UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

VYHĽADÁVANIE PODOZRIVÝCH FIRIEM

Diplomová práca

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	1114 Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci:	prof. RNDr. Pavel Brunovský, DrSc.

Bratislava 2018 **Bc. Matej Hladiš**

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Matej Hladiš
Študijný program: ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Vyhľadávanie podozrivých firiem.
Identification of suspicious firms.

Cieľ: Metódou DEA, použitej na tento účel v tohoročnej bakalárskej práci M. Hladiša preveriť širšie spektrum firiem. Preskúmať možnosti použitia iných matematických metód.

Vedúci: prof. RNDr. Pavel Brunovský, DrSc.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, CSc.

Dátum zadania: 25.01.2017

Dátum schválenia: 27.01.2017 prof. RNDr. Daniel Ševčovič, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie

Touto cestou by som sa rád poďakoval vedúcemu mojej diplomovej práce prof. RNDr. Pavelovi Brunovskému, DrSc. za jeho odborné vedenie a cenné rady. Ďalej ďakujem Mgr. Jane Bockovej a Bc. Jurajovi Druskovi za podnetné diskusie k téme a pomoc pri písaní práce. Veľká vďaka patrí aj Bc. Samuelovi Javorkovi za jeho pomoc pri pracnom prideľovaní koeficientov podozrenia jednotlivým firmám.

Abstrakt

HLADIŠ, Matej. *Vyhľadávanie podozrivých firiem*. [Diplomová práca]. Univerzita Komenského v Bratislave. Fakulta matematiky, fyziky a informatiky. Katedra aplikovanej matematiky a štatistiky. Vedúci: prof. RNDr Pavel Brunovský, DrSc. Bratislava, 2018. 60 s.

Korupcia a rozkrádanie pri verejných zákazkách sa už dlhodobo javí byť problémom v modernej spoločnosti. Táto práca ponúka novú metódu na odhaľovanie nekalých praktík pri verejných zákazkách. Navrhli sme použiť metódu obálky dát (DEA) na odhaľovanie firiem podozrivých z rozkrádania verejných financií. DEA modely slúžia na multikriteriálne porovnávanie jednotiek (decision-making units) na základe vstupov, ktoré spotrebúvajú, a výstupov, ktoré produkujú. Nedostatkom týchto modelov je veľká citlivosť na správnu voľbu vstupov a výstupov. Na odstránenie tohto nedostatku pre aditívny DEA model sme vytvorili metódu na voľbu vstupov a výstupov z daných množín potenciálnych vstupov a potenciálnych výstupov. Navrhli sme aj rozšírenie metódy pre prípad, že je k dispozícii iba množina potenciálnych atribútov a nie je známe, ktoré z nich sú vstupy a ktoré výstupy. Aditívny DEA model s navrhovaným vylepšením sme použili na analyzovanie 188 stavebných firiem, ktoré vyhrali medzi rokmi 2009 a 2016 aspoň jedno verejné obstarávanie s predmetom *Stavebné práce na stavbe budov pre voľný čas, šport, kultúru, ubytovanie a reštauračné stravovanie*. Výsledky sme porovnali s výsledkami získanými expertnou voľbou vstupov a výstupov, rovnako ako s výsledkami voľby vstupov a výstupov metódou group lasso.

Kľúčové slová: Rozkrádanie. Lasso. Výber modelu. Metóda obálky dát.

Abstract

HLADIŠ, Matej. *Identification of suspicious firms*. [Diploma thesis]. Comenius University in Bratislava. Faculty of Mathematics, Physics and Informatics. Department of applied mathematics and statistics. Supervisor: prof. RNDr Pavel Brunovský, DrSc. Bratislava, 2018. 60 pp.

Corruption in public procurements has been considered to be a major problem of modern society over a long period of time. This work offers a new method for detection of illegal machinations in public procurements. Here we suggest using Data Envelopment Analysis (DEA) for the identification of firms suspicious of corruption in public procurements. DEA models are used for multicriterial comparison of decision-making units based on the inputs they use and the outputs they produce. The drawback of these models is high sensitivity with respect to the selection of inputs and outputs. In order to overcome this drawback for additive DEA model, we have developed a method for the selection of inputs and outputs from given sets of potential inputs and potential outputs. We have also enhanced this method for the case when only a set of potential attributes is available, and it is not clear which of them are the inputs and which of them are the outputs. An additive DEA model with the above-mentioned enhancement was used for the analysis of 188 building companies which won at least one public procurement contract comprising building works on buildings for leisure activities, sport, culture, accommodation and restaurants between years 2009 and 2016. The results were compared with the results obtained by expert selection of the inputs and outputs as well as with the results obtained by group lasso method applied in DEA.

Key words: Corruption. Lasso. Model selection. Data envelopment analysis.

Obsah

Úvod

Korupcia a rozkrádanie verejných financií sa stávajú čoraz diskutovanejšími témami na ekonomických fórach, aj v každodenných rozhovoroch medzi ľuďmi, ale spôsobov ako ich odhaľovať je iba veľmi málo. Málokedy je možné bez podnetov od zodpovedných občasnov alebo bez interných informácií z ministerstiev vedeiť, ktorá firma by mohla byť podozrivá z rozkrádania verejných zdrojov, a ktorá nie. Preto sme si v tejto práci dali za cieľ skúmať nástroje na ex-post odhaľovanie firiem podozrivých z rozkrádania verejných financií.

Prvým systematickým pokusom na riešenie tohto problému je postup navrhovaný v bakalárskej práci [1] z roku 2016. V nej bola navrhnutá ucelená dvojfázová metóda na skúmanie podozrivosti firiem, ktoré vyhrali aspoň jedno verejné obstarávanie v ľubovolnom danom odvetví. Prvá fáza metódy slúži na identifikovanie podozrivých firiem z dátovej sady, u ktorých je podľa ich finančných výkazov zvýšené riziko podieľania sa na rozkrádaní verejných financií. Účelom druhej fázy metódy je potom manuálne vyhodnotiť akékoľvek podozrivé prepojenia, obstarávania alebo finančné ukazovatele firiem identifikovaných v prvej fáze.

Myšlienka metódy na identifikovanie podozrivých firiem je založená na predpoklade, že rozkrádajúca firma je vo svojom odvetví vďaka nekalým praktikám výrazne efektívnejšia ako ostatné firmy. V prvej fáze tejto metódy je preto navrhnutý spôsob klasifikácie firiem na podozrivé a nepodozrivé z rozkrádania verejných financií pomocou teórie Data Envelopment Analysis (DEA), ktorá meria relatívnu efektivitu DMU (decision-making units, v našom prípade firiem). Na meranie efektivity používame dáta z verejne dostupných účtovných závierok a verejných obstarávaní. Firma, ktorá je podľa DEA efektívna sa považuje za podozrivú z rozkrádania verejných financií a neefektívna firma sa považuje za nepodozrivú z rozkrádania verejných financií. Predpokladom na použitie DEA je, aby boli výkazy skúmaných firiem porovnateľné, a teda aby firmy pracovali v tom istom odvetví.

DEA meria relatívnu efektivitu DMU na základe vstupov, ktoré spotrebúvajú, a výstupov, ktoré produkujú. Nedostatkom DEA je však vo všeobecnosti voľba týchto vstupov a výstupov. V prípade voľby príliš veľkého počtu vstupov a výstupov do DEA analýzy je efektívny veľký počet DMU a analýza tak stráca výpovednú hodnotu. Podobne vstupy a výstupy by mali byť vhodne volené tak, aby charakterizovali produkčnú funkciu daného odvetvia, a teda aby

bola pri efektívnych firmách zachovaná aj intuitívna ekonomická interpretácia efektívnosti. Ďalším problémom je, že nakoľko porovnávame firmy iba na základe ich finančných výkazov a obstarávaní, ktoré vyhrali, nemôžeme predpokladať, že vstupy a výstupy, ktoré dobre klasifikujú firmy na podozrivé a nepodozrivé z rozkrádania verejných financií budú pre všetky odvetvia rovnaké. Preto sme si v tejto práci dali za cieľ doplniť DEA a vytvoriť všeobecný spôsob na automatizovanú voľbu vstupov a výstupov iba na základe dát a bez dodatočnej ľudskej expertízy.

V prvej kapitole tejto práce sa budeme venovať základnej terminológii týkajúcej sa odhaľovania korupcie, predstavíme skúmané dáta a budeme sa venovať aditívnemu DEA modelu, ktorý je použitý na klasifikáciu firiem na podozrivé a nepodozrivé z rozkrádania verejných financií. V druhej kapitole potom opíšeme metódy na automatizovanú voľbu vstupov a výstupov pre aditívny DEA model a nakoniec v tretej kapitole predložíme výsledky jednotlivých metód s ohľadom na hľadanie firiem podozrivých z rozkrádania verejných financií.

Kapitola 1

Terminológia a dáta

Skôr ako sa budeme venovať samotnej voľbe vstupov a výstupov, či postupom na odhaľovanie korupcie, predstavíme v tejto kapitole používané dáta a nevyhnutnú terminológiu na pochopenie ostatných častí práce. Pracujeme so sadou dát, ktorá bola použitá v [1]. Ide o vzorku 188 firiem, ktoré medzi rokmi 2009 až 2016 vyhrali aspoň jedno verejné obstarávanie s predmetom *Stavebné práce na stavbe budov pre voľný čas, šport, kultúru, ubytovanie a reštauračné stravovanie*. Zdrojom informácie, či nejaká firma vyhrala zákazku v tomto odvetví sú údaje poskytnuté Transparency International Slovensko (TIS) a sú dostupné na webovom portáli [11]. Túto vzorku sme zvolili preto, aby bola jednoducho porovnateľná s predošlými výsledkami v [1] a aby sme tak mohli vyhodnotiť benefity automatizovanej voľby vstupov a výstupov.

1.1 Vstupy a výstupy

Uvažované dáta obsahujú pre každú firmu spolu 17 atribútov:
Priemerný účtovný cashflow; *Priemerné množstvo peňazí na bankových účtoch uvedené v účtovnej závierke*; *Priemerné vlastné imanie*; *Priemerná výška tržieb*; *Priemerné Z-skóre*; *Priemerný počet zamestnancov*; *Priemerné náklady na materiál*; *Maximálny medziročný rast tržieb*; *Priemerné množstvo peňazí v hotovosti uvedené v účtovnej závierke*; *Priemerná pridaná hodnota*; *Počet obstarávaní, v ktorých sa daná firma zúčastnila ako jediná a vyhrala*; *Počet obstarávaní obstarávaných formou neverejnej súťaže, ktoré firma vyhrala* ; *Priemerné ročné príjmy z verejných obstarávaní*; *Priemerný obežný majetok*; *Hodnota najväčšej štátnej zákazky akú firma vyhrala*; *Priemerné náklady na služby*; *Priemer rozdielu nákladov na služby a nákladov na materiál*.[1]

Tieto atribúty boli rozdelené na vstupy a výstupy, pričom vo vyššie uvedenom zozname je prvých 7 určených ako vstupy a zvyšných 10 atribútov sú výstupy.

[1] Pod priemerom sa chápe priemer hodnôt atribútov za sledované obdobie.

1.2 Koeficient podozrenia

Dôležitým pojmom používaným v celej práci je *koeficient podozrenia*. Jeho účelom je kvantifikovať mieru podozrenia firmy získanú pri subjektívnom hodnotení jej finančných výkazov, verejných zákaziek a prepojení štatutárov na iné osoby a firmy. V tejto práci ho budeme používať na vyhodnocovanie správnosti klasifikácie firiem na podozrivé a nepodozrivé z rozkrádania verejných financií. Cieľom práce nie je vytvoriť spôsob na odhad koeficientu podozrenia pre jednotlivé firmy.

Koeficient podozrenia nadobúda diskrétne hodnoty od 0 do 4 s krokom $0,5$, pričom čím je hodnota vyššia, tým vyššia je miera subjektívneho podozrenia z rozkrádania verejných financií. Hodnoty väčšie ako 2 už naznačujú veľkú mieru podozrenia z rozkrádania verejných zdrojov a koeficient podozrenia sa prideľuje subjektívnym uvážením po kontrole troch skupín kritérií:

Zákazky. Prvým kritériom na vyhodnocovanie podozrivosti firiem je kontrola verejných zákaziek, ktoré firma vyhrala. Tieto zákazky sú dohľadateľné na verejne dostupných portáloch [8] a [9]. Pre každú firmu sa sleduje napríklad aký počet uchádzačov bol v jednotlivých zákazkách, koľko bolo vylúčených ponúk, či predmet zákazky súvisí s hlavnou činnosťou firmy a pod. Kontroluje sa tiež, či pre danú firmu neboli všetky zákazky obstarávané rovnakým obstarávateľom.

Finančné výkazy. Ďalej sa kontrolujú finančné výkazy firiem. Zaujímavý je vzťah medzi finančnými ukazovateľmi, ich zmenami a štátnymi zákazkami, ktoré firma vyhrala. Podozrenie vzbudzuje napríklad, ak verejné obstarávania tvoria väčšinu príjmov podniku, ak je podnik neaktívny (ukazovatele aktivity) alebo ak nemala firma podľa finančných výkazov kapacity na realizáciu zákaziek, ktoré vyhrala. Finančné výkazy firiem sú dohľadateľné napríklad na webových stránkach [12] a [13].

Prepojenia. Nakoniec sa sledujú prepojenia medzi danou firmou, jej konateľmi a obstarávateľmi zákaziek, ktoré vyhrala, prípadne inými firmami, ktoré už boli vyhodnotené ako podozrivé z korupcie. Podozrivé je, keď sú konatelia alebo členovia štatutárnych orgánov prepojení s obstarávateľmi alebo keď sú zainteresovaní v podozrivých firmách. Existuje viacero zdrojov, cez ktoré sa dajú informácie o prepojeniach získať a v tejto práci bola použitá sociálna sieť firiem [10].

1.3 Aditívny model

Hlavným predpokladom metódy na hľadanie firiem podozrivých z rozkrádania verejných zdrojov je, že vďaka svojim praktikám sú korumpujúce firmy výrazne efektívnejšie ako

ostatné. Preto je nevyhnutné definovať model na meranie efektivity jednotlivých firiem a v práci na tento účel používame teóriu data envelopment analysis (DEA).

Účelom DEA modelov je meranie relatívnej efektivity DMU (decision-making units, v našom prípade firiem) na základe vstupov, ktoré spotrebúvajú a výstupov, ktoré produkujú. Predpokladom týchto modelov je, že vstupy aj výstupy sú nezáporné a že DMU pracujú v tej istej technológii, čiže sa dajú porovnávať. Hlbšie je teória DEA rozpracovaná v [2], ale intuitívne je zrejmé, že efektívne sú tie DMU, ktoré maximalizujú svoje výstupy pri minimálnych vstupoch. Matematické vyjadrenie efektivity môže byť definované niekoľkými rôznymi DEA modelmi a v tejto práci používame vážený aditívny model s variabilnými výnosmi z rozsahu. Ide o rovnaký model ako v [1] a je zvolený pre jeho invariantnosti na zmenu jednotiek a na posun. Tie sú potrebné, keďže dáta môžu byť v rôznych jednotkách a môžu nadobúdať aj záporné hodnoty.

Označme n celkový počet firiem (DMU) v sledovanej vzorke, M počet vstupov a S počet výstupov. Ďalej označme vektor vstupov pre k-te DMU ako $x_k \in R^M$ a vektor výstupov ako $y_k \in R^S$, $k \in \{1,...,n\}$, dátovú maticu vstupov ako $X = [\ x_1,\ \ x_2,\ \ \ldots,\ \ x_n\ ] \in R^{M\times n}$, dátovú maticu výstupov ako $Y = [\ y_1,\ \ y_2,\ \ \ldots,\ \ y_n\ ] \in R^{S\times n}$ a vektory daných váh váženého aditívneho modelu pre vstupy a výstupy ako ρ^x a ρ^y v tomto poradí. Potom efektivita k-teho DMU v zmysle [1] je podľa váženého aditívneho modelu s variabilnými výnosmi z rozsahu definovaná ako riešenie úlohy

$$\begin{aligned} \min_{\lambda, s^x, s^y} \quad & -(\rho^x)^T s^x - (\rho^y)^T s^y \\ \text{s.t.} \quad & X\lambda + s^x = x_k \\ & Y\lambda - s^y = y_k \\ & \mathbf{1}^T\lambda = 1 \\ & \lambda, s^x, s^y \geq 0, \end{aligned} \tag{1.1}$$

kde $\mathbf{1}$ označuje vektor jednotiek. Dané váhy v tejto práci volíme ako prevrátenú hodnotu rozdielu najväčšej a najmenšej pozorovanej hodnoty daného vstupu, resp. výstupu, a teda

$$\begin{aligned} \rho_i^x &= \tfrac{1}{R_i^x}, \quad R_i^x = \max_k x_{i,k} - \min_k x_{i,k} \\ \rho_r^y &= \tfrac{1}{R_r^y}, \quad R_r^y = \max_k y_{r,k} - \min_k y_{r,k}, \end{aligned} \tag{1.2}$$

kde $x_{i,k}$ je prvok matice X zodpovedajúci i-temu vstupu a k-temu DMU, $i \in \{1,...,M\}$, a $y_{r,k}$ je prvok matice Y zodpovedajúci r-tému výstupu a k-temu DMU, $r \in \{1,...,S\}$. Pri takto formulovanom DEA modeli potom definujeme efektívnosť DMU nasledovne.

Definícia 1

DMU nazývame efektívne v zmysle váženého aditívneho DEA modelu s variabilnými výnosmi z rozsahu práve vtedy, keď je optimálna hodnota účelovej funkcie úlohy (1.1) rovná 0. Ak je optimálna hodnota menšia ako 0, tak DMU nazývame neefektívne.

1.3.1 Posun vstupov a výstupov

Ako bolo naznačené vyššie, pri aplikácii DEA na klasifikovanie firiem z hľadiska podozrenia z rozkrádania verejných financií môže byť v dátach porušený predpoklad nezápornosti vstupov a výstupov. Napríklad firma môže mať záporné vlastné imanie, záporný maximálny rast tržieb alebo ak veľmi zle hospodári, tak aj záporné Z-skóre. Preto je pred samotnou DEA analýzou nevyhnutné dáta vhodne posunúť, aby nadobúdali iba nenulové hodnoty a aby bol tak splnený predpoklad nezápornosti vstupov a výstupov.

Posun vstupov a výstupov je definovaný ako afinná transformácia daná vektormi $\Delta x \geq 0$ a $\Delta y \geq 0$ v tomto poradí, pričom vstupy a výstupy sa pre k-te DMU zobrazia ako

$$x_k \mapsto x_k + \Delta x \tag{1.3}$$

$$y_k \mapsto y_k + \Delta y. \tag{1.4}$$

Vďaka invariantnosti váženého aditívneho modelu s variabilnými výnosmi z rozsahu na posun, ktorá je dokázaná v [2], sa takouto transformáciou nezmení hodnota účelovej funkcie v úlohe (1.1), čo zabezpečí zachovanie rozdelenia DMU na efektívne a neefektívne ako aj ich hodnotu efektivity. V práci volíme posun ako

$$\Delta x_i = |\min_k x_{i,k}| + 1 \quad i \in \{1,...,M\} \tag{1.5}$$

$$\Delta y_j = |\min_k y_{j,k}| + 1 \quad j \in \{1,...,S\}, \tag{1.6}$$

čo zabezpečuje kladnosť dát a teda aj splnenie predpokladu pre DEA.

1.3.2 Zmena jednotiek

Atribúty, ktoré používame na klasifikáciu firiem na podozrivé a nepodozrivé sú rôzne škálované a robia úlohu (1.1) z numerického hľadiska náročnou na riešenie. Preto je potrebné preškálovať vstupy a výstupy tak, aby nadobúdali hodnoty z porovnateľných intervalov a na to využijeme invariantnosť váženého aditívneho modelu s variabilnými výnosmi z rozsahu na zmenu jednotiek. Táto vlastnosť je tiež dokázaná v [2].

Zmena jednotiek je chápaná ako lineárna transformácia vstupov a výstupov diagonálnymi maticami B^x a B^y v tomto poradí, pričom

$$X \mapsto B^x X \qquad s^x \mapsto B^x s^x \tag{1.7}$$

$$Y \mapsto B^y Y \qquad s^y \mapsto B^y s^y \tag{1.8}$$

a diagonálne prvky $B^x_{i,i}$ a $B^y_{j,j}$ sú kladné. Z definície daných váh (1.2) pre vstupy a výstupy tiež vidíme, že

$$\rho^x \mapsto (B^x)^{-1}\rho^x \qquad \rho^y \mapsto (B^y)^{-1}\rho^y. \tag{1.9}$$

V práci volíme diagonálne matice ako $B^x = diag(\rho^x)$ a $B^y = diag(\rho^y)$, čo vedie k špeciálnemu prípadu kedy

$$(\rho^x)^T (B^x)^{-1} = \mathbf{1}^T \tag{1.10}$$

$$(\rho^y)^T (B^y)^{-1} = \mathbf{1}^T \tag{1.11}$$

a pre hodnotu účelovej funkcie (1.1) potom platí

$$-(\rho^x)^T s^x - (\rho^y)^T s^y = -(\rho^x)^T (B^x)^{-1} B^x s^x - (\rho^y)^T (B^y)^{-1} B^y s^y = -\mathbf{1}^T B^x s^x - \mathbf{1}^T B^y s^y. \tag{1.12}$$

1.4 Spôsob klasifikácie firiem

Cieľom tejto práce a práce [1] je vytvoriť ucelenú metódu na odhaľovanie rozkrádania v štátnej správe. Aby sme dosiahli tento cieľ, vytvorili sme spôsob klasifikácie firiem na podozrivé a nepodozrivé z rozkrádania verejných financií. Takto vieme potom určiť z veľkého množstva firiem iba niekoľko, na ktorých preverovanie sa treba sústrediť.

Ako sme naznačili vyššie, tak označenie firmy ako podozrivá z rozkrádania verejných financií je úzko späté s jej relatívnou efektívnosťou v porovnaní s ostatnými firmami v danom odvetví. Z tejto myšlienky pramení aj spôsob klasifikácie firiem na podozrivé a nepodozrivé z rozkrádania verejných financií, ktorý je zhrnutý v nasledujúcej definícii.

Firma je klasifikovaná ako podozrivá z rozkrádania verejných financií, ak je v porovnaní s ostatnými analyzovanými firmami efektívna v zmysle definície 1. Ak je naopak neefektívna, tak je klasifikovaná ako nepodozrivá.

Z definície 1 a z predpokladov váženého aditívneho DEA modelu s variabilnými výnosmi z rozsahu (1.1) je zrejmé, že firmy, pre ktoré chceme robiť klasifikáciu na podozrivé a nepodozrivé z rozkrádania verejných financií, musia byť z toho istého odvetvia podnikania. Toto obmedzenie je prirodzené, lebo porovnávame firmy iba na základe ich finančných výkazov a firmy v rôznych odvetviach sa líšia už iba kvôli špecifickým vlastnostiam jednotlivých odvetví.

Kapitola 2

Metódy voľby vstupov a výstupov

Na klasifikáciu firiem na podozrivé a nepodozrivé z rozkrádania verejných financií používame teóriu DEA a jedným z najväčších úskalí DEA modelov je ich citlivosť na správnu voľbu vstupov a výstupov podľa ktorých meriame efektivitu. Osobitne pre modely s variabilnými výnosmi z rozsahu vyjde pri veľkom počte vstupov a výstupov veľký počet DMU ako efektívnych. Takisto je dôležité zvoliť vstupy a výstupy tak, aby charakterizovali produkčnú funkciu a aby výsledné efektívne DMU korešpondovali s ekonomickou intuíciou. Preto sme si dali za cieľ venovať sa tomuto nedostatku v DEA a vytvoriť metódu na automatizovanú voľbu vstupov a výstupov.

Voľbu vstupov a výstupov chápeme ako jednu z nasledujúcich možností:

I. Dané sú množina všetkých možných vstupov $\mathscr{M}$ a množina všetkých možných výstupov $\mathscr{S}$. Počet ich prvkov označme M a S v tomto poradí. Pod voľbou vstupov a výstupov potom rozumieme voľbu neprázdnych podmnožín $\mathscr{M}^* \subseteq \mathscr{M}$ a $\mathscr{S}^* \subseteq \mathscr{S}$.

II. Daná je množina atribútov $\mathscr{A}$, pričom nie je známe, ktoré atribúty sú vstupy a ktoré výstupy. Počet prvkov množiny $\mathscr{A}$ je $M+S$. Pod voľbou vstupov a výstupov potom rozumieme voľbu neprázdnych disjunktných podmnožín $\mathscr{M}^* \subseteq \mathscr{A}$ a $\mathscr{S}^* \subseteq \mathscr{A}$, $\mathscr{M}^* \cap \mathscr{S}^* = \emptyset$.

V práci uvažujeme dve metódy na voľbu vstupov a výstupov a obe sú aplikované na vážený aditívny DEA model s variabilnými výnosmi z rozsahu. Prvá metóda je metóda group lasso odvodená v [4] a prispôsobená pre DEA v [5] a druhá metóda je metóda rozdielom efektivít navrhnutá v tejto práci.

Aj napriek občasnému používaniu terminológie pre klasifikáciu firiem na podozrivé a nepodozrivé z rozkrádania verejných financií sú myšlienky metód všeobecné a dajú sa použiť pri ľubovoľnej DEA analýze.

2.1 Group lasso

Prvou metódou na voľbu vstupov a výstupov do DEA je výber parametrov modelu pomocou group lasso. Pre DEA bola táto metóda navrhnutá v článku [5] a bude slúžiť ako benchmark pre nami odvodenú metódu rozdielom efektivít. Jej hlavná myšlienka je použiť na voľbu vstupov a výstupov analógiu spôsobu voľby regresorov v lineárnej regresii cez tzv. lasso.

Lasso (least absolute shrinkage and selection operator) bola predstavená v [3] a jej cieľom bolo vybrať podmnožinu z množiny všetkých regresorov v lineárnej regresii a zlepšiť tak predikčné schopnosti a interpretáciu modelov. Označme maticu regresorov X a vektor závislých premenných y pre všetky pozorovania, pričom i-ty riadok X a y zodpovedajú i-temu pozorovaniu. Bez ujmy na všeobecnosti predpokladáme, že prvky X sú normalizované tak, aby výberový priemer prvkov v stĺpcoch X bol 0 a výberová disperzia bola 1.

Úloha na odhad koeficientov metódou najmenších štvorcov je daná

$$\min_{\alpha,\beta} \|y - \alpha - X\beta\|_2^2, \tag{2.1}$$

kde $\|.\|_2$ označuje euklidovskú normu a α je intercept. Myšlienka lasso je v úlohe (2.1) ohraničiť súčet absolútnych hodnôt koeficientov modelu daným parametrom t a znížiť tak rozptyl (variance) modelu pri mierne vyššej výchylke (bias). Odhad koeficientov modelu lineárnej regresie metódou lasso je daný riešením úlohy

$$\begin{aligned} \min_{\alpha,\beta} \quad & \|y - \alpha - X\beta\|_2^2 \\ \text{s.t.} \quad & \|\beta\|_1 \leq t, \end{aligned} \tag{2.2}$$

kde $\|\beta\|_1 = \sum_j |\beta_j|$ je l_1 norma. Geometricky sa na takéto ohraničenie môžeme pozrieť ako na hľadanie optimálneho riešenia vo vnútri l_1 gule $\{\beta \mid \|\beta\|_1 \leq t\}$ s polomerom t. Vďaka vlastnostiam l_1 gule má úloha (2.2) tendenciu vyberať riešenie, kde niektoré $\beta_j = 0$, a teda vplyv na regresiu majú iba regresory pre ktoré $\beta_j \neq 0$. Úloha (2.2) sa dá ekvivalentne zapísať v tzv. Lagrangovej forme

$$\min_{\alpha,\beta} \|y - \alpha - X\beta\|_2^2 + \lambda \|\beta\|_1, \tag{2.3}$$

kde $\lambda > 0$ je hyperparameter vyjadrujúci aká veľká je penalizácia pre normu β.

Analogicky ako pri lasso môžeme aj v prípade DEA modelov pridať do účelovej funkcie l_1 normu premenných (tzv. l_1 regularizáciu) a voliť tak vstupy a výstupy. Avšak rozdielom medzi DEA a lineárnou regresiou je, že kým v lineárnej regresii prislúcha jednému regresoru jedna premenná, v prípade DEA pripadá k jednému vstupu alebo výstupu n premenných, ktoré zodpovedajú jednotlivým DMU. Preto sa na vstupy a výstupy dá pozerať ako na faktorové premenné a autori článku [5] navrhli pre DEA použiť zovšeobecnenie lasso pre prípad faktorových premenných - group lasso.

Group lasso bolo predstavené v [4] a jeho myšlienkou je použiť namiesto l_1 normy súčet euklidovských noriem koeficientov pre jednotlivé faktory. Označme β_j vektor koeficientov pre j-ty faktor, X_j maticu pozorovaní prislúchajúcu j-temu faktoru a Y vektor závislých premenných. Rovnako ako predtým predpokladajme, že stĺpce X_j sú normalizované. Potom je odhad koeficientov v regresii metódou group lasso daný úlohou

$$\min_{\alpha,\beta_1,\ldots,\beta_J} \left\| Y - \alpha - \sum_{j=1}^{J} X_j \beta_j \right\|_2^2 + \lambda \sum_{j=1}^{J} \left\| \beta_j \right\|_2, \tag{2.4}$$

kde $\lambda > 0$ je penalizácia a α je intercept. Podobne ako v lasso majú na regresiu vplyv iba faktory pre ktoré platí $\left\| \beta_j \right\|_2 \neq 0$, a teda faktory pre $\left\| \beta_j \right\|_2 = 0$ môžeme z regresie vynechať. Podrobne je teória lasso a jeho zovšeobecnenia opísaná v [6]. V ďalších častiach práce budeme $\sum_{j=1}^{J} \left\| \beta_j \right\|_2$ nazývať $l_{1,2}$ regularizácia a vrátime sa k označeniam matíc X a Y zavedeným v kapitole 1.

2.1.1 Voľba vstupov a výstupov metódou group lasso

V DEA analýze sa pre každé DMU počíta samostatná úloha lineárneho programovania. Takýto spôsob určovania efektivít DMU je výpočtovo výhodnejší ako počítať jednu mnohorozmernú úlohu pre všetky DMU súčasne. Ak ale chceme do účelovej funkcie pridať $l_{1,2}$ regularizáciu ako v (2.4), je nevyhnutné v úlohe na výpočet efektivít počítať všetky premenné pre všetky vstupy a výstupy súčasne. Preto je potrebné úlohy lineárneho programovania agregovať do jednej mnohorozmernej úlohy, v ktorej sa počítajú efektivity všetkých DMU. Podobne ako v [5] bude pre účely použitia metódy group lasso použitý aditívny model v multiplikatívnom tvare.

Vážený aditívny model s variabilnými výnosmi z rozsahu v multiplikatívnom tvare je duálna úloha k úlohe (1.1) a pre k-te DMU je definovaná ako

$$\begin{aligned} \min_{u,v,w} \quad & e_k^T X^T v - e_k^T Y^T u + w \\ \text{s.t.} \quad & X^T v - Y^T u + \mathbf{1} w \geq 0 \\ & v \geq \rho^x \\ & u \geq \rho^y, \end{aligned} \tag{2.5}$$

kde e_k je jednotkový vektor s jednotkou na k-tom mieste a premenné u a v sú vektory tzv. multiplikatívnych váh. V ďalšom texte budeme potrebovať nematicový zápis úlohy (2.5),

ktorý vyzerá nasledovne

$$\begin{aligned}
\min_{u_1,\ldots,u_S,v_1,\ldots,v_M,w} \quad & \sum_{i=1}^{M} x_{i,k}v_i - \sum_{r=1}^{S} y_{r,k}u_r + w \\
\text{s.t.} \quad & \sum_{i=1}^{M} x_{i,1}v_i - \sum_{r=1}^{S} y_{r,1}u_r + w \geq 0 \\
& \vdots \\
& \sum_{i=1}^{M} x_{i,n}v_i - \sum_{r=1}^{S} y_{r,n}u_r + w \geq 0 \\
& v_i \geq \rho_i^x \qquad i = 1,..,M \\
& u_r \geq \rho_r^y \qquad r = 1,...,S.
\end{aligned} \tag{2.6}$$

Ak potom označíme premenné (multiplikatívne váhy) z úlohy (2.6) pre k-te DMU ako $v_{i,k}$, $u_{r,k}$ a w_k, tak agregovaná úloha pre simultánny výpočet multiplikatívnych váh pre všetky DMU súčasne je daná úlohou

$$\begin{aligned}
\min_{u_{1,1},\ldots,u_{S,n},v_{1,1},\ldots,v_{M,n},w_1,\ldots,w_n} \quad & \sum_{k=1}^{n} \left(\sum_{i=1}^{M} x_{i,k}v_{i,k} - \sum_{r=1}^{S} y_{r,k}u_{r,k} + w_k \right) \\
\text{s.t.} \quad & \sum_{i=1}^{M} x_{i,1}v_{i,k} - \sum_{r=1}^{S} y_{r,1}u_{r,k} + w_k \geq 0 \quad k = 1,...,n \\
& \vdots \\
& \sum_{i=1}^{M} x_{i,n}v_{i,k} - \sum_{r=1}^{S} y_{r,n}u_{r,k} + w_k \geq 0 \quad k = 1,...,n \\
& v_{i,k} \geq \rho_i^x \qquad i = 1,..,M \qquad k = 1,...,n \\
& u_{r,k} \geq \rho_r^y \qquad r = 1,...,S \qquad k = 1,...,n.
\end{aligned} \tag{2.7}$$

Jednou z dobrých vlastností group lasso je, že vedie k riedkym riešeniam úlohy. Preto by sme chceli, aby aj v prípade DEA mohli premenné nadobúdať nulové hodnoty a ťažiť tak z dobrých vlastností group lasso. Preto zavedieme substitúciu

$$\begin{aligned}
\tilde{v}_{i,k} &:= v_{i,k} - \rho_i^x \qquad i = 1,..,M \quad k = 1,...,n \\
\tilde{u}_{r,k} &:= u_{r,k} - \rho_r^y \qquad r = 1,..,S \quad k = 1,...,n.
\end{aligned}$$

Teraz keď označíme riadky matíc X a Y ako x^i a y^r v tomto poradí, $w = [\ w_1, \ \ldots, \ w_n\ ]^T$ a $\tilde{v}^i$ a $\tilde{u}^r$ vektory multiplikatívnych váh pre jednotlivé vstupy a výstupy,

$$\begin{aligned}
\tilde{v}^i &= \left[\ \tilde{v}_{i,1}, \ \tilde{v}_{i,2}, \ \ldots, \ \tilde{v}_{i,n}\ \right]^T \in R^n, \quad i = 1,...,M, \\
\tilde{u}^r &= \left[\ \tilde{u}_{r,1}, \ \tilde{u}_{r,2}, \ \ldots, \ \tilde{u}_{r,n}\ \right]^T \in R^n, \quad r = 1,...,S,
\end{aligned}$$

tak prepísaním úlohy (2.7) dostaneme

$$\begin{aligned}
\min_{\tilde{u}^1,\ldots,\tilde{u}^S,\tilde{v}^1,\ldots,\tilde{v}^M,w} \quad & \sum_{i=1}^{M}(x^i)^T\tilde{v}^i - \sum_{r=1}^{S}(y^r)^T\tilde{u}^r + \mathbf{1}^T w \\
\text{s.t} \quad & \sum_{i=1}^{M} x_{i,1}\tilde{v}^i + \sum_{i=1}^{M} x_{i,1}\rho_i^x\mathbf{1} - \sum_{r=1}^{S} y_{r,1}\tilde{u}^r - \sum_{r=1}^{S} y_{r,1}\rho_r^y\mathbf{1} + w \geq 0 \\
& \vdots \\
& \sum_{i=1}^{M} x_{i,n}\tilde{v}^i + \sum_{i=1}^{M} x_{i,n}\rho_i^x\mathbf{1} - \sum_{r=1}^{S} y_{r,n}\tilde{u}^r - \sum_{r=1}^{S} y_{r,n}\rho_r^y\mathbf{1} + w \geq 0 \\
& \tilde{u}^r \geq 0 \qquad r = 1,\ldots,S \\
& \tilde{v}^i \geq 0 \qquad i = 1,\ldots,M.
\end{aligned} \tag{2.8}$$

Potom pridaním $l_{1,2}$ regularizácie do účelovej funkcie úlohy (2.8) dostaneme voľbu vstupov a výstupov metódou group lasso definovanú nasledovne.

Nech sú dané množiny všetkých možných vstupov $\mathscr{M}$ a výstupov $\mathscr{S}$. Potom pre dané parametere $\lambda_1 > 0$ a $\lambda_2 > 0$ budeme pod voľbou vstupov a výstupov metódou group lasso rozumieť riešenie úlohy

$$\begin{aligned}
\min_{\tilde{u}^1,\ldots,\tilde{u}^S,\tilde{v}^1,\ldots,\tilde{v}^M,w} \quad & \sum_{i=1}^{M}(x^i)^T\tilde{v}^i - \sum_{r=1}^{S}(y^r)^T\tilde{u}^r + \mathbf{1}^T w + \lambda_1\sum_{i=1}^{M}\left\|\tilde{v}^i\right\|_2 + \lambda_2\sum_{r=1}^{S}\|\tilde{u}^r\|_2 \\
s.t \quad & \sum_{i=1}^{M} x_{i,1}\tilde{v}^i + \sum_{i=1}^{M} x_{i,1}\rho_i^x\mathbf{1} - \sum_{r=1}^{S} y_{r,1}\tilde{u}^r - \sum_{r=1}^{S} y_{r,1}\rho_r^y\mathbf{1} + w \geq 0 \\
& \vdots \\
& \sum_{i=1}^{M} x_{i,n}\tilde{v}^i + \sum_{i=1}^{M} x_{i,n}\rho_i^x\mathbf{1} - \sum_{r=1}^{S} y_{r,n}\tilde{u}^r - \sum_{r=1}^{S} y_{r,n}\rho_r^y\mathbf{1} + w \geq 0 \\
& \tilde{u}^r \geq 0 \qquad r = 1,\ldots,S \\
& \tilde{v}^i \geq 0 \qquad i = 1,\ldots,M.
\end{aligned} \tag{2.9}$$

Zvolenými vstupmi a výstupmi sú tie vstupy a výstupy, ktoré zodpovedajú nenulovým vektorom $\tilde{v}^i$ a $\tilde{u}^r$ v optimálnom riešení:

$$\begin{aligned}
\mathscr{M}^* &= \left\{vstup_i \mid \left\|\tilde{v}^i\right\|_2 > 0\right\}, \\
\mathscr{S}^* &= \left\{výstup_r \mid \|\tilde{u}^r\|_2 > 0\right\}.
\end{aligned}$$

Ekvivalentnými úpravami a radom substitúcii, ktoré sú opísané v prílohe A je možné úlohu (2.9) prepísať na úlohu konvexného programovania:

$$\begin{aligned}
\min_{z} \quad & c^T z + \lambda_1\sum_{i=1}^{M}\|\tilde{v}^i\|_2 + \lambda_2\sum_{r=1}^{S}\|\tilde{u}^r\|_2 \\
& Az \geq b,
\end{aligned} \tag{2.10}$$

kde $z \in R^{Mn+Sn+n}$ je vektor obsahujúci všetky premenné $\tilde{u}_{r,k}$, $\tilde{v}_{i,k}$ a w_k[1].

[1]Z definície vektora z uvedenej v prílohe A je zrejmé, že vektory $\tilde{u}^r$ a $\tilde{v}^i$ sú obsiahnuté vo vektore z.

Ako je možné vidieť z definície (2.9), nedostatkom úlohy na výber vstupov a výstupov metódou group lasso je potreba určiť hyperparametre λ_1 a λ_2, ktoré majú podľa našich výsledkov značný vplyv na výslednú voľbu vstupov a výstupov do DEA. Úloha (2.10) je úloha konvexného programovania a na jej riešenie navrhli autori článku [5] použiť metódu ADMM (alternating direction method of multipliers). Nakoľko podľa doterajších výsledkov ide o relatívne rýchlu metódu vzhľadom na počet premenných, rozhodli sme sa ju aj my použiť v tejto práci.

2.2 Metóda rozdielom efektivít pre známe rozdelenie atribútov na vstupy a výstupy

Druhou metódou na voľbu vstupov a výstupov je nami navrhnutá metóda rozdielom efektivít. Predstavené sú dve verzie tejto metódy. Prvá, ktorá predpokladá známe rozdelenie atribútov na vstupy a výstupy (možnosť I. pri definícii voľby vstupov a výstupov na str. 17) a druhá, ktorá pracuje iba s množinou atribútov, bez nutnosti poznať ich rozdelenie na vstupy a výstupy. V tejto podkapitole sa budeme venovať prvému variantu, a teda predpokladáme, že sú dané množiny všetkých možných vstupov $\mathscr{M}$ a všetkých možných výstupov $\mathscr{S}$.

Základnou myšlienkou metódy rozdielom efektivít je pridať do aditívneho modelu diskrétne premenné zodpovedajúce buď zaradeniu alebo nezaradeniu jednotlivých možných vstupov a výstupov do DEA a potom hľadať optimálnu hodnotu týchto premenných vzhľadom na vhodne zvolené kritérium. Na voľbu vstupov a výstupov do DEA sa môžeme pozerať ako na lineárne zobrazenie matíc X a Y a premenných v úlohe (1.1) definované ako

$$X \mapsto D^x X \qquad s^x \mapsto D^x s^x \tag{2.11}$$

$$Y \mapsto D^y Y \qquad s^y \mapsto D^y s^y, \tag{2.12}$$

kde D^x a D^y sú diagonálne matice s vektormi $d^x \in \{0,1\}^M$ a $d^y \in \{0,1\}^S$ na diagonále. $d^x_i = 0$ zodpovedá nezvoleniu i-teho vstupu a $d^y_r = 0$ nezvoleniu r-tého výstupu do DEA. Po takejto transformácii je pre danú voľbu vstupov d^x a voľbu výstupov d^y efektivita k-teho DMU podľa váženého aditívneho modelu s variabilnými výnosmi z rozsahu riešením úlohy

$$\begin{aligned} \min_{\lambda, s^x, s^y} \quad & -(\rho^x)^T (D^x)^+ D^x s^x - (\rho^y)^T (D^y)^+ D^y s^y \\ \text{s.t.} \quad & D^x X\lambda + D^x s^x = D^x x_k \\ & D^y Y\lambda - D^y s^y = D^y y_k \\ & \mathbf{1}^T \lambda = 1 \\ & \lambda, s^x, s^y \geq 0, \end{aligned} \tag{2.13}$$

kde $(D^x)^+$ a $(D^y)^+$ označuje Moore-Penrosove pseudoinverzné matice k maticiam D^x a D^y. Na matice D^x a D^y sa môžme pozerať ako na matice zdopovedajúce zmene jednotiek v

aditívnom modeli, pri ktorých dovoľujeme aj nulové prvky na diagonále. Pre $d^x \in \{0,1\}^M$ a $d^y \in \{0,1\}^S$ sú pseudoinverzné matice k D^x a D^y diagonálne matice s jednotkami a nulami na diagonále, pričom platí $(D^x)^+ = D^x$ a $(D^y)^+ = D^y$. Ekvivalentne môžeme úlohu (2.13) prepísať do tvaru

$$\begin{aligned} \min_{\lambda, s^x, s^y} \quad & -(\rho^x)^T s^x - (\rho^y)^T s^y \\ \text{s.t.} \quad & D^x X\lambda + s^x = D^x x_k \\ & D^y Y\lambda - s^y = D^y y_k \\ & \mathbf{1}^T\lambda = 1 \\ & \lambda, s^x, s^y \geq 0 \end{aligned} \tag{2.14}$$

alebo ak prepíšeme ohraničenia na $s^x = D^x(x_k - X\lambda) \geq 0$ a $s^y = D^y(Y\lambda - y_k) \geq 0$, tak dostaneme

$$\begin{aligned} \min_{\lambda} \quad & -(\rho^x)^T D^x(x_k - X\lambda) - (\rho^y)^T D^y(Y\lambda - y_k) \\ \text{s.t.} \quad & D^x(x_k - X\lambda) \geq 0 \\ & D^y(Y\lambda - y_k) \geq 0 \\ & \mathbf{1}^T\lambda = 1 \\ & \lambda \geq 0. \end{aligned} \tag{2.15}$$

Na model (2.13) sa môžeme pozerať ako na definíciu efektivity ako funkcie od voľby vstupov a výstupov d^x a d^y. Avšak na to, aby sme zostavili metódu schopnú zmysluplnej voľby vstupov a výstupov do DEA, potrebujeme mať ešte nejakú znalosť o ideálnom priestore (danom vstupmi a výstupmi), ktorý najlepšie charakterizuje efektivity DMU v zmysle produkčnej funkcie, resp. ekonomickej intuície. V našom prístupe sa táto znalosť ideálnej voľby vstupov a výstupov opiera o apriórnu znalosť efektívnosti niekoľkých DMU a je zhrnutá v nasledujúcom predpoklade.

Predpoklad 1. *V množine analyzovaných DMU je daných P DMU, o ktorých je známe, že sú v zmysle produkčnej funkcie efektívne a N DMU, o ktorých je známe, že sú neefektívne, pričom* $n > P \geq 1$ *a* $n > N \geq 1$.

Z praktického hľadiska hovorí tento predpoklad o tom, že pri niekoľkých DMU v analyzovanej vzorke vieme povedať, či sú alebo nie sú efektívne. Napríklad pri aplikácii DEA na klasifikáciu firiem na podozrivé a nepodozrivé z rozkrádania verejných financií predpokladáme, že z analyzovanej vzorky firiem poznáme niekoľko príkladov ako vyzerá podozrivá firma a niekoľko príkladov ako vyzerá nepodozrivá firma.

Použitím predpokladu 1 teraz môžeme zostrojiť kritérium definujúce optimálnu voľbu vstupov a výstupov do DEA. Označme efektivity apriórne známych efektívnych DMU dané (2.13) v závislosti od d^x a d^y ako $\theta_o(d^x, d^y)$, $o \in \{1,...,P\}$, a neefektívnych DMU ako $\eta_l(d^x, d^y)$, $l \in \{1,...,N\}$. Potom na základe efektivít týchto DMU volíme vstupy a výstupy tak, aby bola maximalizovaná vzdialenosť skupiny efektívnych a neefektívnych DMU daná

funkciou $U(d^x, d^y)$ a aby boli apriórne známe efekívne DMU čo najefektívnejšie a apriórne známe neefektívne DMU čo najmenej efektívne. V tejto práci sme $U(d^x, d^y)$ zvolili ako rozdiel priemeru efektivít apriórne známych efektívnych DMU a priemeru efektivít apriórne známych neefektívnych DMU[2].

$$U(d^x, d^y) = \sum_{o=1}^{P} \frac{\theta_o(d^x, d^y)}{P} - \sum_{l=1}^{N} \frac{\eta_l(d^x, d^y)}{N}. \tag{2.16}$$

Optimálna voľba vstupov a výstupov je potom daná nasledovnou definíciou.

Nech sú dané množiny všetkých možných vstupov $\mathscr{M}$ a výstupov $\mathscr{S}$. Nech je splnený predpoklad 1. Potom pod voľbou vstupov a výstupov metódou rozdielom efektivít pre známe rozdelenie atribútov na vstupy a výstupy budeme rozumieť riešenie úlohy

$$\max_{d^x, d^y} U(d^x, d^y) = \max_{d^x, d^y} \left(\sum_{o=1}^{P} \frac{\theta_o(d^x, d^y)}{P} - \sum_{l=1}^{N} \frac{\eta_l(d^x, d^y)}{N} \right). \tag{2.17}$$

Zvolenými vstupmi a výstupmi sú tie vstupy a výstupy, ktoré zodpovedajú nenulovým prvkom d_i^x a d_r^y v optimálnom riešení:

$$\mathscr{M}^* = \{vstup_i \mid d_i^x = 1\},$$
$$\mathscr{S}^* = \{výstup_r \mid d_r^y = 1\}.$$

Geometrická interpretácia definície (2.17) je nasledovná: na voľbu vstupov a výstupov sa môžeme pozerať ako na voľbu priestoru, v ktorom odhadujeme produkčnú funkciu a meriame efektivitu. Takto teda dostaneme pre rôzne kombinácie vstupov a výstupov 2^{M+S} rôznych priestorov s odhadmi produkčných funkcií a v každom z nich má k-te DMU inú hodnotu efektivity. Cieľom úlohy (2.17) je potom vybrať taký priestor z 2^{M+S} možností, aby apriórne známe efektívne DMU boli čo najefektívnejšie, čiže čo najbližšie pri odhadnutej produkčnej funkcii v danom priestore, a apriórne známe neefektívne DMU boli čo najmenej efektívne, a teda čo najďalej od odhadnutej produkčnej funkcie.

Z hľadiska aplikácie DEA na klasifikáciu firiem na podozrivé a nepodozrivé z rozkrádania verejných financií je myšlienka metódy rozdielom efektivít taká, že častokrát poznáme už pred samotnou analýzou niekoľko medializovaných príkladov firiem, ktoré sú podozrivé z rozkrádania verejných financií. Relatívne ľahko sa tiež dajú nájsť firmy, ktoré nerozkrádajú. Na základe týchto informácií potom chceme zvoliť také vstupy a výstupy, aby boli podľa DEA známe podozrivé firmy čo najefektívnejšie a nepodozrivé čo najmenej efektívne.

Úloha (2.17) je úloha celočíselného programovania, ktorej účelová funkcia je vážený súčet lineárnych programov (2.13). Z tohto dôvodu je úloha (2.17) ťažko riešiteľná a na jej riešenie sme v tejto práci používali genetické algoritmy[3].

[2] Z tejto definície pochádza aj názov metóda rozdielom efektivít.

[3] Použili sme zabudovanú funkciu ga() v softvéri MATLAB.

2.3 Metóda rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy

V predošlej podkapitole sme predstavili prvý variant metódy rozdielom efektivít, ktorý sa opieral o predpoklad známeho rozdelenia atribútov na vstupy a výstupy. V tejto podkapitole tento predpoklad vynecháme a predpokladáme iba, že je daná množina atribútov $\mathscr{A}$, pričom nie je známe, ktorý atribút je vstup a ktorý výstup.

Metóda rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy je rozšírením úvah variantu známeho rozdelenia atribútov. Podobne ako predtým, je aj teraz jej hlavnou myšlienkou pridať do aditívnho modelu diskrétne premenné, ktoré v tomto prípade zodpovedajú voľbe jednotlivých atribútov ako vstup, voľbe ako výstup alebo ich vynechaniu z DEA analýzy. Potom hľadáme optimálnu hodnotu týchto premenných vzhľadom na vhodne zvolené kritérium.

Uvažujme úlohu (2.14) pre dané rozdelenie atribútov na vstupy a výstupy. Ak vynásobíme druhé ohraničenie tejto úlohy číslom -1, môžeme ju ekvivalentne zapísať ako

$$\begin{aligned} \min_{\lambda, s^x, s^y} \quad & -(\rho^x)^T s^x - (\rho^y)^T s^y \\ \text{s.t.} \quad & D^x X\lambda + s^x = D^x x_k \\ & -D^y Y\lambda + s^y = -D^y y_k \\ & \mathbf{1}^T\lambda = 1 \\ & \lambda, s^x, s^y \geq 0. \end{aligned} \tag{2.18}$$

Keď potom označíme $A := [\ X^T \quad Y^T\ ]^T$, $s := [\ (s^x)^T \quad (s^y)^T\ ]^T$ a maticu

$$D := \begin{bmatrix} D^x & 0 \\ 0 & -D^y \end{bmatrix} \in R^{(M+S)\times(M+S)},$$

môžeme úlohu (2.18) prepísať do tvaru

$$\begin{aligned} \min_{\lambda, s} \quad & -\rho^T s \\ \text{s.t.} \quad & DA\lambda + s = Da_k \\ & \mathbf{1}^T\lambda = 1 \\ & \lambda, s \geq 0, \end{aligned} \tag{2.19}$$

kde a_k je stĺpec A zodpovedajúci k-temu DMU,

$$\rho_j = \frac{1}{R_j}, \quad R_j = \max_k a_{j,k} - \min_k a_{j,k}, \quad j \in \{1, ..., M+S\} \tag{2.20}$$

a $a_{j,k}$ je prvok A prislúchajúci j-temu atribútu a k-temu DMU. Matica D je diagonálna matica s vektorom $d \in \{-1, 0, 1\}^{M+S}$ na diagonále. Úloha (2.19) dáva definíciu efektivity k-teho

DMU ako funkcie od voľby vstupov a výstupov d, pričom A je dátová matica obsahujúca všetky atribúty pre všetky DMU. Z odvodenia úlohy (2.19) vyplýva, že $d_j = -1$ zodpovedá voľbe j-teho atribútu ako výstupu, $d_j = 1$ voľbe ako vstupu a $d_j = 0$ zodpovedá vynechaniu j-teho atribútu z DEA. Ekvivalentne môžeme úlohu (2.19) zapísať do tvaru

$$\begin{aligned} \min_{\lambda,s} \quad & -\rho^T D^2 s \\ \text{s.t.} \quad & D^2 A\lambda + Ds = D^2 a_k \\ & \mathbf{1}^T \lambda = 1 \\ & \lambda, s \geq 0 \end{aligned} \tag{2.21}$$

a ak prepíšeme ohraničenia v (2.19) na $s = D(a_k - A\lambda) \geq 0$, dostávame

$$\begin{aligned} \min_{\lambda} \quad & -(\rho)^T D(a_k - A\lambda) \\ \text{s.t.} \quad & D(a_k - A\lambda) \geq 0 \\ & \mathbf{1}^T \lambda = 1 \\ & \lambda \geq 0. \end{aligned} \tag{2.22}$$

Ak sa pozrieme bližšie na úlohu (2.21), resp. na odvodenie (2.19), vidíme, že o tom, či je daný atribút vstup alebo výstup rozhoduje iba znamienko pri premennej s v ohraničeniach. Ostatné členy potom vyjadrujú zaradenie alebo nezaradenie atribútov do DEA.

Rovnako ako v prípade metódy rozdielom efektivít pre známe rozdelenie atribútov na vstupy a výstupy predpokladáme splnenie predpokladu 1, a teda predpokladáme apriórnu znalosť P efektívnych a N neefektívnych DMU. Keď potom označíme optimálnu hodnotu účelovej funkcie úlohy (2.19) pre o-te apriórne známe efektívne DMU v závislosti od d ako $\bar{\theta}_o(d)$, $o \in \{1,...,P\}$ a pre l-té apriórne známe neefektívne DMU ako $\bar{\eta}_l(d)$, $l \in \{1,...,N\}$, tak kritérium $\bar{U}(d)$ na voľbu vstupov a výstupov metódou rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy je dané ako

$$\bar{U}(d) = \sum_{o=1}^{P} \frac{\bar{\theta}_o(d)}{P} - \sum_{l=1}^{N} \frac{\bar{\eta}_l(d)}{N}. \tag{2.23}$$

Potom môžeme voľbu vstupov a výstupov metódou rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy definovať nasledovne.

Nech je daná množina atribútov $\mathscr{A}$. Nech je splnený predpoklad 1. Potom pod voľbou vstupov a výstupov metódou rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy rozumieme riešenie úlohy

$$\max_d \bar{U}(d) = \max_d \left(\sum_{o=1}^{P} \frac{\bar{\theta}_o(d)}{P} - \sum_{l=1}^{N} \frac{\bar{\eta}_l(d)}{N} \right). \tag{2.24}$$

Zvolenými vstupmi sú atribúty, ktoré zodpovedjú $d_j = 1$ a zvolenými výstupmi sú atribúty,

ktoré zodpovedajú $d_j = -1$ v optimálnom riešení:

$$\mathscr{M}^* = \left\{ atribút_j \mid d_j = 1 \right\},$$
$$\mathscr{S}^* = \left\{ atribút_j \mid d_j = -1 \right\}.$$

Podobne ako predtým je cieľom úlohy (2.24) nájsť taký priestor z 3^{M+S} možností, aby apriórne známe efektívne DMU boli čo najbližšie k odhadnutej produkčnej funkcii v danom priestore, čiže aby boli čo najefektívnejšie, a apriórne známe neefektívne DMU boli čo najďalej od odhadnutej produkčnej funkcie, a teda aby boli čo najmenej efektívne.

Na rozdiel od metódy group lasso a metódy rozdielom efektivít pre známe rozdelenie atribútov na vstupy a výstupy nepotrebuje metóda rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy predpoklad, že vieme rozlíšiť, čo môže byť vstup a čo výstup. Toto je výhodou obzvlášť pri analýzach, kedy nie je úplne jasné, čo sa dá považovať za vstup a čo za výstup. Príkladom takejto analýzy je aj klasifikácia firiem na podozrivé a nepodozrivé z rozkrádania verejných financií, kedy nie je vždy jasné, ktorý atribút zvyšuje riziko rozkrádania a ktorý ho naopak znižuje. Ako bude možné vidieť z výsledkov v tretej kapitole, táto flexibilita pri voľbe čo je vstup a čo je výstup môže viesť k lepším výsledkom v porovnaní s prvými dvomi metódami.

Podobne ako v prípade známeho rozdelenia atribútov na vstupy a výstupy je úloha (2.24) problém celočíselného programovania, ktorého účelová funkcia je vážený súčet lineárnych programov (2.19). Ide o náročnú úlohu, ktorej optimálne riešenie je jedna z 3^{M+S} možností voľby vstupov a výstupov, a rovnako ako v prípade známeho rozdelenia atribútov na vstupy a výstupy, používame v tejto práci na riešenie (2.24) genetické algoritmy[4].

2.4 Diskusia k metóde rozdielom efektivít

Oba varianty metódy rozdielom efektivít predstavujú nový automatizovaný spôsob na voľbu vstupov a výstupov iba na základe dát a bez nutnosti ľudskej expertízy. Predpokladajú apriórnu znalosť niekoľkých efektívnych a niekoľkých neefektívnych DMU, a v prípade metódy rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy dokonca nie je potrebná ani znalosť toho, či je daný atribút vstup alebo výstup. Nevýhodou takejto voľby vstupov a výstupov však je, že úlohy (2.17) a (2.24) sú úlohy celočíselného programovania a pri veľkom počte možných atribútov je ich riešiteľnosť náročná. Preto v nasledujúcej podkapitole načrtneme ako by sa dala úloha na voľbu vstupov a výstupov metódou rozdielom efektivít relaxovať na spojitú úlohu, ktorá je už o niečo ľahšie riešiteľná.

[4]Používame zabudovanú funkciu ga() v softvéri MATLAB.

2.4.1 Relaxácia úlohy na voľbu vstupov a výstupov metódou rozdielom efektivít

Náčrt odvodenia relaxovanej spojitej úlohy na riešenie úlohy na voľbu vstupov a výstupov metódou rozdielom efektivít predstavíme pre metódu rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy. V prípade známeho rozdelenia atribútov sú myšlienky analogické s nižšie uvedeným postupom.

Relaxácia úlohy na voľbu vstupov a výstupov (2.24) spočíva v nahradení diskrétnych premenných $d \in \{-1,0,1\}^{M+S}$ spojitými premennými $d \in [-1,1]^{M+S}$. Uvažujme definíciu efektivity k-teho DMU ako funkcie od voľby vstupov a výstupov d v tvare (2.22). Pre jednoduchosť budeme uvažovať apriórnu znalosť iba jedného efektívneho DMU a jedného neefektívneho DMU. Potom je relaxácia úlohy (2.24) na voľbu vstupov a výstupov metódou rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy, a pre $P = 1$ a $N = 1$ daná úlohou

$$\begin{array}{ll} \max\limits_{d} & \left(\bar{\theta}_o(d) - \bar{\eta}_l(d)\right) \\ \text{s.t.} & d \in [-1,1]^{M+S}, \end{array} \tag{2.25}$$

ktorú môžeme dosadením úlohy (2.22) za $\bar{\theta}_o(d)$ a $\bar{\eta}_l(d)$ napísať ako

$$\max_{d} \left\{ \begin{array}{lll} \min\limits_{\lambda_o} & -(\rho)^T D(a_o - A\lambda_o) & \\ \text{s.t.} & D(a_o - A\lambda_o) \geq 0 & \\ & \mathbf{1}^T\lambda_o = 1 & \\ & \lambda_o \geq 0 & \end{array} \right. + \left. \begin{array}{lll} -\min\limits_{\lambda_l} & -(\rho)^T D(a_l - A\lambda_l) & \\ \text{s.t.} & D(a_l - A\lambda_l) \geq 0 & \\ & \mathbf{1}^T\lambda_l = 1 & \\ & \lambda_l \geq 0 & \end{array} \right\}$$
$$\text{s.t.} \quad d \in [-1,1]^{M+S}, \tag{2.26}$$

kde a_o, vektor premenných λ_o a prvý lineárny program v poradí zodpovedajú apriórne známemu efektívnemu DMU a a_l, vektor premenných λ_l a druhý lineárny program zodpovedajú aporiórne známemu neefektívnemu DMU.

Úloha lineárneho programovania pre apriórne známe neefektívne DMU v (2.26) sa dá pri nezmenených ohraničeniach upraviť na

$$-\min_{\lambda_l} -(\rho)^T D(a_l - A\lambda_l) = \max_{\lambda_l} (\rho)^T D(a_l - A\lambda_l) \tag{2.27}$$

a podobne pre lineárny program pre apriórne známe efektívne DMU platí

$$\min_{\lambda_o} -(\rho)^T D(a_o - A\lambda_o) = \left\{ \begin{array}{ll} \max\limits_{\lambda_o,\tau} & \tau \\ \text{s.t.} & \tau \leq -(\rho)^T D(a_o - A\lambda_o). \end{array} \right. \tag{2.28}$$

Takto získame v účelovej funkcii (2.26) súčet dvoch maximalizačných úloh, a keďže sú

nezávislé, tak ich môžeme agregovať do jednej úlohy. Potom z (2.26) dostávame

$$\max_{d} \left\{ \begin{array}{ll} \max\limits_{\lambda_o,\lambda_l,\tau} & \tau + (\rho)^T D(a_l - A\lambda_l) \\ \text{s.t.} & \tau \leq -(\rho)^T D(a_o - A\lambda_o) \\ & D(a_o - A\lambda_o) \geq 0 \\ & D(a_l - A\lambda_l) \geq 0 \\ & \mathbf{1}^T \lambda_o = 1, \quad \lambda_o \geq 0 \\ & \mathbf{1}^T \lambda_l = 1, \quad \lambda_l \geq 0 \end{array} \right\} \quad \text{s.t.} \quad d \in [-1,1]^{M+S}. \tag{2.29}$$

Úloha (2.29) predstavuje dvojstupňovú maximalizáciu a keď označíme jej hodnotu účelovej funkcie ako $V(d)$, optimálne riešenie ako $\hat{d}$, hodnotu účelovej funkcie lineárneho programu v závislosti od d ako $V_{LP}(d,\lambda_o,\lambda_l,\tau)$ a optimálne riešenie lineárneho programu ako $\hat{\lambda}_o$, $\hat{\lambda}_l$ a $\hat{\tau}$, tak platí

$$V_{LP}(d,\hat{\lambda}_o,\hat{\lambda}_l,\hat{\tau}) \geq V_{LP}(d,\lambda_o,\lambda_l,\tau) \quad \forall \lambda_o,\lambda_l,\tau \tag{2.30}$$

$$V(\hat{d}) \geq V(d) = V_{LP}(d,\hat{\lambda}_o,\hat{\lambda}_l,\hat{\tau}) \quad \forall d. \tag{2.31}$$

Nakoľko λ_o a λ_l môžeme zvoliť ako jednotkové vektory s jednotkou na o-tom a l-tom mieste v tomto poradí, tak lineárny program v úlohe (2.29) má riešenie pre každé d.

Namiesto dvojstupňovej maximalizácie (2.29), ktorá je stále náročná na riešenie, môžeme optimálne hodnoty premenných λ_o, λ_l, τ a d hľadať simultánne ako riešenie úlohy

$$\begin{array}{ll} \max\limits_{d,\lambda_o,\lambda_l,\tau} & \tau + (\rho)^T D(a_l - A\lambda_l) \\ \text{s.t.} & \tau \leq -(\rho)^T D(a_o - A\lambda_o) \\ & D(a_o - A\lambda_o) \geq 0 \\ & D(a_l - A\lambda_l) \geq 0 \\ & \mathbf{1}^T \lambda_o = 1, \quad \lambda_o \geq 0 \\ & \mathbf{1}^T \lambda_l = 1, \quad \lambda_l \geq 0 \\ & d \in [-1,1]^{M+S}. \end{array} \tag{2.32}$$

Ak označíme $V_{Sim}(d,\lambda_o,\lambda_l,\tau)$ hodnotu účelovej funkcie úlohy (2.32) a d^*, λ_o^*, λ_l^*, τ^* jej optimálne riešenie, tak platí

$$V_{Sim}(d^*,\lambda_o^*,\lambda_l^*,\tau^*) \geq V_{Sim}(d,\lambda_o,\lambda_l,\tau) \quad \forall d,\lambda_o,\lambda_l,\tau. \tag{2.33}$$

Každému $d \in [-1,1]^{M+S}$ prislúcha v úlohe (2.29) aj riešenie lineárneho programu $\hat{\lambda}_o, \hat{\lambda}_l, \hat{\tau}$. Z formulácie (2.29) a (2.32) však vyplýva, že d, $\hat{\lambda}_o$, $\hat{\lambda}_l$ a $\hat{\tau}$ zároveň zodpovedajú aj prípustnému riešeniu úlohy (2.32), a teda platí

$$V_{Sim}(d^*,\lambda_o^*,\lambda_l^*,\tau^*) \geq V_{Sim}(d,\hat{\lambda}_o,\hat{\lambda}_l,\hat{\tau}) \quad \forall d. \tag{2.34}$$

Uvedomme si tiež, že účelové funkcie úloh (2.29) a (2.32) sú rovnaké, a teda

$$V_{Sim}(d,\hat{\lambda}_o,\hat{\lambda}_l,\hat{\tau}) = V_{LP}(d,\hat{\lambda}_o,\hat{\lambda}_l,\hat{\tau}) = V(d). \tag{2.35}$$

Vzhľadom na to, že pre d platí v (2.29) iba ohraničenie $d \in [-1,1]^{M+S}$, tak optimálne d^* v (2.32) je prípusným riešením v (2.29). Potom ak do (2.29) dosadíme d^*, tak z ohraničení oboch úloh vyplýva, že v prislúchajúcom lineárnom programe sú premenné λ_o^*, λ_l^* a τ^* prípustné riešenia. Z toho vyplýva, že

$$V(d^*) \geq V_{LP}(d^*,\lambda_o^*,\lambda_l^*,\tau^*) = V_{Sim}(d^*,\lambda_o^*,\lambda_l^*,\tau^*) \geq V_{Sim}(d,\hat{\lambda}_o,\hat{\lambda}_l,\hat{\tau}) = V(d) \quad \forall d, \tag{2.36}$$

z čoho potom dostávame, že ak je d^* optimálne v (2.32), tak je optimálne aj v (2.29).

2.4.2 SOCP formulácia ohraničení úloh (2.32)

V tejto podkapitole sa bližšie pozrieme na ohraničenia úlohy (2.32) a ukážeme, že ohraničenia $D(a_o - A\lambda_o) \geq 0$ a $D(a_l - A\lambda_l) \geq 0$ sa dajú preformulovať do tvaru ohraničení programovania nad kužeľmi druhého rádu (second-order cone programming; SOCP).

Ohraničenie $D(a_o - A\lambda_o) \geq 0$ sa dá po zložkách zapísať ako

$$d_j (a_o - A\lambda_o)_j \geq 0 \quad \forall j, \tag{2.37}$$

kde $(a_o - A\lambda_o)_j$ označuje j-ty prvok vektora $(a_o - A\lambda_o)$. Ak potom vynásobíme (2.37) číslom 2 a pripočítame na obe strany nerovnosti výraz $(d_j)^2 + (a_o - A\lambda_o)_j^2$, tak dostaneme

$$2d_j (a_o - A\lambda_o)_j + (d_j)^2 + (a_o - A\lambda_o)_j^2 \geq (d_j)^2 + (a_o - A\lambda_o)_j^2 \quad \forall j, \tag{2.38}$$

čo sa dá prepísať ako

$$\Big(d_j + (a_o - A\lambda_o)_j\Big)^2 \geq \left\| \begin{bmatrix} d_j, & (a_o - A\lambda_o)_j \end{bmatrix}^T \right\|_2^2 \quad \forall j. \tag{2.39}$$

Pre $D(a_l - A\lambda_l) \geq 0$ potom postupujeme analogicky. Keď teraz definujeme

$$B_j^o := \begin{bmatrix} e_j^T & 0^T & 0^T \\ 0^T & -e_j^T A & 0^T \end{bmatrix} \quad b_{j,o} := \begin{bmatrix} 0 \\ a_{j,o} \end{bmatrix} \quad c_j^o := \begin{bmatrix} e_j^T & -e_j^T A & 0^T \end{bmatrix} \tag{2.40}$$

$$B_j^l := \begin{bmatrix} e_j^T & 0^T & 0^T \\ 0^T & 0^T & -e_j^T A \end{bmatrix} \quad b_{j,l} := \begin{bmatrix} 0 \\ a_{j,l} \end{bmatrix} \quad c_j^l := \begin{bmatrix} e_j^T & 0^T & -e_j^T A \end{bmatrix}, \tag{2.41}$$

kde e_j je jednotkový vektor s jendotkou na j-tom mieste a definujeme

$$z := \begin{bmatrix} d \\ \lambda_o \\ \lambda_l \end{bmatrix}, \tag{2.42}$$

tak ohraničenia $D(a_o - A\lambda_o) \geq 0$ a $D(a_l - A\lambda_l) \geq 0$ sú ekvivalentné ohraničeniam

$$c_j^o z + a_{j,o} \geq \left\| B_j^o z + b_{j,o} \right\|_2 \qquad \forall j \tag{2.43}$$

$$c_j^l z + a_{j,l} \geq \left\| B_j^l z + b_{j,l} \right\|_2 \qquad \forall j. \tag{2.44}$$

Výrazy (2.43) a (2.44) zodpovedajú ohraničeniam v tvare ohraničení SOCP.

Aj napriek možnej formulácii väčšiny ohraničení do tvaru SOCP je úloha (2.32) nekonvexná úloha matematického programovania s bilineárnou účelovou funkciou a s bilineárnou funkciou v ohraničeniach. Z formulácie úlohy (2.32) sa však zdá, že by sa mohla dať preformulovať na úlohu bipartitného bilineárneho programovania opísanú v článku [7]. V tomto článku navrhli autori spôsob SOCP relaxácie úlohy bipartitného bilineárneho programovania a aj branch-and-bound algoritmus na jej riešenie. Formulácia (2.32) podľa [7] ako bilineárne bipartitné programovanie je iba náčrt spôsobu, akým by sa mohla dať úloha (2.32) efektívne riešiť, avšak výsledky takéhoto postupu v tejto práci neuvázdame.

2.4.3 Ohraničenie počtu vstupov a výstupov

Ani v jednej z definícii (2.9), (2.17) a (2.24) nefiguruje ohraničenie na minimálny počet vstupov a výstupov, a teda pripúšťame voľbu 0 vstupov, resp. 0 výstupov. V prípade metód rozdielom efektivít by sa takéto ohraničenia na počet vstupov a výstupov (resp. počet atribútov) dali podľa potreby jednoducho pridať do úloh na voľbu vstupov a výstupov. V prípade voľby vstupov a výstupov metódou rozdielom efektivít pre známe rozdelenie atribútov na vstupy a výstupy by stačilo pridať do úlohy (2.17) ohraničenia $\sum_i d_i^x \geq K^x$ a $\sum_r d_r^y \geq K^y$, kde K^x a K^y vyjadrujú minimálny počet vstupov a výstupov v tomto poradí. Pri voľbe vstupov a výstupov metódou rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy by sme docielili ohraničenie minimálneho počtu atribútov pridaním ohraničenia $\sum_j d_j \geq K$ do úlohy (2.24), kde K vyjadruje minimálny počet atribútov v DEA. Analogicky by bolo možné ohraničiť aj maximálny počet vstupov a výstupov (resp. atribútov). V tejto práci sme však takéto ohraničenia nepridávali a uvažujeme iba pôvodné úlohy v definíciách (2.17) a (2.24).

Kapitola 3

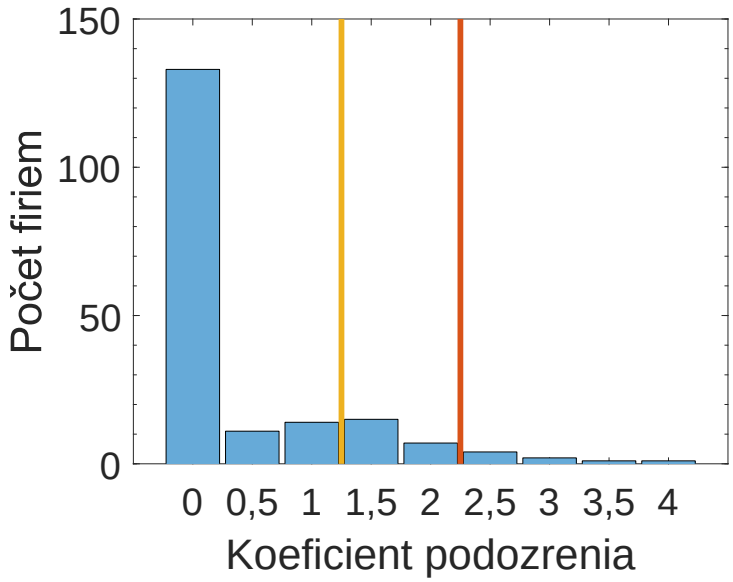
Výsledky analýzy

V poslednej kapitole tejto práce sa venujeme výsledkom klasifikácie firiem na podozrivé a nepodozrivé z rozkrádania verejných financií pomocou DEA. Výsledky sú prezentované formou grafov a tabuliek, ktoré vyjadrujú distribúciu koeficientu podozrenia (KP) vo firmách efektívnych podľa DEA. V prvej časti kapitoly najskôr porovnáme metódy na voľbu vstupov a výstupov vzhľadom na aplikáciu DEA na klasifikáciu firiem na podozrivé a nepodozrivé z rozkrádania verejných financií, a potom sa zaoberáme možnosťou použiť DEA ako klasifikátor viacerých tried.

Predtým ako sa budeme venovať samotným výsledkom, pozrieme sa najskôr na dáta z pohľadu koeficientu podozrenia. Analyzovali sme spolu 188 firiem, ktoré vyhrali aspoň jedno verejné obstarávanie s predmetom *Stavebné práce na stavbe budov pre voľný čas, šport, kultúru, ubytovanie a reštauračné stravovanie*. V grafe na Obr. (3.1) je zobrazená distribúcia koeficientu podozrenia pre všetky analyzované firmy. Koeficienty podozrenia sme firmám prideľovali subjektívnou kontrolou tak, že vždy sme skontrolovali každú firmu samostatne a na základe kritérií z podkapitoly 1.2 sme jej pridelili príslušnú hodnotu KP v rozmedzí 0 – 4 .

Na základe hodnoty koeficientu podozrenia môžeme rozdeliť firmy do troch kategórií. Prvá kategória sú *nepodozrivé* firmy, ktoré majú koeficient podozrenia 0; $0,5$ alebo 1. V grafe na Obr. (3.1) to zodpovedá firmám naľavo od zvislej oranžovej čiary. Druhá kategória sú *mierne podozrivé* firmy, ktoré sa v grafe na Obr. (3.1) nachádzajú medzi červenou a oranžovou čiarou a pripadajú im hodnoty koeficientu podozrenia $1,5$ a 2. Posledná kategória sú *veľmi podozrivé* firmy, ktorých koeficient podozrenia sa nachádza medzi hodnotami $2,5$; 3; $3,5$ a 4. Tieto firmy sa nachádzajú napravo od zvislej červenej čiary v histograme na Obr. (3.1). Ako môžeme vidieť na tomto histograme, až 159 firiem z analyzovaných 188 spadá do kategórie nepodozrivá. Zo zvyšných 29 firiem je väčšina mierne podozrivá (spolu 21) a iba 8 firiem v dátovej sade môžeme označiť ako veľmi podozrivé.

Cieľom metódy na klasifikáciu firiem na podozrivé a nepodozrivé z rozkrádania verej-



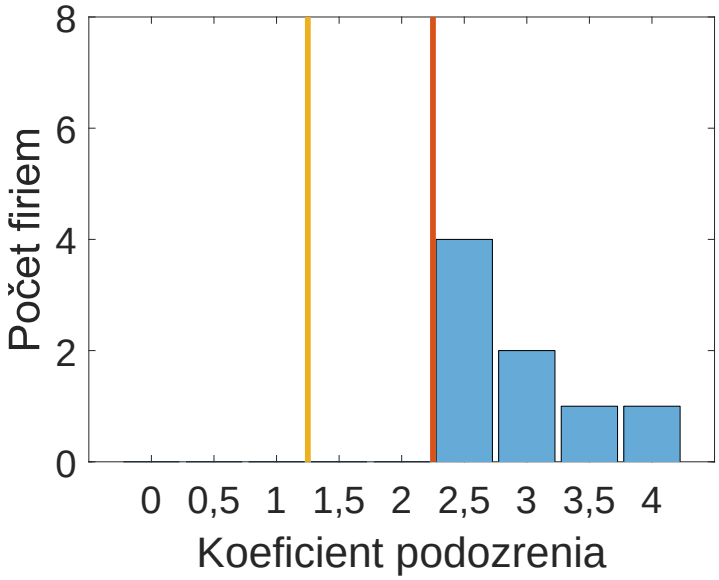
Obr. (3.1) Histogram rozdelenia koeficientu podozrenia pre jednotlivé firmy v dátovej sade.

ných financií pomocou DEA je zvoliť také vstupy a výstupy, aby boli firmy s KP 2,5; 3; 3,5 a 4 (veľmi podozrivé firmy) efektívne podľa DEA a firmy s KP 0; 0,5 a 1 (nepodozrivé firmy) neefektívne. Pre firmy s KP 1,5 a 2 (mierne podozrivé firmy) je prijateľná ľubovoľná efektívnosť, nakoľko koeficient podozrenia bol prideľovaný subjektívnou konrolou na základe nám dostupných informáciách a aj keď boli nejaké podozrenia nájdené, nezdali sa nám dostatočné na zaradenie firiem do kategórie veľmi podozrivých. Ak by však koeficient podozrenia prideľoval niekto iný, tak by týmto firmám mohol prideliť vyšší, či nižší KP a zaradiť ich tak do inej kategórie.

V grefe na Obr. (3.2) môžeme viedieť distribúciu KP pre veľmi podozrivé firmy. Jedná sa o 8 firiem zo 188 a v ďalších podkapitolách tejto práce sa za dobrý výsledok považuje nájdenie takých vstupov a výstupov, aby boli tieto firmy efektívne podľa DEA.

3.1 Porovnanie metód na voľbu vstupov a výstupov

Na klasifikáciu firiem na podozrivé a nepodozrivé z rokrádania verejných financií používame teóriu DEA modelov. DEA rozdelí firmy podľa ich efektívnosti do dvoch kategórií, t.j. pozdozrivé z rozkrádania verejných financií a nepodozrivé z rozkrádania verejných financií. Pojem *efektívny* v zmysle DEA stotožňujeme s pojmom *podozrivý* z rozkrádania verejných financií a pojem *neefektívny* podľa DEA s pojmom *nepodozrivý* z rozkrádania verejných financií. Nevýhodou takéhoto postupu je, že musíme zvoliť také vstupy a výstupy do DEA, ktoré dobre charakterizujú podozrenie z rokrádania. Ako určiť tieto vstupy a výstupy nie je vždy jasné, preto sme v tejto práci vytvorili spôsob na automatizovanú voľbu vstupov a výstupov na základe apriórnej znalosti niekoľkých príkladov nami identifikovaných podozrivých



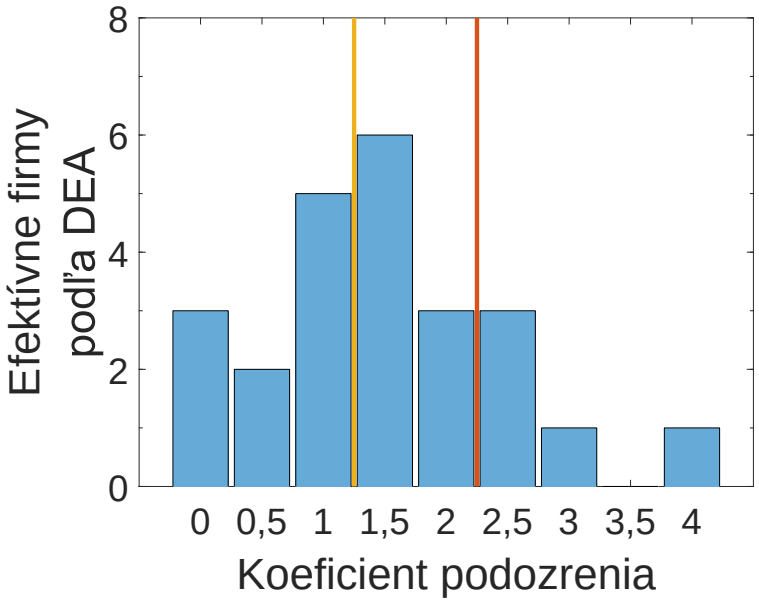
Obr. (3.2) Histogram rozdelenia koeficientu podozrenia pre firmy s $KP > 2$.

a nepodozrivých firiem. Túto metódu ďalej uvádzame pod názvom metóda rozdielom efektivít a v tejto podkapitole porovnáme jej výsledky s výsledkami metódy group lasso [5] a s výsledkami expertnej voľby vstupov a výstupov z [1].

Výsledky sú prezentované formou histogramov vyjadrujúcich distribúciu subjektívne pridelených koeficientov podozrenia firiem, ktoré boli efektívne podľa DEA pri použití vstupov a výstupov volených jednotlivými metódami. Napríklad v histograme na Obr. (3.3) vidíme, že podľa DEA bolo efektívnych 24 firiem, pričom nami pridelené hodnoty KP ukazujú, že 3 z týchto firiem majú KP 0; 2 firmy KP 0,5; 5 firiem KP 1; 6 firiem KP 1,5; atď. Okrem samotného histogramu sledujeme aj priemernú hodnotu koeficientu podozrenia efektívnych firiem.

V grafe na Obr. (3.3) je zobrazený výsledný histogram rozdelenia koeficientov podozrenia pre firmy efektívne podľa DEA pri expertnej voľbe vstupov a výstupov. Spolu bolo efektívnych 24 firiem, z čoho bolo podľa koeficientu podozrenia až 5 firiem veľmi podozrivých. Z hľadiska klasifikácie firiem na podozrivé a nepodozrivé z rozkrádania verejných financii to znamená, že metóda DEA správne vybrala až 5 z 8 veľmi podozrivých firiem. Na druhej strane sa medzi efektívnymi firmami nachádzalo podľa koeficientu podozrenia spolu až 10 nepodozrivých firiem. Priemer KP efektívnych firiem bol 1,48, čo zodpovedá KP nepodozrivej firmy[1].

[1] Výsledky v grafe na Obr. (3.3) sa mierne líšia od výsledkov publikovaných v [1], kvôli chybe v zdrojovom kóde v [1].



Obr. (3.3) Efektívne firmy podľa DEA pri expertnej voľbe vstupov a výstupov. Spolu bolo 24 efektívnych firiem s priemernou hodnotou KP $1,48$. Do DEA bolo zvolených 5 vstupov a 3 výstupy.

3.1.1 Výsledky metódy group lasso

Výsledky metódy group lasso sú zobrazené v grafoch na Obr. (3.4). Grafy zobrazujú štyri situácie zodpovedajúce rôznym hodnotám hyperparametrov λ_1 a λ_2. Úlohu konvexného programovania na voľbu vstupov a výstupov metódou group lasso (2.10) sme pre jednotlivé situácie riešili metódou ADMM navrhnutou v [5]. Z výsledkov môžeme pozorovať, že pri rastúcich hodnotách hyperparametrov λ_1 a λ_2 sa znižuje počet zvolených vstupov, resp. výstupov do DEA. To vyplýva z formulácie úlohy (2.9), nakoľko čím väčšia je hodnota λ_1 a λ_2, tým väčšia je penalizácia za nenulové multiplikatívne váhy v optimálnom riešení. Z toho tiež vyplýva, že čím sú hodnoty λ_1 a λ_2 väčšie, tým menej firiem je efektívncyh podľa DEA a tým pádom DEA vyberie menej firiem ako podozrivé z rozkrádania verejných financií.

Z grafu na Obr. (3.4) sa dá pozorovať aj to, že vzhľadom na priemerný KP efektívnych firiem došlo oproti expertnej voľbe vstupov a výstupv iba k miernemu zlepšeniu. Tento výsledok môžeme vidieť aj v Tab. (3.1), kde sú zhrnuté výsledky všetkých metód. Jediný prípad, kedy došlo vzhľadom na priemerný KP efektívnych firiem k signifikantnému zlepšeniu súladu medzi hodnotami KP a efektívnymi firmami v DEA oproti expertnej voľbe vstupov a výstupov je pre $\lambda_1 = 1,5$ a $\lambda_2 = 1,5$ (graf na Obr. (3.4a)). V tomto prípade však boli efektívne iba 3 firmy, a z toho 1 bola veľmi podozrivá z hľadiska KP. Zvyšných 7 veľmi podozrivých firiem z analyzovanej vzorky bolo neefektívnych podľa DEA, čiže klasifikovaných ako nepodorivé z rozkrádania verejných financií. Dokonca v prípade, že $\lambda_1 = 1$ a $\lambda_2 = 1$ došlo k značnému zhoršeniu súladu medzi efektívnymi firmami podľa DEA

a hodnotám KP oproti expertnej voľbe vstupov a výstupov. Možným vysvetlením výsledkov je to, že lasso má tendenciu vyberať zo skupiny korelovaných premenných iba jednu a je jedno ktorú. Takže metóda group lasso odfiltruje z možných vstupov a výstupov veľmi korelované vstupy, resp. výstupy, ale to nijako nezaručuje správnu klasifikáciu firiem z hľadiska podozrenia z rozkrádania verejných financií.

Za povšimnutie stojí aj fakt, že metóda group lasso je citlivá aj na malé zmeny hyperparametrov λ_1 a λ_2. To je možné pozorovať v grafoch na Obr. (3.4c) a (3.4d), kde je už pri malej zmene λ_1 okolo $0,03$ a rovnakej hodnote λ_2 rozdiel počtu efektívnych firiem 8.

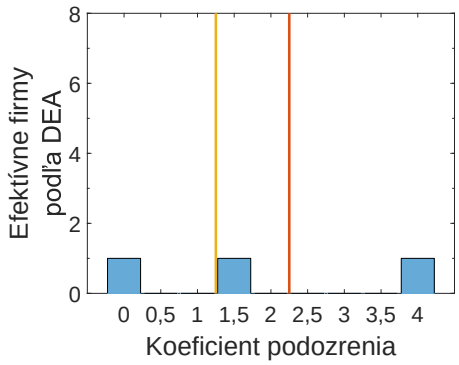
3.1.2 Výsledky metódy rozdielom efektivít pre známe rozdelenie atribútov na vstupy a výstupy

Druhou metódou použitou na voľbu vstupov a výstupov do DEA je metóda rozdielom efektivít pre známe rozdelenie atribútov na vstupy a výstupy. Predpokladáme v nej apriórnu znalosť P príkladov podozrivých firiem a N príkladov nepodozrivých firiem. Vstupy a výstupy volíme tak, aby boli apriórne známe podozrivé firmy čo najefektívnejšie v DEA a apriórne známe nepodozrivé firmy čo najmenej efektívne. V praxi takýto postup znamená, že ešte pred DEA analýzou subjektívne skontrolujeme niekoľko firiem z dátovej sady a rozhodneme, či je vhodné ich použiť na voľbu vstupov a výstupov.

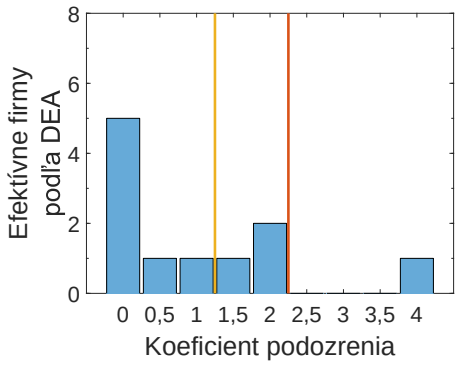
V našej práci sme posudzovali vhodnosť použitia firiem ako príkladov pre voľbu vstupov a výstupov na základe KP. Výsledky pre rôzne počty príkladov P a N a rôzne KP týchto príkladov sú zobrazené v grafoch na Obr. (3.5) a pre jednotlivé možnosti sme úlohy celočíselného programovania na voľbu vstupov a výstupov (2.17) riešili genetickými algoritmami (funkciou ga() v softvéri MATLAB). Pre jednotlivé možnosti sme za riešenie úlohy (2.17) považovali takú voľbu vstupov a výstupov, ktorá bola najčastejším riešením v piatich zbehnutiach funkcie ga() v softvéri MATLAB.

Grafy na Obr. (3.5a) a (3.5c) predstavujú realistické situácie s malým počtom príkladov. V prvom prípade používame na voľbu vstupov a výstupov príklad jednej podozrivej firmy s KP 3 a príklady nepodozrivých firiem s KP 0; 0 a 1. V druhom prípade máme k dispozícii príklady nepodozrivých firiem s KP 0; 0 a $0,5$ a príklad podozrivej firmy s KP $2,5$. Zvyšné grafy na Obr. (3.5b) a (3.5d) predstavujú extrémne prípady. V grafe na Obr. (3.5b) používame na voľbu vstupov a výstupov až 5 príkladov podozrivých firiem s KP $2,5$; $3,5$; $2,5$; $2,5$ a $2,5$ a 5 príkladov nepodozrivých firiem s KP 0 pre všetky. Naopak v grafe na Obr. (3.5d) je medzi dvomi príkladmi podozrivých firiem aj firma s KP 0 a druhá firma má KP $2,5$ a medzi tromi príkladmi nepodozrivých firiem je firma s KP $2,5$ a zvyšné dve firmy majú KP 0.

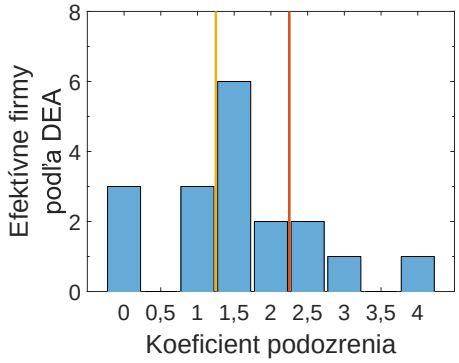
Z výsledkov v Tab. (3.1) vyplýva, že vzhľadom na priemerný KP došlo oproti expertnej voľbe vstupv a výstupov, aj metóde group lasso, v troch prípadoch k signifikantnému zlepše-



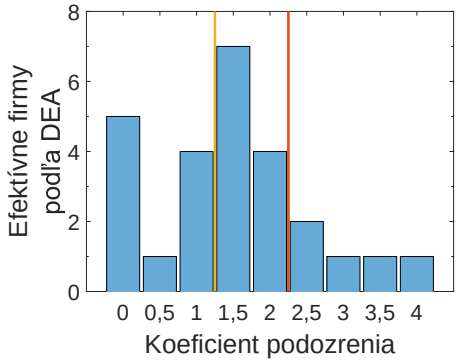
(a) Firmy klasifikované ako podozrivé pri $\lambda_1 = 1,5$ a $\lambda_2 = 1,5$. Spolu boli klasifikované 3 firmy ako podozrivé, pričom priemerná hodnota koeficientu podozrenia pre tieto firmy bola $1,83$. Spolu boli zvolené 2 vstupy a 0 výstupov.



(b) Firmy klasifikované ako podozrivé pri $\lambda_1 = 1$ a $\lambda_2 = 1$. Spolu bolo klasifikovaných 11 firiem ako podozrivé, pričom priemerná hodnota koeficientu podozrenia pre tieto firmy bola $1,0$. Spolu boli zvolené 3 vstupy a 1 výstup.



(c) Firmy klasifikované ako podozrivé pri $\lambda_1 = 10/7$ a $\lambda_2 = 7/10$. Spolu bolo klasifikovaných 18 firiem ako podozrivé, pričom priemerná hodnota koeficientu podozrenia pre tieto firmy bola $1,56$. Spolu boli zvolené 2 vstupy a 3 výstupy.



(d) Firmy klasifikované ako podozrivé pri $\lambda_1 = 1,4$ a $\lambda_2 = 0,7$. Spolu bolo klasifikovaných 26 firiem ako podozrivé, pričom priemerná hodnota koeficientu podozrenia pre tieto firmy bola $1,48$. Spolu boli zvolené 3 vstupy a 3 výstupy.

Obr. (3.4) Histogram rozdelenia koeficientu podozrenia pre firmy klasifikované ako podozrivé z rozkrádania verejných financií pri voľbe vstupov a výstupov metódou group lasso.

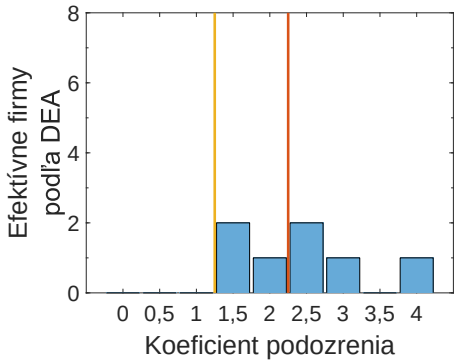
niu súladu medzi efektívnymi firmami v DEA a hodnotami KP. Dva z týchto prípadov sú zobrazené v grafoch na Obr. (3.5a) a (3.5b) a tretí je uvedený v Tab. (3.1). Pozoruhodný výsledok bol dosiahnutý pre realistický scenár v grafe na Obr. (3.5a), ktorý ukazuje, že v prípade apriórnej znalosti $N = 3$ nepodozrivých firiem s hodnotami KP 0; 0 a 1 a iba jedinej podozrivej firmy s KP 3 bolo podľa DEA efektívnych 7 firiem z čoho boli až 4 firmy veľmi podozrivé z hľadiska KP (s KP 2,5; 3 a 4). Zvyšné 3 firmy boli podľa KP mierne podozrivé (s KP 1,5 a 2). Ani jedna firma, ktorá bola efektívna podľa DEA nebola vzhľadom na KP nepodozrivá. To znamená, že pri takejto voľbe vstupov a výstupov sa DEA podarilo správne vybrať 4 z 8 veľmi podozrivých firiem a správne určiť všetkých 159 nepodozrivých firiem ako neefektívne. Podobné výsledky sme dosiahli aj v prípade (3.5b) pri apriórnej znalosti $N = 5$ nepodozrivých a $P = 5$ podozrivých firiem. Zaujímavé je, že na dosiahnutie podobných výsledkov ako v prípade (3.5b) stačilo v prípade (3.5a) poznať príklad iba jednej podozrivej firmy a tri príklady nepodozrivých firiem.

Na druhej strane graf na Obr. (3.5d) ukazuje, čo sa stane ak máme nesprávnu apriórnu znalosť príkladov podozrivých a nepodozrivých firiem. V tomto prípade boli na voľbu vstupov a výstupov použité $N = 3$ nepodozrivé firmy s KP 0; 0; a 2,5; a $P = 2$ podozrivé firmy s KP 0 a 2,5. Podľa priemerného KP firiem efektívnych podľa DEA pri takejto voľbe vstupov a výstupov sú výsledky horšie ako výsledky pri expertnej voľbe vstupov a výstupov a aj ako väčšina výsledkov metódy group lasso. Táto voľba vstupov a výstupov je však stále o čosi lepšia ako voľba vstupov a výstupov metódou group lasso pre $\lambda_1 = 1$ a $\lambda_2 = 1$. Vysvetlením výsledkov v grefe na Obr. (3.5d) je, že metóda rozdielom efektivít volí vstupy a výstupy, ktoré najviac vyhovujú príkladom, ktoré má k dispozícii. Nesprávne určenie efektívnosti príkladov, čiže porušenie predpokladu metódy rozdielom efektivít, potom vedie ku voľbe vstupov a výstupov, ktoré najlepšie vyhovujú takýmto zle určeným príkladom.

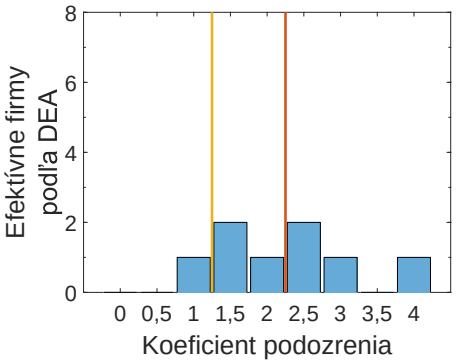
V druhom prípade realistického scenára s $P = 1$ príkladom podozrivej firmy s KP 2,5 a $N = 3$ príkladmi nepodozrivých firiem s KP 0; 0 a 0,5 použitých na voľbu vstupov a výstupov sú výsledky zobrazené v grafe na Obr. (3.5c) porovnateľné s expertnou voľbou vstupov a výstupov. Názvy vstupov a výstupov, ktoré boli do DEA zvolené použitím jednotlivých metód sú vypísané v prílohe B.

3.1.3 Výsledky metódy rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy

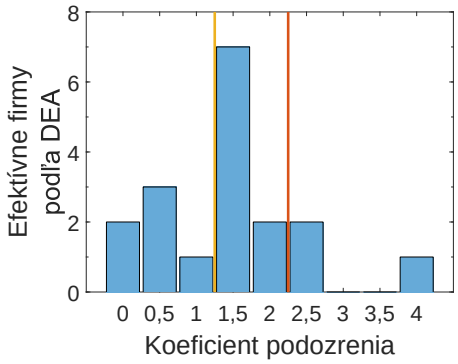
Poslednou uvažovanou metódou na voľbu vstupov a výstupov je metóda rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy. Podobne ako v prípade známeho rozdelenia atribútov, používa metóda rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy predpoklad 1 o znalosti niekoľkých príkladov podozrivých a nepodozrivých



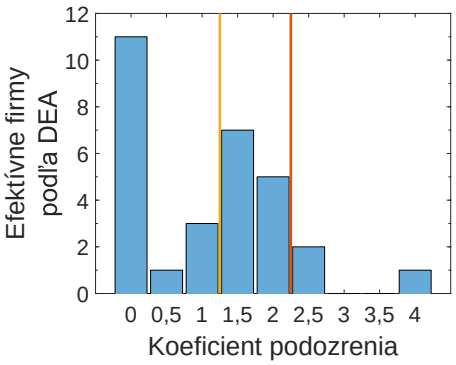
(a) Efektívne firmy podľa DEA pri voľbe vstupov a výstupov danej apriórnou znalosťou $N = 3$ nepodozrivých firiem s priemerným KP rovným $0,33$ a $P = 1$ podozrivej firmy s KP rovným 3. Spolu bolo efektívnych 7 firiem, pričom ich priemerná hodnota KP bola $2,43$.



(b) Efektívne firmy podľa DEA pri voľbe vstupov a výstupov danej apriórnou znalosťou $N = 5$ nepodozrivých firiem s priemerným KP rovným $0,00$ a $P = 5$ podozrivých s priemerným KP rovným $2,70$. Spolu bolo efektívnych 8 firmiem, pričom ich priemerná hodnota KP bola $2,25$.



(c) Efektívne firmy podľa DEA pri voľbe vstupov a výstupov danej apriórnou znalosťou $N = 3$ nepodozrivých firiem s priemerným KP rovným $0,17$ a $P = 1$ podozrivej s KP rovným $2,5$. Spolu bolo efektívnych 18 firiem, pričom ich priemerná hodnota KP bola $1,44$.



(d) Efektívne firmy podľa DEA pri voľbe vstupov a výstupov danej apriórnou znalosťou $N = 3$ nepodozrivých firiem s priemerným KP rovným $0,83$ a $P = 2$ podozrivých s priemerným KP rovným $1,25$. Spolu bolo efektívnych 30 firmiem, pričom ich priemerná hodnota KP bola $1,1$.

Obr. (3.5) Histogram rozdelenia koeficientu podozrenia pre firmy efektívne podľa DEA pri voľbe vstupov a výstupov metódou rozdielom efektivít pre známe rozdelenie atribútov na vstupy a výstupy.

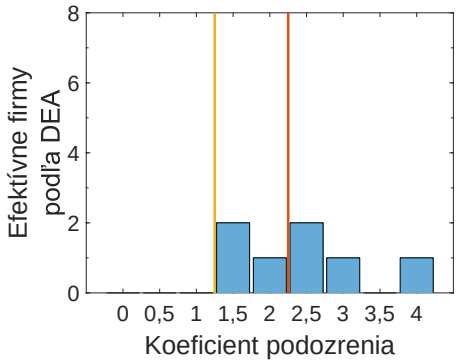
firiem z dátovej sady. Jej výhodou však je, že na rozdiel od predošlých metód nepotrebuje mať dopredu určené, ktorý atribút je vstup a ktorý výstup.

Výsledky metódy rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy sú prezentované v grafoch na Obr. (3.6), pričom príklady podozrivých a nepodozrivých firem použité na voľbu vstupov a výstupov sú identické ako v grafoch na Obr. (3.5). Úlohy celočíselného programovania na voľbu vstupov a výstupov (2.24) sú pre jednotlivé scenáre v grafoch na Obr. (3.5) riešené genetickými algoritmami. Podobne ako v prípade známeho rozdelenia atribútov sa ako voľba vstupov a výstupov považuje najčastejšie riešenie pri piatich zbehnutiach funkcie ga() na riešenie (2.24) v softvéri MATLAB.

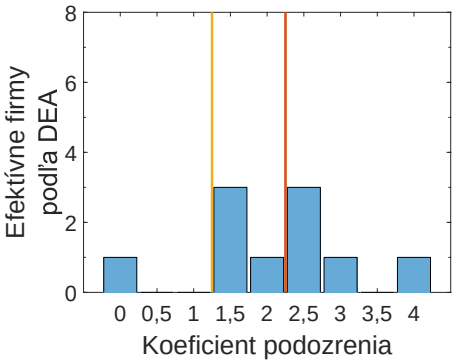
Z grafov, rovnako ako z Tab. (3.1), vyplýva, že z hľadiska priemerného KP efektívnych firiem je metóda rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy lepšia ako pri známom rozdelení vo všetkých scenároch okrem (3.6b). Dokonca aj v prípade nesprávnej apriórnej informácie v scenári v grafe na Obr. (3.6d) je priemerný KP firiem efektívnych podľa DEA vyšší ako pri voľbe vstupov a výstupov metódou rozdielom efektivít pre známe rozdelenie atribútov na vstupy a výstupy v grafe na Obr. (3.5d). V prípade (3.6b) bolo podľa DEA pri voľbe vstupov a výstupov metódou rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy efektívnych až 5 z 8 veľmi podozrivých firiem a iba 1 zo 159 nepodozrivých firiem. Takže aj napriek o niečo menšiemu preimernému KP efektívnych firiem je z praktického hľadiska výsledok lepší ako v prípade (3.5b).

Z Tab. (3.1) vyplýva, že metóda rozdielom efektivít pre neznáme rozdelenia vstupov a výstupov dosiahla podľa priemerného KP najlepší výsledok zo všetkých metód. V takomto prípade stačila na voľbu vstupov a výstupov apriórna znalosť $P = 2$ príkladov podozrivých firiem s KP $2,5$ a $3,5$ a $N = 3$ príkladov nepodozrivých firiem s KP 0; 0 a 1. Zaujímavý je aj výsledok v prípade realistického scenára s $P = 1$ príkladom podozrivej firmy s KP $2,5$ a s $N = 3$ príkladmi nepodozrivých firiem s KP 0; 0; $0,5$ v grafe na Obr. (3.5c). Metóda rozdielom efektivít pre známe rozdelenie atribútov na vstupy a výstupy dosiahla v tomto prípade priemernú hodnotu KP efektívnych firiem $1,44$ a v prípade neznámeho rozdelenia atribútov na vstupy a výstupy sa tento výsledok podarilo značne vylepšiť na priemernú hodnotu KP efektívnych firiem rovnú $1,89$. Na druhej strane však v prípade neznámeho rozdelenia atribútov na vstupy a výstupy nebola v tomto prípade firma s KP 4 efektívna v DEA.

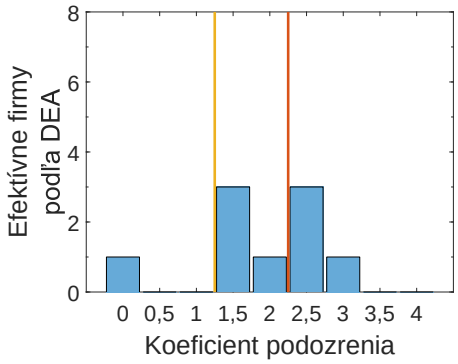
Možným odôvodnením prečo je metóda rozdielom efektivít pre neznáme rodelenie atribútov na vstupy a výstupy lepšia ako zvyšné dve metódy je fakt, že má väčšiu flexibilitu vo voľbe vstupov a výstupov a rozdelenie atribútov na vstupy a výstupy potrebné pri zvyšných dvoch metódach môže byť určené nesprávne.



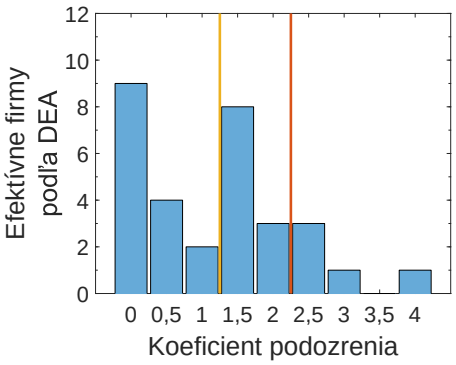
(a) Efektívne firmy podľa DEA pri voľbe vstupov a výstupov danej apriórnou znalosťou $N = 3$ nepodozrivých firiem s priemerným KP rovným $0,33$ a $P = 1$ podozrivej firmy s KP rovným 3. Spolu bolo efektívnych 7 firiem, pričom ich priemerná hodnota KP bola $2,43$.



(b) Efektívne firmy podľa DEA pri voľbe vstupov a výstupov danej apriórnou znalosťou $N = 5$ nepodozrivých firiem s priemerným KP rovným $0,00$ a $P = 5$ podozrivých firiem s priemerným KP rovným $2,70$. Spolu bolo efektívnych 10 firmiem, pričom ich priemerná hodnota KP bola $2,10$.



(c) Efektívne firmy podľa DEA pri voľbe vstupov a výstupov danej apriórnou znalosťou $N = 3$ nepodozrivých firiem s priemerným KP rovným $0,17$ a $P = 1$ podozrivej firmy s KP rovným $2,5$. Spolu bolo efektívncyh 9 firmiem, pričom ich priemerná hodnota KP bola $1,89$.



(d) Efektívne firmy podľa DEA pri voľbe vstupov a výstupov danej apriórnou znalosťou $N = 3$ nepodozrivých firiem s priemerným KP rovným $0,83$ a $P = 2$ podozrivých firiem s priemerným KP rovným $1,25$. Spolu bolo efektívncyh 31 firmiem, pričom ich priemerná hodnota KP bola $1,18$.

Obr. (3.6) Histogram rozdelenia koeficientu podozrenia pre firmy efektívne podľa DEA pri voľbe vstupov a výstupov metódou rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy.

Tab. (3.1) Porovnanie výsledkov metód na voľbu vstupov a výstupov. Pri metódach rozdielom efektivít je pri počte apriórne známych podozrivých firiem (P) a nepodozrivých firiem (N) uvedený v zátvorke priemer KP pre tieto skupiny firiem.

Metóda	Parametre	Počet efektívnych firiem podľa DEA	Priemer KP	Podiel efektívnych firiem s $KP > 2$	Podiel efektívnych firiem s $2 \geq KP > 1$	Podiel efektívnych firiem s $1 \geq KP$
Expertný výber		24	1,48	20,8%	37,5%	41,7%
Group lasso	$[\lambda_1;\lambda_2] = [1,5;1,5]$	3	1,83	33,3%	33,3%	33,3%
	$[\lambda_1;\lambda_2] = [1,0;1,0]$	11	1,00	9,1%	27,3%	63,6%
	$[\lambda_1;\lambda_2] = [\frac{10}{7};0,7]$	18	1,56	22,2%	44.5%	33.3%
	$[\lambda_1;\lambda_2] = [1,4;0,7]$	26	1,48	19,2%	42,3%	38,5%
Metóda rozdielom efektivít so známym rozdelením atribútov	P=1(3,00), N=3(0.33)	7	2.43	57,1%	42,9%	0,0%
	P=2(1.25), N=3(0.83)	30	1,10	10,0%	40,0%	50,0%
	P=5(2,70), N=5(0.00)	8	2,25	50,0%	37,5%	12,5%
	P=1(2,50), N=3(0.17)	18	1,44	16,7%	50,0%	33,3%
	P=2(3,00), N=3(0.33)	5	2,30	40,0%	60,0%	0,0%
Metóda rozdielom efektivít s neznámym rozdelením atribútov	P=1(3,00), N=3(0.33)	7	2.43	57,1%	42,9%	0,0%
	P=2(1,25), N=3(0.83)	31	1,18	16,1%	35,5%	48,4%
	P=5(2,70), N=5(0.00)	10	2,1	50,0%	40,0%	10,0%
	P=1(2,50), N=3(0.17)	9	1,89	44,4%	44,5%	11,1%
	P=2(3,00), N=3(0.33)	6	2,58	66,7%	33,3%	0,00%
Dáta		188	0,42	4,3%	11,7%	84,0%

Tab. (3.2) Výsledky použitia DEA ako klasifikátora viacerých tried. V tabuľke sú prezentované výsledné počty firiem v jednotlivých triedach, ich priemerný KP a podiel počtu veľmi podozrivých, mierne podozrivých a nepodozrivých firiem na celkovom počte firiem v danej triede.

Trieda	Počet firiem v triede	Priemer KP	Podiel firiem s $KP > 2$	Podiel firiem s $2 \geq KP > 1$	Podiel firiem s $1 \geq KP$
1.	7	2,4286	57,14%	42,86%	0%
2.	14	1,0000	7,14%	35,71%	57,14%
3.	18	0,7500	5,56%	16,67%	77,78%
Ostatné	149	0,2349	1,34%	7,38%	91,28%
Dáta	188	0,4229	4,26%	11,70%	84,04%

3.2 DEA ako klasifikátor viacerých tried

Posledná časť tejto práce je venovaná možnosti použiť DEA ako klasifikátor viacerých tried ako iba dvoch: podozrivá a nepodozrivá. Klasifikátor pre viacero tried sme zostavili nasledove: Pre danú voľbu vstupov a výstupov sme pomocou DEA určili efektívne firmy. Následne sme tieto firmy odstránili z dátovej sady a DEA analýzu pri tých istých vstupoch a výstupoch sme zopakovali na zvyšných firmách. Tento postup sme opakovali až dokým sme nevyčerpali všetky firmy z dátovej sady. Takto sme získali niekoľko tried pre jednotlivé firmy a výsledné vlastnosti týchto tried z hľadiska koeficientu podozrenia sú prezentované v Tab. (3.2). Každý riadok Tab. (3.2) obsahuje počet firiem v danej triede, ich priemerný KP a rozloženie veľmi podozrivých, mierne podozrivých a nepodozrivých firiem z pohľadu KP v danej triede . Vstupy a výstupy boli zvolené metódou rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy zodpovedajúce scenáru (3.6a).

Z výsledkov pozorujeme, že ako postupne odoberáme firmy zo vzorky, tak sa znižuje priemerný koeficient podozrenia a rovnako sa znižuje aj podiel počtu firiem s $KP > 2$ na celkovom počte firiem v danej triede. Naopak, podiel firiem s $KP \leq 1$ sa zvyšuje. Z týchto výsledkov vyplýva, že DEA je pri vhodnej voľbe vstupv a výstupv schopná dobre charakterizovať priestor firiem z hľadiska podozrenia z rozkrádania verejných financií. Výsledky skutočne naznačujú, že čím je index triedy nižší, tým je vyšší podiel podozrivých firiem (s vysokým KP) a nižší podiel nepodozrivých firiem (s nízkym KP) v danej triede. Inými slovami, zdá sa, že koeficient podozrenia jednotlivých firiem je nepriamo úmerný ich vzdialenosti od hranice efektívnosti danej DEA.

Záver

Jedným z najväčších úskalí DEA analýzy je voľba vhodných vstupov a výstupov. Cieľom práce bolo preto vytvorenie ucelenej automatizovanej metódy na voľbu vstupov a výstupov do DEA, ako aj jej aplikácia pri hľadaní firiem podozrivých z rozkrádania verejných financií.

Na tento účel sme uvažovali dve metódy. Prvá, ktorá bola publikovaná v článku [5], je voľba vstupov a výstupov metódou group lasso. Ide o úlohu kvadratického programovania, ktorá do DEA analýzy volí také vstupy a výstupy z množiny všetkých možných vstupov a výstupov, ktoré zodpovedajú nenulovým vektorom multiplikatívnych váh v optimálnom riešení. Výhodou takéhoto postupu je jeho efektívna riešiteľnosť a fakt, že na rozdiel od druhej metódy nepotrebuje žiadne predpoklady o efektívnosti DMU. Na druhej strane nevýhody spočívajú v citlivosti na správnu voľby hyperparametrov λ_1 a λ_2 a v potrebe apriórneho rozdelenia atribútov na vstupy a výstupy.

Druhá metóda na voľbu vstupov a výstupov, ktorú sme navrhli v tejto práci, je metóda rozdielom efektivít. Táto metóda má dva varianty, prvý pre známe rozdelenie atribútov na vstupy a výstupy a druhý pre neznáme rozdelenie atribútov na vstupy a výstupy. Ide o postup, ktorý na základe apriórnej znalosti niekoľkých efektívnych a niekoľkých neefektívnych DMU v dátovej sade volí vstupy a výstupy do DEA. Výhoda takéhoto postupu je, že za účelovú funkciu v úlohe na voľbu vstupov a výstupov môžeme zvoliť ľubovoľné vhodne definované kritérium, na základe ktorého potom volíme vstupy a výstupy do DEA. Okrem toho sa do úlohy dajú jednoducho pridávať ohraničenia na maximálny a minimálny počet volených vstupov a výstupov. Nevýhodou tejto metódy je, že vedie na úlohu celočíselného programovania, ktorej riešenie je komplikovanejšie ako v prípade group lasso. Ukazuje sa však, že by bolo možné túto komplikovanú úlohu relaxovať na bipartitné bilineárne programovanie, pre ktoré už existujú efektívnejšie algoritmy ako pre celočíselné programovanie. Na rozdiel od metódy group lasso nepotrebuje variant metódy rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy znalosť, ktorý atribút je vstup a ktorý výstup.

Výhody metódy rozdielom efektivít sa ukazujú pri výsledkoch pre aplikáciu DEA na náročný problém odhaľovania rozkrádania verejných financií. Už pri apriórnej znalosti efektívnosti 5 firiem dokázala metóda rozdielom efektivít pre neznáme rozdelenie atribútov na vstupy a výstupy vybrať z pohľadu KP signifikantne lepšie vstupy a výstupy ako metóda

group lasso. Takto zvolené vstupy a výstupy viedli na priemerný KP efektívnych firiem $2,58$ a až 4 zo 6 efektívnych firiem boli po subjektívnej kontrole veľmi podozrivé z rozkrádania verejných financií.

Odhliadnuc od automatizovanej voľby vstupov a výstupov do DEA môžeme z výsledkov pozorovať, že pre správnu voľbu vstupov a výstupov sú efektívnejšie firmy podľa DEA zároveň podozrivejšie z hľadiska subjektívnej kontroly. Tento súlad výsledkov subjektívnej kontroly a DEA podporuje tvrdenie, že DEA je vhodný nástroj na identifikovanie firiem podozrivých z rozkrádania verejných financií.

Podobné myšlienky, ako v prípade hľadania firiem podozrivých z rozkrádania verejných financií, je možné použiť pri ľubovoľnom probléme na hľadanie podozrivých subjektov, ktoré prijímajú nejaké vstupy a produkujú výstupy. V štátnej správe by napríklad bolo možné identifikovať možných daňových podvodníkov a na základe minulých daňových kontrol by mohla byť použitá aj automatizovaná voľba vstupov a výstupov metódou rozdielom efektivít. Myšlienky a odvodenia metódy rozdielom efektivít tvoria ucelený základ pre vytváranie ďalších metód na voľbu vstupov a výstupov do DEA.

Zoznam použitej literatúry

[1] HLADIŠ, M.: Ekonomika rozkrádania, bakalárska práca, FMFI UK, Bratislava, 2016, dostupné na internete (9.1.2018):
http://opac.crzp.sk/

[2] HALICKÁ, M.: *DEA modely*, učebný text, FMFI UK, Bratislava, 2015, dostupné na internete (5.4.2018):
http://www.iam.fmph.uniba.sk/institute/halicka/teach/DEAskripta.pdf

[3] TIBSHIRANI, R.: *Regression Shrinkage and Selection via the Lasso*, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 58, No. 1 (1996), pp. 267-288, Blackwell Publishers, 1996, dostupné na internete (7.5.2018):
http://www.jstor.org/stable/2346178

[4] YUAN, M., LIN, Y.: *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 68, No. 1 (2006), pp. 49-67, Blackwell Publishing Ltd, 2006, dostupné na internete (7.5.2018):
http://www.jstor.org/stable/3647556

[5] QIN, Z. T., SONG, I.: *Joint Variable Selection for Data Envelopment Analysis via Group Sparsity*, Management Science Manuscript, 2014.

[6] HASTIE, T., TIBSHIRANI, R., WAINWRIGHT, M.: *Statistical Learning with Sparsity: The Lasso and Generalizations*, Chapman & Hall/CRC, 2015.

[7] DEY, S. S., SANTANA, A., WANG, Y.: *New SOCP relaxation and branching rule for bipartite bilinear programs*, ArXiv e-prints, 2018, dostupné na internete (7.5.2018):
https://arxiv.org/pdf/1803.09266.pdf

[8] Úrad vlády SR: *Centrálny register zmlúv*, dostupné na internete (9.5.2018):
https://www.crz.gov.sk/

[9] Úrad pre verejné obstarávanie: *Vestník verejného obstarávania*, dostupné na internete (9.5.2018):

https://www.uvo.gov.sk/verejny-obstaravatel-obstaravatel/vestnik-verejneho-obstaravania-37c.html

[10] foaf: *sociálna siet'firiem*, dostupné na internete (9.5.2018):
http://foaf.sk/

[11] Transparency International Slovensko: *Open Public Procurement*, dostupné na internete (9.5.2018):
https://tender.sme.sk/data/

[12] Finstat: *databáza slovenských firiem*, dostupné na internete (9.5.2018):
http://finstat.sk/

[13] Ministerstvo financií Slovenskej republiky: *Register účtovných závierok*, dostupné na internete (9.5.2018):
http://www.registeruz.sk/

Príloha A: Odvodenia a úpravy výrazov

Uprava vyrazu (2.9)

Uvažujme úlohu na voľbu vstupov a výstupov metódou group lasso danú výrazom (2.9):

$$
\begin{aligned}
\min_{\tilde{u}^1,\ldots,\tilde{u}^S,\tilde{v}^1,\ldots,\tilde{v}^M,w} \quad & \sum_{i=1}^{M}(x^i)^T\tilde{v}^i - \sum_{r=1}^{S}(y^r)^T\tilde{u}^r + \mathbf{1}^T w + \lambda_1\sum_{i=1}^{M}\left\|\tilde{v}^i\right\|_2 + \lambda_2\sum_{r=1}^{S}\|\tilde{u}^r\|_2 \\
\text{s.t} \quad & \sum_{i=1}^{M}x_{i,1}\tilde{v}^i + \sum_{i=1}^{M}x_{i,1}\rho_i^x\mathbf{1} - \sum_{r=1}^{S}y_{r,1}\tilde{u}^r - \sum_{r=1}^{S}y_{r,1}\rho_r^y\mathbf{1} + w \geq 0 \\
& \vdots \\
& \sum_{i=1}^{M}x_{i,n}\tilde{v}^i + \sum_{i=1}^{M}x_{i,n}\rho_i^x\mathbf{1} - \sum_{r=1}^{S}y_{r,n}\tilde{u}^r - \sum_{r=1}^{S}y_{r,n}\rho_r^y\mathbf{1} + w \geq 0 \\
& \tilde{u}^r \geq 0 \qquad r = 1,\ldots,S \\
& \tilde{v}^i \geq 0 \qquad i = 1,\ldots,M,
\end{aligned}
\tag{3.1}
$$

kde $\tilde{v}^i$, $\tilde{u}^r$, w, x^i, y^r a $\mathbf{1}$ sú definované ako

$$
\begin{aligned}
\tilde{v}^i &= \begin{bmatrix} \tilde{v}_{i,1}, & \tilde{v}_{i,2}, & \ldots, & \tilde{v}_{i,n} \end{bmatrix}^T \in R^n, \\
\tilde{u}^r &= \begin{bmatrix} \tilde{u}_{r,1}, & \tilde{u}_{r,2}, & \ldots, & \tilde{u}_{r,n} \end{bmatrix}^T \in R^n, \\
w &= \begin{bmatrix} w_1, & w_2, & \ldots, & w_n \end{bmatrix}^T \in R^n, \\
x^i &= \begin{bmatrix} x_{i,1}, & x_{i,2}, & \ldots, & x_{i,n} \end{bmatrix}^T \in R^n, \\
y^r &= \begin{bmatrix} y_{r,1}, & y_{r,2}, & \ldots, & y_{r,n} \end{bmatrix}^T \in R^n, \\
\mathbf{1} &= \begin{bmatrix} 1, & 1, & \ldots, & 1 \end{bmatrix}^T \in R^n.
\end{aligned}
$$

Označme potom $\bar{A}$, $\bar{B}$, $\bar{I}$, $\tilde{v}$, ρ^v, $\tilde{u}$, ρ^u ako

$$\bar{A} = \begin{bmatrix} x^1 & & & x^2 & & & \dots & x^M & & \\ & x^1 & & & x^2 & & \dots & & x^M & \\ & & \vdots & & & \vdots & \dots & & & \vdots \\ & & & x^1 & & & x^2 & \dots & & & x^M \end{bmatrix} \in R^{(n^2)\times(Mn)}$$

$$\bar{B} = \begin{bmatrix} y^1 & & & y^2 & & & \dots & y^S & & \\ & y^1 & & & y^2 & & \dots & & y^S & \\ & & \vdots & & & \vdots & \dots & & & \vdots \\ & & & y^1 & & & y^2 & \dots & & & y^S \end{bmatrix} \in R^{(n^2)\times(Sn)}$$

$$\bar{I} = \begin{bmatrix} \mathbf{1} & & & \\ & \mathbf{1} & & \\ & & \vdots & \\ & & & \mathbf{1} \end{bmatrix} \in R^{(n^2)\times(n)}$$

$$\tilde{v} = \begin{bmatrix} \left(\tilde{v}^1\right)^T, & \left(\tilde{v}^2\right)^T, & \dots, & \left(\tilde{v}^M\right)^T \end{bmatrix}^T \in R^{Mn},$$

$$\tilde{u} = \begin{bmatrix} \left(\tilde{u}^1\right)^T, & \left(\tilde{u}^2\right)^T, & \dots, & \left(\tilde{u}^S\right)^T \end{bmatrix}^T \in R^{Sn},$$

$$\rho^v = \begin{bmatrix} \left(\rho_1^x \mathbf{1}\right)^T, & \left(\rho_2^x \mathbf{1}\right)^T, & \dots, & \left(\rho_M^x \mathbf{1}\right)^T \end{bmatrix}^T \in R^{Mn},$$

$$\rho^u = \begin{bmatrix} \left(\rho_1^y \mathbf{1}\right)^T, & \left(\rho_2^y \mathbf{1}\right)^T, & \dots, & \left(\rho_S^y \mathbf{1}\right)^T \end{bmatrix}^T \in R^{Sn}.$$

Použitím týchto označení môžeme potom úlohu (3.1) prepísať na úlohu

$$\begin{aligned} \min_{\tilde{u},\tilde{v},w} \quad & \sum_{i=1}^{M} (x^i)^T \tilde{v}^i - \sum_{r=1}^{S} (y^r)^T \tilde{u}^r + \mathbf{1}^T w + \lambda_1 \sum_{i=1}^{M} ||\tilde{v}^i||_2 + \lambda_2 \sum_{r=1}^{S} ||\tilde{u}^r||_2 \\ \text{s.t} \quad & \bar{A}\tilde{v} + \bar{A}\rho^v - \bar{B}\tilde{u} - \bar{B}\rho^u + \bar{I}w \geq 0 \\ & \tilde{u} \geq 0 \\ & \tilde{v} \geq 0. \end{aligned} \tag{3.2}$$

Ak potom definujeme x a y ako

$$x = \begin{bmatrix} \left(x^1\right)^T, & \left(x^2\right)^T, & \dots, & \left(x^M\right)^T \end{bmatrix}^T \in R^{Mn},$$

$$y = \begin{bmatrix} \left(y^1\right)^T, & \left(y^2\right)^T, & \dots, & \left(y^S\right)^T \end{bmatrix}^T \in R^{Sn},$$

tak z (3.2) dostávame

$$\begin{aligned} \min_{\tilde{u},\tilde{v},w} \quad & x^T\tilde{v} - y^T\tilde{u} + \mathbf{1}^T w + \lambda_1 \sum_{i=1}^{M} ||\tilde{v}^i||_2 + \lambda_2 \sum_{r=1}^{S} ||\tilde{u}^r||_2 \\ \text{s.t} \quad & \bar{A}\tilde{v} + \bar{A}\rho^v - \bar{B}\tilde{u} - \bar{B}\rho^u + \bar{I}w \geq 0 \\ & \tilde{u} \geq 0 \\ & \tilde{v} \geq 0. \end{aligned} \tag{3.3}$$

Nakoniec označnme c, z, A, b ako

$$c = \left[\begin{array}{ccc} x^T, & -y^T, & \mathbf{1}^T \end{array} \right]^T \tag{3.4}$$

$$z = \left[\begin{array}{ccc} \tilde{v}^T, & \tilde{u}^T, & w^T \end{array} \right]^T \in R^{Mn+Sn+n} \tag{3.5}$$

$$A = \left[\begin{array}{ccc} \bar{A} & -\bar{B} & \bar{I} \\ & I' & \end{array} \right] \in R^{(n^2+Mn+Sn)\times(Mn+Sn+n)} \tag{3.6}$$

$$b = \left[\begin{array}{c} -\bar{A}\rho^v + \bar{B}\rho^u \\ 0 \end{array} \right] \in R^{n^2+Mn+Sn}, \tag{3.7}$$

kde

$$I' = \left[\begin{array}{cc} I, & 0 \end{array} \right] \in R^{(Mn+Sn)\times(Mn+Sn+n)} \tag{3.8}$$

a $I \in R^{(Mn+Sn)\times(Mn+Sn)}$ je identita. Prepísaním (3.3) pomocou (3.4)–(3.7) potom dostávame úlohu konvexného programovania

$$\begin{aligned} \min_z \quad & c^T z + \lambda_1 \sum_{i=1}^{M} ||\tilde{v}^i||_2 + \lambda_2 \sum_{r=1}^{S} ||\tilde{u}^r||_2 \\ & Az \geq b. \end{aligned} \tag{3.9}$$

Príloha B: Výsledné voľby vstupov a výstupov

Tab. (3.3) Potenciále vstupy.

Vstup	Označenie
Priemerný účtovný cashflow	a)
Priemerné množstvo peňazí na bankových účtoch uvedené v účtovnej závierke	b)
Priemerné vlastné imanie	c)
Priemerná výška tržieb	d)
Priemerné Z-skóre	e)
Priemerný počet zamestnancov	f)
Priemerné náklady na materiál	g)

Tab. (3.4) Potenciále výstupy.

Výstup	Označenie
Maximálny medziročný rast tržieb	h)
Priemerné množstvo peňazí v hotovosti uvedené v účtovnej závierke	i)
Priemerná pridaná hodnota	j)
Počet obstarávaní, v ktorých sa daná firma zúčastnila ako jediná a vyhrala	k)
Počet obstarávaní obstarávaných formou neverejnej súťaže, ktoré firma vyhrala	l)
Priemerné ročné príjmy z verejných obstarávaní	m)
Priemerný obežný majetok	n)
Hodnota najväčšej štátnej zákazky akú firma vyhrala	o)
Priemerné náklady na služby	p)
Priemer rozdielu nákladov na služby a nákladov na materiál	r)

Tab. (3.5) Výsledná voľba vstupov a výstupov jednotlivých metód. Označenia vstupov sú vypísané v Tab. (3.3) a onaženia výstupov v Tab. (3.4).

Metóda	Parametre	Počet efektívnych DMU	Počet vstupov	Vstupy	Počet výstupov	Výstupy
Expertný výber		24	5	a), c), d), e), f)	3	h), m), o)
Group lasso	$[\lambda_1;\lambda_2]=[1,5;1,5]$	3	2	b), f)	0	
	$[\lambda_1;\lambda_2]=[1,0;1,0]$	11	3	b), f), g)	1	i)
	$[\lambda_1;\lambda_2]=[\frac{10}{7};0,7]$	18	2	b), f)	3	h), i), m)
	$[\lambda_1;\lambda_2]=[1,4;0,7]$	26	3	b), f), g)	3	h), i), m)
Metóda rozdielom efektivít so známym rozdelením atribútov	P=1(3,00), N=3(0.33)	7	2	c), e)	7	h), j), k), l), m), n), p)
	P=2(1.25), N=3(0.83)	30	3	d), e), f)	5	i), k), l), p), r)
	P=5(2,70), N=5(0.00)	8	2	c), e)	6	j), k), l), m), o), r)
	P=1(2,50), N=3(0.17)	18	1	g)	7	j), k), l), m), o), p), r)
	P=2(3,00), N=3(0.33)	5	1	e)	7	h), j), k), l), m), n), p)
Metóda rozdielom efektivít s neznámym rozdelením atribútov	P=1(3,00), N=3(0.33)	7	1	e)	14	a), b), c), d), f), g), h), j), k), l), m), n), p), r)
	P=2(1,25), N=3(0.83)	31	4	c), e), h), i)	13	a), b), d), f), g), j), k), l), m), n), o), p), r)
	P=5(2,70), N=5(0.00)	10	2	i), e)	9	b), f), g), h), j), k), l), m), n)
	P=1(2,50), N=3(0.17)	9	1	i)	10	b), d), f), g), h), j), k), l), m), o)
	P=2(3,00), N=3(0.33)	6	1	e)	11	b), d), f), g), h), j), k), l), m), o), r)

Príloha C: Kód programu v MATLAB

Hlavný súbor:

```
%% Metoda group lasso
mu=1;
[Value,Rel_Vst_ind_GL,Rel_Vyst_ind_GL,z_opt]= ...
    My_GLmethod(Vst,Vyst,lambda_1,lambda_2,mu);
Vst_ind_GL=Vstupy(Rel_Vst_ind_GL);
Vyst_ind_GL=Vystupy(Rel_Vyst_ind_GL);
[Pod,effectivity]=WAddModel(Data(:,Vst_ind_GL),Data(:,Vyst_ind_GL));
Pod
% ------------------------------------------------------
%% Metoda rozdielom efektivit pre zname rozdelenie atributov
[Rel_Vst_ind_Dx_Dy,Rel_Vyst_ind_Dx_Dy]= GA_Dx_Dy(Vst,Vyst,op,on);
Vst_ind_Dx_Dy=Vstupy(Rel_Vst_ind_Dx_Dy);
Vyst_ind_Dx_Dy=Vystupy(Rel_Vyst_ind_Dx_Dy);
[Pod,eff,lambda,sx,sy]= ...
    WAddModel(Data(:,Vst_ind_Dx_Dy),Data(:,Vyst_ind_Dx_Dy));
Pod
% ------------------------------------------------------
%% Metoda rozdielom efektivit pre nezname rozdelenie atributov
[Vst_ind_D,Vyst_ind_D]= GA_D(Data,op,on);
Vst_D=Data(:,Vst_ind_D);
Vyst_D=Data(:,Vyst_ind_D);
[Pod,eff,lambda,sx,sy]= WAddModel(Data(:,Vst_ind_D),Data(:,Vyst_ind_D));
Pod
% ------------------------------------------------------
%% Expertny vyber vstupov a vystupov
[Pod,eff,lambda,sx,sy]= ...
    WAddModel(Data(:,[2,5,9,12,13]),Data(:,[1,10,14]));
```

Aditívny DEA model:

```
function[Podozrive,effectivity,lambda,sx,sy]=WAddModel(Vst,Vyst)
% Vst je matica vstupov, DMU su v riadkoch
```

```
% Vyst je matica vzstupov, DMU su v riadkoch
% Podozrive je index DMU, ktore su efektivne
% effectivity su vysledne hodnoty efektivity
% lambda je efektivny vzor
% sx su slacky vstupov
% sy su slacky vystupov
[m,nX]=size(Vst');
[s,nY]=size(Vyst');
if nX≠nY                % Kontrola rozmeru
    error('Rozmer matice X a Y nesedi')
else
    n=nX;
end
Vst=Vst + ones(n,1)*(abs(min(Vst))+ones(1,m));      % Posun vstupov
Vyst=Vyst + ones(n,1)*(abs(min(Vyst))+ones(1,s));   % Posun vystupov
X=Vst';
Y=Vyst';
Rx=zeros(1,m);
for i=1:m
    Rx(i)=max(X(i,:))-min(X(i,:));
end
wx=1./Rx;
Ry=zeros(1,s);
for r=1:s
    Ry(r)=max(Y(r,:))-min(Y(r,:));
end
wy=1./Ry;
X=diag(wx)*X;
Y=diag(wy)*Y;
wx=ones(1,m);
wy=ones(1,s);
Aeq=[X, eye(m,m), zeros(m,s);
    Y, zeros(s,m), -eye(s,s);
    ones(1,n), zeros(1,m), zeros(1,s)];
a=1; effectivity=ones(1,n); lambda=-ones(n,n); sx=-ones(m,n); ...
   sy=-ones(s,n);
for o=1:n
xo=X(:,o);
yo=Y(:,o);
beq=[xo;yo;1];
f= -[zeros(1,n),wx,wy]';
[x(:,o),fval]=linprog(f,[],[],Aeq,beq,zeros(n+m+s,1),[],[]);
effectivity(o)=fval;
lambda(:,o)=x(1:n,o);
sx(:,o)=x((n+1):(n+m),o);
```

```
sy(:,o)=x((n+m+1):(n+m+s),o);
    if effectivity(o)>-10^(-8) && effectivity(o)<10^(-8)
        Podozrive(a)=o;
        a=a+1;
    end
end
```

Metóda group lasso:

```
function ...
    [Value,Vst_ind,Vyst_ind,z_opt]=My_GLmethod(X,Y,lambda_1,lambda_2,mu)
% v_tilde=z(1 : m*n)
% u_tilde=z(m*n+1 : m*n+s*n)
% w=z(m*n+s*n : m*n+s*n+n)
% X \in R^n*m
% Y \in R^n*s
[nX,m]=size(X);
[nY,s]=size(Y);
if nX==nY
    n=nX;
else
    error('nX a nY sa nerovnaju')
end
X=X + ones(n,1)*(abs(min(X))+ones(1,m));      % Posun vstupov
Y=Y + ones(n,1)*(abs(min(Y))+ones(1,s));   % Posun vystupov
Rx=zeros(1,m);
for i=1:m
    Rx(i)=max(X(:,i))-min(X(:,i));
end
wx=1./Rx;
Ry=zeros(1,s);
for r=1:s
    Ry(r)=max(Y(:,r))-min(Y(:,r));
end
wy=1./Ry;
% Normalizacia
X=(X*diag(wx));
Y=(Y*diag(wy));
% Ucelova funkcia
x=X(:);
y=Y(:);
c=sparse([x;-y;ones(n,1)]);
% Ohranicenia
    function [OutMatrix]=BarMatrix(InMatrix)
%     pocet DMU je pocet riadkov
```

```
%     pocet premennych je pocet stlpcov
        NumberOfVariables=size(InMatrix,2);
        NumberOfDMUs=size(InMatrix,1);
        OutMatrix=zeros(NumberOfDMUs^2,NumberOfDMUs*NumberOfVariables);
        for index=1:NumberOfVariables
            OutMatrix(:, ...
                NumberOfDMUs*(index-1)+1:NumberOfDMUs*index)=kron(eye(NumberOfDMUs),InMatri
        end
    end
A_bar=BarMatrix(X);
B_bar=BarMatrix(Y);
I_bar=BarMatrix(ones(n,1));
% Vahy
clear Rx Ry wx wy
Rx=zeros(1,m);
for i=1:m
    Rx(i)=max(X(:,i))-min(X(:,i));
end
wx=1./Rx;
I_rho_v=kron(eye(m),ones(n,1));
rho_v=I_rho_v*wx';
Ry=zeros(1,s);
for r=1:s
    Ry(r)=max(Y(:,r))-min(Y(:,r));
end
wy=1./Ry;
I_rho_u=kron(eye(s),ones(n,1));
rho_u=I_rho_u*wy';
% -----------------------
% Non zero constraint
NonzeroCon=zeros(m*n+s*n,m*n+s*n+n);
NonzeroCon(:,1:m*n+s*n)=eye(m*n+s*n);
% -----------------------
A=sparse([A_bar,-B_bar,I_bar;NonzeroCon]);
b=sparse([-A_bar*rho_v + B_bar*rho_u;zeros(m*n+s*n,1)]);
% -----------------------
C_v=zeros(m*n,m*n+s*n+n);
C_v(:,1:m*n)=eye(m*n);
C_v=sparse(C_v);
% -----------------------
C_u=zeros(s*n,m*n+s*n+n);
C_u(:,m*n+1:m*n+s*n)=eye(s*n);
C_u=sparse(C_u);
% -----------------------
% Inicializacia
```

```
z=sparse(0.001*ones(m*n+s*n+n,1));
slack=sparse(0.001*ones(n^2 + m*n + s*n,1));
v_bar=0.001*ones(m*n,1);
u_bar=0.001*ones(s*n,1);
gamma_s=sparse(0.001*ones(n^2 + m*n + s*n,1));
gamma_v=sparse(0.001*ones(m*n,1));
gamma_u=sparse(0.001*ones(s*n,1));
% ----------------------
% Algoritmus
K=500;
for k=1:K
    z=(A'*A + C_v'*C_v + C_u'*C_u)\(mu*(A'*gamma_s + C_v'*gamma_v + ...
       C_u'*gamma_u - c) + A'*(b+slack) + C_v'*v_bar + C_u'*u_bar);
        Help_s=A*z - b - mu*gamma_s;
    slack=max([Help_s,zeros(size(Help_s,1),1)],[],2);
        Help_v=C_v*z - mu*gamma_v;
    for i=1:m
        v_bar((i-1)*n+1 : i*n)=Help_v((i-1)*n+1 : ...
           i*n)*max(0,1-mu*lambda_1/norm(Help_v((i-1)*n+1 : i*n)));
    end
        Help_u=C_u*z - mu*gamma_u;
    for r=1:s
        u_bar((r-1)*n+1 : r*n)=Help_u((r-1)*n+1 : ...
           r*n)*max(0,1-mu*lambda_2/norm(Help_u((r-1)*n+1 : r*n)));
    end
    gamma_s=gamma_s - (1/mu)*(A*z - (b+slack));
    gamma_v=gamma_v - (1/mu)*(C_v*z - v_bar);
    gamma_u=gamma_u - (1/mu)*(C_u*z - u_bar);
    if mod(k,50)==0
        fprintf('iteration:     %d \n',k);
    end
end
V_bar=reshape(full(v_bar),[n,m]);
U_bar=reshape(full(u_bar),[n,s]);
z_opt=z;
Value=c'*z_opt + lambda_1*sum(sqrt(sum(V_bar.^2, 1))) + ...
    lambda_2*sum(sqrt(sum(U_bar.^2, 1)));
%---------------------------------
% Urcenie ktore Vstupy/Vystupy vybrat
V_sum=sum(abs(V_bar),1)';
if isempty(find(V_sum>10^-6,1)) ...
    NonCorrVst=(find(NonCorrVst==min(NonCorrVst)));
      Vst_ind=[];
else
    Vst_ind=find(V_sum>10^-6);
```

```
end

U_sum=sum(abs(U_bar),1)';
if isempty(find(U_sum>10^-6,1))
 NonCorrVyst=(find(NonCorrVyst==min(NonCorrVyst)));
      Vyst_ind=[];
else
    Vyst_ind=find(U_sum>10^-6);
end
end
```

Metóda rozdielom efektivít pre známe rozdelenie atribútov:

```
function[pEff,nEff]=Diag_f_Dx_Dy_WAddModel(Vst_All,Vyst_All,dx,dy,op,on)
[m,nX]=size(Vst_All');
[s,nY]=size(Vyst_All');
if nX≠nY                % Kontrola rozmeru
    error('Rozmer matice X a Y nesedi')
else
    n=nX;
end
Vst_All=Vst_All + ones(n,1)*(abs(min(Vst_All))+ones(1,m));
Vyst_All=Vyst_All + ones(n,1)*(abs(min(Vyst_All))+ones(1,s));
X=Vst_All';
Y=Vyst_All';
Rx=zeros(1,m);
for i=1:m
    Rx(i)=max(X(i,:))-min(X(i,:));
end
wx=1./(Rx);
Ry=zeros(1,s);
for r=1:s
    Ry(r)=max(Y(r,:))-min(Y(r,:));
end
wy=1./(Ry);
X=diag(wx)*X;
Y=diag(wy)*Y;
Dx=diag(dx);
Dx( ¬any(Dx,2), : ) = [];
Dy=diag(dy);
Dy( ¬any(Dy,2), : ) = [];
pEff=ones(length(op),1);
for p=1:length(op)
oi=op(p);
pEff(p)=f_Dx_Dy_WAddModel(X,Y,Dx,Dy,oi, wx,wy);
```

```
end
nEff=ones(length(on),1);
for n=1:length(on)
oi=on(n);
nEff(n)=f_Dx_Dy_WAddModel(X,Y,Dx,Dy,oi, wx,wy);
end
%---------------------------------------
function[fval]=GA_Dx_Dy_fun(Vst_All,Vyst_All,dx,dy,op,on)
[pEff,nEff]=Diag_f_Dx_Dy_WAddModel(Vst_All,Vyst_All,dx,dy,op,on);
P=length(op);
N=length(on);
fval=-((1/P)*sum(pEff) - (1/N)*sum(nEff));
end
%---------------------------------------
function[Vst_ind_Dx_Dy,Vyst_ind_Dx_Dy,x,fval,exitflag,output,population,scores]=GA_
sizDx=size(Vst,2);
sizDy=size(Vyst,2);
MS=sizDx+sizDy;
nvars=MS;
LB=zeros(MS,1);
UB=ones(MS,1);
IntCon=[1:MS]';
fitnessfcn=@(dxdy) ...
    GA_Dx_Dy_fun(Vst,Vyst,dxdy(1:sizDx),dxdy(sizDx+1:sizDx+sizDy),op,on);
[x,fval,exitflag,output,population,scores] = ...
    ga(fitnessfcn,nvars,[],[],[],[],LB,UB,[],IntCon) ;
Vst_ind_Dx_Dy=find(x(1:sizDx)==1);
Vyst_ind_Dx_Dy=find(x(sizDx+1:sizDx+sizDy)==1);
```

Metóda rozdielom efektivít pre neznáme rozdelenie atribútov:

```
function[pEff,nEff]=Diag_f_D_WAddModel(A,d,op,on)
[n,MS]=size(A);
A=A + ones(n,1)*(abs(min(A))+ones(1,MS));
A(:,find(d==-1))= - A(:,find(d==-1));
A=A';
R=zeros(1,MS);
for k=1:MS
    R(k)=max(A(k,:))-min(A(k,:));
end
w=1./(R);
A=diag(w)*A;
D=diag(d);
D( ¬any(D,2), : ) = [];
pEff=ones(length(op),1);
```

```
for p=1:length(op)
oi=op(p);
pEff(p)=f_D_WAddModel(A,D,oi, w,MS);
end
nEff=ones(length(on),1);
for n=1:length(on)
oi=on(n);
nEff(n)=f_D_WAddModel(A,D,oi, w,MS);
end
%----------------------------------------
function[fval]=GA_D_fun(A,d,op,on)
[pEff,nEff]=Diag_f_D_WAddModel(A,d,op,on);
P=length(op);
N=length(on);
fval=-((1/P)*sum(pEff) - (1/N)*sum(nEff));
end
%----------------------------------------
function[Vst_ind_D,Vyst_ind_D,x,fval,exitflag,output,population,scores]=GA_D(Data,op,on)
MS=size(Data,2);
nvars=MS;
LB=-ones(MS,1);
UB=ones(MS,1);
IntCon=[1:MS]';
fitnessfcn=@(d) GA_D_fun(Data,d',op,on);
[x,fval,exitflag,output,population,scores] = ...
    ga(fitnessfcn,nvars,[],[],[],[],LB,UB,[],IntCon) ;
Vst_ind_D=find(x==1);
Vyst_ind_D=find(x==-1);
```