf{p}_* = \mathbf{a},\, \mathbf{p}^* = \mathbf{b} \right\}.$$

Bez újmy na obecnosti se dále omezíme pouze na vektory $\mathbf{a}$ a $\mathbf{b}$ s kladnými souřadnicemi a budeme v celém textu předpokládat $I \leq J$.

Platí $T_{\mathbf{ab}} \subset T$, tedy i $\vec{T}_{\mathbf{ab}} \subset \vec{T}$, množina $\vec{T}_{\mathbf{ab}}$ je konvexní a její dimenze je $(I-1)(J-1)$. Prvky $\vec{T}_{\mathbf{ab}}$ je proto možné vyjádřit podobně jako v případě množiny $\vec{T}$ pomocí některých (ne libovolných) $(I-1)(J-1)$ souřadnic.

Označíme T_0 lineární prostor všech matic řádu $I \times J$, jejichž oba marginální součty jsou rovny nulovým vektorům, tj. $T_0 = T_{\mathbf{00}}$ a

$$\mathbf{q} \in T_0 \quad \Longleftrightarrow \quad \mathbf{q}_* = \mathbf{0} \quad \text{a zároveň} \quad \mathbf{q}^* = \mathbf{0}.$$

Je zřejmé, že množiny $T_{\mathbf{ab}}$ jsou konvexními podmnožinami množin

$$\tilde{T}_{\mathbf{ab}} = \{\mathbf{q}_0 + \tilde{\mathbf{q}} \,:\, \mathbf{q}_0 \in T_{\mathbf{ab}} \text{ lib. pevné a } \tilde{\mathbf{q}} \in T_0\},$$

přičemž jejich hranice jsou určeny lineárními omezeními danými požadavkem na nezápornost všech souřadnic.

Označme $\mathcal{J}_k = \left\{J_1, \ldots, J_{\binom{J}{k}}\right\}$ množinu všech k-prvkových podmnožin množiny sloupcových indexů $\mathcal{J} = \{1, 2, \ldots, J\}$ a $\mathcal{I}_k = \left\{I_1, \ldots, I_{\binom{I}{k}}\right\}$ množinu všech k-prvkových podmnožin množiny řádkových indexů $\mathcal{I} = \{1, 2, \ldots, I\}$. Pro danou matici $\mathbf{p}$ řádu $I \times J$ a množiny $I_r \in \mathcal{I}_{k_1}$ a $J_s \in \mathcal{J}_{k_2}$ označíme symbolem $\mathbf{p}_{I_r J_s}$ submatici řádu $k_1 \times k_2$, která vznikne z $\mathbf{p}$ vybráním řádků, jejichž indexy patří do I_r a sloupců, jejichž indexy patří do J_s, v pořadí, v jakém po sobě následují v matici $\mathbf{p}$.

Definice 1. Nechť $J \geq I \geq 2$, $P_I = \{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_{I!}\}$ je množina všech permutačních matic řádu I, $\mathcal{I} = \mathcal{I}_I = \{1, 2, \ldots, I\}$ a

$$\mathcal{J}_I = \left\{J_1, J_2, \ldots, J_{\binom{J}{I}}\right\}$$

je systém všech I-prvkových podmnožin množiny sloupcových indexů $\{1, 2, \ldots, J\}$.

Řekneme, že matice $\mathbf{Q}$ je *rozšířená permutační matice,* jestliže existuje množina indexů $J_k \in \mathcal{J}_I$ a permutační matice $\mathbf{P}_\ell$ řádu I tak, že $\mathbf{Q}$ má nulové sloupce na místech, jejichž indexy nejsou v množině J_k, zatímco po jejich vynechání zbývající sloupce tvoří permutační matici $\mathbf{P}_\ell$, tj.

$$\mathbf{Q}_{\mathcal{I}, J_k} = \mathbf{P}_\ell, \tag{1}$$

$$\mathbf{Q}_{\mathcal{I}, \mathcal{J} \setminus J_k} = \mathbf{0}_{I \times (J-I)}. \tag{2}$$

Pokud pro dvě rozšířené permutační matice $\mathbf{Q}_1$ a $\mathbf{Q}_2$ platí

$$\mathbf{Q}_1 - \mathbf{Q}_2 \in T_0,$$

pak řekneme, že jsou *stejně rozšířené.* Množinu všech matic $\mathbf{R}_n$, které vzniknou jako rozdíl dvou stejně rozšířených permutačních matic označíme T_0^G.

3 Množina $T_{\mathbf{ab}}$ a její extremální body

Strukturu prostoru T_0 je možné popsat pomocí permutačních matic řádu I. Platí následující lemma.

Lemma 2. *Nechť matice* $\mathbf{q_0}$ *řádu* $I \times J$, $I \leq J$, *je prvkem lineárního prostoru* T_0, *tj. má nulové marginální součty. Pro pevně danou množinu* $J_k \in \mathcal{J}_I$ *a pevnou permutační matici* $\mathbf{P}_\ell \in P_I$ *označíme* $\mathbf{Q}^{k\ell}$ *rozšířenou permutační matici, která splňuje podmínky* (1) *a* (2) *z definice 1.*

Potom platí následující tvrzení.

(i) Matici $\mathbf{q}_0$ *lze vyjádřit jako lineární kombinaci rozšířených permutačních matic ve tvaru*

$$\mathbf{q}_0 = \sum_{k=1}^{\binom{J}{I}} \sum_{\ell=1}^{I!} \lambda_{k\ell}\ \mathbf{Q}^{k\ell} \tag{3}$$

a pro každý takový rozklad (3) *platí*

$$\sum_{k=1}^{\binom{J}{I}} \sum_{\ell=1}^{I!} \lambda_{k\ell} = 0. \tag{4}$$

(ii) Matici $\mathbf{q}_0$ *lze vyjádřit jako kladnou lineární kombinaci matic* $\mathbf{R}_n \in T_0^G$, *tj.*

$$\mathbf{q}_0 = \sum_n \alpha_n \mathbf{R}_n, \quad \alpha_n > 0. \tag{5}$$

Proof. Lineární prostor $\vec{T}_0$ (a tedy i prostor T_0) má dimenzi $(I-1)(J-1)$, stačí tedy nalézt množinu $\vec{G}_0$ alespoň $(I-1)(J-1)$ lineárně nezávislých vektorů z $\vec{T}_0$ (matic z T_0), které jsou zároveň lineárními kombinacemi některých vektorizovaných rozšířených permutačních matic $\mathbf{Q}^{k\ell}$. Tato množina je potom množinou generátorů prostoru $\vec{T}_0$ a rozklad (3) existuje.

Budiž $\mathbf{R}^{ij}$ matice, která má jedničku na pozicích (i,j) a (I,J), mínus jedničku na pozicích (i,J) a (I,j) a nulu na všech ostatních pozicích. Zřejmě $\mathbf{R}^{ij} \in T_0$ a existuje dvojice stejně rozšířených permutačních matic $\mathbf{Q}^{k_1\ell_1}$, $\mathbf{Q}^{k_2\ell_2}$ tak, že

$$\mathbf{R}^{ij} = \mathbf{Q}^{k_1\ell_1} - \mathbf{Q}^{k_2\ell_2}.$$

Položme

$$\vec{G}_0 = \left\{ \vec{\mathbf{R}}^{ij} \ : \ 1 \leq i \leq I-1,\ 1 \leq j \leq J-1 \right\}.$$

Množina $\vec{G}_0$ obsahuje $(I-1)(J-1)$ lineárně nezávislých vektorů z $\vec{T}_0$ a je proto bazí tohoto prostoru. Každý vektor $\vec{\mathbf{q}}_0 \in \vec{T}_0$ je tak možné (jednoznačně)

vyjádřit jako lineární kombinaci vektorů $\vec{\mathbf{R}}^{ij}$, z nichž každý je rozdílem nějakých (již ne jednoznačně určených) dvou vektorů $\vec{\mathbf{Q}}^{k_1\ell_1}$ a $\vec{\mathbf{Q}}^{k_2\ell_2}$. Odtud plyne existence rozkladu (3) včetně rovnosti (4). Kladné koeficienty potřebné k důkazu rozkladu (5) dostaneme například tak, že pro dvojici (i, j), jejíž koeficient v rozkladu by byl záporný, vyměníme v bazi $\vec{G}_0$ vektor $\vec{\mathbf{R}}^{ij}$ vektorem $-\vec{\mathbf{R}}^{ij}$. □

Význam matic z množiny T_0^G naznačuje následující lemma.

Lemma 3. *Nechť $J \geq I \geq 2$, a nechť $\mathbf{q} \in T_{\mathbf{ab}}$ obsahuje alespoň $K \geq I + J$ nenulových prvků. Pak existují $\mathbf{q}^0 \in T_0^G$ a $\alpha_1, \alpha_2 > 0$ tak, že*

$$\mathbf{q} - \alpha_1\mathbf{q}^0 = \mathbf{q}_1 \in T_{\mathbf{ab}},$$
$$\mathbf{q} + \alpha_2\mathbf{q}^0 = \mathbf{q}_2 \in T_{\mathbf{ab}}$$

a matice $\mathbf{q}_{1,2}$ obsahují méně než K nenulových prvků.

Proof. Najdeme dvě stejně rozšířené permutační matice $\mathbf{q}'$ a $\mathbf{q}''$ tak, že matice $\mathbf{q}^0 = \mathbf{q}' - \mathbf{q}''$ má všechny své nenulové prvky na místech některých nenulových prvků matice $\mathbf{q}$. Každá nenulová matice $\mathbf{q}^0$ z T_0^G obsahuje $J - I$ nulových sloupců a $2k$, $2 \leq k \leq I$, nenulových prvků (k jedniček a k mínus jedniček). Jedničky odpovídají jedničkám permutační matice $\mathbf{q}'$, mínus jedničky odpovídají jedničkám permutační matice $\mathbf{q}''$, pokud mají obě matice $\mathbf{q}'$ a $\mathbf{q}''$ některou jedničku na stejném místě, obsahuje matice $\mathbf{q}^0$ nulový řádek a sloupec, který této pozici odpovídá.

Nejprve vybereme $J - I$ sloupců, které budou nulové. Vybereme takové sloupce, aby počet nenulových prvků matice $\mathbf{q}$ v ostatních sloupcích byl alespoň $2I$. Takové sloupce existují, neboť předpokládáme, že $\mathbf{q}$ obsahuje alespoň $K \geq I + J$ nenulových prvků. Pokud vybereme $J - I$ sloupců, které obsahují nejmenší počet nenulových prvků, pak mohou nastat jen dvě situace. Buď má alespoň jeden z vybraných sloupců nejméně dva nenulové prvky. Pak ale každý ze zbylých sloupců obsahuje alespoň dva nenulové prvky a celkem jich je ve zbytku matice $\mathbf{q}$ alespoň $2I$. Druhou možností je, že v každém z vybraných sloupců byl právě jeden nenulový prvek. V takovém případě jich ve zbytku matice $\mathbf{q}$ je alespoň $I + J - (J - I) = 2I$.

Rozšířené permutační matice $\mathbf{q}'$ a $\mathbf{q}''$ tak mohou mít nulové sloupce například na místech $J - I$ sloupců matice $\mathbf{q}$ s nejmenším počtem nenulových prvků. Definujme matici $\mathbf{V}$ řádu $I \times J$ tak, že obsahuje nulové sloupce na stejných pozicích jako matice $\mathbf{q}'$ a $\mathbf{q}''$ a na ostatních místech má nuly a jedničky rozmístěné podle předpisu

$$v_{ij} = \begin{cases} 1, & \text{pokud } \mathbf{q}_{ij} \neq 0; \\ 0, & \text{jinak.} \end{cases}$$

Matice $\mathbf{V}$ odpovídá matici hran neorientovaného bipartitního grafu s $I + J$ uzly. Jedna skupina je tvořena I vrcholy, které odpovídají řádkům matice $\mathbf{V}$,

druhá J vrcholy, které odpovídají sloupcům matice $\mathbf{V}$. Do $J-I$ sloupcových vrcholů nevede žádná hrana a nemusíme se jimi proto dál zabývat. Mezi zbylými I řádkovými a zbylými I sloupcovými vrcholy existuje nejméně $2I$ hran, z nichž každá je reprezentovaná jednou jedničkou v matici $\mathbf{V}$. Z toho plyne, že tento podgraf nemůže být strom a pokud je souvislý, určitě obsahuje kružnici. Pokud není souvislý, určitě musí obsahovat komponentu, která není stromem, jinak by jeho počet hran musel být $2I$ mínus počet komponent. V grafu reprezentovaném maticí $\mathbf{V}$ proto určitě existuje kružnice. Jelikož jde o bipartitní graf, má tato kružnice sudou délku (nejméně 4, nejvýše $2I$) a pravidelně se v ní střídají řádkové a sloupcové vrcholy. Nechť je délka nějaké takové kružnice $2k$, $2 \le k \le I$. Projděme tuto kružnici od některého řádkového vrcholu a označme čísla vrcholů postupně $i_1, j_1, i_2, j_2, \ldots, i_k, j_k$. Z konstrukce matice $\mathbf{V}$ plyne, že všechny prvky

$$q_{i_1 j_1}, q_{i_2 j_2}, \ldots, q_{i_k j_k}$$
$$q_{i_2 j_1}, q_{i_3 j_2}, \ldots, q_{i_1 j_k}$$

jsou kladné. Položme nyní

$$q'_{i_\ell j_\ell} = 1,\ 1 \le \ell \le k; \quad q''_{i_{\ell+1} j_\ell} = 1,\ 1 \le \ell \le k-1, \quad q''_{i_1 j_k} = 1.$$

Zbylých $I-k$ jedniček můžeme v maticích $\mathbf{q}'$ a $\mathbf{q}''$ rozmístit libovolně tak, aby se jednalo o stejně rozšířené permutační matice s nulovými sloupci na předem zvolených pozicích, v obou maticích však musí být takto rozmístěné jedničky na stejných místech.

Našli jsme tedy dvě stejně rozšířené permutační matice $\mathbf{q}'$ a $\mathbf{q}''$ takové, že matice $\mathbf{q}^0 = \mathbf{q}' - \mathbf{q}''$ má všechny své nenulové prvky na místech některých nenulových prvků matice $\mathbf{q}$. Nyní stačí položit

$$\alpha_1 = \min\left\{q_{ij} \,:\, q^0_{ij} = 1\right\},$$
$$\alpha_2 = \min\left\{q_{ij} \,:\, q^0_{ij} = -1\right\}.$$

Zřejmě je $\alpha_1 > 0$, $\alpha_2 > 0$ a $\pm\alpha_{1,2}\mathbf{q}^0 \in T_0$. Také je snadno vidět, že všechny prvky matic

$$\mathbf{q}_1 = \mathbf{q} - \alpha_1 \mathbf{q}^0,$$
$$\mathbf{q}_2 = \mathbf{q} + \alpha_2 \mathbf{q}^0$$

jsou nezáporné a proto patří do množiny $T_{\mathbf{ab}}$. Matice $\mathbf{q}_{1,2}$ mají nuly na všech místech, kde má nuly matice $\mathbf{q}$ a navíc ještě nejméně na jednom místě, kde má matice $\mathbf{q}$ hodnoty $\alpha_{1,2}$. □

Důsledkem lemmat 2 a 3 je následující tvrzení.

Lemma 4. *Pro každou dvojici navzájem různých matic* $\mathbf{q}$, $\mathbf{r}$ *z* $T_{\mathbf{ab}}$ *existuje konečná množina matic* $\{\alpha_n \mathbf{R}_n\}$, *kde* $\alpha_n > 0$ *a* $\mathbf{R}_n \in T_0^G$, *pro kterou platí*

$$\mathbf{r} = \mathbf{q} + \sum_n \alpha_n \mathbf{R}_n. \tag{6}$$

Připustíme-li i záporné hodnoty koeficientů α_n, *pak matice* $\alpha_n \mathbf{R}_n$ *a jejich pořadí lze volit tak, že všechny částečné součty*

$$\mathbf{q}^{(N)} = \mathbf{q} + \sum_{n=1}^{N} \alpha_n \mathbf{R}_n \tag{7}$$

leží v množině $T_{\mathbf{ab}}$.

Proof. K důkazu tvrzení (6) si stačí uvědomit, že $\mathbf{r} - \mathbf{q} = \mathbf{q}_0 \in T_0$, tvrzení (6) je pak přímou aplikací lemmatu 2.

Důkaz tvrzení (7) je podobný důkazu lemmatu 3. Matice $\mathbf{q}_0 = \left(q_{ij}^0\right)$ obsahuje sudý počet nenulových prvků, nejméně pak čtyři. Definujme matici $\mathbf{V}_0 = \left(v_{ij}^0\right)$, která bude mít nuly a jedničky rozmístěné podle předpisu

$$v_{ij}^0 = \begin{cases} 1, & \text{pokud } q_{ij}^0 \neq 0; \\ 0, & \text{jinak.} \end{cases}$$

Matice $\mathbf{V}_0$ reprezentuje hrany bipartitního grafu s I řádkovými a J sloupcovými vrcholy. Všechny vrcholy mají buď stupeň nula nebo stupeň alespoň dvě, tj. buď z nich nevychází žádná hrana nebo alespoň dvě hrany. Pokud vynecháme vrcholy stupně nula, obsahuje proto zbylý podgraf alespoň tolik hran, kolik má vrcholů. To znamená, že ze stejných důvodů jako v důkazu lemmatu 3 tento podgraf obsahuje kružnici sudé délky $2k$, $1 \leq k \leq I$, v níž se pravidelně střídají řádkové a sloupcové vrcholy. Tentokrát však nestačí libovolná kružnice, ale hledáme kružnici, jejíž hrany odpovídají střídavě kladným a záporným prvkům matice $\mathbf{q}_0$, říkejme ji *střídavá kružnice*.

Taková střídavá kružnice určitě existuje, neboť z každého vrcholu nenulového stupně vychází alespoň jedna hrana odpovídající kladnému a alespoň jedna hrana odpovídající zápornému prvku matice $\mathbf{q}_0$, nazývejme je dále *kladné a záporné hrany*. Stačí vyjít z některého řádkového vrcholu nenulového stupně a kladnou hranou přejít do některého sloupcového vrcholu. Odtud vede alespoň jedna záporná hrana do některého řádkového vrcholu (různého od vrcholu výchozího). Z tohoto vrcholu vede alespoň jedna kladná hrana do některého sloupcového vrcholu (různého od předchozího sloupcového vrcholu). Odtud určitě vede alespoň jedna záporná hrana do některého řádkového vrcholu. Pokud je to vrchol výchozí, nalezli jsme hledanou střídavou kružnici, pokud je to jiný vrchol, postup znovu opakujeme tak dlouho, dokud se některou hranou nedostaneme do již navštíveného vrcholu a vytvoříme tak (po případném

vynechání určité počáteční sekvence) požadovanou kružnici. Vzhledem ke skutečnosti, že $I \leq J < \infty$, má nalezená kružnice délku nejvýše $2I$, nejkratší kružnice bipartitního grafu má délku 4.

Projděme nějakou takto nalezenou střídavou kružnici od některého řádkového vrcholu a označme čísla vrcholů postupně $i_1, j_1, i_2, j_2, \ldots, i_k, j_k$. Z konstrukce matice $\mathbf{V}_0$ plyne, že všechny prvky

$$q^0_{i_1 j_1}, q^0_{i_2 j_2}, \ldots, q^0_{i_k j_k}$$

jsou kladné a všechny prvky

$$q^0_{i_2 j_1}, q^0_{i_3 j_2}, \ldots, q^0_{i_1 j_k}$$

jsou záporné. Položme nyní

$$q'_{i_\ell j_\ell} = 1,\ 1 \leq \ell \leq k; \quad q''_{i_{\ell+1} j_\ell} = 1,\ 1 \leq \ell \leq k-1, \quad q''_{i_1 j_k} = 1.$$

Zbylých $I - k$ jedniček můžeme v maticích $\mathbf{q}'$ a $\mathbf{q}''$ rozmístit libovolně tak, aby se jednalo o stejně rozšířené permutační matice s nulovými sloupci na pozicích, jejichž indexy nepatří do množiny $\{j_1, \ldots, j_k\}$. V obou maticích však musí být takto rozmístěné jedničky na stejných místech. Matici $\mathbf{R}_1$ definujeme jako

$$\mathbf{R}_1 = (R^1_{ij}) = \mathbf{q}'' - \mathbf{q}'$$

a absolutní hodnotu koeficientu

$$|\alpha_1| = \min\left\{|q^0_{ij}| \,:\, R^1_{ij} \neq 0\right\},$$

jeho znaménko pak můžeme zvolit tak, aby matice $\mathbf{q}^1 = \mathbf{r} - (\mathbf{q} + \alpha_1 \mathbf{R}_1)$ obsahovala alespoň o jednu nulu více než matice $\mathbf{q}^0$. Protože všechny prvky $q^{(1)}_{ij}$ matice

$$\mathbf{q}^{(1)} = \mathbf{q} + \alpha_1 \mathbf{R}_1$$

jsou buď rovny odpovídajícím prvkům q_{ij} matice $\mathbf{q}$ nebo leží v intervalu ohraničeném q_{ij} a příslušným prvkem r_{ij} matice $\mathbf{r}$. Proto jsou všechny prvky $\mathbf{q}^{(1)}$ nezáporné a tudíž je $\mathbf{q}^{(1)}$ prvkem množiny $T_{\mathbf{ab}}$.

Pokud je $\mathbf{q}^{(1)} \neq \mathbf{r}$, celý postup zopakujeme a najdeme matici $\mathbf{R}_2 \in T^G_0$ a koeficient α_2 tak, že matice

$$\mathbf{q}_2 = \mathbf{r} - (\mathbf{q} + \alpha_1 \mathbf{R}_1 + \alpha_2 \mathbf{R}_2)$$

obsahovala alespoň o jednu nulu více než matice $\mathbf{q}_1$ a matice

$$\mathbf{q}^{(2)} = \mathbf{q} + \alpha_1 \mathbf{R}_1 + \alpha_2 \mathbf{R}_2$$

ležela v množině $T_{\mathbf{ab}}$. Tento postup opakujeme, dokud pro nějaké N nedostaneme $\mathbf{q}^{(N)} = \mathbf{r}$, k čemuž díky zaručenému přibývání nul v maticích $\mathbf{q}_n$ dojde po konečném počtu kroků. □

Zabývejme se nyní extremálními body množiny $T_{\mathbf{ab}}$, tj. takovými maticemi $\mathbf{q} \in T_{\mathbf{ab}}$, které nemohou být vyjádřeny jako konvexní lineární kombinace jiných prvků množiny $T_{\mathbf{ab}}$. Množinu všech extremálních bodů množiny $T_{\mathbf{ab}}$ označíme $E_{\mathbf{ab}}$. Vzhledem k nezápornosti všech matic z $T_{\mathbf{ab}}$ jsou extremální body charakterizovány svým rozložením nul.

Označme $\mathbf{1}_k$ sloupcový vektor, který se skládá z k jedniček. Každý prvek $\mathbf{q} \in T_{\mathbf{ab}}$ vyhovuje soustavě rovnic

$$\begin{pmatrix} \mathbf{1}_J \otimes \mathbf{I}_I \\ \mathbf{I}_J \otimes \mathbf{1}_I \end{pmatrix} \mathsf{Vec}\, \mathbf{q} = \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix}, \tag{8}$$

která vyjadřuje podmínky na marginální součty. Matice soustavy (8) má IJ sloupců a $I+J$ řádků. Součet prvních I řádků je stejný jako součet zbývajících řádků. Hodnost matice soustavy (8) je proto nejvýše $I + J - 1$. Mezi IJ neznámými je tak možné nejméně $IJ - (I + J - 1) = (I - 1)(J - 1)$ z nich pevně zvolit a ostatní dopočítat. Platí následující lemma.

Lemma 5. *Každý prvek množiny $E_{\mathbf{ab}}$ obsahuje alespoň $(I - 1)(J - 1)$ nul a je (nezáporným) řešením soustavy* (8)*. Naopak je-li $\mathbf{q}$ nezáporné řešení soustavy* (8) *za podmínky, že právě K, $K \geq (I - 1)(J - 1)$, pevně zvolených neznámých je rovno nule a že je $\mathbf{q}$ za této podmínky jediné řešení soustavy* (8)*, pak je $\mathbf{q} \in E_{\mathbf{ab}}$.*

Proof. První tvrzení dokážeme snadno. Pokud některý prvek $\mathbf{q}$ množiny $T_{\mathbf{ab}}$ obsahuje méně než $(I - 1)(J - 1)$ nul, pak obsahuje alespoň $I + J$ nenulových prvků a podle lemmatu 3 je konvexní lineární kombinací jiných dvou prvků množiny $T_{\mathbf{ab}}$. Konkrétně při označení lemmatu 3 je

$$\mathbf{q} = \frac{\alpha_2}{\alpha_1 + \alpha_2}\ \mathbf{q}_1 + \frac{\alpha_1}{\alpha_1 + \alpha_2}\ \mathbf{q}_2.$$

Bod $\mathbf{q}$ tedy nemůže patřit do množiny extremálních bodů $E_{\mathbf{ab}}$.

Naopak nechť $\mathbf{q}$ je nezáporné řešení soustavy (8) za podmínky, že právě K, kde $K \geq (I - 1)(J - 1)$, pevně zvolených neznámých je rovno nule a nechť je $\mathbf{q}$ za této podmínky jediné řešení soustavy (8). Potom neexistuje žádné řešení této soustavy, které by mělo více nul než $\mathbf{q}$ a mělo nuly na všech místech, kde je má $\mathbf{q}$. Protože jsou všechny matice z množiny $T_{\mathbf{ab}}$ nezáporné, nemůže být $\mathbf{q}$ konvexní lineární kombinací jiných prvků množiny $T_{\mathbf{ab}}$. Platí proto $\mathbf{q} \in E_{\mathbf{ab}}$. □

Lemma 5 dává návod, jak hledat prvky množiny $E_{\mathbf{ab}}$ pomocí řešení soustavy (8). Jejich počet však s rostoucími dimenzemi I a J prudce roste. Je zřejmé, že pro počet prvků $E_{\mathbf{ab}}$ platí horní odhad

$$\mathsf{Card}\, E_{\mathbf{ab}} \leq \binom{IJ}{(I-1)(J-1)} = \frac{(IJ)!}{[(I-1)(J-1)]!\quad (I+J-1)!}.$$

Tento odhad lze za předpokladu, že vektory $\mathbf{a}$, $\mathbf{b}$ mají všechny složky kladné, zlepšit odečtením všech konfigurací, které dávají nějaký nulový sloupec. Označme

$$C_{IJ}^{(k)} = \binom{I(J-k)}{I+J-1} = \frac{[I(J-k)]!}{[(I-1)(J-1)-kI]! \quad (I+J-1)!}$$

počet všech možností, jak rozmístit $I+J-1$ nenulových prvků do matice s I řádky a $J-k$ sloupci. Snadno lze odvodit nerovnost

$$\mathsf{Card}\, E_{\mathbf{ab}} \leq \; C_{IJ} = \sum_{k=0}^{K_{IJ}} (-1)^k \binom{J}{k} C_{IJ}^{(k)}, \tag{9}$$

kde

$$K_{IJ} = \left\lfloor \frac{(I-1)(J-1)}{I} \right\rfloor$$

a $\lfloor \cdot \rfloor$ označuje celou část. Také odhad (9) je možné zlepšit, odečteme-li konfigurace, které jsou sice bez nulových sloupců, ale s nulovými řádky. Označme D_{IJ} počet všech matic řádu $I \times J$, které obsahují právě $I+J-1$ nenulových prvků, mají nějaký nulový řádek a zároveň neobsahují žádný nulový sloupec. Podobně jako výše buď $D_{IJ}^{(k)}$ počet všech možností, jak rozmístit $I+J-1$ nenulových prvků do matice s $(I-k)$ řádky a J sloupci tak, aby v žádném sloupci nebyly samé nuly. Podobně jako ve vzorci (9) lze odvodit

$$D_{IJ}^{(k)} = \sum_{\ell=0}^{K_J^{(k)}} (-1)^\ell \binom{J}{\ell} \binom{(I-k)(J-\ell)}{I+J-1},$$

kde

$$K_J^{(k)} = \left\lfloor \frac{(I-1)(J-1)-kJ}{I-k} \right\rfloor,$$

a protože je

$$D_{IJ} = \sum_{k=1}^{K_{JI}} (-1)^{k+1} \binom{I}{k} D_{IJ}^{(k)},$$

dostaneme nerovnost

$$\begin{aligned} \mathsf{Card}\;\; E_{\mathbf{ab}} &\leq C_{IJ} - D_{IJ}, \qquad (10) \\ &\leq \sum_{k=0}^{K_{IJ}} (-1)^k \binom{J}{k} C_{IJ}^{(k)} + \sum_{k=1}^{K_{JI}} (-1)^k \binom{I}{k} D_{IJ}^{(k)}, \\ &\leq \sum_{k=0}^{K_{IJ}} (-1)^k \binom{J}{k} \binom{I(J-k)}{I+J-1} + \sum_{k=1}^{K_{JI}} \sum_{\ell=0}^{K_J^{(k)}} (-1)^{k+\ell} \binom{I}{k} \binom{J}{\ell} \binom{(I-k)(J-\ell)}{I+J-1}. \end{aligned}$$

Jak ukazují výsledky simulací, také odhad (10) může být velmi nadsazený. Pro velký počet možností, které je třeba projít při hledání prvků množiny $E_{\mathbf{ab}}$ podle lemmatu 5, je v současné době možné vypočítat všechny extremální body množiny $T_{\mathbf{ab}}$ pouze pro relativně malé dimenze I, J. I přes to jsou extremální body množiny $T_{\mathbf{ab}}$ užitečné při užití iterativních algoritmů pro hledání přibližného řešení úlohy adaptace nějakého daného dvojrozměrného rozdělení na dané marginály metodou minimalizace divergencí.

4 Implementace algoritmu

Pro implementaci v aplikačním prostředí *Google Docs* byl zvolen algoritmus Náhodných tětiv odvozený v článku [1] Nechť $\mathbf{q}^{(0)} \in T_{\mathbf{ab}}$ je libovolný výchozí bod, $\mathbf{q}^{(n)} \in T_{\mathbf{ab}}$ je nejlepší nalezené řešení po n krocích a $\mathbf{q}_{n+1} \in T_{\mathbf{ab}}$ je náhodně zvolený bod různý od $\mathbf{q}^{(n)}$.

(i) Vypočteme průsečíky $\mathbf{q}_D^{(n+1)}$ a $\mathbf{q}_H^{(n+1)}$ přímky určené body $\mathbf{q}^{(n)}$, $\mathbf{q}_{n+1}$ s hranicí množiny $T_{\mathbf{ab}}$.

(ii) Nový bod $\mathbf{q}^{(n+1)}$ nalezneme jednorozměrnou minimalizací divergence $D_\phi(\mathbf{q}, \mathbf{p})$ přes všechna $\mathbf{q}$ z úsečky určené krajními body $\mathbf{q}_D^{(n+1)}$ a $\mathbf{q}_H^{(n+1)}$.

(iii) Postup opakujeme, dokud nedosáhneme předem zvoleného maximálního počtu iterací nebo maximálního počtu iterací, při nichž nedošlo ke změně divergence o předem danou velikost.

Algoritmus MIDIA je momentálně zařazen ke sdílení prostřednictvím platformy *Google Docs* na adrese

`https://docs.google.com/spreadsheet/ccc?key=`
`0AgWXbTuCD2XOdC1qSWlvSDA3clZSMzQ0czRHd212OUE`

Zde je implementován jako funkce tabulkového kalkulátoru:

`=MiDiAsolve(K,a,b)`

s parametry definovanými jako

- Kontingenční tabulka K velikosti $I \times J$ s nezápornými prvky, reprezentující diskrétní dvojrozměrné rozdělení.
- Sloupcový vektor a délky I s nezápornými prvky, určující marginální rozdělení pomocí požadovaných řádkových součtů.
- Řádkový vektor b délky J s nezápornými prvky, určující marginální rozdělení pomocí požadovaných sloupcových součtů.

Součty prvků vektorů a a b musí být rovny ($\sum a_i = \sum b_j$). Výstupem z funkce je tabulka R velikosti $I \times J$ reprezentující hledané diskrétní dvojrozměrné rozdělení pomocí pravděpodobnostní matice ($R_{i,j} \geq 0; \sum R_{i,j} = 1$), která minimalizuje vzdálenost mezi K a R ve smyslu ϕ-divergence. Snadné použití funkce vracející více hodnot ve tvaru matice je umožněno aplikačním prostředím *Google Spreadsheet*, které na rozdíl od běžných tabulkových procesorů, tento typ výstupu podporuje díky interní funkci `CONTINUE`.

Pokud řešení úlohy neexistuje vrací finkce `MiDiAsolve` hodnotu -1. V implemetovaném algoritmu jsou ošetřeny pomocí standardní interní funkce `throw` následující omezení na vstupní parametry:

- Neshodné rozměry tabulky K a vektorů a a b.
- Záporné prvky ve vstupních parametrech.
- Nerovnost součtu prvků ve vektorech a a b.

Tyto chyby jsou během výpočtu oznámeny uživateli pomocí prostředků obvyklých pro tabulkové procesory.

5 Ilustrační příklad

Pro demonstraci algoritmu byla použita následující data:

$$K = \begin{pmatrix} 3 & 15 & 25 & 10 & 3 \\ 12 & 25 & 30 & 12 & 8 \\ 5 & 10 & 33 & 3 & 1 \\ 2 & 7 & 19 & 1 & 1 \end{pmatrix}$$

$$a = (53, 90, 50, 32)'$$

$$b = (17, 60, 110, 22, 16)$$

Na obrázku je příklad užití funkce `MiDiAsolve` v prostředí *Google Spreadsheet*. Modře, žlutě a červeně jsou vyznačeny části tabulky K, a a b vstupující jako parametry. Zelená tabulka je vypočtená výsledná pravděpodobnostní matice reprezentující hledané optimální rozdělení. Ve třetí části výstupu můžeme vidět, že po přepočtu výsledné matice R na rozsah výběru určený zvolenými marginálami (v tomto případě opět rovný 225) skutečně dostáváme požadované marginální součty.

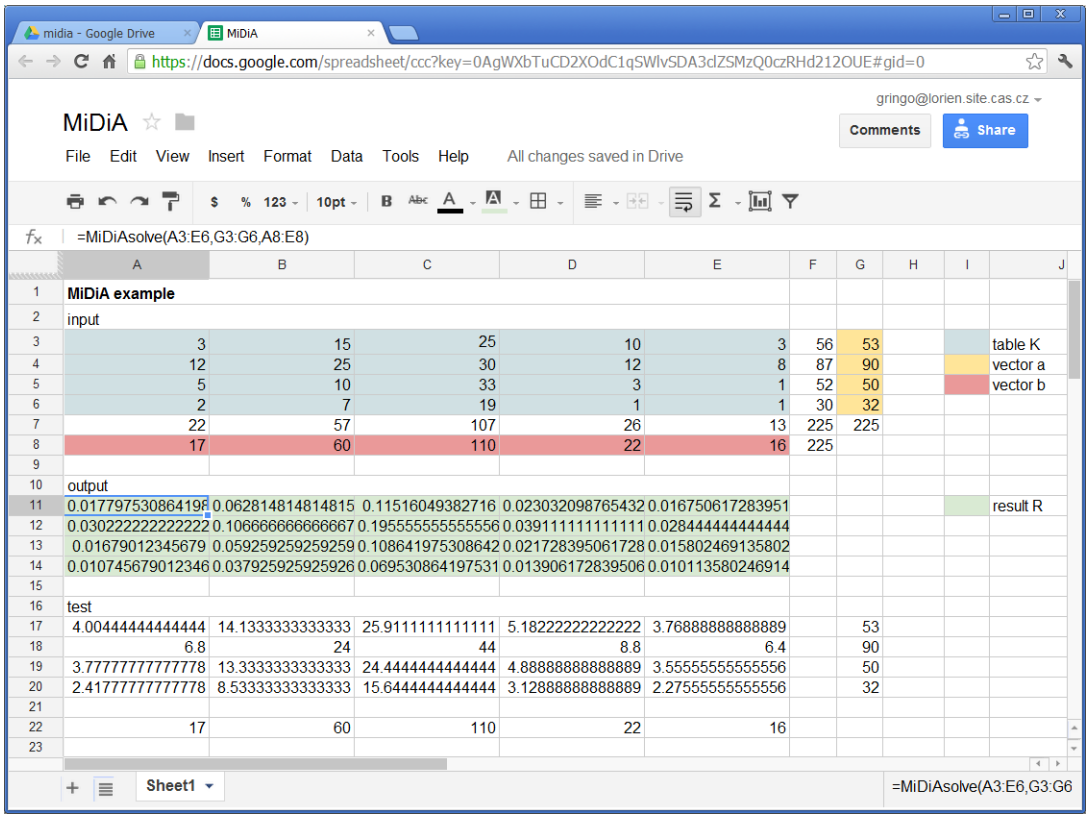

Figure 1: Příklad užití funkce `MiDiAsolve` v prostředí *Google Spreadsheet*

Literatura

[1] Marek, T., Vrbenský, K. : Algoritmy adaptace dvojrozmerných rozdelení metodou minimalizace divergencí, ÚTIA AV CR, (Praha 2007) Research Report 2007/12 (2007)

[2] Marek, T., Vajda, I., Vrbenský, K. (2005): Minimum divergence adaptation of bivariate distributions. Interní publikace DAR - ÚTIA 2005/44. ÚTIA AV ČR, Praha 2005, 35 pp.