UNIVERZITA PARDUBICE
FAKULTA EKONOMICKO – SPRÁVNÍ

BAKALÁŘSKÁ PRÁCE

2008 PETR ŠIMER

Univerzita Pardubice

Fakulta ekonomicko-správní

Předzpracování ekonomických dat pomocí analýzy hlavních komponent

Petr Šimer

Bakalářská práce

2008

Univerzita Pardubice
Fakulta ekonomicko-správní
Ústav systémového inženýrství a informatiky
Akademický rok: 2007/2008

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení: **Petr ŠIMER**

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

Studijní obor: **Informatika ve veřejné správě**

Název tématu: **Předzpracování ekonomických dat pomocí analýzy hlavních komponent**

Zásady pro vypracování:

Vícerozměrné statistické metody
Analýza hlavních komponent
Předzpracování dat pro ohodnocování bonity obcí
Analýza výsledků

Rozsah grafických prací:

Rozsah pracovní zprávy:

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

Seznam odborné literatury:

OLEJ, V. Modelovanie ekonomických procesov na báze výpočtovej inteligencie. Hradec Králové : M&V, 2003.

MELOUN, M., MILITKÝ, J. Statistická analýza experimentálních dat. Praha : Academia, 2004.

MELOUN, M., MILITKÝ, J., HILL, M. Počítačová analýza vícerozměrných dat v příkladech. Praha : Academia, 2005.

HEBÁK, P. Vícerozměrné statistické metody 3. Praha : Informatorium, 2007.

HALÁSEK, D., PILNÝ, J., TOMÁNEK, P. Určování bonity obcí. Ostrava : VŠB - Technická univerzita Ostrava, 2002.

Vedoucí bakalářské práce: **Ing. Petr Hájek, Ph.D.**
Ústav systémového inženýrství a informatiky

Datum zadání bakalářské práce: **30. října 2007**

Termín odevzdání bakalářské práce: **19. května 2008**

prof. Ing. Jan Čapek, CSc.
děkan

L.S.

doc. Ing. Pavel Petr, Ph.D.
vedoucí ústavu

V Pardubicích dne 30. října 2007

SOUHRN

Práce se zabývá předzpracováním ekonomických dat pomocí analýzy hlavních komponent. Nejprve jsou uvedeny základní pojmy a vysvětleny jednotlivé metody pro pochopení využití vícerozměrných statistických analýz a dále se podrobně věnuji analýze hlavních komponent. Předzpracování ekonomických dat je provedeno za použití sady statistických programových nástrojů v prostředí softwarového produktu Matlab. V závěru jsou vyhodnoceny výstupy a interpretovány výsledky, které poslouží jako vstupní data pro hodnocení bonity obcí.

KLÍČOVÁ SLOVA

vícerozměrné statistické metody, PCA, hlavní komponenty, předzpracování dat, hodnocení bonity obcí

TITLE

Preprocessing of economic data by principal components analysis.

ABSTRACT

The subject of this work is preprocessing of economic data using the principal components analysis.First, the basic concepts are mentioned, and the particular methods for understanding the usage of multidimensional statistical analyses are explained. Then follows the principal components analysis in detail. Preprocessing of economic data is implemented using statistical toolboxes in Matlab software environment. In the conclusion, the outputs are analyzed and the results are interpreted, which will serve as input data for municipality value assesment.

KEYWORDS

multidimensional statistical analyses, PCA, principal components, data preprocessing, municipality value assesment

OBSAH

1 Úvod

Současnou etapu vývoje společnosti lze chápat jako etapu přemíry informací, kterými jsme doslova zahlcováni. V tomto světě je nutné se naučit pohybovat a na jejich základě činit správná rozhodnutí. Jednou z vědních disciplín, která může být v této orientaci nápomocna, je právě matematická statistika, přesněji řečeno metody, vyplívající z jejího teoretického základu. Zpracování vícerozměrných dat v praxi využívá poznatků přírodních věd, matematické statistiky a informatiky v kombinaci se speciálními počítačově orientovanými postupy. Současné výkonné osobní počítače umožňují interaktivnost při zpracování vícerozměrných dat a interpretaci získaných výsledků. Aplikace statistických metod nachází široké uplatnění jak v klasických, tak i ekonomických a technických oborech.

Cílem této práce je přiblížit problematiku vícerozměrných statistických metod, kde je důraz kladen na analýzu hlavních komponent a s pomocí této metody provést předzpracování ekonomických dat, čímž lze dosáhnou lepší vypovídací schopnosti dat a je možné je vhodněji využít pro ohodnocování bonity obcí.

Práce je tématicky rozdělena na tři kapitoly. První kapitola je zaměřena na všeobecný popis vícerozměrných statistických metod, jejich klasifikaci, podstatě a možnostech aplikace. Po přiblížení jednotlivých metod se následně podrobně věnuji analýze hlavních komponent, které je tato práce věnována. Snažím se zde přiblížit její podstatu a zaměření, jaké problémy se touto metodou nejčastěji řeší a v neposlední řadě také grafické pomůcky, které se nejčastěji při realizaci této metody využívají. Další kapitola je věnována již samotnému předzpracování dat a jejich přípravě pro ohodnocování bonity obcí. Jako vstupní data byly použity údaje obsahující informace o 452 obcích, které jsou popsány pomocí 18 různých ekonomických kategorií vyjadřující různorodou kvalitu života a ekonomickou úroveň v těchto obcích. Kapitola dále obsahuje postup výpočtu hlavních komponent, který byl proveden pomocí sady statistických nástrojů v programovém prostředí Matlab. Pro přehlednější znázornění výsledků jsou uvedeny grafické diagramy a další tabulky s výsledky. V závěru potom následuje interpretace dosažených výsledků a vyhodnocení výstupů.

2 Vícerozměrné statistické metody

S využitím metod vícerozměrné statistické analýzy se lze setkat v oborech technického, ekonomického, demografického a sociologického charakteru, kde je cílem výzkumu poznání závislosti mezi proměnnými [4,9]. Při systematických kontrolních měřeních i při jednorázových pozorováních se zpravidla u jednotlivých statistických jednotek (země, regiony, podniky, pracovníci, výrobky) analyzuje větší počet jejich vlastností (statistických znaků). Vícerozměrná statistická analýza je poměrně mladá disciplína, její teoretické základy byly položeny ve 30. a 40. letech 20. století. Její rozvoj a uplatnění v praxi bylo ale podmíněno rozvojem výpočetní techniky a programového vybavení, protože řešení většiny úloh vícerozměrné statistiky vyžaduje provést velké množství výpočtů.

Zdrojová data bývají uspořádána do matice X má rozměr n * m. Řádky matice X často představují jisté objekty, na kterých se provádí zkoumání. Sloupce matice X představují zkoumané znaky, respektive vlastnosti (charakteristiky objektů), které se na objektech zkoumají. Matice X má následující tvar

$$\mathbf{X} = \begin{bmatrix} x_{1,1} \cdots & x_{1,i} \cdots & x_{1,m} \\ \vdots & \vdots & \vdots \\ x_{j,1} \cdots & x_{j,i} \cdots & x_{j,m} \\ \vdots & \vdots & \vdots \\ x_{n,1} \cdots & x_{n,i} \cdots & x_{n,m} \end{bmatrix}, \qquad (1)$$

Před vlastní aplikací vhodné metody vícerozměrné statistické analýzy je třeba vždy provést průzkumovou (exploratorní) analýzu dat, která umožňuje [6]:

a) posoudit podobnost objektů pomocí rozptylových a symbolových grafů,
b) nalézt vybočující objekty, resp. jejich znaky,
c) stanovit,zda lze použít předpoklad lineárních vazeb,
d) ověřit předpoklady o datech (normalitu, nekorelovanost, homogenitu, atd.).

Jednotlivé techniky k určení vzájemných vazeb se dále dělí podle toho, zda hledají strukturu a vazby ve znacích nebo v objektech. Jedná se o následující metody [6]:

a) hledání struktury ve znacích v metrické škále[1] – faktorová analýza, analýza hlavních komponent a shluková analýza,
b) hledání struktury v objektech v metrické škále – shluková analýza,
c) hledání struktury v metrické i v nemetrické škále[2] – vícerozměrné škálování,
d) hledání struktury v objektech v nemetrické škále – korespondenční analýza.

Většina metod vícerozměrné statistické analýzy umožňuje zpracování lineárních vícerozměrných modelů, kde se závisle proměnné uvažují jako lineární kombinace nezávisle proměnných, resp. vazby mezi proměnnými jsou lineární. V řadě případů se také uvažuje normalita metrických proměnných[6].

2.1 Určení struktury ve znacích a objektech

Určením struktury a vzájemných vazeb mezi znaky ale i mezi objekty se zabývají techniky redukce znaků na latentní proměnné, metoda hlavních komponent a metoda faktorové analýzy [5]. Důležitou metodou určení vzájemných vazeb mezi znaky je i kanonická korelační analýza, která se používá ke zkoumání závislosti mezi dvěma skupinami znaků, přičemž jedna ze skupin se považuje za proměnné nezávislé a druhá za skupinu proměnných závislých.

2.1.1 Metoda hlavních komponent

Metoda hlavních komponent (PCA) [6] je jedna z nejstarších a nejvíce používaných metod vícerozměrné analýzy. Poprvé byla zavedena Pearsonem již v roce 1901 jako popisná statistická metoda, sloužící především k redukci vícerozměrných dat. H. Hotelling zobecnil v roce 1933 postup aplikací komponentní analýzy na náhodné vektory a navrhl použití analýzy hlavních komponent pro rozbor kovarianční struktury proměnných. Cílem analýzy

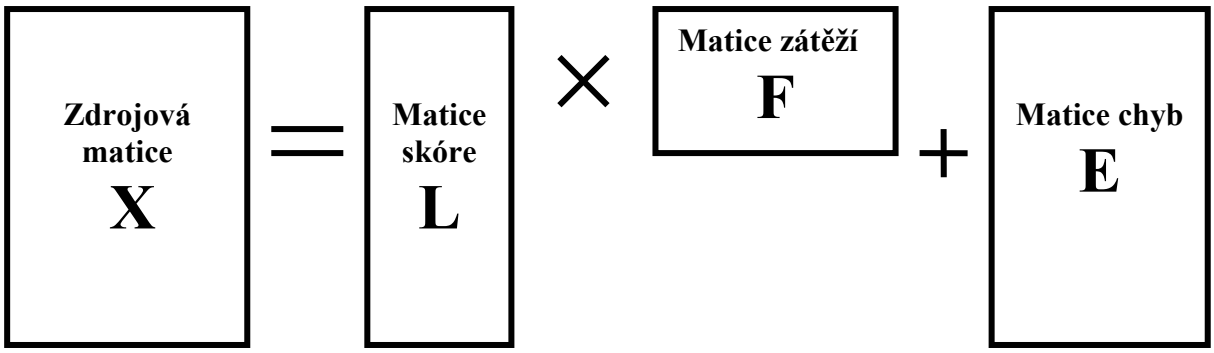
[1] Metrická škála vyjadřuje kvantitativní hodnotu posuzovaného znaku, kde proměnné jsou měřeny v číselné škále.
[2] Nemetrická škála vyjadřuje kvalitativní hodnotu posuzovaného znaku, proměnné jsou v pořadové nebo znakové škále.

hlavních komponent je především zjednodušení popisu skupiny vzájemně lineárně závislých čili korelovaných znaků.V analýze hlavních komponent nejsou znaky děleny na závisle a nezávisle proměnné jako v regresi. Techniku lze popsat jako metodu lineární transformace původních znaků na nové, nekorelované proměnné, nazvané hlavní komponenty. Základní charakteristikou každé hlavní komponenty je její míra variability čili rozptyl. Hlavní komponenty jsou seřazeny dle důležitosti, tj. dle klesajícího rozptylu, od největšího k nejmenšímu. Většina informace o variabilitě původních dat je přitom soustředěna do první komponenty a nejméně informace je obsaženo v poslední komponentě. Platí pravidlo, že má-li nějaký původní znak malý či dokonce nulový rozptyl, není schopen přispívat k rozlišení mezi objekty. Této metodě je věnována podrobněji následující kapitola.

2.1.2 Faktorová analýza

Faktorová analýza je vícerozměrná k vyšetření vnitřních souvislostí a vztahů čili korelací a odhalení základní struktury zdrojové matice dat [5]. Týká se analýzy struktury vnitřních vztahů mezi velkým počtem původních znaků pomocí souboru menšího počtu latentních proměnných zvaných faktory. Nejprve jsou identifikovány faktory, a pak je každému faktoru přidělen obsahový, obvykle fyzikální, význam, pomocí kterého je každý původní znak vysvětlen vybraným faktorem. Jde o dva primární cíle faktorové analýzy, a to jednak o sumarizaci a jednak redukci dat. V sumarizaci dat využívá faktorová analýza faktorů tak, aby data vysvětlila a usnadnila jejich pochopení daleko menším počtem latentních proměnných, než je počet původních znaků. Redukce dat je dosaženo vyčíslením skóre pro každý faktor a následnou náhradou původních znaků novými latentními proměnnými – faktory.

Podobně jako metoda hlavních komponent patří faktorová analýza mezi metody snížení dimenze čili redukce počtu původních znaků. Ve faktorové analýze předpokládáme, že každý vstupující znak můžeme vyjádřit jako lineární kombinaci nevelkého počtu skrytých společných faktorů a jediného specifického faktoru. Na rozdíl od PCA se faktová analýza věnuje vysvětlení závislosti znaků. K nevýhodám metody patří zejména nutnost zvolit počet společných faktorů ještě před prováděním vlastní analýzy[5,6]. Schematicky lze výpočet pomocí faktorové analýzy znázornit na Obrázku 1.



Obrázek 1 - Schéma maticových výpočtů faktorové analýzy (zdroj:[6])

Dle dosažených cílů faktorové analýzy lze ukončit analýzu interpretací nalezených faktorů. Je-li cílem nalézt logickou kombinaci znaků a lépe tak pochopit vnitřní vazbu mezi znaky, pak faktorová analýza postačí a celá vícerozměrná analýza je u konce. Tím se provede posouzení struktury znaků a nástin této struktury poslouží k vysvětlení i závěrů ostatních vícerozměrných technik. Je-li však cílem nalézt vhodné znaky pro následnou aplikaci ostatních statistických technik, pak se užije některá z forem redukce dat. Postup zahrnuje [6]:

- vyšetření faktorové matice a výběr znaků s největší faktorovou zátěží jako reprezentativní náhrady za dotyčnou faktorovou dimenzi,
- nahrazení původní skupiny znaků novou menší skupinou znaků, vytvořenou z faktorových skóre.

Cílem je zestručnění informace obsažené ve značném počtu původních znaků zdrojové matice dat do menšího souboru latentních proměnných, a to s minimální ztrátou informace. Jedná se o vyhledání konstrukce nebo rozměrů takové matice, která bude obsahovat informaci všech původních znaků. Nejprve je identifikována struktura dat, a to vyšetřením korelací mezi znaky (ve sloupcích) nebo korelací mezi objekty (v řádcích). Je-li cílem zestručnění popisných charakteristik, aplikuje se faktorová analýza na korelační matici znaků a analýza se nazývá R-faktorová analýza. Faktorová analýza se může aplikovat také na korelační matici objektů, a pak se nazývá Q-faktorová analýza. Zhušťuje velký počet objektů do různě velkých shluků. Výhodnější je užít analýzu shluků. Pak následuje redukce dat. Faktorová analýza může z velkého počtu původních znaků, vedle identifikace reprezentativní

skupiny souboru některých znaků, vytvořit naprosto novou skupinu znaků o daleko menším počtu a tou nahradit původní znaky. Cílem je zachovat povahu a charakter původních znaků při redukci jejich počtu za účelem zjednodušení vícerozměrné analýzy dat. Příspěvky každého znaku do faktoru, zvaného zátěž, je totiž vše, co je potřebné k této analýze.Užití faktorové analýzy s ostatními technikami vícerozměrného analýzy dat poskytuje přímý pohled do vnitřních vztahů mezi znaky a objekty. Zestručnění popisu dat poskytuje uživateli jasné vysvětlení, jak velký význam má ten který znak, s kterými vlastnostmi je spojen a kolik znaků má vliv na vlastní analýzu.Redukce dat a jejich sumarizace může být provedena buď s původní skupinou znaků, nebo se znaky vytvořenými novou analýzou. Když bude do analýzy zahrnut nekriticky velký počet všech původních znaků, je malá šance na dobrý výsledek[6,7].

2.1.3 Kanonická korelační analýza

Kanonická korelační analýza byla navržena v roce 1935 Hotellingem v souvislosti s hledáním lineární kombinace jedné skupiny znaků $x = (x_1,\ldots,x_q)$, která nejlépe koreluje s lineární kombinací druhé skupiny znaků $y = (y_1,\ldots,y_p)$ [6]. Vychází z předpokladu společného rozdělení obou skupin znaků. Podobně jako u PCA a faktorové analýzy se hledá lineární kombinace znaků obou skupin, tj. hypotetických kanonických proměnných, které vedou k maximálním vzájemným korelacím. Tyto kanonické proměnné tvoří nový souřadnicový systém vzájemně ortogonálních složek. Přesněji řečeno, jde o krokový proces, podobně jako v PCA, kdy v prvním kroku hledáme lineární kombinace x a lineární kombinace y, jejichž korelace je maximální. Tyto lineární kombinace x a lineární kombinace y, jejichž korelace je maximální. Tyto lineární kombinace tvoří první složky souřadnicových systémů kanonických proměnných pro x a y. V dalších krocích hledáme další lineární kombinace x a y, tj. kanonické proměnné takové, které mají maximální vzájemnou korelaci a přitom jsou nekorelované s kanonickými proměnnými (složkami obou nových souřadnicových systémů) nalezenými v předchozích krocích. Přímé využití této metody je při snižování dimenze, kdy jsou skupiny původních znaků veliké a účelem je nalézt malý počet kanonických proměnných (lineární kombinace původních znaků), které postihují v maximální míře korelace mezi původními skupinami znaků [5,6,7].

Kanonická korelační analýza úzce souvisí s chováním vícenásobného korelačního koeficientu R mezi jednou náhodnou veličinou a lineární kombinací jiných proměnných.

Tento koeficient nabývá maxima pro případ, kdy jsou koeficienty lineární kombinace přímo koeficienty regresními. Interpretace kanonických proměnných je ještě problematičtější než u faktorové analýzy. Slouží proto v řadě případů jako předzpracování dat pro další analýzu (například diskriminační) [5].

Kanonická korelační analýza se často využije v situacích , ve kterých se tvoří regresní modely a v nichž existuje více než jedna závisle proměnná. Zvláště je užitečná v situacích, kdy závisle proměnné jsou vnitřně korelovány, takže nemá cenu je vyhodnocovat odděleně, protože by se zanedbala jejich vzájemná vnitřní korelace. Užitečnou vlastností kanonické korelace je možnost ověřit nezávislosti mezi skupinami znaků x a y [6].

2.2 Klasifikace objektů

Hledáním struktury a vzájemných vazeb v objektech se zabývají klasifikační metody vícerozměrné statistické analýzy [5]. Klasifikační metody jsou postupy, pomocí kterých se jeden objekt zařadí do existující třídy (diskriminační analýza), nebo pomocí nichž lze neuspořádanou skupinu objektů uspořádat do několika vnitřně sourodých tříd či shluků (analýza shluků). Postup klasifikace je založen na určitých předpokladech o vlastnostech klasifikovaných objektů, například, když rozdělení náhodného vektoru charakterizující objekty je normální, pak hovoříme o parametrických klasifikačních metodách [5]. Ne-ní li klasifikace založena na znalostech rozdělení náhodného vektoru, mluví se o neparametrických klasifikačních metodách [5]. Významnou roli při hledání struktury a vazeb mezi objekty na základě jejich podobnosti tvoří také vícerozměrné škálování .

Obecně patří tyto úlohy do kategorie statistického učení (statistical learning) [5], kdy se na základě výstupů konstruuje predikce, která je založena na skupině znaků (vstupní data). Pro vybrané výstupy a znaky jdou k dispozici tzv. trénovaní data, kde je pro každý objekt určen jak výstup (y), tak i hodnoty všech znaků (x). Na základě těchto dat se sestavuje predikční model y = f(x), který umožňuje předpovědět výstupy $\hat{y}_0$ pro nový objekt charakterizovaný znaky x_0. Je zřejmé, že takto definovaná úloha statistického učení souvisí velmi úzce s regresní analýzou. Protože se na základě trénovacích dat učí predikční model předvídat y, označuje se tento postup jako učení s učitelem (supervised learning). Při učení bez učitele (unsuperwised learning) jsou k dispozici určité znaky pro objekty a nikoliv výstupy. Úlohou je pak pouze stanovit organizaci dat (shluky).

2.2.1 Diskriminační analýza

Diskriminační analýza umožňuje hodnocení rozdílů mezi dvěma nebo více skupinami objektů charakterizovaných více znaky. Obyčejně se dělí na techniky, které interpretují rozdíly mezi předem stanovenými skupinami objektů, a techniky, kde je cílem klasifikace objektů do skupin. Jsou porovnávány znaky objekty se znaky ostatních objektů. Na základě podobnosti nebo rozdílů se provede klasifikace objektů buď čistě subjektivně na základě zkušeností, nebo objektivními metodami [5].

Klasický klasifikační diskriminační metoda, zavedená Donaldem Fischerem v roce 1936, patří mezi metody zkoumání vztahu mezi skupinou p nezávislých znaků, zvaných diskriminátory (sloupců zdrojové matice), a jednou kvalitativní závisle proměnnou – výstupem. Výstupem je v nejednodušším případě binární proměnná y, nabývající hodnotu 0 pro případ, že objekt je v první třídě, respektive hodnotu 1 pro případ, že objekt je ve druhé třídě. O třídách je známé, že jsou zřetelně odlišené a každý objekt patří do jedné z nich. Účelem může být také identifikace, které znaky přispívají do procesu klasifikace. Ve vstupních datech trénovací skupiny jsou svými hodnotami diskriminátorů a výstupů všechny objekty zařazené do tříd. Účelem je nalézt predikční model umožňující zařadit nové objekty do tříd. Diskriminační analýza tedy [6]:

a) určuje, zda existují statisticky významné rozdíly mezi profily průměrného skóre diskriminátorů pro dvě či více předem definovaných tříd,
b) určuje, který z diskriminátorů se projevuje nejvíce v rozdílových profilech průměrného skóre dvou či více tříd,
c) stanoví postupy k zařazování objektů (jednotlivců,firem, výrobků atd.) do tříd, a to na základě jejich skóre v souboru diskriminátorů,
d) stanoví počet a složení dimenzí diskriminace mezi třídami tvořenými ze souboru diskriminátorů.

Z cílů diskriminační analýzy je zřejmé, že jsou důležité rozdíly mezi jednotlivými třídami nebo správné zařazování objektů do tříd. Technika pracuje nejlépe, když jde o jedinou závisle proměnnou a několik metrických nezávislých znaků diskriminátorů. Jako speciální postup analýzy profilu poskytuje diskriminační analýza ve výběru diskriminátorů

objektivní vyčíslení rozdílů mezi třídami. V tomto směru je diskriminační analýza dosti podobná vícerozměrné analýze rozptylu, k pochopení rozdílů mezi třídami umožňuje pohled do jednotlivých diskriminátorů a definuje rozměry diskriminace mezi třídami. Konečně pro klasifikační účely poskytuje diskriminační analýza základ k zařazení objektu a určuje diskriminační funkce k zařazování objektů do předem nadefinovaných tříd [6,7].

2.2.2 Shluková analýza

Jednou z možností využití informace obsažené ve vstupní datové matici je roztřídění množiny objektů do několika poměrně stejnorodých shluků [3]. Aplikací vhodných algoritmů můžeme odhalit strukturu datového souboru a jednotlivé objekty klasifikovat. Pojem klasifikace se tudíž ve statistické analýze používá ve dvou významech. Buď klasifikujeme objekty tak, že pro ně odhadujeme hodnotu nominální vysvětlované proměnné (například pomocí diskriminační analýzy), nebo objekty zařazujeme do skupin bez využití vysvětlované proměnné (například pomocí shlukové analýzy). Protože jde o různé úlohy, používá se odlišná symboliku. Počet shluků nejčastěji značíme písmenem k. Obvykle tento počet není znám a zjišťujeme jeho optimální hodnotu.

Pojem shluková analýza zahrnuje celou řadu metod a přístupů [3], jejichž cílem je nalézt skupiny podobných objektů (kromě shlukové analýzy lze ke stejnému účelu použít i metody patřící k jiným typům analýz, například k vícerozměrnému škálování). Uplatnění metod shlukové analýzy vede k příznivým výsledkům zejména tam, kde se množina objektů reálně rozpadá do tříd, tj. objekty mají tendenci se seskupovat do přirozených shluků. Zbývá pak již pouze najít vhodnou interpretaci pro popsaný rozklad, tj. charakterizovat vzniklé třídy.

Shlukovat můžeme nejen objekty, ale také proměnné. Pokud najdeme skupinu proměnných, jejichž hodnoty jsou si podobné, pak tuto skupinu může zastoupit jediná proměnná, čímž lze snížit rozměr úlohy. Další možností využití shlukové analýzy je zjišťování podobností kategorií nominální proměnné na základě dvourozměrné tabulky četností, tj. sdružených četností pro dva kategoriální znaky. Získaného poznatku můžeme využít pro sloučení kategorií, čímž získáme vyšší sdružené četnosti v kontingenční tabulce. Kromě výše uvedených přístupů existují metody, které umožňují shlukovat současně objekty i proměnné, případně současně kategorie dvou proměnných [3].

Shluková analýza může sloužit též pouze jako pomocný postup pro výběr objektů při analýze velkých datových souborů. Je-li vytvořen potřebný počet shluků objektů, pak lze analyzovat pouze data zjištěná u zástupců těchto shluků.

Stejně jako v mnoha ostatních úlohách z vícerozměrné analýzy máme k dispozici datovou matici X typu n × p, kde n je počet objektů a p je počet proměnných. Uvažujeme různé rozklady $S^{(k)}$ množiny n objektů do k shluků. Hledáme takový rozklad, který by byl z určitého hlediska výhodnější. Zde připouštíme pouze rozklady s disjunktními shluky (v širším slova smyslu chápaná shluková analýza řeší i úlohy spojené s pokrytím množiny objektů překrývajícími se shluky). Cílem je v podstatě dosáhnout stavu, kdy objekty uvnitř shluku jsou si podobné co nejvíce a objekty z různých shluků co nejméně [3].

2.2.3 Vícerozměrné škálování

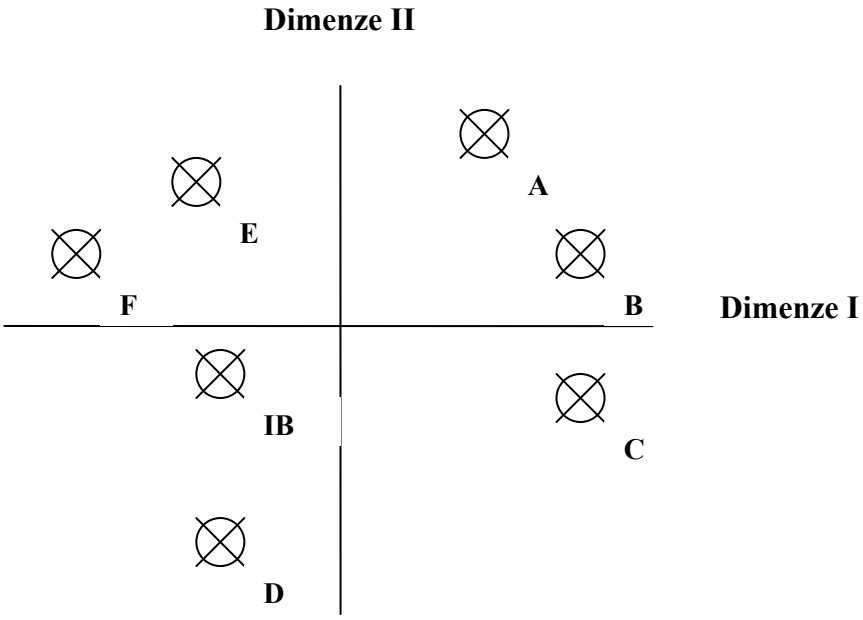
Vícerozměrné škálování (MDS) [3] je název pro skupinu exploratorních statistických metod, založených na redukci vícerozměrného prostoru objektů a průzkumové analýze vztahů mezi nimi. Objekty rozumíme řádky (pozorování) nebo sloupce (proměnné) datové matice. Vícerozměrné škálování pracuje s různými typy relací mezi objekty, přičemž nejčastěji jde o číselně vyjádřenou vzájemnou vzdálenost (blízkost) či nepodobnost (podobnost). Jsou však možné i jinak vyjádřené vztahy, např. korelace, asociace apod.

Polohu jednoho objektu ve dvourozměrném prostoru vyjádříme souřadnicemi $[x_1,x_2]$, polohu téhož objektu v p-rozměrném prostoru pak souřadnicemi $[x_1,x_2,\ldots,x_p]$. Vzdálenost dvou objektů v p-rozměrném prostoru vypočteme např. pomocí euklidovské míry. Vícerozměrné škálování řeší opačný postup. Na základě vzdáleností nebo měr nepodobnosti hledáme v prostoru odpovídající souřadnice objektů [3].

V typické aplikaci je každý z n objektů popsán p-rozměrným vektorem hodnot, jejichž přímé vzájemné porovnání může být komplikované. Smyslem MDS je optimálně snížit rozměr dat a zkoumat relace objektů v redukovaném prostoru. Ačkoli jsou výstupy MDS i číselné, jde hlavně o vizuální techniku. Objekty zobrazujeme v redukovaném prostoru, který označujeme konfigurace bodů (mapa objektů), a který bývá základním vodítkem pro interpretaci vztahů mezi objekty.

Ve své podstatě řeší MDS obdobné úlohy jako jiné vícerozměrné metody, např. faktorová analýza, korespondenční analýza, shluková analýza a analýza hlavních komponent. Na rozdíl od nich však nevyžaduje přímé určení matice pozorování – tu je možné určit

nepřímo z matice relací mezi objekty [3]. Každý objekt je popsán svými dimenzemi zvanými také znaky (Obrázek 2), a to znaky subjektivními (znaky vnímané člověkem) a znaky objektivními (znaky měřitelné fyzikálně).



Obrázek 2 - Subjetivní mapa relativního umístění objektů a znaků (zdroj:[6])

Zatímco subjektivní znaky jsou nemetrické, postavené na názoru respondenta (např. kvalitní, nekvalitní, laciný, drahý), objektivní znaky jsou metrické veličiny přístroji měřitelné. Mezi objektivními a subjektivními znaky objektu jsou určité rozdíly, jimiž jsou individuální rozdíly a vzájemná závislost [5,6].

2.2.4 Korespondenční analýza

Korespondenční analýza (CA) [3] je metoda založená na rozboru struktury vzájemných závislostí dvou a více proměnných uspořádaných do kontingenční tabulky. Ve své podstatě řeší obdobný problém jako faktorová analýza nebo metoda hlavních komponent, ve kterých se variability a závislosti původních proměnných vysvětlují pomocí menšího počtu latentních veličin (faktorů, komponent). V korespondenční analýze je obdobným způsobem sledování vliv jednotlivých kategorií, jejich vzájemná podobnost či asociace s kategoriemi ostatních proměnných. Latentní veličiny si lze představit jako osy redukovaného souřadného systému (korespondenční mapy), ve kterém jsou kategorie graficky reprezentovány.

Metoda je oblíbeným nástrojem zejména při zpracování rozsáhlejších kontingenčních tabulek, které obsahují mnohočetné kategorie, a kdy se grafické metody stávají ve srovnání s číselnými přehlednější. V uvedených případech mohou být výsledky analýzy rovněž vhodným návodem k tomu, které kategorie sloučit, a které ponechat samostatně. Vzhledem k tomu, že korespondenční analýza umožňuje v zásadě výzkum závislostí nominálních[3] nebo ordinálních[4] proměnných, je třeba případné spojité proměnné nejprve kategorizovat. Dále upozorněme, že jde hlavně o popisnou a průzkumovou metodu, která neobsahuje nástroje pro testování statistické významnosti získaných modelů. Pro tyto účely existují jiné metody – např. logaritmicko lineární modely kontingenčních tabulek [3].

V kvantitativním výzkumu může být korespondenční analýza součástí všech fází procesu zpracování kategoriálních[5] proměnných – od přípravy dat po vlastní prezentaci výsledků. Uplatňuje se hlavně v oblasti marketingových výzkumů při hodnocení vlastností výrobků, značek, postojů zákazníků, apod. [3] .

[3] Nominální proměnná je taková, o jejíž dvou hodnotách můžeme pouze říci, zda jsou stejné či různé (škola, fakulta, obor). Hodnotami mohou být texty (písmena), případně i číselné kódy. Lze u nich zjišťovat jen rozdělení četností, nemůžeme provádět aritmetické operace.

[4] Ordinální (pořadová), u jejíž dvou hodnot můžeme navíc určit pořadí (úroveň spokojenosti, vzdělání). Jako hodnoty lze použít text, datum, číslo. Pro statistické analýzy (s výjimkou zjišťování četností) je třeba texty převést na čísla.

[5] Proměnná, kterou není možno měřit, kvantifikovat, ale jen zařadit do tříd (např. svobodný, ženatý, rozvedený, vdovec.

3 Analýza hlavních komponent

V kapitole je popsána podstata PCA, rozdíly oproti faktorové analýze, možnosti vizualizace získaných výsledků a možné problémy, se kterými se lze při aplikaci této metody setkat.

3.1 Zaměření metody PCA

U mnoha výzkumných úloh se lze setkat se situací, kdy výchozí počet proměnných, sledovaných u zkoumaných jevů a procesů, je značný a pro interpretaci nepřehledný. Pro zjednodušení analýzy a snadnější hodnocení výsledků je často vhodné zkoumat, zda by studované vlastnosti pozorovaných objektů nebylo možné nahradit menším počtem jiných proměnných, shrnujících poznatky o výchozích proměnných získaných z dat, aniž by při tom došlo k větší ztrátě informace.

K řešení tohoto problému byly vytvořeny dvě příbuzné metody, a to metoda hlavních komponent a její obsahové, výpočetní a hlavně interpretační rozšíření – faktorová analýza. Pro metodu hlavních komponent je při stejných měřicích jednotkách a relativně podobné variabilitě všech proměnných výhodnější vycházet z analýzy kovarianční matice, zatímco faktorová analýza se téměř výhradně opírá o korelační matici. Obě metody se pokoušejí nalézt v pozadí stojící a tedy skryté veličiny, označované za hlavní komponenty nebo faktory, vysvětlující variabilitu a závislost uvažovaných proměnných. Tyto nově vytvořené proměnné nejsou ničím jiným než lineární kombinace původních měřitelných proměnných.
Od nových proměnných se v obou metodách požaduje, aby co nejlépe reprezentovaly původní proměnné. Konkretizace tohoto požadavku není však v obou metodách úplně stejná. Požadujeme-li, aby nové proměnné co nejvíce vysvětlovaly variabilitu původních proměnných, docházíme k metodě analýzy hlavních komponent. Požadujeme-li, aby soubor vytvořených proměnných co nejlépe reprodukoval vzájemné lineární vztahy původních proměnných, docházíme k metodě faktorové analýzy [3].

Metoda PCA je výhodná především pro možnost zobrazení vícerozměrných dat. Předpokládá se, že nevyužité hlavní komponenty obsahují malé množství informace, protože jejich rozptyl je příliš malý. Tato metoda je atraktivní především z důvodu, že hlavní komponenty jsou nekorelované. Namísto vyšetřování velkého počtu původních znaků s komplexními vnitřními vazbami analyzuje uživatel pouze malý počet nekorelovaných

hlavních komponent. Dále lze vybrané hlavní komponenty využít také k testu vícerozměrné normality. Analýza hlavních komponent je rovněž součástí průzkumové analýzy dat. Snížení rozměrnosti je často využíváno při konstrukci komplexních ukazatelů jako lineárních kombinací původních znaků. Například první hlavní komponenta je vlastně vhodným ukazatelem jakosti, pokud původní znaky charakterizují její složky. Využití první hlavní komponenty jako komplexního ukazatele je běžné v oblasti ekonomie, sociologie a medicíny. První dvě respektivě první tři hlavní komponenty se využívají především jako techniky zobrazení vícerozměrných dat v projekci do roviny nebo do prostoru.. Výhodou je, že tato projekce zachovává vzdálenosti a úhly mezi jednotlivými objekty. V řadě případů jsou hlavní komponenty pouze jednou z fází komplexnější analýzy. Například regrese využitím hlavních komponent umožňuje odstranění problémů s multikolinearitou[6] a přebytečným počtem vysvětlujících proměnných. Také hlavní komponenty, kterým odpovídá malý rozptyl mohou být v kontextu regrese důležité. Oblíbené je také využití hlavních komponent v oblasti řízení jakosti [5,6].

3.2 Základní vztahy metody PCA

Základním cílem PCA je transformace původních znaků, x_j, j=1,…m, do menšího počtu latentních proměnných y_j. Tyto latentní proměnné mají vhodnější vlastnosti: je jich výrazně méně, vystihují téměř celou proměnlivost původních znaků a jsou vzájemně nekorelované. Latentní proměnné jsou nazvány hlavními komponentami a jsou to lineárně kombinace původních proměnných: první hlavní komponenta y_1 popisuje největší část proměnlivosti čili rozptylu původních dat, druhá hlavní komponenta y_2 zase největší část rozptylu neobsaženého v y_1 atd. Matematicky řečeno, první hlavní komponenta je takovou lineární kombinací vstupních znaků, která má největší rozptyl mezi všemi ostatními lineárními kombi nacemi [5]. Má tvar

$$y_1 = \sum_{j=1}^{m} \mathbf{V}_{1j} \mathbf{x}_{Cj} = \mathbf{V}_1^T \mathbf{x}_C , \tag{2}$$

[6] Multikolinearita je silná (avšak nikoli funkční) vzájemná lineární závislost všech nebo některých vysvětlujících proměnných. Je-li vysoká, pak jsou odhadnuté výsledky nestabilní a nespolehlivé, v extrémních případech model nelze vůbec odhadnout.

kde sloupcový vektor původních znaků x_C obsahuje původní znaky v odchylkách od středních hodnot čili centrované hodnoty

$$\mathbf{x}_C = \left(x_1 - \mu_1, x_2 - \mu_2, \ldots, x_m - \mu_m,\right)^T, \tag{3}$$

Je zřejmé že rozptyl

$$D(y_1) = D\left(\mathbf{V}_1^T \mathbf{x}_C\right) = E\left[\left(V_1^T X_C\right)\left(V_1^T X_C\right)^T\right] = \mathbf{V}_1^T \mathbf{E}\left(\mathbf{x}_C \mathbf{x}_C^T\right)\mathbf{V}_1 = \mathbf{V}_1^T \mathbf{C} \mathbf{V}_1, \tag{4}$$

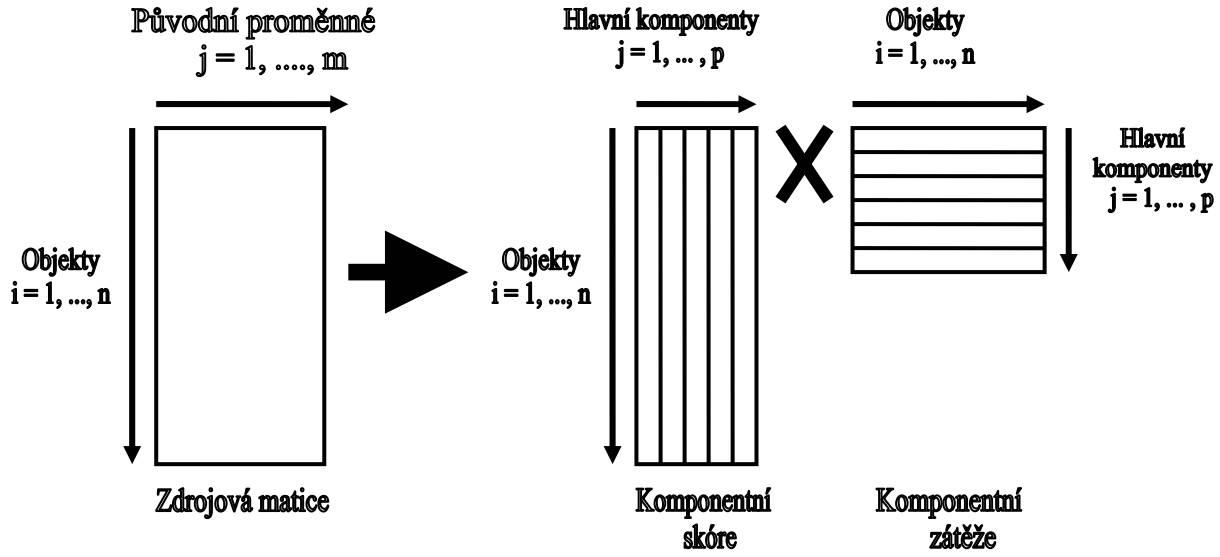
je závislý na velikosti vektoru koeficientů V_1. Symbol C označuje kovarianční matici. Je tedy třeba zavést vhodné omezení velikosti V_1. Standardním je použití normalizace $\mathbf{V}_1^T \mathbf{V}_1 = 1$. Pro vektor koeficientů

$$\mathbf{V}_1^T = \left(V_{11}, \ldots, V_{1m}\right)^T, \tag{5}$$

pak platí, že proměnlivost vyjádřená rozptylem $D(y_1)$ je maximální. Rozdíl mezi souřadnicemi objektů v původních znacích a v hlavních komponentách čili ztráta informace projekcí do menšího počtu rozměrů se nazývá mírou těsností proložení modelu PCA nebo také chybou modelu PCA. Na následujícím obrázku je tato situace schematicky znázorněna spolu s použitým označením [5].

I při velkém počtu původních znaků m může být k velmi malé, běžně 2 až 5. Volba počtu užitých komponent k vede k modelu hlavních komponent PCA. Vysvětlení užitých hlavních komponent, jejich pojmenování a vysvětlení vztahu původních znaků x_j, j = 1, …, m, k hlavním komponentám y_j, j = 1, …, k, tvoří dominantní součásti modelu hlavních komponent PCA [5,6].

Z Obrázku 3 je zřejmé, že zdrojová matice X_C se rozkládá na matici komponentních skóre T rozměru n * k a matici komponentních zátěží P^T rozměru k * m.



Obrázek 3 - Schéma maticových výpočtů v PCA (zdroj:[6])

Model PCA odpovídá aproximaci zdrojové matice dat X, který užijeme místo původní zdrojové matice dat X. Aproximace má řadu výhod v interpretaci dat. Nejde zde pouze o změnu systému souřadnic, ale především o nalezení a vypuštění šumu. PCA má proto dvojí cíl: transformaci do nového systému os a snížení rozměrnosti úlohy užitím několika prvních hlavních komponent, které vystihují strukturu v datech. Problémem zůstává, kolik hlavních komponent je nutné použít. Existuje horní mez počtu hlavních komponent, které mohou být odvozeny ze zdrojové matice dat X. Největší počet hlavních komponent se buď rovná číslu n – 1, nebo m v závislosti na tom, které z těchto dvou čísel je menší. Je-li X složena například z n = 40 spekter měřených při m = 2000 vlnových délkách, bude maximální počet hlavních komponent 39. Počet efektivních hlavních komponent se rovná hodnosti zdrojové matice X [6].

3.3 Grafické pomůcky analýzy hlavních komponent

Výsledek analýzy hlavních komponent lze zobrazit v několika různých typech grafů [6].

3.3.1 Cattelův indexový graf úpatí vlastních čísel

Je vlastně sloupcovým diagramem vlastních čísel $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m$ zdrojové matice dat **X** v závislosti na indexu i. Zobrazuje relativní velikost jednotlivých vlastních čísel. Využívá se k určení významných hlavních komponent. Úpatí je zlomové místo mezi kolmou

stěnou a vodorovným dnem. Nevýznamné hlavní komponenty nebo faktory představují vodorovné dno. Významné komponenty jsou tak odděleny zřetelným zlomovým místem, úpatím, a hodnota indexu i tohoto zlomu udává počet významných komponent [6].

3.3.2 Graf komponentních vah

Zobrazuje komponentní váhy čili zátěže pro první dvě hlavní komponenty. Každý bod v grafu odpovídá jednomu znaku a v grafu se porovnávají vzdálenosti mezi znaky. Krátká vzdálenost mezi dvěma znaky znamená silnou korelaci. Lze nalézt i shluk podobných znaků, jež spolu korelují. Graf můžeme považovat za most mezi původními znaky a hlavními komponentami, protože ukazuje, jakou měrou přispívají jednotlivé původní znaky do hlavních komponent. Někdy se podaří hlavní komponenty pojmenovat, vysvětlit a přidělit jim fyzikální, chemický nebo biologický význam. Pak lze názorně vysvětlit, jak jednotlivé původní znaky x_j, j=1,…,m, blízko sebe v prostorovém shluku, jde o silnou pozitivní kovarianci. Kovariance ještě však nemusí nutně znamenat korelaci [5,6].

3.3.3 Rozptylový diagram komponentního skóre

Zobrazuje komponentní skóre t_1, t_2 obyčejně pro první dvě hlavní komponenty u všech objektů. Body v tomto grafu jsou t_1, t_2, i = 1, …, n. Dokonalé rozptýlení objektů v rovině obou hlavních komponent ukazuje na dokonalé rozlišení objektů. Lze snadno nalézt shluk vzájemně podobných objektů a dále objekty odlehlé a silně odlišné od ostatních objektů. Diagram komponentního skóre však může být i ve třech či více hlavních komponentách a v rovinném grafu se pak sleduje pouze jeho průmět do roviny [6]. Tento diagram se užívá k identifikaci odlehlých objektů, identifikaci trendů, identifikaci tříd, shluku objektů, k objasnění podobnosti objektů atd. Nelze analyzovat všechny možné rozptylové diagramy, protože jich je velmi mnoho.

Výklad rozptylového diagramu komponentního skóre se týká [5]:

d) Umístění objektů – objekty daleko od počátku si jsou podobné, respektive centra jsou extrémy. Objekty nejblíže počátku jsou nejtypičtější.
e) Podobnosti objektů – objekty blízko sebe si jsou podobné, objekty daleko od sebe jsou si nepodobné.

f) Objektů v shluku – objekty umístěné zřetelně v jedno shluku jsou si podobné a přitom nepodobné objektům v ostatních shlucích. Jsou-li blízko sebe znamená to značnou podobnost objektů.
g) Osamělých objektů – izolované objekty mohou být odlehlými objekty, které jsou silně nepodobné ostatním objektům.
h) Odlehlých objektů – v ideálním případě bývají objekty rozptýlené po celé ploše diagramu. v opačném případě je něco špatného v modelu, obyčejně je přítomen silně odlehlý objekt. Odlehlé objekty jsou totiž schopny zbortit celý diagram.
i) Pojmenováním objektů – výstižná jména objektů slouží k hledání hlubších souvislostí mezi objekty a mezi pojmenovanými hlavními komponenty.
j) Vysvětlení místa objektu – umístění objektu v diagramu může být porovnáváno s komponentními vahami původních znaků ve dvojném grafu a pomocí původních znaků pak i vysvětleno.

3.3.4 Dvojný graf

Kombinuje předchozí dva grafy. Úhel mezi průvodiči dvou znaků je nepřímo úměrný velikosti korelace mezi těmito dvěma znaky. Čím je menší úhel, tím je větší korelace. Každý průvodič má své souřadnice na první a na druhé hlavní komponentě. Kombinace obou grafů do jediného společného přináší cenné srovnání, jeden graf zde působí jako doplněk vůči druhému. Když se ve dvojném grafu nachází objekt v blízkosti určitého znaku, znamená to, že tento objekt obsahuje velký podíl právě tohoto znaku a je s ním v interakci. Modifikací je 2D-graf s osami odpovídajícími prvním dvěma hlavním komponentám. Body zde identifikují objekty a vektory jsou projekce znaků. Jednotlivá skóre t_1 a t_2 obsahují souřadnice bodů, které se obvykle označují čísly objektů, stejně jako v rozptylovém grafu komponentních skóre. Jako měřítko na osách se volí měřítko odpovídající komponentním skóre, takže body leží uvnitř elipsy ohraničené křivkou [5].

3.3.5 Graf reziduí jednotlivých objektů

Rozptyl reziduí jednoho í-tého objektu představuje vzdálenost mezi tímto objektem a modelem. Je vhodné zobrazovat tuto veličinu pro všechny objekty v jednom společném grafu a odhalit tak odlehlé objekty či jiné anomálie objektů. Toto zobrazení je zvláště výhodné když potřebujeme porovnat rezidua jednotlivých objektů mezi sebou. Dosahuje-li reziduum

určitého objektu na ose y proti ostatním objektům abnormálně velké hodnoty, půjde pravděpodobně o odlehlý objekt [6].

3.3.6 Graf celkového reziduálního rozptylu všech objektů

Výstavba modelu hlavních komponent se provádí postupným přidáváním hlavních komponent. Graf celkového reziduálního rozptylu je obdobou grafu úpatí vlastních čísel a nalezení zlomu na sestupné křivce čili úpatí analogicky vystihuje optimální počet A využitelných hlavních komponent [6].

3.4 Nevýhody metody PCA

V analýze hlavních komponent se často lze setkat např. s těmito problémy [6]:

a) Data neobsahují předpokládanou informaci.Vysvětlení grafu diagramu metody PCA nemá smysl, protože data neobsahují informaci popisující studovaný problém.
b) Je užito příliš málo hlavních komponent V modelu PCA bylo použito příliš málo hlavních komponent.Nedostatečné vysvětlení dat vede ke ztrátě informace.Problém se může vyřešit opětovným rozborem grafu úpatí vlastních čísel.
c) Je užito příliš mnoho hlavních oponent.V modelu PCA bylo zahrnuto příliš mnoho hlavních komponent, což může vyvolat vážnou chybu, protože šum je zahrnut do modelu tečka.
d) Nejsou odstraněny odlehlé objekty. Odlehlé objekty mohou být důvodem hrubých chyb v datech. Do modelu jsou vtahovány spíše hrubé chyby než zajímavé proměnlivosti v datech.
e) Odstraněné odlehlé objekty obsahovaly důležitou informaci. Ztrátou určitých objektů se vytratila důležitá informace z dat a nalezený model je proto zkreslený.
f) Komponentní skóre je nedostatečně analyzováno. Nedostatečným rozborem rozptylového diagramu byly zanedbány důležité rysy v datech.
g) Vysvětlení komponentních vah se špatným počtem hlavních komponent může vést k vážnému zkreslení interpretace. Může totiž dojít k vyjmutí důležitých znaků, protože se zdají být odlehlými.

h) Je třeba hodně rozvažovat a přemýšlet o úloze samé a specifickém problému řešeném před pohodlným přebíráním počítačových výsledků.
i) S užitím špatného předzpracování dat. Chybná předúprava dat může vést ke zkresleným závěrům a neporozumění úloze.
i) Chybné předzpracování dat může vést ke zkresleným závěrům a neporozumění úloze. Způsob předpravy dat je obecně dán typem úlohy a druhem instrumentálních dat a může vést ke zkreslení informace.

3.5 Shrnutí metody PCA

Jako vhodnou situaci je možné označit případ, ve kterém máme k dispozici slušně velký náhodný výběr z vícerozměrného normálního rozdělení hodnot příliš velkého počtu vzájemně silně korelovaných veličin. Metoda hlavních komponent se především využívá pro zlepšení úrovně průzkumové analýzy dat, ale je užitečná rovněž v regresní, shlukové i faktorové analýze,jakož i v některých speciálních postupech jiných metod. Analýza hlavních komponent umožňuje odhalit případná narušení dat,jako jsou odlehlá pozorování, nestejná homogenita přirozených skupin v datech, odchylky od podmínky nezávislosti jednotlivých pozorování a nebo narušení předpokladu vícerozměrného normálního rozdělení [3]. Pokud by pomohla odhalit v pozadí stojící pojmenované příčiny korelovanosti a variability studovaných veličin, bylo by to asi více než se od metody hlavních komponent očekává.

Prvním krokem analýzy po formální inspekci dat je posouzení stupně korelovanosti sledovaných proměnných. K tomu je možné využít některou z variant Bartlettova testu diagonální korelační matice. Za uspokojivý stav se pro metodu hlavních komponent považuje silná vzájemná lineární závislost proměnných, kdy na rozdíl od vysokých párových korelačních koeficientů jsou dílčí korelační koeficienty téměř nulové, což ukazuje na nepodstatnost vlivu činitelů, které v dané úloze nejsou uvažovány. Další kroky metody hlavních komponent směřují k nalezení skutečného rozměru R dat a k určeních nejmenšího počtu hlavních komponent takových, aby zbývajících p – R komponent už nepředstavovalo užitečný přínos. Nástrojem pro rozhodnutí může být podíl vysvětlovaného součtu rozptylu původních proměnných, počet potřebných komponent podle subjektivního dojmu na základě o grafu charakteristických čísel, anebo v případě nutnosti použití korelační matice místo kovarianční matice i počet charakteristických čísel větších než jedna. Dává se sice přednost analýze hlavních komponent odvozených z kovarianční matice, ale v případě nestejných

měřicích jednotek nebo při velkých rozdílech mezi rozptyly nezbývá nic jiného než vyjít z normovaných proměnných a analýzu hlavních komponent založit na charakteristických číslech a vektorech korelační matice. Uživatel bude určitě spokojen, podaří-li se relativně velký počet proměnných nahradit uspokojivě nahradit jednou až dvěma hlavními komponentami, ale ani mírně větší počet hlavních komponent neznamená neúspěšnost metody hlavních komponent [3].

Silná závislost sledovaných proměnných, optimální volba relativně malého počtu hlavních komponent, silné korelace mezi výchozími proměnnými a ortogonálními komponentami, jsou důležité podmínky užitečnosti hodnot hlavních komponent u sledovaných objektů. Kromě posouzení kvality dat nabízejí nové proměnné příležitost využití i k jiným účelům a v optimistických případech dovolují formulovat závěry o interpretaci nově vzniklých proměnných. Pro analýzu dat je výhodnější komponentní skóre vycházející z kovarianční matice v nenormovaném tvaru, zatímco k některým dalším účelům, jako je třeba využití proměnných pro klasifikaci objektů, je lepší vycházet z normovaných komponent a z normovaného rozdělení.

V současné statistické literatuře [3] je analýza hlavních komponent doporučována především jako význačný nástroj průzkumové analýzy dat, dále jako velmi užitečná podpora některých dalších metod analýzy vícerozměrných pozorování, ale do jisté míry i jako samostatný nástroj rozboru struktury vztahů v množině vzájemně závislých proměnných.

4 Předzpracování dat pro ohodnocování bonity obcí

Cílem analýzy hlavních komponent je redukce původního počtu popisovaných proměnných novými veličinami (umělými), označenými jako komponenty, které shrnují informaci o původních proměnných za cenu minimální ztráty informace. Tyto komponenty jsou vzájemně nezávislé a jsou seřazeny podle svého příspěvku k vysvětlení celkového rozptylu pozorovaných proměnných. Analýza hlavních komponent může být chápána jako transformace z původního do nového souřadnicového systému, jehož osy jsou tvořeny hlavními komponentami. Osy procházejí směry maximálního rozptylu, protože podmínka nezávislosti komponent vede ke kolmosti os [1].

4.1 Zdrojová data

Uvažujeme příklad obsahující informace o 452 obcích, které jsou popsány pomocí 18 různých ekonomických kategorií vyjadřující různorodou kvalitu života a ekonomickou úroveň těchto obcí. Tyto znaky vycházejí z množiny znaků označených jako statisticky významných činitelů pro ohodnocování bonity obcí[7] experty v dané oblasti [2]. Ukázka dat je uvedena v Příloze A. Korelační matice dat je uvedena v Příloze B. Z ní je zřejmé, že data jsou vhodná pro předzpracování metodou PCA (tj. existují mezi nimi statisticky významné závislosti). Cílem je tato data zpracovat metodou PCA, pomocí které budou data analyzována a původní proměnné budou nahrazeny hlavními komponentami.

Data jsou popsána těmito kategoriemi (znaky):

1. 'Počet obyvatel'
2. 'Růst počtu obyvatel'
3. 'Míra nezaměstnanosti'
4. 'Míra koncentrace ekonomiky'
5. 'Index stáří'
6. 'Podíl vysokoškolsky vzdělaných obyvatel'
7. 'Průměrná mzda / Průměrná mzda v kraji'
8. 'Vybavenost infrastrukturu'
9. 'Opakující se příjmy / Běžné výdaje'
10. 'Vlastní příjmy / Celkové příjmy'
11. 'Kapitálové výdaje / Celkové výdaje'

[7] Finanční metoda hodnocení bonity obcí je založena na hodnocení nejvýznamnějších parametrů hospodaření obcí. Hlavním cílem je posouzení dlouhodobé stability, velikosti rozpočtu obce, efektivnosti a kvality hospodaření obce a v širším smyslu i zhodnocení schopnosti obcí zajistit splacení přijatého dluhu [2].

12. 'Investiční příjmy / Celkové příjmy'
13. 'Likvidní majetek na obyvatele'
14. 'Příjmy / Výdaje'
15. 'Příjmy / Počet obyvatel'
16. 'Dluhová služba'
17. 'Dluh na obyvatele'
18. 'Krátkodobé dluhy / Celkové dluhy

Nejprve jsou data v programovém prostředí Matlab načtena, dále je (pomocí příkazu *whos*) vygenerována tabulka s informacemi o všech proměnných.

Tato tabulka obsahuje 3 proměnné :

- categories – obsahuje názvy 18 kategorií (znaků),
- names - obsahuje názvy 452 obcí,
- ratings - tabulka se základními daty o velikosti 452 řádků a 18 sloupců.

Pro získání představy o datech je vytvořen diagram box plot (Příloha C), z něhož je patrné, že u kategorie likvidní majetek na obyvatele a dluh na obyvatele je větší variabilita než například u kategorie počet obyvatel a příjmy/počet obyvatel.

Jestliže jsou proměnné ve stejných jednotkách, lze spočítat hlavní komponenty přímo z hrubých, nestandardizovaných dat. Standardizace dat se často doporučuje, když jsou proměnné v různých jednotkách nebo když rozdíl v hodnotách mezi jednotlivými sloupci je značný. Proto je provedena standardizace dat. Nyní jsou data připravena pro nalezení hlavních komponent.

4.2 Analýza hlavních komponent

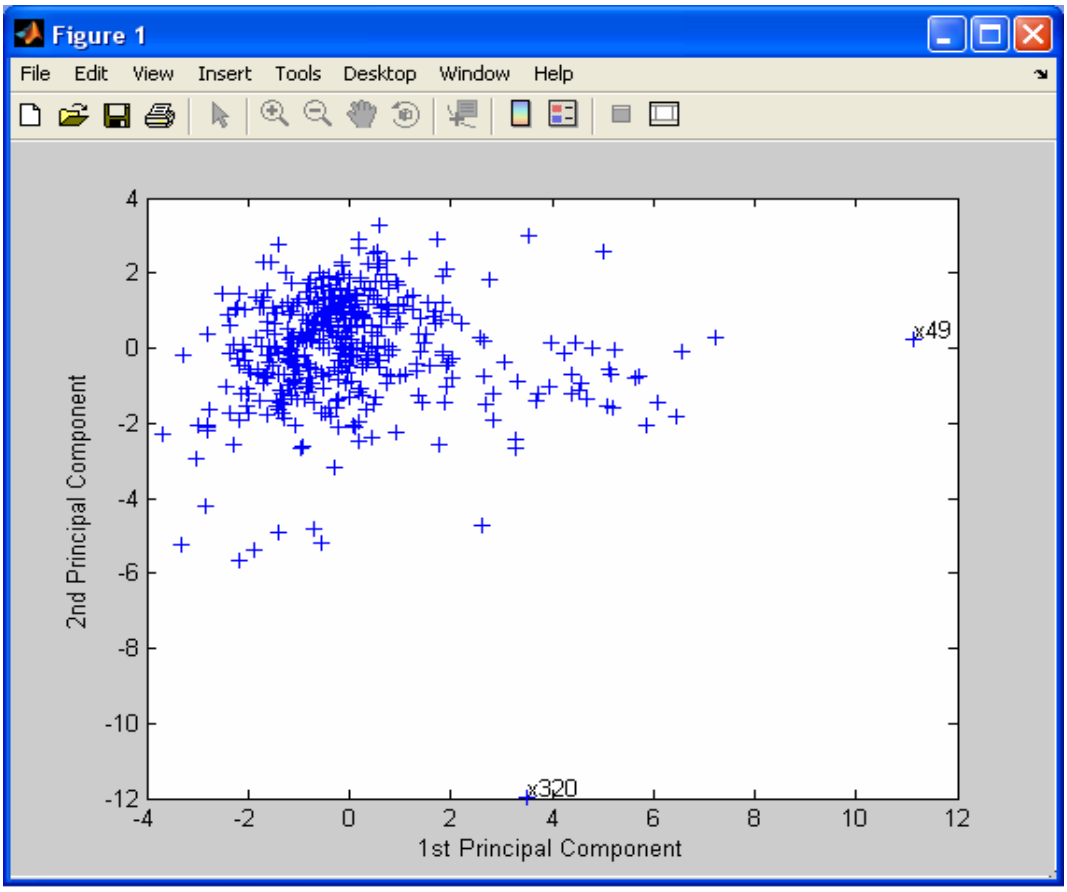
Výstupem realizované analýzy hlavních komponent (pomocí funkce princomp) jsou koeficienty pro 18 hlavních komponent. Jejich ukázka je uvedena v Tabulce 1.

Tabulka 1 - Ukázka koeficientů pro 8 hlavních komponent

	Eigenvectors - Charakteristické vektory matice							
Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8
Počet obyvatel	0,30773	-0,020452	0,25561	-0,010392	0,089452	0,16872	-0,08478	0,22034
Růst počtu obyvatel	0,19555	0,27713	-0,26869	-0,19686	-0,23758	-0,40826	-0,11637	-0,0233
Míra nezaměstnanosti	-0,14187	-0,46328	-0,012318	-0,049727	-0,06944	-0,0735	-0,20459	0,15915
Míra koncentrace ekonomiky	-0,14809	-0,2003	0,045755	0,34096	-0,14084	-0,13482	-0,39958	-0,3945
Index stáří	-0,19969	-0,1369	0,25707	0,081204	0,42858	0,41986	-0,016612	-0,26928
Podíl vysokoškolsky vzdělaných obyvatel	0,2879	0,21662	0,18376	-0,095412	0,091381	0,20782	0,26371	0,17542
Průměrná mzda / Průměrná mzda v kraji	0,07586	0,28779	-0,12516	-0,324	0,28166	0,084337	-0,17572	-0,42238
Vybavenost infrastrukturu	0,38319	0,1256	0,0001379	0,074397	0,010858	0,015113	0,030625	-0,091186
Opakující se příjmy / Běžné výdaje	-0,29732	0,18074	-0,44111	-0,024787	-0,027328	0,36798	0,019607	0,12398
Vlastní příjmy / Celkové příjmy	-0,16216	-0,31131	-0,075294	-0,34085	-0,0056808	-0,0042159	0,13106	0,33332
Kapitálové výdaje / Celkové výdaje	0,077132	-0,06402	-0,44912	0,15492	-0,39913	0,52004	0,10643	-0,099228
Investiční příjmy / Celkové příjmy	0,42789	-0,16141	-0,070254	0,17206	-0,10837	0,21009	-0,2779	-0,0071162
Likvidní majetek na obyvatele	0,13146	-0,42971	-0,17418	-0,42314	0,068613	-0,0049135	0,037277	-0,067367
Příjmy / Výdaje	-0,068776	0,18363	-0,23144	-0,0893	0,39796	0,083845	-0,58361	0,36459
Příjmy / Počet obyvatel	0,43243	-0,26406	0,015998	-0,095697	0,006472	0,10579	-0,2152	0,02747
Dluhová služba	0,10017	-0,066248	-0,3665	0,33965	0,32593	-0,24363	-0,010303	0,073936
Dluh na obyvatele	0,10951	-0,19267	-0,28392	-0,18998	0,28974	-0,10667	0,27507	-0,40193
Krátkodobé dluhy / Celkové dluhy	-0,12203	0,12009	0,19675	-0,44515	-0,33806	0,12443	-0,32547	-0,19579

Například největší hodnoty koeficientů v první hlavní komponentě mají znaky x_{12} a x_{15}, které odpovídají znakům 'Investiční příjmy/ Celkové příjmy' a 'Příjmy / Počet obyvatel'.

Dalším výstupem jsou komponentní skóre. Původní data jsou namapována do nového souřadnicového systému definovaného hlavními komponentami. Tento výstup má stejný rozměr jako původní matice dat. Ukázka hodnot komponentního skóre 1. a 2. hlavní komponenty pro obce je uvedena na Obrázku 4.



Obrázek 4 - Hodnoty komponentního skóre pro obce 1. a 2. komponenty

Tento rozptylový diagram komponentního skóre ukazuje celou vyšetřovanou strukturu objektů, tzn. shluky objektů, izolované objekty, odlehlé objekty, anomálie atd. Objekty mohou být označeny textovým popisem nebo číselně, indexem. Jak lze vidět na Obrázku 4, leží kromě objektů 49 a 320, zbývající objekty v několika shlucích. Objekty 49 a 320 jsou tedy odlehlé body. Objekt 49 je nejvíce odlehlý na první hlavní komponentě a popisuje většinu jejího rozptylu. První hlavní komponenta popisuje rozdíl mezi objektem 49 a ostatními objekty. Na druhé straně objekt 320 je extrém na druhé hlavní komponentě. Ostatní objekty tvoří v rovině prvních dvou hlavních komponent jeden velký a několik menších shluků.

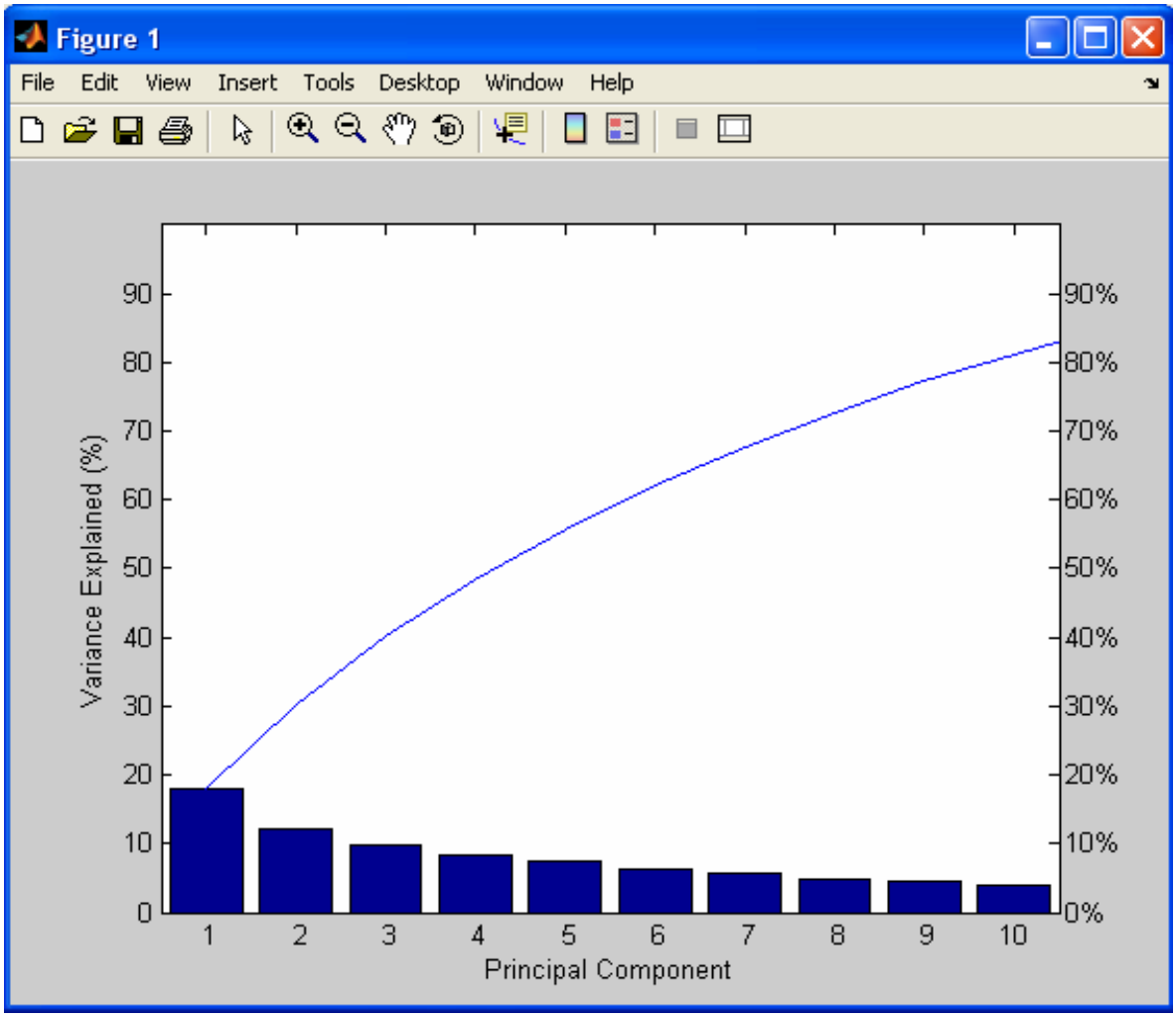
Jednotlivé komponenty lze charakterizovat pomocí vlastních čísel (Eigenvalue), kdy hodnota vlastního čísla větší než jedna slouží jako rozhodovací kriterium pro významnost hlavní komponenty. Dalším parametrem hlavních komponent je jejich podíl na celkovém rozptylu dat. Cílem je vysvětlit pomocí zvolených hlavních komponent maximální velikost celkového rozptylu. Příspěvky jednotlivých komponent k celkovému vysvětlenému rozptylu se načítají. Hodnoty vlastních čísel a vysvětlených rozptylů jsou uvedeny v Příloze D. Z té je zřejmé, že je vhodné zvolit prvních sedm hlavních komponent, jejich hodnoty vlastních čísel

jsou větší než jedna. Pomocí těchto sedmi komponent lze vysvětlit 67.78% celkového rozptylu v datech. V následující Tabulce 2 jsou uvedeny charakteristiky těchto sedmi hlavních komponent. První hlavní komponenta vysvětluje 18.06% celkového rozptylu, další hlavní komponenta 12.14%, atd.

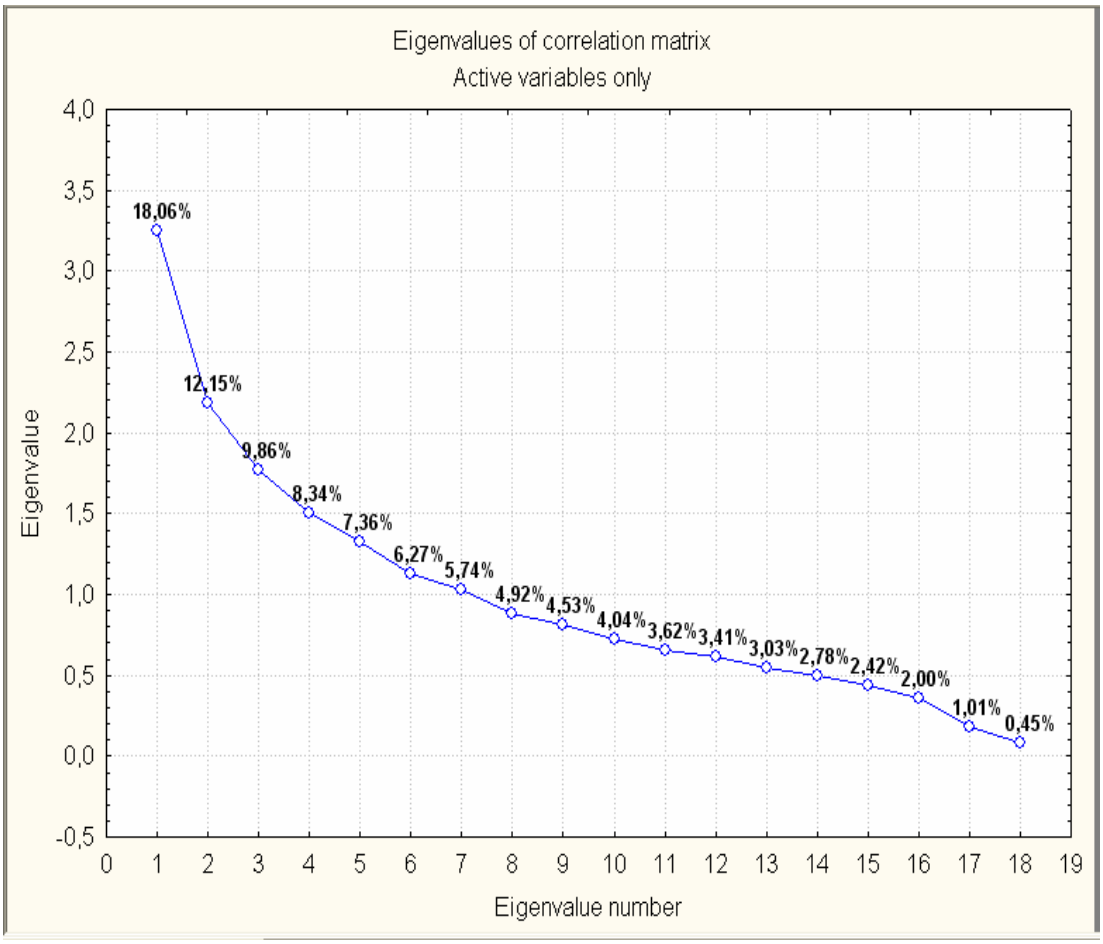
Tabulka 2 - Podíl jednotlivých komponent na celkovém rozptylu

KOMPONENTA	HODNOTA VLASTNÍHO ČÍSLA	PODÍL NA VYSVĚTLENÍ ROZPTYLU V %
I.	3.250466	18.05
II.	2.186827	12.15
III.	1.775603	9.86
IV.	1.501522	8.34
V.	1.325092	7.36
VI.	1.128484	6.26
VII.	1.032455	5.73

Graficky lze tyto výsledky prezentovat na Obrázku 5. Pro znázornění podílu jednotlivých komponent na celkovém rozptylu bylo zvoleno prvních deset komponent. Pokud by došlo ke zřetelnému zlomu v křivce kumulovaného rozptylu, bylo by možno zvolit takový počet hlavních komponent, který by odpovídal místu tohoto zlomu. V tomto případě roste vysvětlený podíl celkového rozptylu stabilně, k žádnému zlomu nedochází. Takto je tedy 18 znaků redukováno do 10 vzájemně nezávislých komponent a seřazeno podle toho, jaký podíl mají na vysvětlení celkového rozptylu. Vlastnosti hlavních komponent jsou takové, že 1. komponenta vysvětluje největší množství rozptylu, 2.menší a podíly vysvětleného rozptylu se u dalších komponent zpravidla rychle snižují. Na dalším Obrázku 6 jsou uvedeny podíly rozptylů společně s hodnotami vlastních čísel jednotlivých komponent.



Obrázek 5 - Podíl komponent na celkovém rozptylu



Obrázek 6 - Procentuální vyjádření a hodnoty vlastních čísel komponent

4.3 Interpretace hlavních komponent

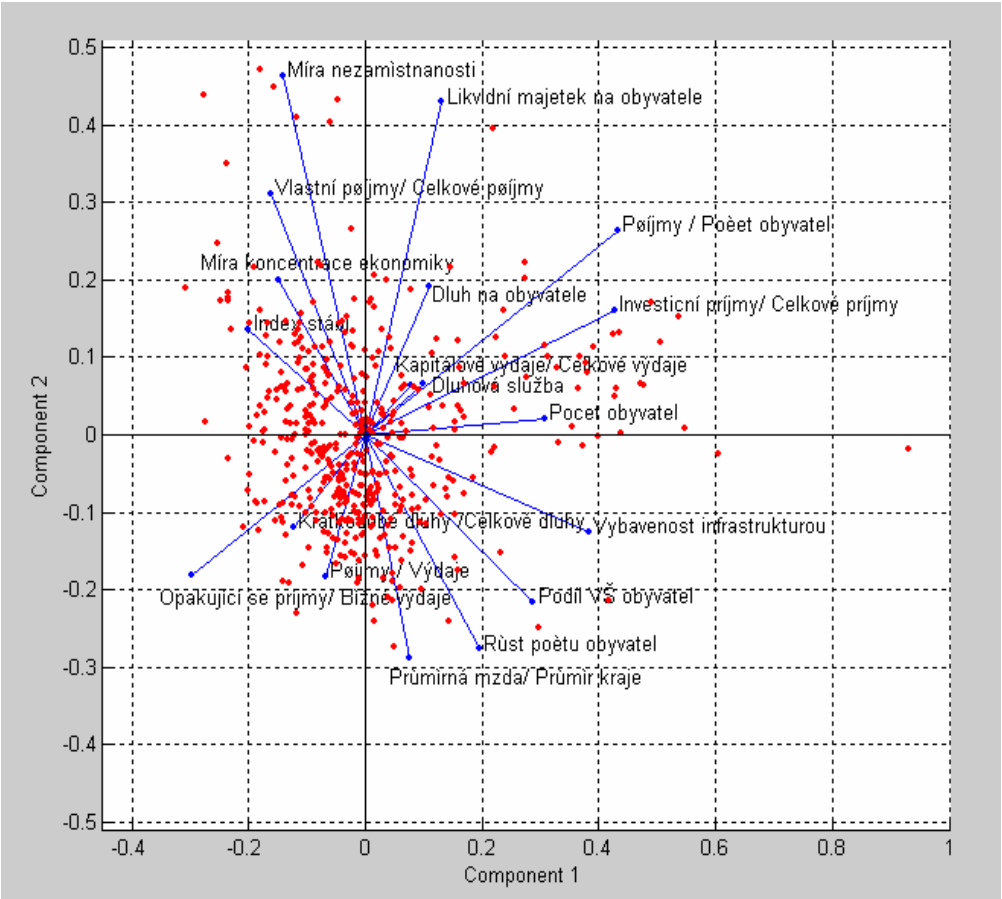
Po zvolení optimálního počtu hlavních komponent následuje interpretace těchto komponent, zpravidla je cílem nazvat skupinu znaků nejvíce zastoupených v každé hlavní komponentě jedním názvem.

I. KOMPONENTA – obsahuje několik vysvětlujících znaků s podobnými komponentními skóre. Je-li hodnota komponentního skóre záporná, znak má negativní vliv na komponentu. V této komponentě jsou nejvíce zastoupeny znaky uvedené v Tabulce 3. Graficky jsou hodnoty komponentních skóre uvedeny na Obrázku 7 (pro první dvě hlavní komponenty).

Tabulka 3 - Znaky a jejich hodnota komponentního skóre zastoupené v první komponentě

Vysvětlující znak	Hodnota komponentního skóre
1) Příjmy / Počet obyvatel	0.4323
2) Investiční příjmy / Celkové příjmy	0.4278
3) Vybavenost infrastrukturou	0.3832
4) Opakující se příjmy / Běžné výdaje	-0.2973

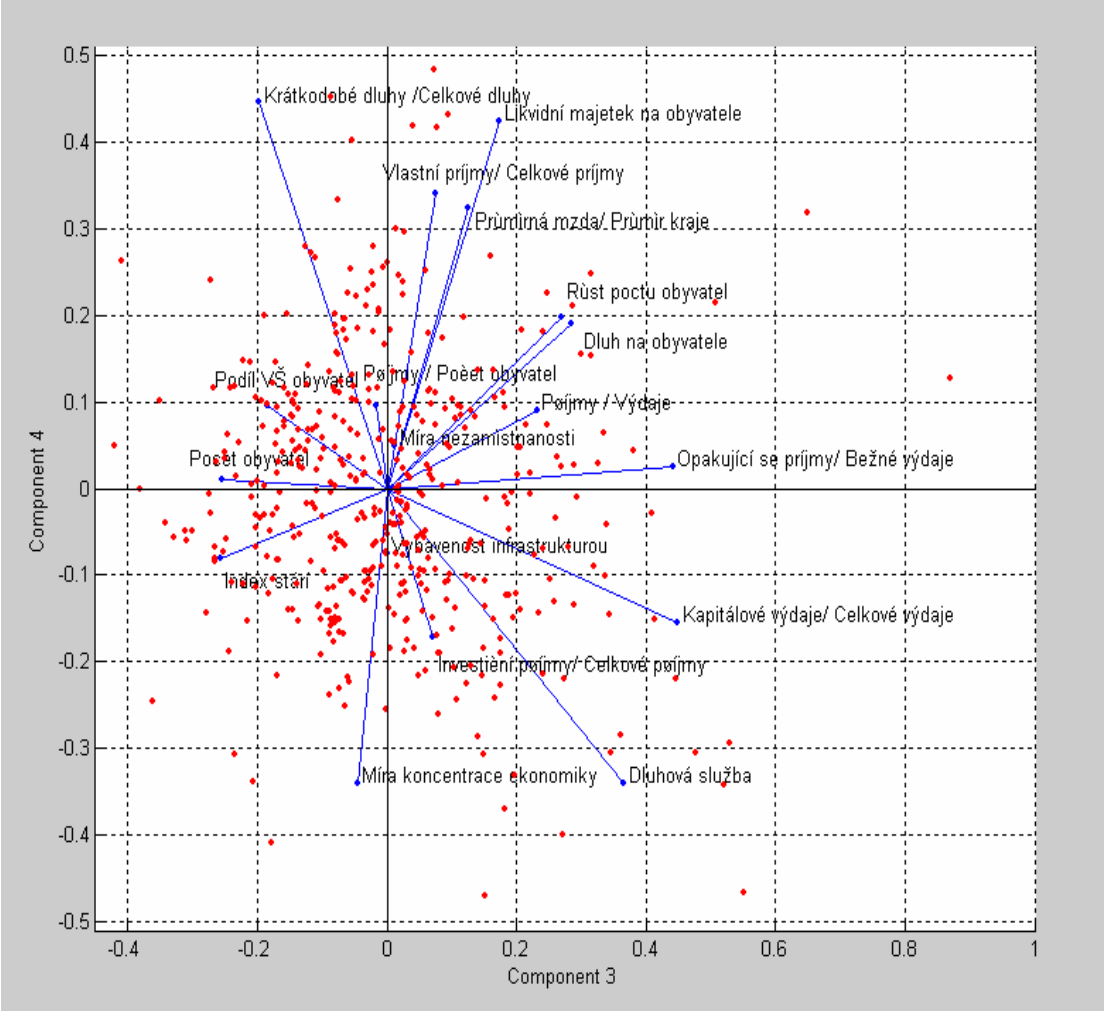
Lze ji interpretovat jako **Komponentu příjmů a infrastruktury**. Vysoká hodnota této komponenty ukazuje na vysoké příjmy obce (celkové i investiční), dobrou vybavenost infrastrukturou (plyn, vodovod, atd.), ale také na špatné rozpočtové hospodaření obce způsobené pravděpodobně vysokou investiční aktivitou.



Obrázek 7 - Komponentní skóre pro 1. a 2. hlavní komponentu

II. KOMPONENTA – největší koeficient korelace s komponentou (tj. vysoké hodnoty komponentního skóre) vykazuje Míra nezaměstnanosti (0.4632), Likvidní majetek na obyvatele (0.4297), Vlastní příjmy / Celkové příjmy (0.31131) a Průměrná mzda / Průměr kraje (-0.28779). Lze ji nazývat **Komponentou bohatství**. Vysoké hodnoty druhé komponenty dosahují obce s vysokou mírou nezaměstnanosti, s množstvím majetku (nejčastěji budov a pozemků), vysokým podílem vlastních příjmů (plynoucích obvykle z vlastnictví tohoto majetku) a nízkými mzdami.

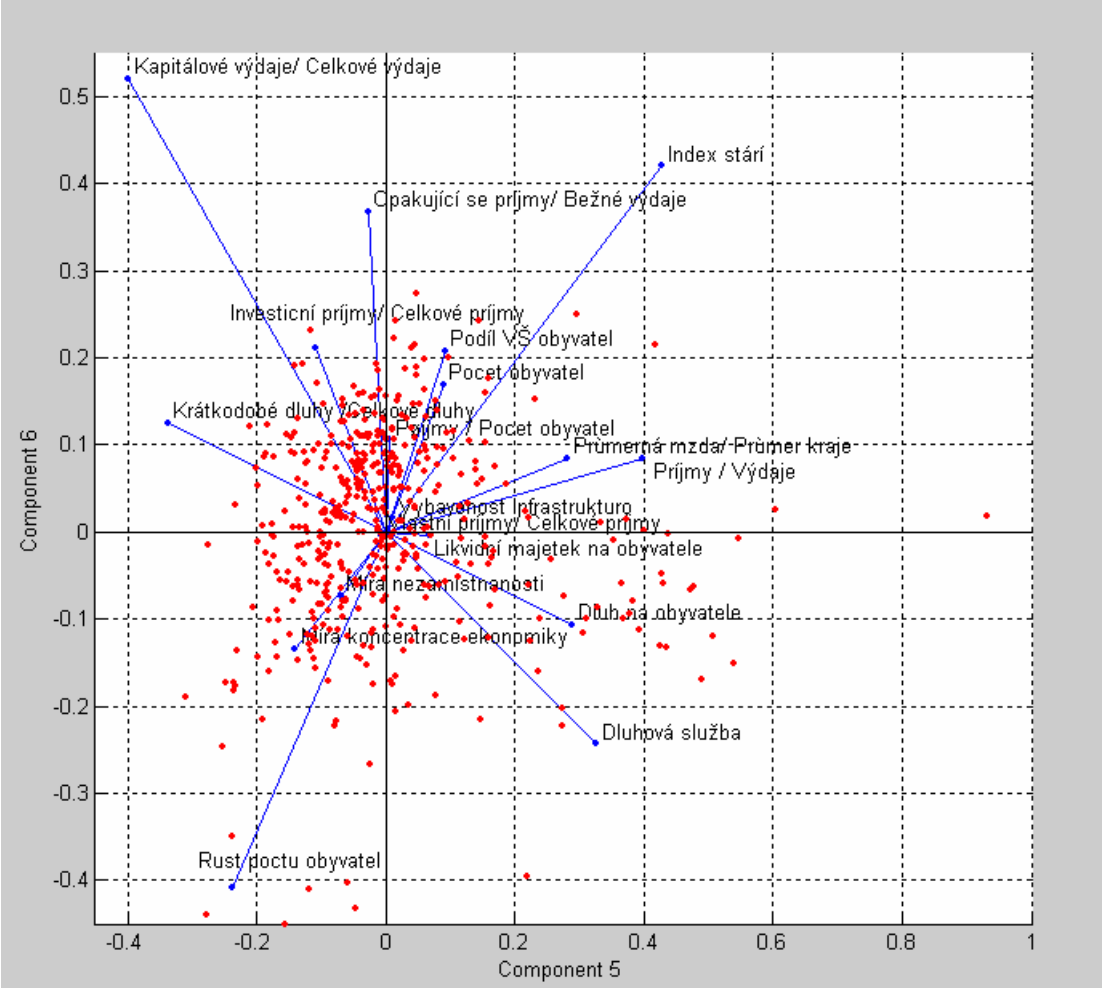
III. KOMPONENTA – nejvíce s ní korelují Kapitálové výdaje / Celkové výdaje (0.44912), Opakující se příjmy / Běžné výdaje (0.44111) a Dluhová služba (0.3665). Lze ji nazvat jako **Výdajová komponenta**. Hodnotu této komponenty zvyšuje podíl kapitálových výdajů, přebytek opakujících se příjmů nad výdaji a podíl výdajů na dluhovou službu. Grafické znázornění pro 3. a 4. hlavní komponentu je na Obrázku 8.



Obrázek 8 - Komponentní skóre pro 3. a 4. hlavní komponentu

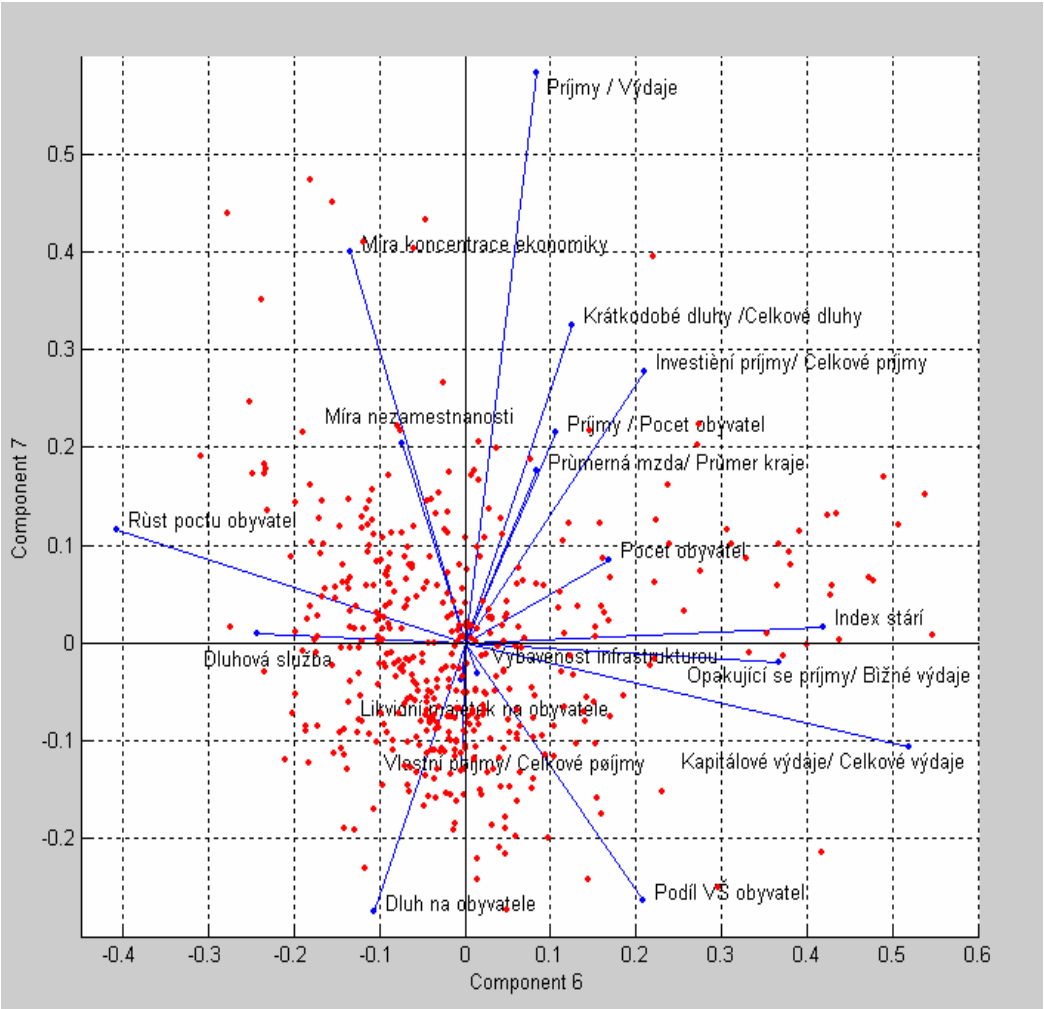
IV.KOMPONENTA – nejvíce s ní korelují Krátkodobé dluhy / Celkové dluhy (0.44515), Likvidní majetek na obyvatele (0.42314), Dluhová služba (-0.33965) a Míra koncentrace ekonomiky (-0.34096). Lze ji nazvat **Dluhovou komponentou.** Hodnotu této komponenty zvyšuje podíl krátkodobých dluhů, množství likvidního majetku obce a naopak snižuje koncentrace ekonomiky obce a dluhová služba obce.

V.KOMPONENTA - nejvíce s ní korelují Index stáří (0.42858), Příjmy / Výdaje (0.39796) a Kapitálové výdaje / Celkové výdaje (-0.39913). Lze ji nazvat **Komponentou investičních výdajů**. Podíl obyvatel nad 65 let zvyšuje výdaje na sociální služby, zároveň však obyvatelé takových obcí nemají takovou kupní sílu, snižuje se tak potřeba ostatních veřejných služeb a statků (a tím i kapitálových výdajů, které jsou tvořeny zejména výdaji na infrastrukturu). Nižší potřeba kapitálových výdajů pak vede ke zvýšení rozpočtového salda (Příjmy / Výdaje). Na Obrázku 9 jsou znázorněny hodnoty komponentního skóre pro 5. a 6. hlavní komponentu.



Obrázek 9 - Komponentní skóre pro 5. a 6. hlavní komponentu

VI. KOMPONENTA – nejvíce s ní korelují Kapitálové výdaje / Celkové výdaje (0.52004) Index stáří (0.41986) a Růst počtu obyvatel (-0.40826). Lze ji nazvat **Demografickou komponentou**. Podíl obyvatel nad 65 let značí obce se snižujícím se počtem obyvatel. S tímto ukazatelem souvisí negativní vliv znaku Růst počtu obyvatel.



Obrázek 10 - Komponentní skóre pro 6. a 7. hlavní komponentu

VII. KOMPONENTA - nejvíce s ní koreluje znak Příjmy / Výdaje (0.58361) jak je patrné z Obrázku 10. Lze ji nazvat **Rozpočtovou komponentou**. Komponenta je pozitivně ovlivněna přebytkem celkových příjmů nad celkovými výdaji.

Z uvedených výsledků vyplývá, že tyto komponenty pokrývají veškeré ekonomické a finanční oblasti rozvoje obce. Jsou zde zastoupeny komponenty spojené s rozpočtem (tj. s příjmy, výdaji a jejich saldem), dluhem obce, jejím majetkem i ekonomickým a demografickým rozvojem. Nalezené komponenty snižují počet původních znaků (18) na výsledných 7 komponent při zachování vysokého procenta vysvětleného celkového rozptylu v datech.

5 Závěr

První část této práce je zaměřena na všeobecný popis vícerozměrných statistických metod a následně věnována jedné z nejvýznamnějších metod - analýze hlavních komponent.

Po prostudování materiálů jsem dospěl k závěru, že většina učebnic o vícerozměrných statistických metodách uvádí jako hlavní cíle analýzy hlavních komponent redukci rozměru množiny dat a identifikaci nových smysluplných proměnných. První cíl však není úplně pravdivý, protože ve skutečnosti se nesnažíme snížit, ale nalézt správný rozměr souboru dat. Hlavní komponenty mohou tedy usnadnit určení rozměru úlohy, a to bez výrazné ztráty informace zlepšit kvalitu analýzy. Pokud jde o druhý cíl, metoda hlavních komponent za běžně splněných podmínek vždy vede k novým proměnným, ale nelze nijak zaručit, že tyto nově vzniklé proměnné budou smysluplné. Avšak i v takovém případě je analýza hlavních komponent velice užitečná při identifikaci přirozených shluků objektů či proměnných, a ovšem z nejrůznějších důvodů i při snaze o snížení uvažovaných proměnných. Pokud je možné nové proměnné i rozumně interpretovat, je to spíše méně častý případ, a tedy něco navíc. Je však velice důležité si uvědomit, že metoda hlavních komponent pomáhá výzkumníkům, a to i v případě, že hlavní komponenty nelze přímo rozumně interpretovat.

Pro téměř každou vícerozměrnou analýzu dat lze doporučit metodu hlavních komponent jako první krok pro ověření předpokladů a pro odhalení případných odlehlých pozorování či jiných datově podezřelých okolností. Metoda hlavních komponent je i z jiného hlediska užitečným pomocníkem některých statistických metod. Například pomáhá regresní analýze v případě velkého počtu vzájemně závislých vysvětlujících proměnných, diskriminační analýze při malém počtu pozorování a velkém počtu proměnných, shlukové analýze při klasifikaci objektů do homogenních skupin na základě velkého počtu proměnných, ale i faktorové analýze či jiným vícerozměrným metodám jako jedno z možných prvních řešení.

Analýza hlavních komponent může být velmi dobře využita nejen k nalezení odpovědí na vyslovené otázky, ale je to obecně vynikající diagnostický nástroj pro identifikaci a hodnocení zvláštností posuzovaných a analyzovaných údajů. Je také považována za jeden z nejužitečnějších nástrojů pro posouzení a prověření kvality vícerozměrných dat, a proto se doporučuje pro poznání a pochopení dat téměř u každé vícerozměrné úlohy začít výpočtem a zobrazením hlavních komponent [6].

V další části práce jsem provedl předzpracování dat pomocí analýzy hlavních komponent v programovém prostředí produktu Matlab a jeho statistických nástrojů, zejména Statistical Toolbox. Vstupními daty pro analýzu hlavních komponent byla matice obsahující informace o 452 obcích Pardubického kraje, které jsou popsány pomocí 18 odlišných ekonomických kategorií vyjadřující různorodou ekonomickou úroveň těchto obcí (tzv. bonitu obcí).

Po provedení výpočtu mi jako rozhodující kritérium pro výběr počtu hlavních komponent posloužila hodnota vlastního čísla komponenty, kdy hodnota větší než jedna byla hraniční. S původních 18 proměnných byla hodnota vlastního čísla komponenty větší než jedna u prvních 7. Dalším významným parametrem hlavních komponent je jejich podíl na celkovém rozptylu dat. Cílem je vysvětlit pomocí zvolených hlavních komponent maximální velikost celkového rozptylu. Pomocí těchto sedmi komponent lze vysvětlit 67.78% celkového rozptylu v datech, kde první hlavní komponenta vysvětluje 18.06% celkového rozptylu, další hlavní komponenta 12.14%.

Z uvedených výsledků tedy po jejich interpretaci vyplynulo, že jsou zde zastoupeny komponenty spojené s rozpočtem, dluhem, majetkem i ekonomickým a demografickým rozvojem obce a pokrývají tak veškeré ekonomické a finanční oblasti rozvoje obce. S původních 18 znaků nalezené komponenty snížili jejich počet na 7 hlavních komponent, a to při zachování vysokého procenta vysvětleného celkového rozptylu v datech a tímto považuji cíle práce za splněné.

Použitá literatura

[1] *Analýza hlavních komponent v programu SAS* [online]. 2005 [cit. 2008-07-20]. Dostupný z WWW:<http://info.lu2.name/soubory/Principal_components_analysis_481.pdf>.

[2] HALÁSEK , Dušan, PILNÝ, Jaroslav, TOMÁNEK, Petr. *Určování bonity obcí*. Ostrava : Technická univerzita Ostrava, 2002. ISBN 80-248-0159-0. s. 130.

[3] HEBÁK, Petr, et al. *Vícerozměrné statistické metody (3)*. 2. dopl. vyd. Praha : Informatorium, 2007. ISBN 978-80-7333-0. s. 262.

[4] KUBANOVÁ, Jana. *Statistické metody pro ekonomickou a technickou praxi*. 2. vyd. Bratislava : Statis, 2004. ISBN 80-85659-37-9. s. 248.

[5] MELOUN, Milan, MILITNÝ, Jiří. *Statistická analýza experimentálních dat*. Praha : Academia, 2004. ISBN 80-200-1254-0. s. 941.

[6] MELOUN, Milan, MILITNÝ, Jiří, HILL, Martin. *Počítačová analýza vícerozměrných dat v příkladech*. Praha : Academia, 2005. ISBN 80-200-1335-0. s. 445.

[7] MELOUN , Milan, MILITNÝ, Jiří. *Kompendium statistického zpracování dat*. Praha : Academia, 2006. ISBN 80-200-1396-2. s. 982.

[8] OLEJ, Vladimír. *Modelovanie ekonomických procesov na báze výpočtovej inteligencie*. 1. vyd. Hradec Králové : Miloš Vognar - M&V, 2003. ISBN 80-903024-9-1. s. 160

[9] VÍŠEK, Jan. *Statistická analýza dat.* Skriptum, Praha : ČVUT , 1998. s. 187

Seznam obrázků

Seznam tabulek

Seznam příloh

Přílohy

PŘÍLOHA A – Ukázka tabulky zdrojových dat

Tabulka 4 - Zdrojová data

	A	B	C	D	E	F	G	
1		Pocet obyvatel	Růst počtu obyvatel	Míra nezaměstnanosti	Míra koncentrace ekonomiky	Index stáří	Podíl VŠ obyvatel	Průměrn
2	x1	296	1.260	10.256	0.124	14.286	4.815	
3	x2	589	2.572	8.784	0.153	5.986	2.651	
4	x3	154	1.055	6.329	0.207	15.385	6.289	
5	x4	239	0.860	9.924	0.196	15.789	4.231	
6	x5	3280	1.439	5.499	0.153	11.881	9.085	
7	x6	225	1.160	10.680	0.125	11.607	1.000	
8	x7	236	1.040	10.000	0.203	17.241	2.609	
9	x8	936	1.051	8.958	0.196	13.646	5.139	
10	x9	212	1.019	7.619	0.156	14.612	1.852	
11	x10	71	1.029	12.821	0.183	16.438	1.389	
12	x11	395	1.018	13.636	0.135	15.250	3.741	
13	x12	1200	1.504	6.289	0.131	12.556	9.598	
14	x13	970	1.136	5.837	0.170	11.802	5.959	
15	x14	400	0.990	8.213	0.144	10.922	2.600	
16	x15	1847	1.047	9.324	0.189	13.411	3.620	
17	x16	327	0.962	6.557	0.181	15.031	2.410	
18	x17	1937	1.015	9.692	0.166	13.217	3.043	
19	x18	788	1.012	9.406	0.190	12.594	3.636	
20	x19	282	1.160	10.833	0.129	13.011	4.472	
21	x20	252	0.969	9.091	0.190	12.302	0.405	
22	x21	6299	0.980	7.468	0.178	15.276	4.475	
23	x22	1821	1.208	10.147	0.174	13.418	4.234	
24	x23	811	1.056	7.254	0.180	15.711	3.861	
25	x24	234	1.153	8.257	0.161	13.559	4.018	
26	x25	182	1.117	11.458	0.268	16.757	5.405	
27	x26	912	1.016	9.131	0.188	13.706	4.772	

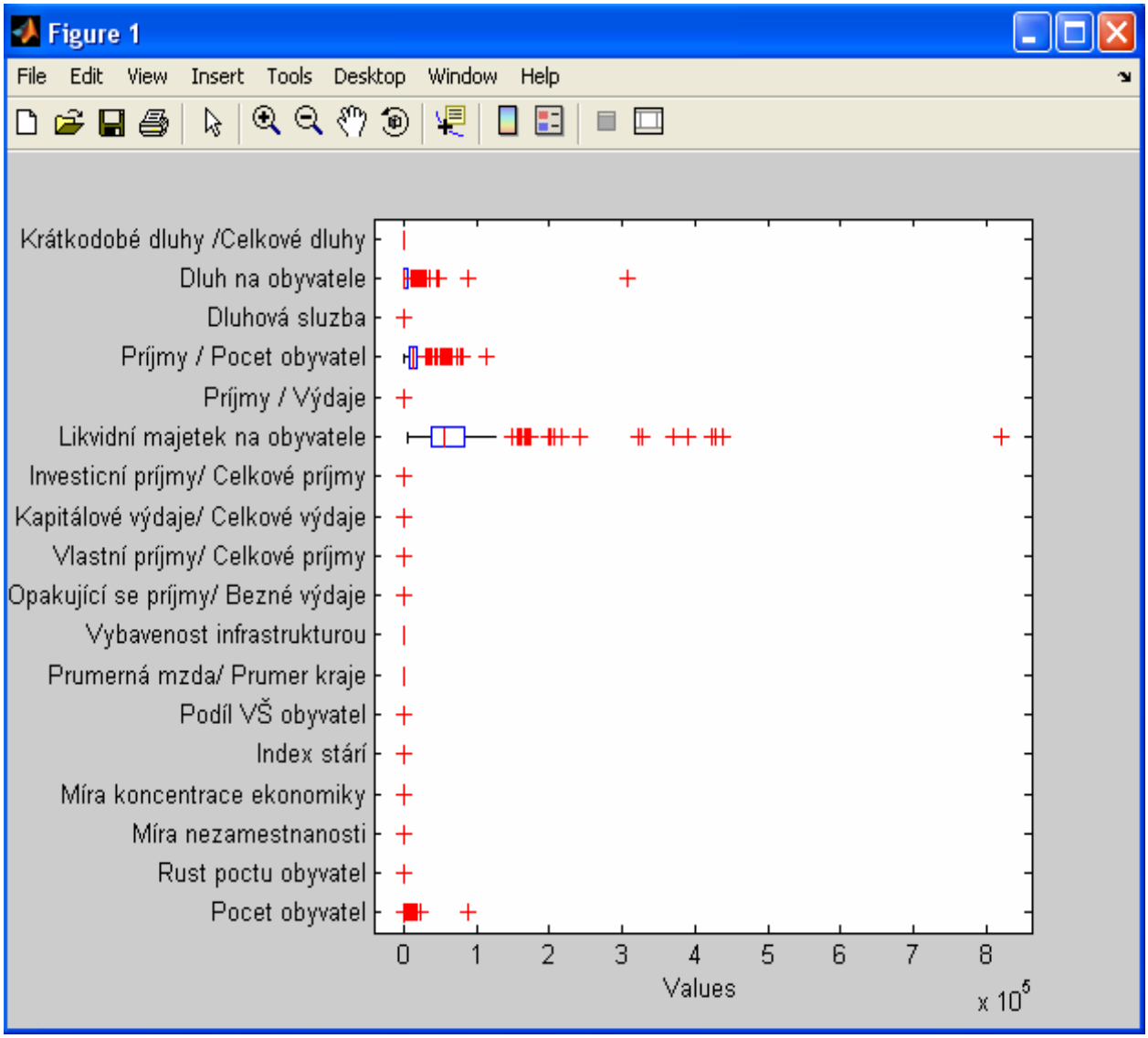
PŘÍLOHA B – Matice korelačních vztahů mezi všemi znaky

Tabulka 5 - Matice korelačních vztahů

Variable	Correlations (data) Marked correlations are significant at p < ,05000 N=452 (Casewise deletion of missing data)																	
	Pocet ob	Růst počtu ob	Míra nezaměst	Míra koncentrace ekonomiky	Index stáří	Podíl VŠ ob	Průměrná mzda/ Průměr kraje	Vybavenost infrastruk	Opakující se příjmy/ Běžné výdaje	Vlastní příjmy/ Celkové příjmy	Kapitálové výdaje/ Celkové výdaje	Investiční příjmy/ Celkové příjmy	Likvidní majetek na ob	Příjmy / Výdaje	Příjmy / Počet obyvatel	Dluhová služba	Dluh na ob	Krátkodobé dluhy /Celkové dluhy
Pocet obyvatel	**1,00**	0,03	-0,08	-0,10	-0,05	0,34	0,04	0,28	-0,31	-0,11	-0,08	0,36	0,05	-0,07	0,38	-0,03	0,01	-0,06
Růst počtu obyvatel	0,03	**1,00**	-0,21	-0,16	-0,49	0,15	0,26	0,20	-0,02	-0,15	0,08	0,13	-0,00	0,06	0,09	0,11	0,05	0,08
Míra nezaměstnanosti	-0,08	-0,21	**1,00**	0,20	0,14	-0,35	-0,25	-0,24	-0,02	0,30	0,00	-0,04	0,31	-0,07	0,06	-0,01	0,03	-0,01
Míra koncentrace ekonom	-0,10	-0,16	0,20	**1,00**	0,11	-0,27	-0,21	-0,17	-0,01	0,03	0,01	-0,02	-0,08	-0,05	-0,10	0,06	-0,03	0,02
Index stáří	-0,05	-0,49	0,14	0,11	**1,00**	-0,07	0,04	-0,25	0,03	0,07	-0,15	-0,17	-0,03	0,02	-0,16	-0,05	-0,05	0,01
Podíl VŠ obyvatel	0,34	0,15	-0,35	-0,27	-0,07	**1,00**	0,09	0,33	-0,20	-0,16	-0,04	0,21	-0,05	-0,04	0,26	-0,05	0,02	0,00
Průměrná mzda/ Průměr l	0,04	0,26	-0,25	-0,21	0,04	0,09	**1,00**	0,18	0,09	-0,07	-0,04	-0,02	0,00	0,19	0,00	0,02	0,06	0,11
Vybavenost infrastrukturo	0,28	0,20	-0,24	-0,17	-0,25	0,33	0,18	**1,00**	-0,27	-0,25	0,08	0,42	0,01	-0,07	0,40	0,11	0,05	-0,15
Opakující se příjmy/ Běžn	-0,31	-0,02	-0,02	-0,01	0,03	-0,20	0,09	-0,27	**1,00**	0,09	0,42	-0,42	-0,12	0,36	-0,47	0,02	-0,00	0,06
Vlastní příjmy/ Celkové př	-0,11	-0,15	0,30	0,03	0,07	-0,16	-0,07	-0,25	0,09	**1,00**	-0,00	-0,17	0,34	-0,03	-0,07	-0,04	0,05	0,07
Kapitálové výdaje/ Celkov	-0,08	0,08	0,00	0,01	-0,15	-0,04	-0,04	0,08	0,42	-0,00	**1,00**	0,38	0,09	-0,15	0,12	0,11	0,06	-0,08
Investiční příjmy/ Celkové	0,36	0,13	-0,04	-0,02	-0,17	0,21	-0,02	0,42	-0,42	-0,17	0,38	**1,00**	0,16	-0,02	0,72	0,19	0,09	-0,17
Likvidní majetek na obyva	0,05	-0,00	0,31	-0,08	-0,03	-0,05	0,00	0,01	-0,12	0,34	0,09	0,16	**1,00**	-0,08	0,48	0,06	0,38	0,02
Příjmy / Výdaje	-0,07	0,06	-0,07	-0,05	0,02	-0,04	0,19	-0,07	0,36	-0,03	-0,15	-0,02	-0,08	**1,00**	-0,04	0,12	0,02	0,02
Příjmy / Počet obyvatel	0,38	0,09	0,06	-0,10	-0,16	0,26	0,00	0,40	-0,47	-0,07	0,12	0,72	0,48	-0,04	**1,00**	0,07	0,17	-0,12
Dluhová služba	-0,03	0,11	-0,01	0,06	-0,05	-0,05	0,02	0,11	0,02	-0,04	0,11	0,19	0,06	0,12	0,07	**1,00**	0,15	-0,40
Dluh na obyvatele	0,01	0,05	0,03	-0,03	-0,05	0,02	0,06	0,05	-0,00	0,05	0,06	0,09	0,38	0,02	0,17	0,15	**1,00**	-0,17
Krátkodobé dluhy /Celkov	-0,06	0,08	-0,01	0,02	0,01	0,00	0,11	-0,15	0,06	0,07	-0,08	-0,17	0,02	0,02	-0,12	-0,40	-0,17	**1,00**

Pozn.: Matice korelačních koeficientů r všech kategorií (znaků) zdrojové matice dat. Korelačním koeficientem je vypočtena hladina významnosti p, je-li $p < 0.05$, je r statisticky významný, což je v tabulce vyznačeno červeně.

PŘÍLOHA C – Diagram Box plot ukazující na rozptyl v datech



Obrázek 11 - Diagram Box plot ukazující na rozptyl v datech

PŘÍLOHA D – Charakteristiky jednotlivých komponent vzhledem k celkovému rozptylu

Tabulka 6 - Charakteristiky komponent vzhledem k celkovému rozptylu

	Charakteristiky jednotlivých komponent vůči celkovému rozptylu			
Value number	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	3,250466	18,05814	3,25047	18,0581
2	2,186827	12,14904	5,43729	30,2072
3	1,775603	9,86446	7,21290	40,0716
4	1,501522	8,34179	8,71442	48,4134
5	1,325092	7,36162	10,03951	55,7751
6	1,128484	6,26936	11,16799	62,0444
7	1,032455	5,73586	12,20045	67,7803
8	0,886076	4,92264	13,08653	72,7029
9	0,814928	4,52738	13,90145	77,2303
10	0,727306	4,04059	14,62876	81,2709
11	0,652359	3,62422	15,28112	84,8951
12	0,614273	3,41263	15,89539	88,3077
13	0,544761	3,02645	16,44015	91,3342
14	0,500697	2,78165	16,94085	94,1158
15	0,434874	2,41597	17,37572	96,5318
16	0,360078	2,00043	17,73580	98,5322
17	0,182687	1,01493	17,91849	99,5472
18	0,081512	0,45284	18,00000	100,0000

PŘÍLOHA E – Seznam použitých příkazů v Matlabu

```
>> load matlab

>> whos

>> categories

>> boxplot(ratings,'orientation','horizontal','labels',categories)

>> stdr = std(ratings);

sr = ratings./repmat(stdr,452,1);


>> [coefs,scores,variances,t2] = princomp(sr);

>> c7 = coefs(:,1:7)

>> plot(scores(:,1),scores(:,2),'+')

xlabel('1st Principal Component');

ylabel('2nd Principal Component');


>> variances

>> percent_explained = 100*variances/sum(variances)

>> pareto(percent_explained)

xlabel('Principal Component')

ylabel('Variance Explained (%)')


>> biplot(coefs(:,1:2), 'scores',scores(:,1:2),...

'varlabels',categories);

axis([-.46 1 -.51 .51]);
```