

Technická univerzita v Liberci

Hospodářská fakulta

DIPLOMOVÁ PRÁCE

1999

Dana Nejedlová

Technická univerzita v Liberci

Hospodářská fakulta

Obor

Podniková ekonomika

Analýza ekonomických dat pomocí neuronové sítě

Economic Data Analysis Using Neural Network

DP - PE - KIN - 901

Dana Nejedlová

Vedoucí práce:

Ing. Jan Skrbek, Dr., Katedra informatiky

Konzultant:

Ing. Jaromír Kukal, Katedra informatiky

Ing. Petra Rydvalová, Katedra podnikové ekonomiky

Počet stran: 84

Počet příloh: 15

Datum odevzdání

28. 5. 1999

TECHNICKÁ UNIVERZITA V LIBERCI

Hospodářská fakulta

Katedra informatiky

Školní rok 1998/99

ZADÁNÍ DIPLOMOVÉ PRÁCE

pro

Bc. Dana Nejedlová

obor č.: 62 – 68 – 8

Podniková ekonomika

Vedoucí katedry Vám ve smyslu zákona č. 172/1990 Sb. o vysokých školách a navazujících předpisech určuje tuto diplomovou práci:

Název tématu: **Analýza ekonomických dat pomocí neuronové sítě**

Zásady pro vypracování:

Literární a informační průzkum v oblasti neuronových sítí

Význam faktorové analýzy (PCA) v hodnocení ekonomických subjektů

Kategorizace ekonomických subjektů pomocí neuronové sítě

Aplikace zjištěných poznatků na ekonomický výzkum

V 108/99 H

KIN/POE
85. 18. 9. 1999

DIPLOMOVÁ PRÁCE

TÉMA: ANALÝZA EKONOMICKÝCH DAT POMOCÍ NEURONOVÉ SÍTĚ

ABSTRAKT:

Práce se zabývá využitím neuronové sítě pro rozpoznání a kategorizaci různých objektů. Obsahuje rovněž základní průzkum literatury o strukturách, vlastnostech a technikách učení neuronových sítí. Práce ukazuje, že existuje úzký vztah mezi chováním samoorganizujících se neuronových sítí a statistickou metodou analýzy hlavních komponent (PCA), neboli faktorové analýzy. Neparametrická robustní technika Jackknife je použita ke zdokonalení PCA. Hlavním výsledkem práce je použití konfidenčních elipsoidů pro pochopení podobností mezi ekonomickými objekty. Celá metoda analýzy je testována na třech úlohách z oblasti ekonomie.

THEME: ECONOMIC DATA ANALYSIS USING NEURAL NETWORK

ABSTRACT:

The thesis deals with the use of neural network for recognition and categorization of various objects. The work also contains basic research of literature which deals with neural network structures, properties, and learning techniques. The thesis shows that there is a close relationship between the behavior of self-organized neural networks and the statistical method of principal components analysis (PCA), also called factor analysis. Non-parametric and robust technique of Jackknife is used to improve PCA. The main result of the thesis is the use of confidence ellipsoids for understanding the similarities between economic objects. The general method of analysis is tested on three economic tasks.

Obsah diplomové práce

	Strana
Úvod	8
1. Literární a informační průzkum v oblasti neuronových sítí	10
1.1 Lineární klasifikátory	10
1.2 Vícevrstvá neuronová síť	12
1.3 Kohonenovy samoorganizující se sítě	15
1.4 Hopfieldovy sítě	17
2. Význam faktorové analýzy (PCA) v hodnocení ekonomických objektů	19
3. Kategorizace ekonomických objektů pomocí neuronové sítě	24
3.1 Statistická analýza vícerozměrných dat	25
3.2 Analýza hlavních komponent (PCA)	32
3.3 Zobecněný Hebbův algoritmus (GHA)	43
3.4 Jacobiho metoda nalezení vlastních čísel a vlastních vektorů symetrické matice	58
3.5 Jackknife	67
3.6 Kombinace metody Jackknife s metodou PCA	68
3.7 Konstrukce konfidenčních elipsoidů	69
4. Aplikace zjištěných poznatků na ekonomický výzkum	71
4.1 Výstupy analýzy	72
4.1.1 Sloupcové grafy 1. a 2. pohledu	72
4.1.2 Sloupcové grafy vlastních čísel	73
4.1.3 Rozptylové grafy	73
4.1.4 Seznam minimálních kvantilů a mapa podobnosti objektů	74
4.2 Ekonomické úlohy	74
4.2.1 Profese	74
4.2.2 Měny	77
4.2.3 Podniky	79
Závěr	82
Přílohy	

Seznam použitých zkratek

GHA	Generalized Hebbian Algorithm, tj. zobecněný Hebbův algoritmus
PCA	Principal Components Analysis, tj. analýza hlavních komponent

Úvod

Podnik jako hospodářská organizace je uměle vytvářený systém. Probíhají v něm hospodářské procesy, které se realizují prostřednictvím trhu. Trh zprostředkovává oběh zboží a služeb mezi jednotlivými ekonomickými subjekty (podniky a domácnostmi).

Termín „ekonomické subjekty“ je třeba nezaměňovat s termínem „ekonomické objekty“, který používám v dalším textu této práce. Ekonomický objekt může znamenat podnik, zaměstnance, domácnost, pojištěnce, měnu, výrobek a další širší kategorie patřící do ekonomie.

Cílem podniku je uspokojování potřeb zákazníků, se kterými se setkává na trhu, a zároveň dosahování zisku nebo zvětšování bohatství. Konkurenční prostředí, které se na trhu vytváří, stále více ztěžuje praktické naplňování tohoto cíle. Ukazuje se, že mezi úspěšnější patří ty podniky, které dobře zvládají práci s informacemi.

Při všech hospodářských procesech vznikají ekonomická data. Ta jsou následně zpracovávána s cílem získat podklady pro správné rozhodování lidí o dalším postupu při realizaci podnikových cílů. Moderní výpočetní technika nabízí mnoho způsobů zpracování ekonomických dat a k nejnovějším trendům v této oblasti patří využívání neuronových sítí.

Je obecně známo, že některé úlohy lépe řeší automaty, jiné lidé. K těm prvním patří například sečíst sloupec čísel. Mezi ty druhé lze zařadit například odhad rizika, které plyne z uzavření hospodářského kontraktu. Proč je automat (počítač) lepší v prvním případě a člověk ve druhém? Odpověď tkví ve způsobu zpracování informací. Úlohy prvního typu je možno řešit sériově, to znamená provádět jednu operaci za druhou v jedné výkonné jednotce, kterou je v počítači procesor. Úlohy druhého typu vyžadují takzvané paralelní zpracování. To znamená, že je třeba zpracovat mnoho různých druhů informací, které se vzájemně ovlivňují. Znalosti potřebné k řešení těchto úloh procházejí z mnoha různých

oblastí, kde každá přispívá ke konečnému výsledku svým zvláštním způsobem. Lidský mozek má potřebnou strukturu pro řešení paralelních úloh, počítač s jedním procesorem ne.

Existují úlohy, které je nutno zpracovávat paralelně a přitom je postup řešení snadno algoritmizovatelný. Pro řešení takových úloh byly vyvinuty neuronové sítě. Neuronovou síť můžeme chápat jako automat s mnoha vstupy, který připouští paralelní zpracování informací v reálném čase. Navíc má hierarchickou strukturu, čímž se blíží uspořádání nervové soustavy inteligentních bytostí. V aplikacích se setkáváme jak s elektronickým řešením neuropočítače jako paralelní jednotky pro zpracování dat, tak se simulací neuronových sítí na jednoprocesorových počítačích běžného typu.

Neuronové sítě nejsou schopny řešit tak široké spektrum problémů jako inteligentní člověk a ani do budoucna nehrozí jejich dominance nad lidmi. Na druhé straně, specializovaná neuronová síť předčí již nyní člověka při řešení konkrétní třídy úloh. Příkladem takové úlohy může být kategorizace dodavatelů nebo jejich výběr podle mnohorozměrných dat kvantitativní povahy. Pokud jsou hodnocených objektů stovky, stává se úloha pro člověka nepřehlednou, vyčerpávající, a přitom má rutinní charakter. Řešení úloh tohoto i jiných typů pomocí různých druhů neuronových sítí naznačuji v 1. kapitole této práce.

Neuronové sítě však pouze neodstraňují nudnou práci, jako to dělalo mnoho technických vymožeností vyvinutých před nimi. Současné tržní prostředí se rychle vyvíjí a mnoho tradičních metod analýzy informací tím rychle zastarává. Neuronové sítě jako učící se systémy dokáží pružně reagovat na měnící se podmínky a proto bude jejich použití i nadále získávat v ekonomické praxi na významu.

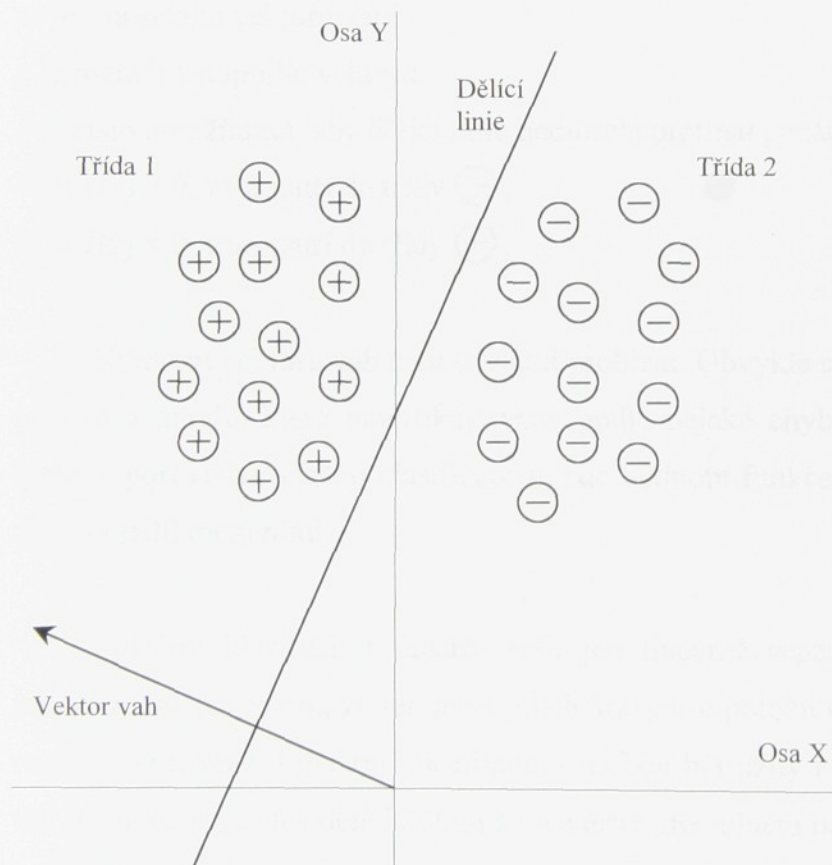
Další kapitoly mé práce se zabývají neuronovou sítí specializovanou na kategorizaci ekonomických objektů. Cílem práce bylo dosáhnout výsledků srovnatelných s vícekritériálním (paralelním) úsudkem člověka – specialisty v oblasti ekonomie.

1. Literární a informační průzkum v oblasti neuronových sítí

Zdrojem informací pro tuto kapitulu je literatura [3] shrnující všechny hlavní přístupy k neuronovým sítím a jejich významné výsledky spolu s mnoha odkazy na další literaturu. Z této knihy zde přiblížím čtyři techniky neuronového zpracování dat.

1.1 Lineární klasifikátory

Lineární klasifikace je jednoduchá technika rozpoznání vzorů a jejich zatřídění do kategorií. Problém objasňuje obrázek 1.1.1.



Obrázek 1.1.1
Lineární klasifikátor

Pramen: [3]

Vstupními vzory neboli daty jsou vektory mířící od počátku souřadnic do polohy vzorů $\oplus$ nebo $\ominus$. Úkolem je najít takový směr vektoru vah $\mathbf{w}$, aby dělicí linie, která je na něj kolmá, rozdělovala dvě kategorie vzorů. Na rozdíl od úlohy, kterou řeším v této diplomové práci, musí být o vstupních vzorech předem známo, do které kategorie patří.

Rovnice, která rozděluje vzory do kategorií, je založena na skalárním součinu:

$$f(x) = \sum_{i=1}^p w_i x_i - \theta = |\mathbf{w}| \cdot |\mathbf{x}| \cdot \cos \varphi - \theta, \quad (1.1.1)$$

kde

x_i je i -tá složka vstupního vektoru,

w_i je i -tá složka vektoru vah,

p je rozměr vstupního vektoru,

θ je číslo umožňující, aby dělicí linie nemusela protínat počátek souřadnic.

Je-li $f(x) > 0$, vzor patří do třídy $\oplus$,

Je-li $f(x) < 0$, vzor patří do třídy $\ominus$.

Nalezení vektoru vah není triviální problém. Obvykle se hledá iterativními metodami pokusu a omylu, které modifikují váhy podle nějaké chybové funkce. Chybová funkce typicky porovnává výstup klasifikátoru, zde hodnotu funkce $f(x)$, se správnou odpovědí a udává rozdíl mezi nimi.

Lineární klasifikátor dokáže řešit jen lineárně separovatelné problémy. To jsou taková rozložení vzorů, že jde mezi jejich kategorie položit dělicí přímku, obecně $(p - 1)$ -rozměrnou rovinu. Lineární klasifikátory mohou být užity i k rozlišení více než dvou tříd tím, že se použije více dělicích linií a více testů pro splnění podmínek pro každou třídu.

Další informace podává například literatura [9].

1.2 Vícevrstvá neuronová síť

Vícevrstvá neuronová síť (*multilayer perceptron*) je na rozdíl od lineárního klasifikátoru použitelná pro řešení lineárně neseparabilních problémů. Je postavena na tom, jak funguje mozek. Jejímí komponentami jsou neurony tvořící několik vzájemně propojených vrstev.

Jednoduchý model neuronu je funkce, která sečte své vstupy a produkuje výstup, pokud je součet vstupů vyšší než určitá hodnota známá jako *prahová hodnota* (*threshold value*). V mozku platí, že vstupní podněty do neuronu přicházejí po množině synapsí, které jsou různě průchodné, to jest mají různé váhy. Výstupní synapse neuronu mají také různé váhy a jsou napojeny na další neurony. Pokud synapsí prochází signál často, její průchodnost (váha) se zvyšuje. To je varianta pravidla učení, které v roce 1949 navrhl Donald Hebb jako výsledek svých studií přírodních neuronových systémů.

Matematicky může být výstup neuronu zapsán pomocí skalárního součinu jako funkce

$$y = f_h \left[\sum_{i=1}^p w_i x_i - \theta \right], \quad (1.2.1)$$

kde f_h je takzvaná *schodovitá funkce* (*thresholding/step/Heaviside function*) vracející hodnotu

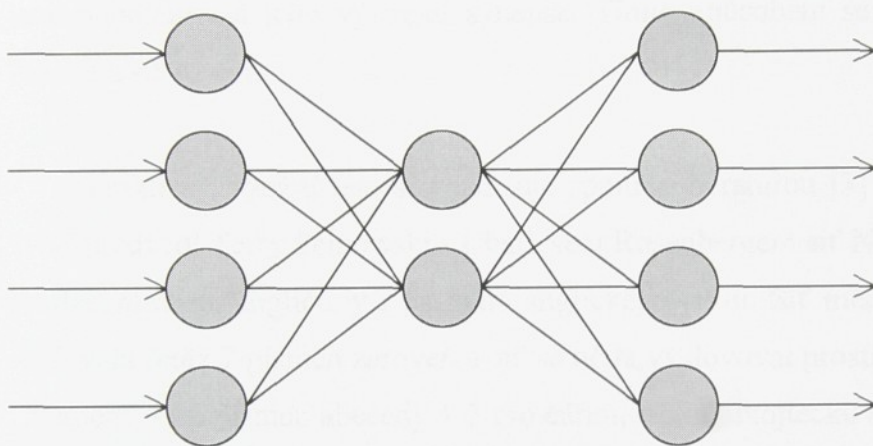
$$\begin{aligned} f_h(x) &= 1, & \text{když } x > 0, \\ f_h(x) &= 0, & \text{když } x \leq 0. \end{aligned} \quad (1.2.2)$$

Argument funkce f_h je v podstatě funkce lineárního klasifikátoru z předchozí kapitoly.

Výzkumy chování takových neuronů v modelech vícevrstevných sítí vedly k závěru, že síť bude schopna se učit (upravovat své váhy žádoucím způsobem), pokud bude schodovitá funkce nahrazena *esovitou/sigmoidální/logistickou funkcí*. Ta nevrací pouze hodnoty 0 nebo 1 ale i hodnoty mezi nimi. To znamená, že výstup neuronu ponese více informace

o jeho vstupech. Navíc je tento model bližší chování biologického neuronu. Frank Rosenblatt v roce 1962 nazval modelové neurony spojené do jednoduché sítě „perceptrony“.

Sít' na obrázku 1.2.1 má vrstvu vstupní, výstupní a mezi nimi takzvanou skrytou vrstvu. Jednotky ve vstupní vrstvě nemají prahovou funkci, ale pouze distribuují vstupní hodnoty následující vrstvě.



Obrázek 1.2.1
Třívrstvá neuronová síť

Pramen: [3]

Pravidlo učení pro vícevrstvé perceptrony se nazývá *zobecněné delta pravidlo* (*generalized delta rule*) nebo *pravidlo zpětného šíření* (*backpropagation rule*). Bylo navrženo v roce 1986 Rumelhartem, McClellandem a Williamsem v časopise [10]. Později bylo zjištěno, že Parker publikoval podobné výsledky v roce 1982 a Werbos již v roce 1974.

Při učení síť přijímá na vstupní vrstvě vzory (data) a výstupní vrstva produkuje výstup. Zpočátku jsou váhy na všech synapsích mezi neurony nastaveny náhodně, takže i výstupy sítě jsou náhodné. Potřebujeme definovat chybovou funkci reprezentující rozdíl mezi výstupem sítě a jeho žádoucí hodnotou. Protože musíme předem znát správný výstup a sdělovat jej síti, nazývá se tento typ učení *učení s dohledem* (*supervised learning*). Učení

spočívá v upravování vah takovým způsobem, aby se výstup blížil správné hodnotě, což znamená, že hodnota chybové funkce se bude snižovat. *Zobecněné delta pravidlo* vypočte hodnotu chybové funkce pro každý vstupní vzor a potom jsou podle ní upravovány váhy v síti tak, aby při příštím čtení stejného vzoru byla hodnota chybové funkce nižší. Jako první se musí upravit váhy na synapsích vedoucích do výstupních neuronů. S úpravami se postupuje směrem k synapsím vedoucím od vstupních neuronů. Oprava se tak zpětně šíří od konce sítě k jejímu počátku, odtud název *pravidlo zpětného šíření*. Z matematického hlediska musí být váhy určitého neuronu upraveny přímo úměrně chybě v neuronech, které jsou napojeny na jeho výstupní synapse. Tímto způsobem se snižuje hodnota chybové funkce a síť se učí.

Praktické využití této sítě ukazuje spolu s literaturou [3] také literatura [11]: Roku 1987 předvedl Terry Sejnowski s Charlesem Rosenbergem síť NETalk. Jejím úkolem byla tvorba mluvené angličtiny z psaného anglického textu. Síť měla 3 vrstvy. Vstupní vrstva sledovala řetěz 7 písmen zároveň a síť se učila vyslovovat prostřední písmeno. Každé z 29 „písmen“ – 26 písmen abecedy + 3 pro čárku, tečku, dvojtečku a hranice slov – mělo svoji jednotku. Z toho plyne nutnost $29 \times 7 = 203$ jednotek vstupní vrstvy. Výstupní vrstva má jednotku pro každý z 21 „artikulačních znaků“ požadovaných fonémů (základních zvuků v jazyce, ze kterých se tvoří všechna slova) plus 5 jednotek nutných ke zpracování hranic jednotlivých slabik a přízvuků, tedy celkem 26 jednotek výstupních. Skrytá střední vrstva původní sítě měla 28 jednotek, výkon pozdější verze, se 120 jednotkami, byl o něco lepší. Stroj musel přizpůsobovat účinnost asi 20 000 synapsí. Účinnost mohla být pozitivní nebo negativní. Autoři za tím účelem nepostavili skutečnou vysoce paralelní síť, ale simulovali síť na tehdy středně rychlé počítači VAX 11/780 FPA. Sejnowski s Rosenbergem rovněž ukázali, že síť je značně odolná vůči náhodnému poškození svých spojů a že její výkon za těchto okolností „klesal s půvabem“ (*graceful degradation in performance* – síť se nezhroutila naráz). Učinili rovněž pokus se skupinami 11 písmen (namísto 7). Výkon se tím značně zlepšil. Přidání další skryté vrstvy výkon nezlepšilo, zato síť lépe zobecňovala (vyslovovala dosud pro ni neznámá slova).

Vícevrstvé neuronové sítě se používají ve finančních aplikacích například jako nástroj pro předvídání vývoje trhů a pro posuzování rizikovosti půjček a pojistek. Výhodou neuronových sítí je, že se dokáží rychle učit z mnoha tisíců případů (vzorů) z podnikových záznamů a extrahovat z nich podstatné faktory naznačující, co se pravděpodobně stane.

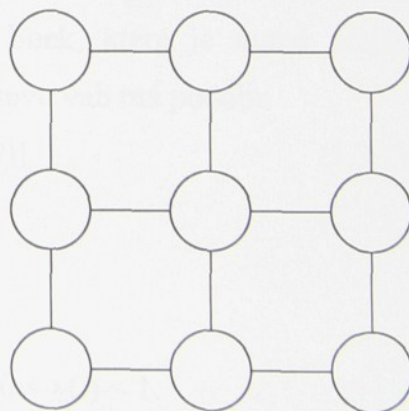
1.3 Kohonenovy samoorganizující se sítě

Kohonenovy samoorganizující se sítě nebo také mapy patří mezi metody *učení bez dohledu* (*unsupervised learning*). To znamená, že síť si vytvoří své vlastní klasifikace (třídění) trénovacích dat. Aby to bylo možné, musí síť splňovat dva základní předpoklady. Za prvé, příslušnost k určité třídě vzorů je jasně definována jako vstupní vzory, které sdílí společné základní rysy. Druhým předpokladem je, že síť bude schopna identifikovat společné rysy v celém rozmezí vstupních vzorů.

Předobrazem této metody v přírodě je fakt, že buňky určitých oblastí mozku vytvářejí struktury reprezentující interní mapování reakcí smyslových orgánů.

Kohonenova síť provádí redukci vícerozměrných datových vektorů, aby mohly být reprezentovány v prostoru o nižší dimenzi, typicky v rovině. Z tohoto důvodu je tato metoda nejbližší způsobu analýzy dat popisovaném v této práci.

Jednovrstvá dvourozměrná Kohonenova síť je tvořena mřížkou s neurony, viz obrázek 1.3.1. Každý z těchto neuronů je přímo napojen na p vstupních uzlů, kterými jsou síti prezentovány p -rozměrné datové vektory neboli vzory (trénovací data). Neurony v mřížce komunikují mezi sebou přes spojovací linky na obrázku 1.3.1 tím způsobem, že každý neuron nejvíce ovlivňuje nejbližší neurony ve svém okolí. Po procesu učení je tato mřížka výstupní vrstvou sítě.



Obrázek 1.3.1 Pramen: [3]
Síť neuronů Kohonenovy mapy

Učící algoritmus organizuje neurony v mřížce do mapy základních rysů dat. Na počátku jsou na všech spojnicích mezi neurony v mřížce a vstupními uzly náhodné váhy. Každý neuron tak získá odlišný p -rozměrný vektor vah. Během učení je síti ukázána množina trénovacích vzorů, která je reprezentativní podmnožinou celé množiny dat. Každý trénovací vzor je porovnán s vektory vah u všech neuronů. Zvítězí neuron, který má svůj vektor vah trénovacímu vzoru nejpodobnější. Tento neuron upraví své váhy tak, aby byly tomuto vzoru ještě bližší. Podobnost se určuje jako euklidovská vzdálenost rovná

$$\sum_{i=1}^p [x_i(t) - w_{ij}(t)]^2, \quad (1.3.1)$$

kde

$x_i(t)$ je i -tá složka vstupního vektoru v čase t ,

$w_{ij}(t)$ je i -tá složka vektoru vah j -tého neuronu mřížky v čase t ,

p je rozměr vstupního vektoru,

t je počet průchodů trénovací množinou.

Vítězný neuron má tuto vzdálenost nejnižší. Po úpravě svých vah je tento neuron na daný vzor citlivý. Neurony v jeho okolí také upravují své váhy o stejnou hodnotu jako vítěz. To má za následek, že kolem vítězného neuronu se vytvoří oblast neuronů s vahami o něco více odlišnými od daného vzoru. Síť se tak snaží vytvořit oblasti, které budou odpovídat na určitý rozsah hodnot trénovacích dat. To umožní síti správně kategorizovat i vzory, které jsou blízké určitému vzoru z trénovacích dat přesto, že nebyly v trénovací dávce. To demonstruje zobecňovací (generalizační) vlastnosti této sítě. Za neurony v *okolí* určitého neuronu se považují neurony s nízkým počtem linek, které je nutné přejít, abychom se od daného neuronu k nim dostali. Vzorec pro úpravu vah má podobu

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)[x_i(t) - w_{ij}(t)], \quad (1.3.2)$$

kde

$i = 1, 2, \dots, p$,

$j = 1, 2, \dots$, počet neuronů v okolí,

$\eta(t)$ je parametr rychlosti učení klesající s časem t takový, že $0 < \eta(t) < 1$.

S poklesem velikosti parametru η se zpomaluje adaptace vah a tedy i citlivé okolí vítězného neuronu. Tím se lokalizuje oblast maximální aktivity sítě.

Vektory vah se v každém kroku normalizují, protože vstupní vektory neklasifikujeme podle velikosti ale podle orientace. Vítězný neuron by mohl být určen také jako neuron s maximální hodnotou skalárního součinu svých vah se vzorem podle vzorce

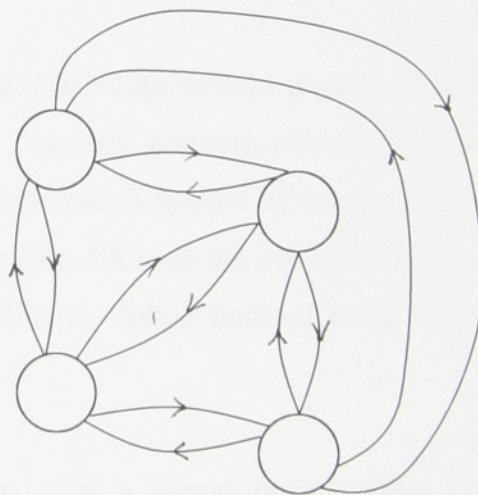
$$\sum_{i=1}^p x_i(t) \cdot w_{ij}(t). \quad (1.3.3)$$

Hlubší studium problému umožní literatura [12].

1.4 Hopfieldovy sítě

John Hopfield studující autoasociativní síť, která je v něčem podobná perceptronům, ale má i některé významné odlišnosti, významně obohatil oblast neuronových sítí na počátku 80. let.

Hopfieldova síť se skládá z uzlů, které jsou propojeny každý s každým. Je to tedy *plně propojená (fully-connected)* síť a ukazuje ji obrázek 1.4.1. Tato síť má váhy na spojnicích mezi uzly stejné v obou směrech, je proto tzv. *symetricky vyvážená (symmetrically-weighted)*.



Obrázek 1.4.1
Hopfieldova síť

Pramen: [3]

Každý uzel má stejně jako perceptron prahovou hodnotu a schodovitou funkci. Uzly počítají vážený součet svých vstupů minus prahová hodnota. Výsledek je argumentem schodovité funkce vracející hodnotu (0, 1). V tomto případě jsou však tyto výstupy transformovány na (−1, +1), kvůli jednoduššímu matematickému zpracování.

Hlavním rysem Hopfieldovy sítě je, že nemá žádné vstupní a výstupní synapse. Všechny uzly jsou si rovnocenné. Vstupy do sítě jsou dány všem uzlům najednou a mohou mít buď hodnotu $+1$ nebo -1 .

Ve fázi učení se síti předloží trénovací dávka vzorů v podobě p -rozměrných vektorů, kde p je rovno počtu uzlů sítě. Z této dávky síť vypočte váhy synapsí mezi všemi svými uzly. Váha mezi uzlem i a j je rovna

$$w_{ij} = w_{ji} = \sum_{s=1}^n x_{is} x_{js} \quad \text{pro } i \neq j, \quad (1.4.1)$$

kde n je počet vzorů v trénovací množině. Pro $i = j$ je w_{ii} nebo w_{jj} rovno nule, protože žádný uzel nemá synapsi vedoucí od něho přímo k němu. Hodnotou složek vstupních vektorů x_i, x_j je $(-1, +1)$. Výsledkem jejich součinu je také -1 nebo $+1$. Váha w_{ij} je součtem posloupnosti n hodnot -1 nebo $+1$. Její hodnota je tedy maximálně $\pm n$. Po fázi učení se váhy již nemění.

Ve druhé fázi do uzlů vložíme neznámý vzor μ . Následuje iterační proces, ve kterém složky vektoru neznámého vzoru v každém kroku dostanou hodnotu schodovité funkce s argumentem rovným skalárnímu součinu vah a předchozích hodnot složek μ . Změněný vektor μ opět vstupuje do iterace, tj. je znovu sítí přečten. Síť tedy čte své vlastní výstupy, počítá z nich nové vstupy a tento proces trvá tak dlouho, dokud dochází mezi starým a novým výstupem ke změnám.

Praktickým využitím Hopfieldovy sítě je rozpoznání a oprava poškozeného vzoru, který se síť předtím naučila podle vzoru nepoškozeného. Může to být využito například při rozpoznání písma. Jednu síť je možno naučit větší počet vzorů, které pak jednotlivě dokáže obnovovat z poškozených vstupů. Při velkém počtu vzorů uchovávaných v síti však roste pravděpodobnost, že vzor bude obnoven nesprávně.

Hopfieldova originální práce je publikována v časopise [13].

2. Význam faktorové analýzy (PCA) v hodnocení ekonomických objektů

Ekonomické objekty jsou tradičně hodnoceny ve smyslu určování jejich pořadí od nejlepšího po nejhorší. Pokud jsou popsány množinou parametrů, lze tyto parametry dosadit do nějakého vzorce, jako je například *Altmanův Z-score model* pro určování finančního zdraví podniků [8], a získat tak hodnotu, která objekt (v tomto případě ekonomický subjekt – podnik) zařadí do nějaké škály. Jinou možností, jak zpracovat vícerozměrná data o ekonomických objektech, je jejich zařazení do kategorií, a to je i cílem této diplomové práce.

Výhodou metody, kterou v této práci popisuji je jednoduchost formy a libovolnost obsahu vstupních dat. Vstupními údaji pro zpracování je tabulka čísel. Jejímí řádky jsou hodnocené ekonomické objekty, které budu dále nazývat *vzory*. Jejímí sloupky jsou *vlastnosti* vzorů. Není podmínkou, aby vzorem byl právě ekonomický objekt. Tato metoda je použitelná na vše, co se dá popsat čísly, a to čísla, která vyjadřují kvantitativní typ informace.

Soubor n vzorů, z nichž každý je vyjádřen pomocí p stejných druhů vlastností, je z matematického hlediska množinou n p -rozměrných vektorů. Tyto vektory si můžeme představit jako p -rozměrný shluk jejich koncových bodů za předpokladu, že všech n vektorů vychází z počátku souřadného systému p -rozměrného prostoru. Takováto představa je pouze abstraktní, protože dokážeme vnímat nanejvýš trojrozměrné předměty. Musíme tedy použít nějakou techniku, která dané vzory reprezentované vektory v p -rozměrném prostoru promítne do prostoru o nižší dimenzi, obvykle do dvourozměrného souřadného systému, jako je to i v této práci.

Takovéto zobrazení umožňuje [5]

- a) identifikaci vektorů, které se jeví jako vybočující,
- b) identifikaci různých struktur v datech, jako jsou shluky, jež ukazují na heterogenitu použitého výběru nebo přítomnost různých dílčích výběrů s odlišným chováním.

Musíme zvolit takový způsob projekce, aby se při něm ztratilo minimum informací obsažených v původních p -rozměrných datových vektorech. V případě, že sloupce vstupní tabulky jsou silně korelovány, což je u ekonomických dat běžné, je vhodným způsobem projekce takzvaná *analýza hlavních komponent* (*principal components analysis*), dále jen PCA, kterou podrobně popisují v kapitole 3.2. Kapitola 3.3 potom ukazuje, že PCA je nezávisle realizovatelná s využitím neuronové sítě.

Korelace znamená určitý stupeň lineární funkční závislosti. Její význam si můžeme představit ve trojrozměrném prostoru. Body na koncích trojrozměrných vektorů, jejichž složky jsou nekorelované, budou tvořit shluk ve tvaru pravidelné koule. Korelovaná data budou oproti tomu vyplňovat elipsoid. To znamená, že koule se v jednom směru protáhne a v jiném směru zploští. Kdybychom si chtěli tento elipsoid promítnout do dvourozměrného prostoru, bude pro nás nejužitečnější pohled, který nám ukáže elipsoid z takového úhlu, že směr, podél kterého je elipsoid nejvíce zploštělý, je kolmý na projekční plochu. Čím více bude elipsoid zploštělý, tím méně informací se takovouto projekcí ztratí. V extrémním případě se elipsoid zploští natolik, že ve směru, ve kterém je nejtenčí, bude jeho tloušťka nulová, a potom se nám výše zvoleným způsobem projekce neztratí žádná informace o datech. Nyní dvourozměrný elipsoid bude prostě položen na rovině, kterou pozorujeme.

Při takovéto filozofii operování s daty považují za užitečné ztotožňovat pojem „informace“ s pojmem „rozptylu“ pozorování. Kdybychom zvolili souřadný systém jehož osy by splývaly s nejdelší, se druhou nejdelší a nejkratší osou elipsoidu, a původní trojrozměrné vektory bychom vyjádřili v tomto souřadném systému, zjistili bychom, že složka vektorů vyjadřující souřadnici na nejdelší ose elipsoidu má nejvyšší rozptyl a složka vektorů patřící souřadnici na nejkratší ose elipsoidu má rozptyl nejnižší. Promítneme-li elipsoid výše popsáním způsobem do dvourozměrného prostoru, budou mít nově vzniklé souřadnice zobrazovaných vektorů rozptyl ve směru jedné osy projekční roviny stejný jako rozptyl souřadnic na nejdelší ose elipsoidu a ve druhém na něj kolmém směru bude rozptyl stejný jako rozptyl souřadnic na druhé nejdelší ose elipsoidu. Při projekci se ztratila informace o rozptylu na nejkratší ose elipsoidu. Přesněji řečeno, ztratila se informace o souřadnici na nejkratší ose elipsoidu u všech zpracovávaných vektorů.

Čím jsou složky zpracovávaných vektorů korelovanější, tím je p -rozměrný shluk, který tvoří, v určitých směrech zploštělejší, a tím je účinnější i použití metody PCA pro jejich projekci do méně než p -rozměrného prostoru.

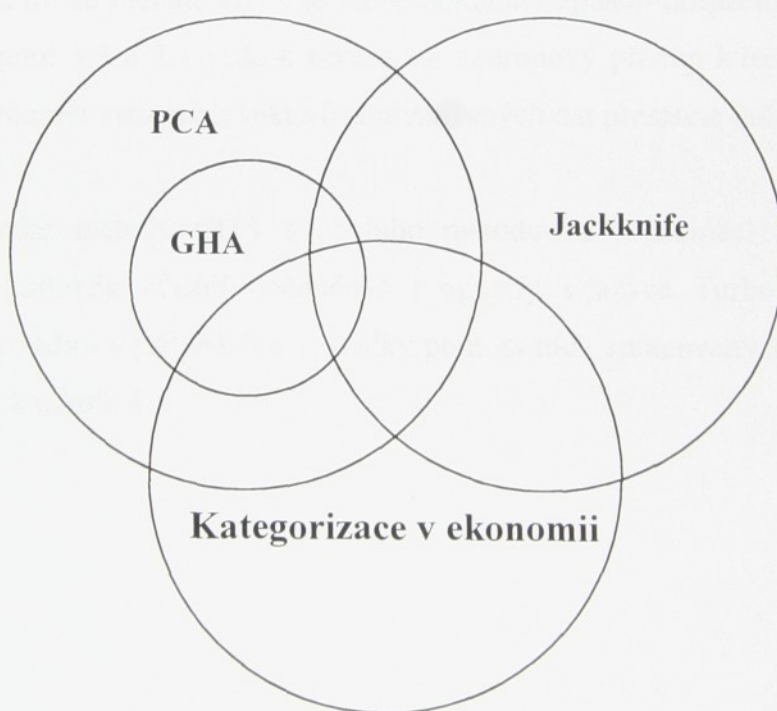
Promítáme-li n p -rozměrných vektorů patřících n vzorům do dvourozměrného prostoru metodou PCA, dostaneme pro každý vzor dvě souřadnice. Podle těchto souřadnic můžeme umístit n vzorů do takzvaného *rozptylového grafu*. Zde můžeme pozorovat různé shluky a odlehlá pozorování. Můžeme očekávat, že vzory, které zde leží blízko sebe, toho mají společného více než vzory ležící na opačných stranách grafu. Vzory tvořící shluky mohou být zařazeny do kategorií.

Chceme-li však činit výroky o míře podobnosti mezi vzory s větší určitostí, musíme data podrobit dalšímu zpracování. Tradiční PCA je bodový odhad polohy vzoru v prostoru o nižší dimenzi než p , například 2. Vzájemné rozdíly mezi polohami jsou zavádějící z důvodu neurčitosti ve vstupních datech a to je důvod, proč je pro vnímání podobnosti vzorů nutno zavést intervalové odhady. Ve dvourozměrném prostoru je vyjadřují pomocí takzvaných *konfidenčních elipsoidů*, což jsou v tomto případě obyčejné dvourozměrné elipsy. Konfidenční elipsoid je termín ze statistiky označující obecně několikarozměrnou oblast spolehlivosti hodnoty určitého parametru. V této práci je oním parametrem poloha vzoru na rozptylovém grafu.

Intervalové odhady jsem získala pomocí metody *Jackknife* popisované v kapitolách 3.5, 3.6 a 3.7. Použití metody Jackknife umožňuje konstruovat nestranné odhady polohy vzoru. Každému vzoru bude náležet jeden konfidenční elipsoid. Vzory potom patří do stejné kategorie, překrývají-li se jejich elipsoidy. To přináší nový pohled na podobnost vzorů, kterými mohou být ekonomické objekty.

Shrnutí

PCA promítá data ve formě p -rozměrných vektorů z p -rozměrného prostoru do m -rozměrného prostoru, kde $m < p$. Při tom minimalizuje ztrátu informací obsažených v původních datech a zachovává rozdíly a poměry rozdílů mezi polohami koncových bodů vektorů, protože osy výsledného m -rozměrného prostoru jsou oproti těm původním z p -rozměrného prostoru pouze pootočené. Nová metoda analýzy ekonomických dat, kterou v této práci popisují, používá tradiční metodu PCA obohacenou o intervalový přístup pomocí metody Jackknife. Problematiku této práce přibližuje obrázek 2.1.



Obrázek 2.1
Schéma řešeného problému

Kategorizaci ekonomických objektů lze provádět mimo jiné tradiční metodou PCA. V 80. letech se podařilo výsledky PCA reprodukovat pomocí neuronového přístupu k řešení takzvaným *zobecněným Hebbovým algoritmem (generalized Hebbian algorithm)*, dále jen GHA, viz kapitola 3.3.

Ekonomické úlohy jsem nejprve zpracovávala metodou GHA, tj. s využitím neuronové sítě, a získané výsledky jsem porovnávala s výsledky alternativní *Jacobiho metody* výpočtu PCA, viz kapitola 3.4. Zjistila jsem, že tyto výsledky se pro mé účely dostatečně shodují tak, že obě metody mohu považovat za vzájemně zaměnitelné, a že Jacobiho metoda při zvoleném rozměru vstupních vektorů konverguje rychleji. Abych mohla zkombinovat PCA s metodou Jackknife, zvolila jsem rychlejší Jacobiho metodu.

Neznamená to, že metoda GHA je méně dokonalý způsob dosažení výsledků PCA. Prostudování kapitol 3.3 a 3.4 vede k závěru, že neuronový přístup k řešení problému je vhodnější, když rozměr vstupních vektorů analyzovaných dat přesáhne určitou hodnotu.

Pro porovnání metody GHA s Jacobiho metodou a kombinací PCA s metodou Jackknife jsem sestavila vlastní počítačové programy v jazyce Turbo Pascal, jejichž ukázky výstupů uvádím v přílohách a výsledky pomocí nich zpracovaných ekonomických úloh rozebírám v kapitole 4.

3. Kategorizace ekonomických objektů pomocí neuronové sítě

Tato kapitola obsahuje teoretickou část diplomové práce. Pro její lepší přehlednost zde uvádím stručný obsah jejích podkapitol.

3.1 Statistická analýza vícerozměrných dat: Sumarizuje obecná fakta, která musí být brána v úvahu při zpracovávání vícerozměrných dat, a objasňuje základní pojmy této oblasti.

3.2 Analýza hlavních komponent (PCA): Matematicky popisuje základní metodu PCA, kterou jsem použila pro zpracování ekonomických dat.

3.3 Zobecněný Hebbův algoritmus (GHA): Zde jsou uvedeny matematické důkazy o tom, že metoda PCA je nezávisle realizovatelná pomocí modelu lineárního neuronu. V jejím závěru udávám, kdy je metoda GHA vhodnější pro realizaci PCA než alternativní Jacobiho metoda.

3.4 Jacobiho metoda nalezení vlastních čísel a vlastních vektorů symetrické matice: Matematickou teorií zde popisují konkrétní postup, který jsem použila pro získání výsledků analýzy hlavních komponent. Kapitola přibližuje algoritmus z minulého století, který je dnes výborně uplatnitelný v počítačovém zpracování dat.

3.5 Jackknife: Zde je obsažena teorie metody konstrukce nestranných odhadů, kterou používám pro následné zpracování výsledků metody PCA.

3.6 Kombinace metody Jackknife s metodou PCA: Je o praktické realizaci metody Jackknife v této diplomové práci.

3.7 Konstrukce konfidenčních elipsoidů: Přibližuje, jakým způsobem vznikají intervalové odhady z kombinace metody Jackknife a PCA. V této kapitole uvedenou rovnici elipsoidu jsem nevzala z literatury. Je to můj vlastní návrh, který si nečiní nároky na korektnost podle pravidel platících pro konfidenční elipsoidy ve statistické literatuře. Záměrem mého elipsoidu je prostě vyznačit oblast spolehlivosti polohy obrazu vzoru.

3.1 Statistická analýza vícerozměrných dat [5]

Tato kapitola vychází z poznatků o jednorozměrném normálním rozdělení, které v této práci nepopisují. Lze je nastudovat v mnoha učebnicích statistiky, například [14].

V ekonomické praxi se vedle jednorozměrných analytických informací, obsažených v náhodném skaláru ξ , vyskytují i vícerozměrné analytické informace, obsažené v náhodném vektoru ξ s p složkami $\xi_1, \xi_2, \dots, \xi_p$. Příklady vícerozměrných informací, které jsou v této diplomové práci zpracovány v kapitole 4 a přílohách, jsou

- a) míra naplnění určitých aspektů vybraných profesí,
- b) meziroční pohyby měnových kurzů,
- c) složení aktiv a pasiv a meziroční pohyb určitých tokových veličin zjištěných z účetní závěrky podniků vybraného odvětví.

Na základě provedených analýz je k dispozici *náhodný výběr* o velikosti n . Výběr je tvořen n -ticí vektorů $\mathbf{x}_i^T = (x_{i1}, x_{i2}, \dots, x_{ip})$, které lze chápat jako souřadnice n bodů v p -rozměrném prostoru. Tento náhodný výběr lze vyjádřit maticí o rozměru $n \times p$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \quad (3.1.1)$$

pro kterou platí, že počet bodů (velikost výběru) n je větší než počet složek p náhodného vektoru ξ .

Vícerozměrná náhodná veličina ξ je jednoznačně určena svou sdruženou distribuční funkcí $F(\mathbf{x})$, která je definována jako pravděpodobnost P , že všechny složky ξ_i vektoru ξ budou menší než složky x_i zadaného (nenáhodného) vektoru $\mathbf{x}$.

$$F(\mathbf{x}) = P(\xi_1 \leq x_1 \cap \xi_2 \leq x_2 \cap \dots \cap \xi_p \leq x_p). \quad (3.1.2)$$

Symbol $\cap$ označuje logický součin a vyjadřuje současnou platnost uvedených podmínek.

Sdružená (simultánní) distribuční funkce $F(\mathbf{x})$ má stejné vlastnosti jako distribuční funkce jedné náhodné veličiny. Je neklesající funkcí svých argumentů, je nezáporná a maximálně rovna jedné.

Marginální (okrajová) distribuční funkce $F(x_i)$ složky ξ_i je pak zvláštním případem simultánní distribuční funkce $F(\mathbf{x})$, u které jsou všechny ostatní složky náhodného vektoru na horní mezi svého definičního intervalu; obvykle $\xi_j = \infty$ pro $j \neq i$.

Speciálním typem jsou jednoduché *podmíněné distribuční funkce* $F(\mathbf{x}/x_i)$, vyjadřující pravděpodobnost, že všechny složky vektoru ξ kromě i -té budou menší než odpovídající složka vektoru $\mathbf{x}$. Pro složku ξ_i platí, že je přibližně konstantní, tj. leží v nekonečně malém intervalu $x_i \leq \xi_i \leq x_i + dx_i$. Lze tedy psát

$$F(\mathbf{x}/x_i) = P(\xi_1 \leq x_1 \cap \xi_2 \leq x_2 \cap \dots \cap x_i \leq \xi_i \leq (x_i + dx_i) \cap \dots \cap \xi_p \leq x_p). \quad (3.1.3)$$

V případě, že jsou složky vektoru ξ *nezávislé*, nezávisí podmíněné distribuční funkce na podmínce. Sdružená distribuční funkce se dá pak vyjádřit v jednoduchém tvaru

$$F(\mathbf{x}) = \prod_{i=1}^p F(x_i). \quad (3.1.4)$$

Derivací distribučních funkcí získáme hustoty pravděpodobnosti $f(x_i)$, $f(\mathbf{x})$, respektive $f(\mathbf{x}/x_i)$ označované jako *frekvenční funkce*.

Centrální roli mezi vícerozměrnými náhodnými rozděleními má *vícerozměrné normální rozdělení*, jehož sdružená hustota pravděpodobnosti má tvar

$$f(\mathbf{x}) = (2\pi)^{-\frac{p}{2}} \cdot (\det \mathbf{C})^{-\frac{1}{2}} \cdot e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})}. \quad (3.1.5)$$

Symbol $\det \mathbf{C}$ označuje determinant matice $\mathbf{C}$, definované vzorcem (3.1.18), a symbol $\mathbf{x}^T$ označuje transponovaný vektor $\mathbf{x}$. Parametry tohoto rozdělení jsou vektor středních hodnot $\boldsymbol{\mu}$ a kovarianční matice $\mathbf{C}$ s prvky $c_{ij} = \text{cov}(\xi_i, \xi_j)$ definovanými vzorcem (3.1.10). Kovarianční matice je obvykle pozitivně definitní, takže existuje její inverze. K označení vícerozměrného normálního rozdělení se používá symbol $N(\boldsymbol{\mu}, \mathbf{C})$.

Pokud vektor $\mathbf{x}$ pochází z rozdělení $N(\boldsymbol{\mu}, \mathbf{C})$, platí, že veličina

$$Q(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \quad (3.1.6)$$

má χ^2 rozdělení s p stupni volnosti. Pro případ dvou náhodných veličin ξ_1, ξ_2 lze určit, že

$$\det \mathbf{C} = \sigma_1^2 \sigma_2^2 (1 - \rho_{12}^2), \quad (3.1.7)$$

kde σ_1^2, σ_2^2 jsou rozptyly veličin ξ_1 a ξ_2 a ρ_{12} je párový korelační koeficient definovaný vzorcem (3.1.13). Platí, že

$$\mathbf{C}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2(1-\rho_{12}^2)} & \frac{-\rho_{12}}{\sigma_1\sigma_2(1-\rho_{12}^2)} \\ \frac{-\rho_{12}}{\sigma_1\sigma_2(1-\rho_{12}^2)} & \frac{1}{\sigma_2^2(1-\rho_{12}^2)} \end{bmatrix}. \quad (3.1.8)$$

Po dosazení do rovnice (3.1.5) dostáváme sdruženou hustotu pravděpodobnosti $f(\mathbf{x})$ v jednoduchém tvaru

$$f(x_1, x_2) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1-\rho_{12}^2}} \cdot e^{-\frac{1}{2(1-\rho_{12}^2)} \left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho_{12}(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \right]}, \quad (3.1.9)$$

kde symboly μ_1 a μ_2 označují střední hodnoty náhodné veličiny ξ_1 a ξ_2 .

Mezi důležité vlastnosti vícerozměrného normálního rozdělení patří:

- a) Odpovídající marginální i podmíněná rozdělení jsou také normální.
- b) Jsou-li všechny složky vektoru ξ vzájemně *nekorelované* (tj. všechny párové korelační koeficienty jsou nulové), znamená to, že složky ξ_i , $i = 1, 2, \dots, p$ jsou *nezávislé*.
- c) Pokud má vektor ξ vícerozměrné normální rozdělení, mají libovolné lineární kombinace jeho složek ξ_i také normální rozdělení.

Z uvedeného plyne, že předpoklad normality zde usnadňuje analýzu a umožňuje poměrně jednoduché zpracování úloh souvisejících s náhodným vektorem ξ .

Obdobně jako u jednorozměrných náhodných veličin lze jednotlivé složky náhodného vektoru ξ charakterizovat pomocí momentů. K charakterizaci polohy i -té složky ξ_i se používá *střední hodnota* $E(\xi_i) = \mu_i$ a pro charakterizaci rozptýlení *rozptyl* $D(\xi_i) = \sigma_i^2$. Navíc je třeba definovat *míru intenzity* vztahu mezi složkami ξ_i a ξ_j , $i = j$. Vhodnou charakteristikou je druhý smíšený centrální moment, nazývaný *kovariance* $\text{cov}(\xi_i, \xi_j)$, pro který lze psát

$$\text{cov}(\xi_i, \xi_j) = E(\xi_i \cdot \xi_j) - E(\xi_i) \cdot E(\xi_j) \quad (3.1.10)$$

a který má tyto základní vlastnosti:

- a) Kovariance je vždy nezáporná.
- b) Kovariance je v absolutní hodnotě shora ohraničená součinem $\sigma_i \cdot \sigma_j$, to jest $|\text{cov}(\xi_i, \xi_j)| \leq \sigma_i \cdot \sigma_j$.
- c) Kovariance je symetrickou funkcí svých argumentů.
- d) Kovariance se nemění posunem počátku, ale změna měřítka se projeví úměrně jeho velikosti. Pro čísla $\alpha_1, \alpha_2, \beta_1, \beta_2$ pak platí, že

$$\text{cov}(\alpha_1 \xi_i + \beta_1, \alpha_2 \xi_j + \beta_2) = \alpha_1 \alpha_2 \text{cov}(\xi_i, \xi_j) \quad (3.1.11)$$

e) Pro nekorelované náhodné veličiny je $\text{cov}(\xi_i, \xi_j) = 0$ a mohou nastat dva případy:

1. $E(\xi_i, \xi_j) = 0$ a zároveň $E(\xi_i) = E(\xi_j) = 0$, což je případ *centrovaných ortogonálních* náhodných veličin.
2. $E(\xi_i \cdot \xi_j) = E(\xi_i) \cdot E(\xi_j)$, což je případ *nezávislých* náhodných veličin.

Je tedy patrné, že nekorelovanost neznamená vždy nezávislost.

f) Kovariance je mírou intenzity lineární závislosti. Limitní je případ, kdy jsou ξ_i, ξ_j lineárně závislé, takže $\xi_j = \alpha \xi_i + \beta$, a z rovnice (3.1.10) vychází

$$\text{cov}(\xi_i, \xi_j) = E(\alpha \xi_i^2 + \beta \xi_i) - E(\xi_i) \cdot E(\alpha \xi_i + \beta) = \alpha \sigma_i^2. \quad (3.1.12)$$

Kovariance nepostihuje míru libovolné závislosti, ale pouze lineární. I v případě nulové kovariance mohou být ξ_i a ξ_j nelineárně funkčně závislé.

Nevýhodou kovariance je fakt, že její hodnoty závisí na měřítku, ve kterém jsou vyjádřeny ξ_i a ξ_j . Její velikost lze hodnotit vzhledem k součinu $\sigma_i \cdot \sigma_j$. Proto je přirozené provést standardizaci dělením tímto součinem. Vzniklá veličina $\rho_{ij} = \rho(\xi_i, \xi_j)$ se nazývá *párový korelační koeficient*

$$\rho(\xi_i, \xi_j) = \rho_{ij} = \frac{\text{cov}(\xi_i, \xi_j)}{\sigma_i \sigma_j}. \quad (3.1.13)$$

Je zřejmé, že korelační koeficient leží v rozmezí $-1 \leq \rho_{ij} \leq 1$. Pokud je $\rho_{ij} > 0$, jde o *pozitivně korelované* náhodné veličiny, a pokud je $\rho_{ij} < 0$, jde o *negativně korelované* náhodné veličiny.

Korelační koeficient má následující základní vlastnosti:

- a) Rovnost $|\rho_{ij}| = 1$ ukazuje, že mezi ξ_i a ξ_j existuje přesně lineární vztah.
- b) Pokud jsou náhodné veličiny ξ_i a ξ_j vzájemně nekorelované, je $\rho_{ij} = 0$.
- c) V případě, že ξ_i a ξ_j pocházejí z vícerozměrného normálního rozdělení a $\rho_{ij} = 0$, znamená to, že jsou *vzájemně nezávislé*.
- d) I pro nelineárně závislé náhodné veličiny může být $\rho_{ij} = 0$.

- e) Korelační koeficient ρ_{ii} náhodné veličiny ξ_i samotné se sebou je roven jedné.
- f) Korelační koeficient je invariantní vůči lineární transformaci náhodných proměnných ξ_i, ξ_j . Pro čísla $\alpha_1, \alpha_2, \beta_1, \beta_2$ platí vztah

$$\rho(\alpha_1 \xi_i + \beta_1, \alpha_2 \xi_j + \beta_2) = \text{sign}(\alpha_1 \alpha_2) \rho(\xi_i, \xi_j), \quad (3.1.14)$$

kde $\text{sign}(x)$ je znaménková funkce, pro kterou platí

$$\text{sign}(x) = \begin{cases} -1 & \text{pro } x < 0 \\ 0 & \text{pro } x = 0 \\ 1 & \text{pro } x > 0 \end{cases} \quad (3.1.15)$$

Standardizace čili *normování* náhodné veličiny znamená její převedení na náhodnou veličinu s jednotkovým rozptylem a nulovou střední hodnotou. Standardizované náhodné veličiny ξ_i^*, ξ_j^* se vypočítají ze vzorců

$$\xi_i^* = \frac{\xi_i - E(\xi_i)}{\sqrt{D(\xi_i)}} \quad \text{a} \quad \xi_j^* = \frac{\xi_j - E(\xi_j)}{\sqrt{D(\xi_j)}}. \quad (3.1.16)$$

Protože rozptyly musí být nezáporné, plyne z rovnice (3.1.14), že $\rho_{ij}^* = \rho_{ij}$. Z definice korelačního koeficientu (3.1.13) pro standardizované náhodné veličiny však plyne $\rho_{ij}^* = E(\xi_i^* \cdot \xi_j^*)$, což znamená, že korelační koeficient standardizovaných náhodných veličin je roven střední hodnotě jejich součinu. Při použití standardizovaných náhodných veličin se tedy vztah pro korelační koeficient značně zjednoduší.

Statistické chování náhodného vektoru se charakterizuje pomocí vektoru středních hodnot

$$\boldsymbol{\mu}^T = [E(\xi_1), E(\xi_2), \dots, E(\xi_p)] \quad (3.1.17)$$

a kovarianční matice řádu $p \times p$

$$\mathbf{C} = \begin{bmatrix} D(\xi_1) & \text{cov}(\xi_1, \xi_2) & \cdots & \text{cov}(\xi_1, \xi_p) \\ \text{cov}(\xi_1, \xi_2) & D(\xi_2) & \cdots & \text{cov}(\xi_2, \xi_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\xi_1, \xi_p) & \text{cov}(\xi_2, \xi_p) & \cdots & D(\xi_p) \end{bmatrix}. \quad (3.1.18)$$

Místo kovarianční matice můžeme použít také její normovanou verzi, tj. korelační matici

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix}. \quad (3.1.19)$$

Korelační matice má na diagonále samé jedničky a mimodiagonální prvky jsou jednotlivé párové korelační koeficienty. Kovarianční matice $\mathbf{C}$ i korelační matice $\mathbf{R}$ jsou symetrické.

Na tomto místě bych chtěla upozornit na rozdíl v pojmech mezi literaturou [5] a [1]. Zatímco podle knížky [5] je korelační matice sestavena z normovaných náhodných proměnných, podle knížky [1] je sestavena z pouze centrovaných náhodných proměnných, viz kapitola 3.2 „Analýza hlavních komponent (PCA)“, rovnice (3.2.9). Počítačový program, který jsem vyhotovila, abych mohla zpracovávat data, provádí standardizaci proměnných podle literatury [5]. Standardizace proměnných odstraňuje závislost na jednotkách proměnných.

Z vícerozměrného výběru velikosti n , definovaného n -ticí p -rozměrných vektorů $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$, $i = 1, 2, \dots, n$, který lze vyjádřit maticí měřených hodnot $\mathbf{X}$ rozměru $n \times p$, je možno určit výběrový vektor středních hodnot $\hat{\boldsymbol{\mu}}$ vztahem

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i. \quad (3.1.20)$$

Podobně pro odhad kovarianční matice $\mathbf{S}^0$ platí rovnice

$$\mathbf{S}^0 = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T. \quad (3.1.21)$$

3.2 Analýza hlavních komponent (PCA) [1]

Analýza hlavních komponent v angličtině zvaná *principal components analysis* (PCA), známá také jako *Karhunen–Loève transformation*, je asi nejstarší a nejznámější metodou *multivariační analýzy* (*multivariate analysis*). Byla poprvé použita Pearsonem v roce 1901 v biologii jako nová forma lineární regresní analýzy. Potom byla vyvinuta Hotellingem v roce 1933 v práci o psychometrii. Objevila se znova zcela nezávisle při pokládání základů teorie pravděpodobnosti u Karhunena v roce 1947 a byla následně zobecněna Loèvem v roce 1963.

V této části práce vysvětlím metodu PCA a v následující kapitole i její souvislost s tzv. *Hebbovým učením* (*Hebbian learning*).

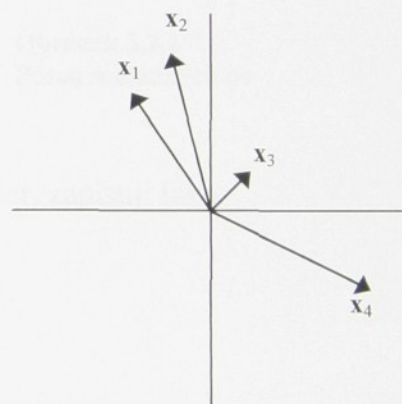
$\mathbf{x}$ je p -rozměrný náhodný vektor dat, která nás zajímají. Tento vektor může reprezentovat například následující 4 vektory:

$$\mathbf{x}_1 = (-1; 1,5)$$

$$\mathbf{x}_2 = (-0,5; 2)$$

$$\mathbf{x}_3 = (0,5; 0,5)$$

$$\mathbf{x}_4 = (2; -1)$$



Obrázek 3.2.1
Příklad 4 vektorů

Aritmetický průměr 1. souřadnic: $(-1 - 0,5 + 0,5 + 2) / 4 = 0,25$.

Aritmetický průměr 2. souřadnic: $(+ 1,5 + 2 + 0,5 - 1) / 4 = 0,75$.

$$p = 2$$

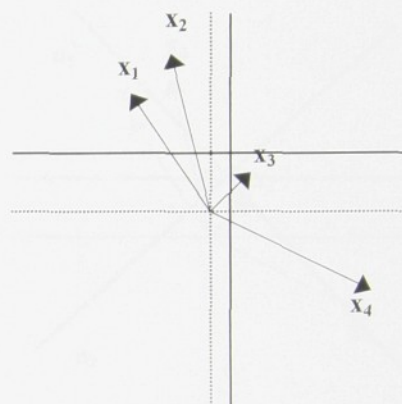
$n = 4$, n je počet zkoumaných vektorů.

Vektor $\mathbf{x}$, který zde chápu jako skupinu p -rozměrných vektorů, znormuji tak, že od jeho souřadnic odečtu jejich aritmetické průměry. Výsledkem je to, že aritmetický průměr všech nových souřadnic je roven nule.

$$0 = \sum_{i=1}^n x_{i1}, \quad 0 = \sum_{i=1}^n x_{i2}, \quad \dots, \quad 0 = \sum_{i=1}^n x_{ip},$$

kde x_{ij} je j -tý prvek i -tého vektoru $\mathbf{x}$.

Geometricky lze tuto operaci interpretovat jako horizontální posunutí souřadných os o aritmetický průměr 1. souřadnic a vertikální posunutí souřadných os o aritmetický průměr 2. souřadnic, aniž by došlo k pohybu koncových bodů vektorů $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$, pokud uvažujeme pouze dvourozměrný prostor.



Obrázek 3.2.2
Posun souřadných os

Skutečnost, že vektor $\mathbf{x}$ má nulový aritmetický průměr, zapisuji jako

$$E(\mathbf{x}) = 0. \quad (3.2.1)$$

Dále budu hledat takové směry vektoru $\mathbf{u}$, které, když se na ně promítnou skalárním součinem zkoumané vektory $\mathbf{x}$, dají extrémní (lokálně nebo globálně minimální a maximální) hodnoty rozptylu průmětů.

Rozptyl σ^2 má vzorec

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad (3.2.2)$$

kde $\bar{x}$ je aritmetický průměr zkoumané proměnné a n je počet proměnných.

Pokud $\bar{x} = 0$, jako v našem případě, potom má vzorec (3.2.2) podobu

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n x_i^2, \quad (3.2.3)$$

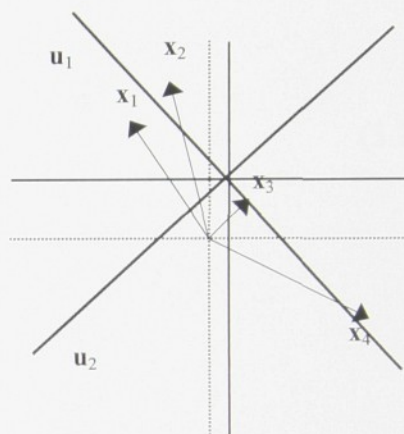
neboli rozptyl σ^2 je aritmetickým průměrem druhých mocnin proměnné x_i .

$$\sigma^2 = \overline{x^2} = E(x^2) \quad (3.2.4)$$

Směr vektoru $\mathbf{u}_1$ je takový, že rozptyl průmětů vektorů $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ na něj je maximální.

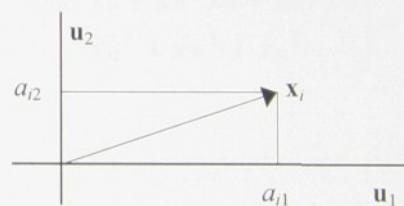
Směr vektoru $\mathbf{u}_2$ je takový, že rozptyl průmětů vektorů $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ na něj je minimální.

Vektor $\mathbf{u}_1$ je kolmý na vektor $\mathbf{u}_2$.



Obrázek 3.2.3
Vektory $\mathbf{u}_1$ a $\mathbf{u}_2$

Průměty a se nazývají *hlavní komponenty*.



Obrázek 3.2.4
Hlavní komponenty

Délka průmětů a může být vyjádřena skalárním součinem

$$a = \mathbf{x}^T \mathbf{u} = \mathbf{u}^T \mathbf{x}, \quad (3.2.5)$$

protože vektor $\mathbf{u}$ je jednotkový.

$$|\mathbf{u}| = (\mathbf{u}^T \mathbf{u})^{1/2} = 1 \quad (3.2.6)$$

Je-li aritmetický průměr vektoru $\mathbf{x}$ roven nule, znamená to, že aritmetický průměr jeho průmětů na konstantní vektor $\mathbf{u}$ je také roven nule.

$$E(a) = \mathbf{u}^T E(\mathbf{x}) = 0 \quad (3.2.7)$$

Rozptyl průmětů a na jeden z vektorů $\mathbf{u}$ je potom roven

$$\sigma^2 = E[a^2] = E[(\mathbf{u}^T \mathbf{x})(\mathbf{x}^T \mathbf{u})] = \mathbf{u}^T E[\mathbf{x}\mathbf{x}^T] \mathbf{u} = \mathbf{u}^T \mathbf{R} \mathbf{u}, \quad (3.2.8)$$

kde

$$\mathbf{R} = E(\mathbf{x}\mathbf{x}^T) \quad (3.2.9)$$

je korelační matice datového vektoru $\mathbf{x}$.

Příklad korelační matice $\mathbf{R}$ pro $p = 2$ a $n = 3$:

$$\mathbf{x}_1 = (x_{11}; x_{12})^T, \quad \mathbf{x}_2 = (x_{21}; x_{22})^T, \quad \mathbf{x}_3 = (x_{31}; x_{32})^T$$

$$\mathbf{x}\mathbf{x}^T = \begin{bmatrix} x_{11} & x_{21} & x_{31} \\ x_{12} & x_{22} & x_{32} \end{bmatrix} \cdot \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix} = \begin{bmatrix} x_{11}^2 + x_{21}^2 + x_{31}^2 & x_{11} \cdot x_{12} + x_{21} \cdot x_{22} + x_{31} \cdot x_{32} \\ x_{12} \cdot x_{11} + x_{22} \cdot x_{21} + x_{32} \cdot x_{31} & x_{12}^2 + x_{22}^2 + x_{32}^2 \end{bmatrix} = n \cdot \mathbf{R}$$

$\mathbf{R}$ je symetrická matice rozměru $p \times p$ vzniklá vynásobením normovaných p -rozměrných datových vektorů $\mathbf{x}$, jejíž každý prvek je podělen počtem n zkoumaných vektorů $\mathbf{x}$.

Rozptyl $\sigma^2[a]$ můžeme chápat jako funkci jednotkového vektoru $\mathbf{u}$.

$$\sigma^2[a] = \psi(\mathbf{u}) = \mathbf{u}^T \mathbf{R} \mathbf{u}. \quad (3.2.10)$$



Obrázek 3.2.5

Rozptyl průmětů a jako funkce směru vektoru $\mathbf{u}$

Dále budu řešit úlohu, pro jaké hodnoty $\mathbf{u}$ nabývá funkce $\psi(\mathbf{u})$ extrémní hodnoty. Funkce má extrém tam, kde její 1. derivace je rovna nule. V místě extrému funkce $\psi(\mathbf{u})$ přibližně platí

$$\psi(\mathbf{u} + \delta\mathbf{u}) - \psi(\mathbf{u}) = 0,$$

kde $\delta\mathbf{u}$ je malý vektor. Proto v místě extrému platí přibližná rovnice

$$\psi(\mathbf{u} + \delta\mathbf{u}) = \psi(\mathbf{u}). \quad (3.2.11)$$

Do vzorce (3.2.10) dosadím $\mathbf{u} + \delta\mathbf{u}$ místo $\mathbf{u}$ a dostanu rovnici

$$\psi(\mathbf{u} + \delta\mathbf{u}) = (\mathbf{u} + \delta\mathbf{u})^T \mathbf{R} (\mathbf{u} + \delta\mathbf{u}),$$

kterou budu dále upravovat.

$$\psi(\mathbf{u} + \delta\mathbf{u}) = \mathbf{u}^T \mathbf{R} \mathbf{u} + \mathbf{u}^T \mathbf{R} \delta\mathbf{u} + (\delta\mathbf{u})^T \mathbf{R} \mathbf{u} + (\delta\mathbf{u})^T \mathbf{R} \delta\mathbf{u}$$

$$\psi(\mathbf{u} + \delta\mathbf{u}) = \mathbf{u}^T \mathbf{R} \mathbf{u} + (\delta\mathbf{u})^T \mathbf{R} \mathbf{u} + (\delta\mathbf{u})^T \mathbf{R} \mathbf{u} + (\delta\mathbf{u})^T \mathbf{R} \delta\mathbf{u}$$

$$\psi(\mathbf{u} + \delta\mathbf{u}) = \psi(\mathbf{u}) + 2(\delta\mathbf{u})^T \mathbf{R} \mathbf{u} \quad (3.2.12)$$

Při úpravě na rovnici (3.2.12) jsem použila následující pravidla:

Člen $\mathbf{u}^T \mathbf{R} \delta \mathbf{u}$ je skalár rovný své transpozici $(\delta \mathbf{u})^T \mathbf{R}^T \mathbf{u} = (\delta \mathbf{u})^T \mathbf{R} \mathbf{u}$, neboť $\mathbf{R} = \mathbf{R}^T$, (matice $\mathbf{R}$ je symetrická),

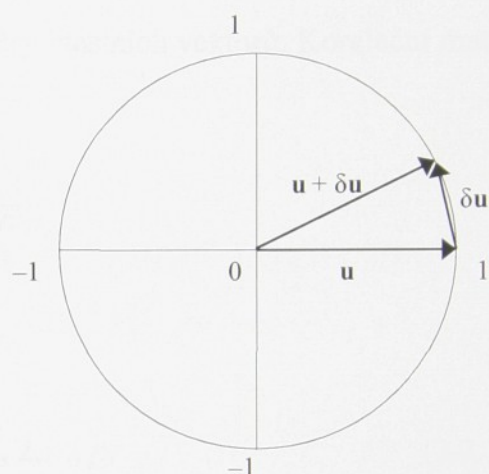
člen $(\delta \mathbf{u})^T \mathbf{R} \delta \mathbf{u}$ je možno vynechat, neboť je v něm dvakrát zastoupen malý vektor $\delta \mathbf{u}$.

S ohledem na vzorec (3.2.11) mohu rovnici (3.2.12) napsat ve tvaru

$$0 = 2(\delta \mathbf{u})^T \mathbf{R} \mathbf{u}$$

a podělit ji číslem 2.

Malý vektor $\delta \mathbf{u}$ způsobující drobnou odchylku směru jednotkového vektoru $\mathbf{u}$ je na vektor $\mathbf{u}$ téměř kolmý. Mohu stanovit, že délka jeho kolmého průmětu na $\mathbf{u}$ je nulová. Tedy, že skalární součin $(\delta \mathbf{u})^T \mathbf{u} = 0$.



Obrázek 3.2.6

Demonstrace kolmosti vektorů $\mathbf{u}$ a $\delta \mathbf{u}$

Při hledání extrému funkce $\psi(\mathbf{u})$ jsem tedy dostala tyto dva vzorce:

$$(\delta \mathbf{u})^T \mathbf{R} \mathbf{u} = 0, \quad (3.2.13)$$

$$(\delta \mathbf{u})^T \mathbf{u} = 0, \quad (3.2.14)$$

které teď vzájemně zkombinuji.

Jelikož vektor $\mathbf{u}$ je bezrozměrný, zatímco matice $\mathbf{R}$ je složena z datových vektorů $\mathbf{x}$, jejichž souřadnice mohou mít nějaký význam (například 1. souřadnice – počet zaměstnanců, 2. souřadnici – míra zisku, atd.), musím členu $(\delta \mathbf{u})^T \mathbf{u}$ přiřadit číslo λ představující nějaký fiktivní fyzický rozměr. Výsledkem je rovnice

$$(\delta \mathbf{u})^T \mathbf{R} \mathbf{u} = \lambda (\delta \mathbf{u})^T \mathbf{u},$$

kterou mohu podělit vektorem, $(\delta \mathbf{u})^T$ a dostanu rovnici

$$\mathbf{R} \mathbf{u} = \lambda \mathbf{u}. \quad (3.2.15)$$

Rovnici (3.2.15) vyhovují nulové vektory $\mathbf{u}$ jako triviální řešení a nenulové vektory $\mathbf{u}$, které se nazývají *vlastní vektory* nebo *eigenvektory* náležející určitým *vlastním číslům* λ matice $\mathbf{R}$. Různým vlastním číslům náleží různé směry vlastních vektorů. Korelační matice $\mathbf{R}$ má vlastní čísla reálná nezáporná.

Vlastní čísla matice $\mathbf{R}_{p \times p}$ označím jako $\lambda_1, \lambda_2, \dots, \lambda_p$,

a jim odpovídající vlastní vektory jako $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$.

Potom mohu rovnici (3.2.15) zapsat jako

$$\mathbf{R} \mathbf{u}_j = \lambda_j \mathbf{u}_j, \quad j = 1, 2, \dots, p. \quad (3.2.16)$$

Vlastní čísla srovnám sestupně tak, aby λ_1 bylo nejvyšší a λ_p nejnižší.

$$\lambda_1 > \lambda_2 > \dots > \lambda_j > \dots > \lambda_p$$

Z odpovídajících vlastních vektorů sestavím matici $\mathbf{U}$ o rozměrech $p \times p$.

$$\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_j, \dots, \mathbf{u}_p] \quad (3.2.17)$$

Nyní mohu množinu rovnic (3.2.16) zapsat jako jedinou maticovou rovnici

$$\mathbf{R} \mathbf{U} = \mathbf{U} \mathbf{\Lambda}, \quad (3.2.18)$$

kde $\mathbf{\Lambda}$ je diagonální matice s vlastními čísly.

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix}$$

Matice $\mathbf{U}$ je *ortonormální*, což znamená, že vektory $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$ jsou na sebe kolmé a mají jednotkovou délku.

Nyní uvedu důkaz, že vektory $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$ jsou na sebe kolmé. Vlastní vektory $\mathbf{v}$, $\mathbf{w}$ a vlastní čísla λ obecné (nejen symetrické) matice $A_{p \times p}$ vyhovují rovnicím

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v},$$

$$\mathbf{w}^T \mathbf{A} = \lambda\mathbf{w}^T.$$

Druhá rovnice $\mathbf{w}^T \mathbf{A} = \lambda\mathbf{w}^T$ se dá zapsat také ve tvaru $\mathbf{A}^T \mathbf{w} = \lambda\mathbf{w}$, což není ekvivalentní s první rovnicí $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$, protože $\mathbf{A} \neq \mathbf{A}^T$. Dokáži, že vlastní vektory (pravý = $\mathbf{v}$ a levý = $\mathbf{w}$) matice $\mathbf{A}$ jsou na sebe kolmé, pokud odpovídají navzájem odlišným vlastním číslům:

$$\begin{array}{llll} \mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{v}_i & / \cdot \mathbf{w}_j^T & \rightarrow & \mathbf{w}_j^T \mathbf{A}\mathbf{v}_i = \lambda_i \mathbf{w}_j^T \mathbf{v}_i \\ \mathbf{w}_j^T \mathbf{A} = \lambda_j \mathbf{w}_j^T & / \cdot \mathbf{v}_i & \rightarrow & \mathbf{w}_j^T \mathbf{A}\mathbf{v}_i = \lambda_j \mathbf{w}_j^T \mathbf{v}_i \end{array}$$

$$\lambda_i \mathbf{w}_j^T \mathbf{v}_i = \lambda_j \mathbf{w}_j^T \mathbf{v}_i$$

$$(\lambda_i - \lambda_j) \mathbf{w}_j^T \mathbf{v}_i = 0.$$

Pokud jsou vlastní čísla λ_i a λ_j navzájem odlišná a tedy $\lambda_i - \lambda_j$ není rovno nule, musí se rovnat nule skalární součin $\mathbf{w}_j^T \mathbf{v}_i$. Levé a pravé vlastní vektory odpovídající různým vlastním číslům jsou tedy skutečně na sebe kolmé. Je-li matice $\mathbf{A}$ symetrická a tedy $\mathbf{A} = \mathbf{A}^T$, potom platí ekvivalence rovnic $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ a $\mathbf{A}^T \mathbf{w} = \lambda\mathbf{w}$, což znamená, že levé a pravé vlastní vektory symetrické matice $\mathbf{A}$ odpovídající stejnému vlastním číslu λ splývají. Jelikož i u symetrické matice platí vlastnost $\mathbf{w}_j^T \mathbf{v}_i = 0$, když $j \neq i$, znamená to, že vlastní vektory symetrické matice odpovídající různým vlastním číslům jsou na sebe kolmé, což jsem chtěla dokázat.

To, že vektory $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_p$ jsou jednotkové, dokazovat nemusím. Postačí konstatovat, že vlastní vektor mohu upravit na libovolnou délku (například jednotkovou), neboť platí $\mathbf{A}\alpha\mathbf{v} = \lambda\alpha\mathbf{v}$, kde $\alpha \in \mathbb{R}$.

Ortonormalita vektorů $\mathbf{u}$ se projevuje těmito rovnicemi:

$$\begin{aligned}\mathbf{u}_i^T \mathbf{u}_j &= 1, & j &= i, \\ \mathbf{u}_i^T \mathbf{u}_j &= 0, & j &\neq i.\end{aligned}\tag{3.2.19}$$

Rovnice (3.2.19) vyžadují, aby matice $\mathbf{R}$ měla vlastní čísla všechna odlišná, nenásobná. Ekvivalentně mohu rovnice (3.2.19) zapsat jako

$$\mathbf{U}^T \mathbf{U} = \mathbf{E},$$

kde $\mathbf{E}$ je jednotková matice o rozměrech $p \times p$.

$$\mathbf{E} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Z toho vyplývá, že transpozice matice $\mathbf{U}$ se rovná její inverzi.

$$\mathbf{U}^T = \mathbf{U}^{-1}\tag{3.2.20}$$

Rovnici (3.2.20) mohu nyní zkombinovat s rovnicí (3.2.18) a dostanu

$$\begin{aligned}\mathbf{R}\mathbf{U} &= \mathbf{U}\mathbf{\Lambda} & / \cdot \mathbf{U}^{-1} \\ \mathbf{U}^{-1} \mathbf{R}\mathbf{U} &= \mathbf{E}\mathbf{\Lambda} \\ \mathbf{U}^T \mathbf{R}\mathbf{U} &= \mathbf{\Lambda},\end{aligned}\tag{3.2.21}$$

což je možno zapsat jako

$$\begin{aligned} \mathbf{u}_i^T \mathbf{R} \mathbf{u}_j &= \lambda_i, & j &= i, \\ \mathbf{u}_i^T \mathbf{R} \mathbf{u}_j &= 0, & j &\neq i. \end{aligned} \quad (3.2.22)$$

Z rovnic (3.2.10) a (3.2.22) vyplývá, že rozptyl hlavních komponent je roven vlastním číslům.

$$\sigma^2(a) = \lambda_j = \psi(\mathbf{u}_j) \quad (3.2.23)$$

Dosavadní výsledky lze shrnout takto: Vlastní vektory korelační matice $\mathbf{R}$ náležející datovému vektoru $\mathbf{x}$ s nulovým aritmetickým průměrem jeho jednotlivých složek (souřadnic) mají takový směr, ve kterém nabývá rozptyl průmětů vektorů $\mathbf{x}$ na jednotlivé vlastní vektory matice $\mathbf{R}$ extrémních hodnot. Vlastní čísla odpovídající vlastním vektorům matice $\mathbf{R}$ jsou rovna rozptylu průmětů datového vektoru $\mathbf{x}$ na jednotlivé směry odpovídajících vlastních vektorů matice $\mathbf{R}$.

Dále se budu zabývat reprezentací dat. Existuje-li p možných netriviálních řešení rovnice $\mathbf{R}\mathbf{u} = \lambda\mathbf{u}$, uvažujeme potom o p možných projekcích (průmětech) datového vektoru $\mathbf{x}$ do p vlastních směrů vektoru $\mathbf{u}$ matice $\mathbf{R}$.

$$a_j = \mathbf{u}_j^T \mathbf{x} = \mathbf{x}^T \mathbf{u}_j, \quad j = 1, 2, \dots, p \quad (3.2.24)$$

Hlavní komponenty a_j mají stejné fyzické rozměry jako datový vektor $\mathbf{x}$.

Dále budu zpětně rekonstruovat datový vektor $\mathbf{x}$ z jeho projekcí a_j . Vektor $\mathbf{a}$ je sestaven z projekcí jednoho datového vektoru $\mathbf{x}$ na vlastní vektory matice $\mathbf{R}$.

$$\mathbf{a} = [a_1, a_2, \dots, a_p]^T = [\mathbf{x}^T \mathbf{u}_1, \mathbf{x}^T \mathbf{u}_2, \dots, \mathbf{x}^T \mathbf{u}_p]^T = \mathbf{U}^T \mathbf{x} \quad (3.2.25)$$

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} \mathbf{u}_1^T \\ \mathbf{u}_2^T \\ \vdots \\ \mathbf{u}_p^T \end{bmatrix} \cdot \mathbf{x} = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1p} \\ u_{21} & u_{22} & \dots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{p1} & u_{p2} & \dots & u_{pp} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

Nyní vynásobím rovnici (3.2.25) z levé strany maticí $\mathbf{U}$ a využiji přitom vztah (3.2.20) a dostanu:

$$\mathbf{U}\mathbf{a} = \mathbf{x} = \sum_{j=1}^p a_j \mathbf{u}_j \quad (3.2.26)$$

V tomto smyslu představují vlastní vektory $\mathbf{u}_j$ bázi datového prostoru.

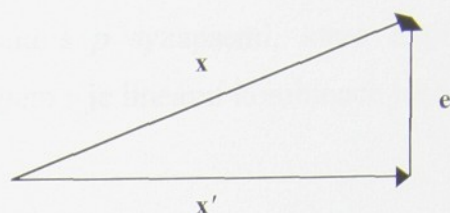
Dále budu redukovat počet rozměrů datového prostoru. PCA umožňuje snížit počet rozměrů potřebných pro dostatečně přesnou reprezentaci datového vektoru $\mathbf{x}$ tím, že v rovnici (3.2.26) odstraníme ty členy $a_j \mathbf{u}_j$, které mají nízkou hodnotu. Když $\lambda_1, \lambda_2, \dots, \lambda_m$, je m nejvyšších vlastních čísel matice $\mathbf{R}$, potom můžeme aproximovat datový vektor $\mathbf{x}$ odříznutím rozvoje rovnice (3.2.26) po m prvních členech následovně:

$$\mathbf{x}' = \sum_{j=1}^m a_j \mathbf{u}_j, \quad m < p. \quad (3.2.27)$$

Chyba této aproximace je vyjádřena chybovým vektorem $\mathbf{e}$.

$$\mathbf{e} = \mathbf{x} - \mathbf{x}' = \sum_{j=m+1}^p a_j \mathbf{u}_j \quad (3.2.28)$$

Vektor $\mathbf{e}$ je kolmý na vektor $\mathbf{x}'$, neboť máme-li množinu ortogonálních vektorů a tu rozdělíme na dvě disjunktní podmnožiny (nemající společné prvky), potom budou všechny lineární kombinace vektorů z jedné podmnožiny kolmé na lineární kombinace vektorů z druhé podmnožiny.



Obrázek 3.2.7
Demonstrace kolmosti vektorů $\mathbf{e}$ a $\mathbf{x}'$

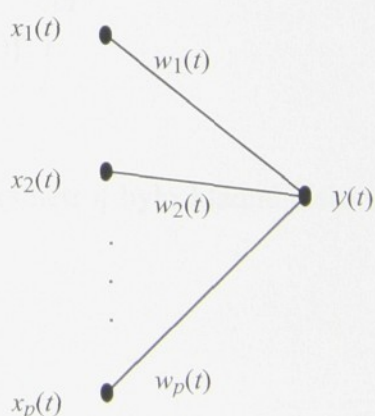
$$\mathbf{e}^T \mathbf{x}' = 0 \quad (3.2.29)$$

Čím jsou vlastní čísla vlastních vektorů $\mathbf{u}$ vyřazených z aproximace blíže nule, tím efektivnější je redukce rozměrů datového prostoru v zachování vstupní informace.

Metoda ortogonální projekce dat na podprostor tvořený vlastními vektory náležejícími nejvyšším vlastním číslům korelační matice vstupních dat se nazývá *podprostorová dekompozice* (*subspace decomposition*). PCA nám umožňuje odhalit charakter shluků dat skrytých ve vícerozměrném prostoru tím, že poskytne pohled na data z takových úhlů, ze kterých je rozptýl dat nejvyšší.

3.3 Zobecněný Hebbův algoritmus (GHA) [1]

Lineární model neuronu funguje jako filtr pro vlastní vektor odpovídající maximálnímu vlastnímu číslu. Mezi chováním samoorganizujících se neuronových sítí a statistickou metodou PCA je úzká souvislost. V této části práce budu demonstrovat, jak osamocený lineární neuron adaptující své synaptické váhy podle Hebbovy strategie se může chovat jako filtr pro první hlavní komponentu rozdělení vstupních dat. Tento objev publikoval Oja v roce 1982.



Obrázek 3.3.1 Pramen: [1]
Model lineárního neuronu

Na obrázku 3.3.1 je model jednoduchého neuronu s p synapsemi, které mají synaptické váhy w . Neuron je lineární, protože jeho výstupem y je lineární kombinace jeho vstupů x .

Výstup y je definován jako

$$y = \sum_{i=1}^p w_i x_i. \quad (3.3.1)$$

V souladu s *Hebbovým učením*, zmíněným již v kapitole 1.2, se synaptické váhy mění v čase t . Zvyšují se, když presynaptický signál x_i a postsynaptický signál y jsou oba vysoké (velké). To je vyjádřeno vzorcem

$$w_i(t+1) = w_i(t) + \eta y(t) x_i(t), \quad i = 1, 2, \dots, p, \quad (3.3.2)$$

kde t označuje diskrétní čas a η [éta] je parametr rychlosti učení. To je základní forma Hebbova pravidla učení. V této podobě však vede k neomezenému růstu synaptických vah w_i . V reálném světě musí synapse soutěžit o omezené zdroje. Proto zavedu do vzorce pro úpravu vah (3.3.2) normalizaci:

$$w_i(t+1) = \frac{w_i(t) + \eta y(t) x_i(t)}{\left(\sum_{i=1}^p [w_i(t) + \eta y(t) x_i(t)]^2 \right)^{1/2}}. \quad (3.3.3)$$

Pro další úpravu vzorce (3.3.3) je nezbytné, aby parametr η bylo kladné číslo blížíící se nule. Potom přibližně platí.

$$w_i(t+1) = \frac{w_i(t)}{\left(\sum_{i=1}^p [w_i(t)]^2 \right)^{1/2}}$$

a z toho plyne

$$\left(\sum_{i=1}^p [w_i(t)]^2 \right)^{-1/2} = 1, \quad (3.3.4)$$

což je normalizace vektoru $\mathbf{w}$ o složkách $[w_1, w_2, \dots, w_p]$ rovnajících se synaptickým vahám.

Za proměnnou funkce (3.3.3) zvolím η a budu ji aproximovat pomocí Taylorova polynomu v okolí bodu $\eta = 0$. To je možné, přestože funkce (3.3.3) není spojitá ale diferenční rovnice. V okolí bodu $\eta = 0$ má totiž v jednotlivých krocích tak malé změny, že se blíží spojitě funkci, kterou je možno derivovat.

Hledaný polynom má obecný tvar

$$\frac{f(0)}{0!} \cdot \eta^0 + \frac{f'(0)}{1!} \cdot \eta^1 + \frac{f''(0)}{2!} \cdot \eta^2 + \frac{f'''(0)}{3!} \cdot \eta^3 + \dots + \frac{f^{(k)}(0)}{k!} \cdot \eta^k = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} \cdot \eta^k$$

Členy polynomu obsahující druhou a vyšší mocniny proměnné η jsou pro hledanou aproximaci zanedbatelně malé a proto je vypustím. Postačí tedy udělat pouze první derivaci funkce (3.3.3).

Pro větší přehlednost následujících výpočtů zavedu substituci:

$$r_i = w_i(t),$$

$$s_i = y_i(t) x_i(t).$$

Vzorec (3.3.3) má potom podobu

$$w_i(t+1) = (r_i + \eta s_i) \cdot \left(\sum_{i=1}^p [r_i + \eta s_i]^2 \right)^{-1/2}.$$

První derivace funkce (3.3.3) podle η je

$$[w_i(t+1)] = s_i \cdot \left(\sum_{i=1}^p [r_i + \eta s_i]^2 \right)^{-1/2} + (r_i + \eta s_i) \cdot \left(-\frac{1}{2} \right) \cdot \left(\sum_{i=1}^p [r_i + \eta s_i]^2 \right)^{-3/2} \cdot \sum_{i=1}^p [2(r_i + \eta s_i) \cdot s_i]$$

$$[w_i(t+1)] = s_i \cdot \left(\sum_{i=1}^p [r_i + \eta s_i]^2 \right)^{-1/2} - (r_i + \eta s_i) \cdot \left(\sum_{i=1}^p [r_i + \eta s_i]^2 \right)^{-3/2} \cdot \sum_{i=1}^p [s_i(r_i + \eta s_i)]$$

Nyní mohu napsat koeficienty hledaného polynomu, jimiž je hodnota funkce (3.3.3) a její první derivace pro $\eta = 0$.

$$f(0) = r_i \cdot \left(\sum_{i=1}^p r_i^2 \right)^{-1/2}$$

$$[f(0)] = s_i \cdot \left(\sum_{i=1}^p r_i^2 \right)^{-1/2} - r_i \cdot \left(\sum_{i=1}^p r_i^2 \right)^{-3/2} \cdot \sum_{i=1}^p s_i r_i$$

Po odstranění substituce mají tyto koeficienty podobu

$$f(0) = w_i(t) \cdot \left(\sum_{i=1}^p [w_i(t)]^2 \right)^{-1/2} = w_i(t) \cdot 1 = w_i(t)$$

$$[f(0)] = y(t) x_i(t) \cdot \left(\sum_{i=1}^p [w_i(t)]^2 \right)^{-1/2} - w_i(t) \cdot \left(\sum_{i=1}^p [w_i(t)]^2 \right)^{-3/2} \cdot \sum_{i=1}^p y(t) x_i(t) w_i(t)$$

$$[f(0)] = y(t) x_i(t) \cdot 1 - w_i(t) \cdot 1 \cdot y(t) \cdot \sum_{i=1}^p x_i(t) w_i(t) = y(t) x_i(t) - w_i(t) \cdot y(t) \cdot y(t)$$

Při těchto úpravách jsem použila vztahů (3.3.4) a (3.3.1). Po dosazení koeficientů do prvních dvou členů obecného Taylorova polynomu má výsledná aproximace podobu

$$w_i(t+1) = w_i(t) + \eta y(t) [x_i(t) - y(t) w_i(t)]. \quad (3.3.5)$$

Člen $\eta y(t) x_i(t)$ v rovnici (3.3.5) představuje obvyklé hebbovské úpravy synaptické váhy w_i . Záporný člen $\eta y(t)^2 w_i(t)$ umožňuje stabilizaci synaptických vah a je negativní zpětnou vazbou pro omezení růstu váhy. Představuje tak faktor zapominání v neuronové síti.

Výraz

$$x_i(t)' = x_i(t) - y(t) w_i(t) \quad (3.3.6)$$

může být chápán jako efektivní vstup i -té synapse. Rovnice (3.3.5) s tímto výrazem má tvar

$$w_i(t+1) = w_i(t) + \eta y(t) x_i(t)'. \quad (3.3.7)$$

Rovnice (3.3.7) je svým tvarem velmi blízká rovnici Hebbova učení (3.3.2).

Nyní budu zkoumat vlastnosti tohoto modelu.

$\mathbf{x}(t) = [x_1(t), x_2(t), \dots, x_p(t)]^T$ je vektor vstupních signálů jdoucích na p synapsí neuronu.

$\mathbf{w}(t) = [w_1(t), w_2(t), \dots, w_p(t)]^T$ je vektor vah na p synapsích náležejících neuronu.

Rovnici (3.3.1) potom mohu přepsat jako skalární součin vektorů. Srovnajte s rovnicí (3.2.24).

$$y(t) = \mathbf{w}(t)^T \mathbf{x}(t) = \mathbf{x}^T(t) \mathbf{w}(t) \quad (3.3.8)$$

Rovnici (3.3.5) mohu přepsat do tvaru

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \eta y(t) [\mathbf{x}(t) - y(t) \mathbf{w}(t)]. \quad (3.3.9)$$

Dosadím-li rovnici (3.3.8) do rovnice (3.3.9), dostanu vztah

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \eta [\mathbf{x}(t) \mathbf{x}(t)^T \mathbf{w}(t) - \mathbf{w}(t)^T \mathbf{x}(t) \mathbf{x}(t)^T \mathbf{w}(t) \mathbf{w}(t)]. \quad (3.3.10)$$

Tento algoritmus učení je rekurzivní stochastická diferenční rovnice, jejíž hodnoty se mění s časem. Řešení této rovnice konverguje k určitému vektoru $\mathbf{q}_1$, když se počet iterací t blíží nekonečnu.

$$\mathbf{w}(t) \rightarrow \mathbf{q}_1, \quad \text{když } t \rightarrow \infty \quad (3.3.11)$$

Potom také

$$\mathbf{w}(t+1) \rightarrow \mathbf{w}(t); \quad |\Delta \mathbf{w}(t)| \rightarrow 0. \quad (3.3.12)$$

Když z obou stran rovnice (3.3.10) udělám aritmetický průměr, dostanu rovnici

$$E[\mathbf{w}(t+1)] = E[\mathbf{w}(t)] + \eta [E[\mathbf{x}(t) \mathbf{x}(t)^T] \mathbf{w}(t) - \mathbf{w}(t)^T E[\mathbf{x}(t) \mathbf{x}(t)^T] \mathbf{w}(t) \mathbf{w}(t)].$$

Vzhledem ke konvergenci (3.3.12) mohu vypustit členy $E[\mathbf{w}(t+1)]$ a $E[\mathbf{w}(t)]$. Dále použiji vztah pro korelační matici $\mathbf{R}$ (3.2.9) a dostanu

$$0 = \eta [\mathbf{R} \mathbf{w}(t) - \mathbf{w}(t)^T \mathbf{R} \mathbf{w}(t) \mathbf{w}(t)], \quad t \rightarrow \infty. \quad (3.3.13)$$

Dále využiji konvergenci (3.3.11) a rovnici podělím koeficientem η . Výsledkem je rovnice

$$0 = \mathbf{R} \mathbf{q}_1 - (\mathbf{q}_1^T \mathbf{R} \mathbf{q}_1) \mathbf{q}_1, \quad t \rightarrow \infty. \quad (3.3.14)$$

Kvadratický funkcionál $\mathbf{q}_1^T \mathbf{R} \mathbf{q}_1$ označím jako nějaké nezáporné číslo λ_1 . Výsledkem je známá rovnice pro vlastní vektory matice

$$\mathbf{R} \mathbf{q}_1 = \lambda_1 \mathbf{q}_1, \quad (3.3.15)$$

kde

$$\lambda_1 = \mathbf{q}_1^T \mathbf{R} \mathbf{q}_1. \quad (3.3.16)$$

$\mathbf{q}_1$ je tedy vlastním vektorem korelační matice $\mathbf{R}$.

Dosadím-li rovnici (3.3.15) do rovnice (3.3.16), dostanu vztah

$$\lambda_1 = \mathbf{q}_1^T \lambda_1 \mathbf{q}_1 = \lambda_1 \mathbf{q}_1^T \mathbf{q}_1 = \lambda_1 |\mathbf{q}_1|,$$

kde $|\mathbf{q}_1|$ je délka vlastního vektoru $\mathbf{q}_1$, pro kterou z tohoto vztahu vyplývá, že je jednotková, neboli že vektor $\mathbf{q}_1$ je normalizovaný.

$$|\mathbf{q}_1| = 1 \quad (3.3.17)$$

$\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_p$ označím jako p normalizovaných vlastních vektorů korelační matice $\mathbf{R}_{p \times p}$. Všechny tyto vektory jsou řešením rovnice (3.3.14). Ale pouze vlastní vektor náležící nejvyššímu vlastnímu číslu $\lambda_1 = \lambda_{\max}$ je stabilním řešením rekursivní rovnice (3.3.10), ze které je rovnice (3.3.14) odvozena. Tuto skutečnost budu nyní demonstrovat.

Nejdříve budu předpokládat, že po velkém počtu iterací t je vektor vah $\mathbf{w}(t)$ v okolí vlastního vektoru $\mathbf{q}_1$, což vyjádřím vztahem

$$\mathbf{w}(t) = \mathbf{q}_1 + \boldsymbol{\varepsilon}, \quad t \rightarrow \infty, \quad (3.3.18)$$

kde $\boldsymbol{\varepsilon}$ je malý vektor. Na rozdíl od rovnice (3.3.13) budu předpokládat, že počet iterací ještě není tak velký, aby se mohlo psát $E[\mathbf{w}(t+1)] = E[\mathbf{w}(t)]$, ale že $E[\Delta \mathbf{w}(t)] = E[\mathbf{w}(t+1) - \mathbf{w}(t)]$, kde $E[\Delta \mathbf{w}(t)]$ je střední hodnota změny vektoru $\mathbf{w}$ v čase t . Při této demonstraci nám totiž jde právě o chování vektorů působících změnu $\mathbf{w}$.

Vztah (3.3.18) dosadím do rovnice (3.3.13) a dostanu (s přihlédnutím k předchozímu odstavci):

$$E [\Delta \mathbf{w}(t)] = \eta [\mathbf{R} (\mathbf{q}_j + \boldsymbol{\varepsilon}) - (\mathbf{q}_j + \boldsymbol{\varepsilon})^T \mathbf{R} (\mathbf{q}_j + \boldsymbol{\varepsilon}) (\mathbf{q}_j + \boldsymbol{\varepsilon})], \quad t \rightarrow \infty. \quad (3.3.19)$$

Tuto rovnici mohu aproximovat tak, že roznásobím závorky a vypustím členy se dvěma a více výskyty zanedbatelně malého vektoru $\boldsymbol{\varepsilon}$ a využiji toho, že matice $\mathbf{R}$ je symetrická.

$$\begin{aligned} E [\Delta \mathbf{w}(t)] &= \eta [\mathbf{R} \mathbf{q}_j + \mathbf{R} \boldsymbol{\varepsilon} - \mathbf{q}_j^T \mathbf{R} \mathbf{q}_j \mathbf{q}_j - \mathbf{q}_j^T \mathbf{R} \mathbf{q}_j \boldsymbol{\varepsilon} - \mathbf{q}_j^T \mathbf{R} \boldsymbol{\varepsilon} \mathbf{q}_j - \mathbf{q}_j^T \mathbf{R} \boldsymbol{\varepsilon} \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^T \mathbf{R} \mathbf{q}_j \mathbf{q}_j - \boldsymbol{\varepsilon}^T \mathbf{R} \mathbf{q}_j \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^T \mathbf{R} \boldsymbol{\varepsilon} \mathbf{q}_j - \boldsymbol{\varepsilon}^T \mathbf{R} \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}] \\ E [\Delta \mathbf{w}(t)] &= \eta [\mathbf{R} \mathbf{q}_j + \mathbf{R} \boldsymbol{\varepsilon} - \mathbf{q}_j^T \mathbf{R} \mathbf{q}_j \mathbf{q}_j - \mathbf{q}_j^T \mathbf{R} \mathbf{q}_j \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^T \mathbf{R} \mathbf{q}_j \mathbf{q}_j - \boldsymbol{\varepsilon}^T \mathbf{R} \mathbf{q}_j \mathbf{q}_j] \end{aligned}$$

$$E [\Delta \mathbf{w}(t)] = \eta [\mathbf{R} \mathbf{q}_j + \mathbf{R} \boldsymbol{\varepsilon} - \mathbf{q}_j^T \mathbf{R} \mathbf{q}_j \mathbf{q}_j - \mathbf{q}_j^T \mathbf{R} \mathbf{q}_j \boldsymbol{\varepsilon} - 2 \boldsymbol{\varepsilon}^T \mathbf{R} \mathbf{q}_j \mathbf{q}_j], \quad t \rightarrow \infty. \quad (3.3.20)$$

Dále využiji známých vztahů mezi vlastními vektory matice $\mathbf{R}$, o ortogonalitě a normalizaci jejích vlastních vektorů:

$$\lambda_j = \mathbf{q}_j^T \mathbf{R} \mathbf{q}_j, \quad \mathbf{R} \mathbf{q}_j = \lambda_j \mathbf{q}_j \quad \text{a také} \quad \mathbf{q}_j^T \mathbf{R} = \lambda_j \mathbf{q}_j^T, \quad j = 1, 2, \dots, p, \quad (3.3.21)$$

$$\begin{aligned} \mathbf{q}_i^T \mathbf{q}_j &= 1, & \text{když} & \quad i = j & \quad (\text{normalizace}), \\ \mathbf{q}_i^T \mathbf{q}_j &= 0, & \text{když} & \quad i \neq j & \quad (\text{ortogonalita}). \end{aligned} \quad (3.3.22)$$

Nyní mohu zjednodušit rovnici (3.3.20) následovně:

$$E [\Delta \mathbf{w}(t)] = \eta [\lambda_j \mathbf{q}_j + \mathbf{R} \boldsymbol{\varepsilon} - \lambda_j \mathbf{q}_j - \lambda_j \boldsymbol{\varepsilon} - 2 \boldsymbol{\varepsilon}^T \lambda_j \mathbf{q}_j \mathbf{q}_j]$$

$$E [\Delta \mathbf{w}(t)] = \eta [\mathbf{R} \boldsymbol{\varepsilon} - \lambda_j \boldsymbol{\varepsilon} - 2 \lambda_j \mathbf{q}_j \mathbf{q}_j^T \boldsymbol{\varepsilon}], \quad t \rightarrow \infty, \quad (3.3.23)$$

neboť $\boldsymbol{\varepsilon}^T \mathbf{q}_j = \mathbf{q}_j^T \boldsymbol{\varepsilon}$, což je skalár (číslo) a číslo mohu násobit s vektorem $\mathbf{q}_j$ z obou stran. Rovnici (3.3.23) dále vynásobím z levé strany vektorem $\mathbf{q}_j^T$, což je jiný vlastní vektor korelační matice $\mathbf{R}$.

$$\mathbf{q}_i^T E [\Delta \mathbf{w}(t)] = \eta [\mathbf{q}_i^T \mathbf{R} \boldsymbol{\varepsilon} - \lambda_j \mathbf{q}_i^T \boldsymbol{\varepsilon} - 2 \lambda_j \mathbf{q}_i^T \mathbf{q}_j \mathbf{q}_j^T \boldsymbol{\varepsilon}] \quad (3.3.24)$$

V případě, že $i = j$, má pravá strana rovnice (3.3.24) hodnotu

$$\mathbf{q}_j^T E [\Delta \mathbf{w}(t)] = \eta [\lambda_j \mathbf{q}_j^T \boldsymbol{\varepsilon} - \lambda_j \mathbf{q}_j^T \boldsymbol{\varepsilon} - 2 \lambda_j \cdot 1 \cdot \mathbf{q}_j^T \boldsymbol{\varepsilon}] = -2 \eta \lambda_j \mathbf{q}_j^T \boldsymbol{\varepsilon}.$$

V případě, že $i \neq j$, má pravá strana rovnice (3.3.24) hodnotu

$$\mathbf{q}_i^T E [\Delta \mathbf{w}(t)] = \eta [\lambda_i \mathbf{q}_i^T \boldsymbol{\varepsilon} - \lambda_j \mathbf{q}_i^T \boldsymbol{\varepsilon} - 2 \lambda_j \cdot 0 \cdot \mathbf{q}_j^T \boldsymbol{\varepsilon}] = \eta (\lambda_i - \lambda_j) \mathbf{q}_i^T \boldsymbol{\varepsilon}.$$

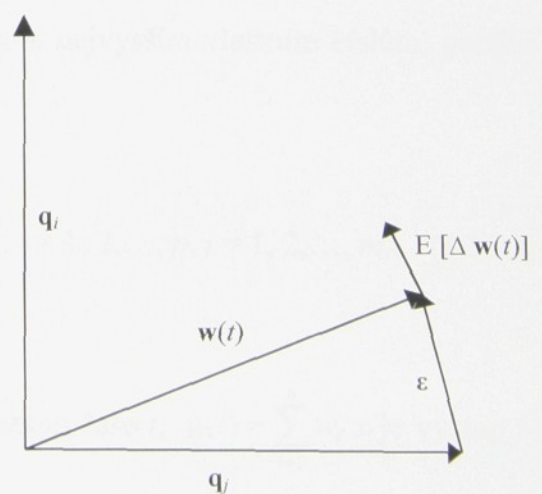
Při odvozování jsem využila rovnice (3.3.21) a (3.3.22).

Nejdříve vyšetřím stav, kdy $i \neq j$ a zároveň $\lambda_i > \lambda_j$. Rovnici

$$\mathbf{q}_i^T E [\Delta \mathbf{w}(t)] = \eta (\lambda_i - \lambda_j) \mathbf{q}_i^T \boldsymbol{\varepsilon} \quad (3.3.25)$$

mohu interpretovat tak, že průmět (daný skalárním součinem) vektoru střední hodnoty změny vektoru $\mathbf{w}$ na $\mathbf{q}_i$ má stejné znaménko jako průmět chybového vektoru $\boldsymbol{\varepsilon}$ na $\mathbf{q}_i$. Parametr η je totiž kladné číslo a také $(\lambda_i - \lambda_j) > 0$ a proto mají oba průměty souhlasné znaménko.

Situaci snáze objasní geometrická interpretace na obrázku 3.3.2. Rovnice $\mathbf{w}(t) = \mathbf{q}_j + \boldsymbol{\varepsilon}$ znamená, že vektor $\boldsymbol{\varepsilon}$ míří směrem od $\mathbf{q}_j$. Mají-li skalární součiny vektorů $\boldsymbol{\varepsilon}$ a $E [\Delta \mathbf{w}(t)]$ s vektorem $\mathbf{q}_i$ stejná znaménka, znamená to, že vektor $E [\Delta \mathbf{w}(t)]$ míří přibližně stejným směrem jako $\boldsymbol{\varepsilon}$. Vektor $\mathbf{w}(t)$ se tak odchyluje od $\mathbf{q}_j$ stále více a to směrem k vektoru $\mathbf{q}_i$, který patří vyššímu vlastnímu číslu a je kolmý na vektor $\mathbf{q}_j$.



Obrázek 3.3.2
Divergence od vektoru $\mathbf{q}_j$

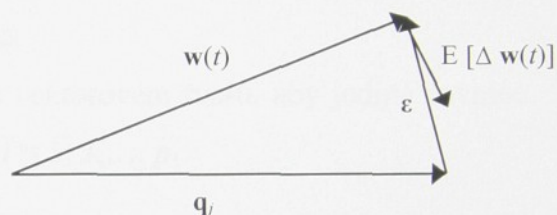
Pokud bude parametr η záporné číslo blízké nule a λ_j nejnižší vlastní číslo, vektor $\mathbf{w}(t)$ bude konvergovat k vektoru $\mathbf{q}_j$, protože vektor $E[\Delta \mathbf{w}(t)]$ bude mířit opačným směrem.

Nakonec vyšetřím stav, kdy $i = j$. V takové situaci má rovnice (3.3.24) podobu

$$\mathbf{q}_j^T E[\Delta \mathbf{w}(t)] = -2\eta \lambda_j \mathbf{q}_j^T \boldsymbol{\varepsilon}. \quad (3.3.26)$$

Skalární součiny vektorů $\boldsymbol{\varepsilon}$ a $E[\Delta \mathbf{w}(t)]$ s vektorem $\mathbf{q}_j$ mají opačné znaménko. Vektor střední hodnoty změny vektoru $\mathbf{w}$ a chybový vektor $\boldsymbol{\varepsilon}$ tedy míří přibližně opačným směrem.

Z obrázku 3.3.3 je zřejmé, že vektor $\mathbf{w}(t)$ se díky skládání s vektorem $E[\Delta \mathbf{w}(t)]$ přibližuje k vektoru $\mathbf{q}_j$. Pokud λ_j bude nejvyšším vlastním číslem $\lambda_1 = \lambda_{\max}$, bude rovnice (3.3.24) konvergovat i v případě, že $i \neq j$, protože výraz $(\lambda_i - \lambda_j)$ bude záporný.



Obrázek 3.3.3
Konvergence k vektoru $\mathbf{q}_j$

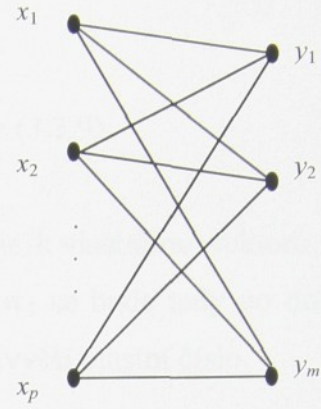
Tím jsem dokázala, že rovnice (3.3.24) a tedy i (3.3.10) konverguje k vlastnímu vektoru odpovídajícímu maximálnímu vlastnímu číslu $\lambda_j = \lambda_{\max}$ korelační matice $\mathbf{R}$.

Abych našla m vlastních vektorů náležejících m nejvyšším vlastním číslům, použiji vzorec pro zobecněnou formu Hebbova učení

$$\Delta w_{ji}(t) = \eta \left[y_j(t) x_i(t) - y_j(t) \sum_{k=1}^j w_{ki}(t) y_k(t) \right], \quad i = 1, 2, \dots, p, j = 1, 2, \dots, m, \quad (3.3.27)$$

kde $\Delta w_{ji}(t)$ je změna synaptické váhy $w_{ji}(t)$ v diskrétním čase t , $y_j(t) = \sum_{i=1}^p w_{ji} x_i$ je výstup j -tého neuronu a η , parametr rychlosti učení, je kladné číslo blízké nule. Počáteční hodnoty synaptických vah w_{ji} jsou náhodné. Vzorec (3.3.27) publikoval Sanger v roce 1989.

V následující části práce ukáži, proč iterativní použití vzorce (3.3.27) na p -rozměrný vektor $\mathbf{x}$, (což je ve skutečnosti množina vektorů znormovaných tak, aby aritmetický průměr jejich souhlasných složek byl roven nule), vede k nalezení m vlastních vektorů náležejících m nejvyšším vlastním číslům korelační matice $\mathbf{R} = E(\mathbf{x}\mathbf{x}^T)$. Obrázek 3.3.4 schematicky vyjadřuje tuto úlohu. Na rozdíl od obrázku 3.3.1 je zde m výstupních uzlů (neboli neuronů), z nichž každý je napojen na p složek vstupního vektoru $\mathbf{x}$.



Obrázek 3.3.4 Pramen: [1]
Jednovrstevná síť neuronů y

Z rovnice (3.3.27) vytknu $y_j(t)$ a přepíšu ji ve vektorovém tvaru, aby jediná rovnice vyjadřovala úpravu vah $\mathbf{w}_j$ pro vektor $\mathbf{x}$ se složkami $i = 1, 2, \dots, p$.

$$\Delta \mathbf{w}_j(t) = \eta y_j(t) \left[\mathbf{x}(t) - \sum_{k=1}^j \mathbf{w}_k(t) y_k(t) \right], \quad j = 1, 2, \dots, m. \quad (3.3.28)$$

V dalším kroku uvedu rovnici (3.3.28) do tvaru blízkého rovnici (3.3.9), ze které nyní vyjádřím člen $\Delta \mathbf{w} = \mathbf{w}(t+1) - \mathbf{w}(t)$.

$$\Delta \mathbf{w}_j(t) = \eta y_j(t) [\mathbf{x}'(t) - \mathbf{w}_j(t) y_j(t)], \quad (3.3.29)$$

kde

$$\mathbf{x}'(t) = \mathbf{x}(t) - \sum_{k=1}^{j-1} \mathbf{w}_k(t) y_k(t). \quad (3.3.30)$$

Vektor $\mathbf{x}'(t)$ v rovnici (3.3.30) je pozměněným vstupním vektorem a funkcí indexu j . Když se index j rovná 1, potom rovnice (3.3.30) přejde do tvaru

$$j = 1: \quad \mathbf{x}'(t) = \mathbf{x}(t) \quad (3.3.31)$$

a rovnice (3.3.29) bude identická s hodnotou $\Delta \mathbf{w}$ z rovnice (3.3.9).

Jak jsem výše dokázala, rovnice (3.3.9) konverguje k vlastnímu vektoru patřícímu nejvyššímu vlastnímu číslu korelační matice $\mathbf{R}$. Vektor $\mathbf{w}_1$ se bude tedy po dostatečném počtu iterací blížit vlastnímu vektoru matice $\mathbf{R}$ pro její nejvyšší vlastní číslo.

Další hodnoty indexu j dávají:

$$\begin{aligned} j = 2: \quad & \mathbf{x}''(t) = \mathbf{x}(t) - \mathbf{w}_1(t) y_1(t) = \mathbf{x}'(t) - \mathbf{w}_1(t) y_1(t) \\ j = 3: \quad & \mathbf{x}'''(t) = \mathbf{x}(t) - \mathbf{w}_1(t) y_1(t) - \mathbf{w}_2(t) y_2(t) = \mathbf{x}''(t) - \mathbf{w}_2(t) y_2(t) \\ j = 4: \quad & \mathbf{x}''''(t) = \mathbf{x}(t) - \mathbf{w}_1(t) y_1(t) - \mathbf{w}_2(t) y_2(t) - \mathbf{w}_3(t) y_3(t) = \mathbf{x}'''(t) - \mathbf{w}_3(t) y_3(t) \\ & \vdots \\ j = m: \quad & \mathbf{x}^{(m)}(t) = \mathbf{x}(t) - \mathbf{w}_1(t) y_1(t) - \mathbf{w}_2(t) y_2(t) - \dots - \mathbf{w}_{m-1}(t) y_{m-1}(t) = \mathbf{x}^{(m-1)}(t) - \mathbf{w}_{m-1}(t) y_{m-1}(t) \end{aligned} \quad (3.3.32)$$

Nyní dokáži, že modifikovaná matice $\mathbf{R}'' = E(\mathbf{x}''\mathbf{x}''^T)$ se liší od matice $\mathbf{R} = E(\mathbf{x}\mathbf{x}^T)$, viz rovnice (3.2.9), tím, že namísto nejvyššího vlastního čísla má nulu a ostatní vlastní čísla má stejné jako v $\mathbf{R}$.

Modifikovaná matice $\mathbf{R}''$, poté co vektor $\mathbf{w}_1$ již dosáhl hodnoty prvního vlastního vektoru pro $\lambda_1 = \lambda_{\max}$, má hodnotu

$$\begin{aligned} \mathbf{R}'' &= E[\mathbf{x}''\mathbf{x}''^T] = E[(\mathbf{x} - \mathbf{w}_1 y_1)(\mathbf{x} - \mathbf{w}_1 y_1)^T] = E[\mathbf{x}\mathbf{x}^T - \mathbf{x}(\mathbf{w}_1 y_1)^T - \mathbf{w}_1 y_1 \mathbf{x}^T + \mathbf{w}_1 y_1 (\mathbf{w}_1 y_1)^T] = \\ &= E[\mathbf{x}\mathbf{x}^T - \mathbf{x}(\mathbf{w}_1 \mathbf{w}_1^T \mathbf{x})^T - \mathbf{w}_1 \mathbf{w}_1^T \mathbf{x} \mathbf{x}^T + \mathbf{w}_1 \mathbf{w}_1^T \mathbf{x} (\mathbf{w}_1 \mathbf{w}_1^T \mathbf{x})^T], \end{aligned}$$

protože podle vzorce (3.3.8) je $y_1 = \mathbf{w}_1^T \mathbf{x} = \mathbf{x}^T \mathbf{w}_1$. Následují další úpravy vztahu pro matici $\mathbf{R}''$:

$$\begin{aligned} \mathbf{R}'' &= E[\mathbf{x}\mathbf{x}^T - \mathbf{x} \mathbf{x}^T \mathbf{w}_1 \mathbf{w}_1^T - \mathbf{w}_1 \mathbf{w}_1^T \mathbf{x} \mathbf{x}^T + \mathbf{w}_1 \mathbf{w}_1^T \mathbf{x} \mathbf{x}^T \mathbf{w}_1 \mathbf{w}_1^T], \\ \mathbf{R}'' &= \mathbf{R} - \mathbf{R} \mathbf{w}_1 \mathbf{w}_1^T - \mathbf{w}_1 \mathbf{w}_1^T \mathbf{R} + \mathbf{w}_1 \mathbf{w}_1^T \mathbf{R} \mathbf{w}_1 \mathbf{w}_1^T = \mathbf{R} - 2\mathbf{R} \mathbf{w}_1 \mathbf{w}_1^T + \mathbf{w}_1 \mathbf{w}_1^T \mathbf{R} \mathbf{w}_1 \mathbf{w}_1^T. \end{aligned}$$

Abych zjistila vlastní číslo matice $\mathbf{R}''$ pro vlastní vektor $\mathbf{w}_1$ matice $\mathbf{R}$, nejdříve ji vynásobím zleva i zprava tímto vlastním vektorem $\mathbf{w}_1$ pro maximální vlastní číslo matice $\mathbf{R}$:

$$\begin{aligned}\mathbf{w}_1^T \mathbf{R}'' \mathbf{w}_1 &= \mathbf{w}_1^T (\mathbf{R} - 2\mathbf{R} \mathbf{w}_1 \mathbf{w}_1^T + \mathbf{w}_1 \mathbf{w}_1^T \mathbf{R} \mathbf{w}_1 \mathbf{w}_1^T) \mathbf{w}_1 = \\ &= \mathbf{w}_1^T \mathbf{R} \mathbf{w}_1 - 2 \mathbf{w}_1^T \mathbf{R} \mathbf{w}_1 \mathbf{w}_1^T \mathbf{w}_1 + \mathbf{w}_1^T \mathbf{w}_1 \mathbf{w}_1^T \mathbf{R} \mathbf{w}_1 \mathbf{w}_1^T \mathbf{w}_1 = \\ &= \lambda_1 - 2 \lambda_1 + \lambda_1 = 0.\end{aligned}\tag{3.3.33}$$

Násobím-li matici $\mathbf{R}''$ nějakým jiným vlastním vektorem $\mathbf{w}_j$, kde $j \neq 1$, matici $\mathbf{R}$, dostanu

$$\begin{aligned}\mathbf{w}_j^T \mathbf{R}'' \mathbf{w}_j &= \mathbf{w}_j^T (\mathbf{R} - 2\mathbf{R} \mathbf{w}_1 \mathbf{w}_1^T + \mathbf{w}_1 \mathbf{w}_1^T \mathbf{R} \mathbf{w}_1 \mathbf{w}_1^T) \mathbf{w}_j = \\ &= \mathbf{w}_j^T \mathbf{R} \mathbf{w}_j - 2 \mathbf{w}_j^T \mathbf{R} \mathbf{w}_1 \mathbf{w}_1^T \mathbf{w}_j + \mathbf{w}_j^T \mathbf{w}_1 \mathbf{w}_1^T \mathbf{R} \mathbf{w}_1 \mathbf{w}_1^T \mathbf{w}_j = \\ &= \lambda_j.\end{aligned}\tag{3.3.34}$$

Při těchto výpočtech jsem využila vztahů (3.2.22) a (3.2.19), ($\mathbf{u}_1 = \mathbf{w}_1$).

Tím jsem dokázala, že korelační matice $\mathbf{R}''$ z pozměněného vektoru $\mathbf{x}''$ je stejná jako matice $\mathbf{R}$ až na to, že má oproti matici $\mathbf{R}$ vynulováno nejvyšší vlastní číslo λ_1 .

Vzorec (3.3.29) tedy provádí takové úpravy vektoru $\mathbf{w}_2$, že ten se blíží vlastnímu vektoru náležejícímu druhému nejvyššímu vlastnímu číslu matice $\mathbf{R}$, když index $j = 2$. Analogicky, když $j = 3$, potom příslušný modifikovaný vektor $\mathbf{x}'''$ tvoří korelační matici $\mathbf{R}'''$. Podle vztahu (3.3.32) vektor $\mathbf{x}''' = \mathbf{x}'' - \mathbf{w}_2 y_2$. V matici $\mathbf{R}'''$ oproti $\mathbf{R}$ chybí dvě nejvyšší vlastní čísla matice $\mathbf{R}$. Z toho plyne, že vzorec (3.3.29) potom hledá vlastní vektor $\mathbf{w}_3$ patřící třetímu nejvyššímu vlastnímu číslu matice $\mathbf{R}$ poté co se $\mathbf{w}_2$ přiblížil ke druhému nejvyššímu vlastnímu vektoru matice $\mathbf{R}$ a $\mathbf{w}_1$ nejvyššímu vlastnímu číslu matice $\mathbf{R}$. A tak dále pro vyšší hodnoty indexu j .

Tím jsem dokázala, že vzorec (3.3.27) upravuje váhy w_{ji} tak, že konvergují ke složkám m vlastních vektorů pro m nejvyšších vlastních čísel matice $\mathbf{R}$.

Vzorce (3.3.33) a (3.3.34) můžeme nahradit následující úvahou: Vektor $\mathbf{x}''$ vznikne tak, že se od vektoru $\mathbf{x}$ odečte $\mathbf{w}_1 y_1$, tedy součin prvního vlastního vektoru a délky průmětu vektoru $\mathbf{x}$ na něj. Výše jsem ukázala že vektor $\mathbf{x}$ je sumou součinů $\mathbf{w}_j y_j$, kde $\mathbf{w}_j$ jsou vlastní vektory patřící několika nejvyšším vlastním číslům. To je vystiženo vztahem (3.2.27), ($a_j = y_j$, $\mathbf{u}_j = \mathbf{w}_j$). Když tedy od vektoru $\mathbf{x}$ jeden z těchto součinů odečtu, znamená to, že jej tak zbavuji jednoho z rozměrů a to ve směru odčítaného vlastního vektoru.

Dále se budu zabývat některými zvláštnostmi algoritmu GHA.

Jak vyplývá z analýzy rovnice (3.3.25), pokud bude parametr η malé záporné číslo, dalo by se očekávat, že vzorec (3.3.27) najde m vlastních vektorů patřících m nejnižším vlastním číslům. To je takzvané *antihebbiánské učení*. Při praktickém provádění GHA však všech m vektorů konverguje k jedinému vlastnímu vektoru pro nejnižší vlastní číslo. Tento jev si vysvětluji následovně:

1. Vektory konvergují postupně. Druhý vektor pro druhé nejnižší vlastní číslo se začíná formovat dříve, než je dostatečně přesně vypočítán vlastní vektor pro nejnižší vlastní číslo a tak se od vektoru $\mathbf{x}$ zpočátku odečítá něco velmi nepřesného, což má ten samý účinek, jako by se neodečítalo nic.
2. Rozdíly ve vlastních číslech korelační matice jsou většinou značné. $\lambda_{\max}$ může být řádově 10^2 a $\lambda_{\min}$ řádově 10^{-1} . Vlastní vektor pro nejnižší vlastní číslo je směr, na který mají datové vektory nejnižší rozptyl svých průmětů. To znamená, že jsou na tento směr v průměru nejvíce kolmé. Po odečtení součinu nejnižšího vlastního vektoru a průmětu vektoru $\mathbf{x}$ na něj se vektor $\mathbf{x}$ zbaví svého rozměru ve směru nejnižšího vlastního vektoru, což znamená, že začne být na tento směr kolmý. Ale již předtím byl na tento směr téměř kolmý a tak se příliš nezmění.

Pokud GHA při antihebbiánském učení vychází ze stavu, kde 1. vlastní vektor je pro nejnižší vlastní číslo, 2. vlastní vektor je pro 2. nejnižší vlastní číslo a tak dále až k m -tému vlastnímu vektoru pro m -té nejnižší vlastní číslo, metoda GHA tento stav při svých

iteracích zachová, pokud parametr η není ve své absolutní hodnotě příliš vysoký, aby tento stav rozkolísal a začal konvergovat pouze k nejnižšímu vlastnímu vektoru.

Když je $\eta > 0$, vektory pro nejvyšší vlastní čísla se diferencují správně z jakýchkoli počátečních pozic.

Shrnutí a porovnání GHA s Jacobiho metodou

Zobecněný Hebbův algoritmus (generalized Hebbian algorithm), dále jen GHA, v praxi najde prvních m vlastních vektorů srovnaných sestupně podle jim náležejících velikostí vlastních čísel korelační matice $\mathbf{R}$. Tyto vektory konvergují současně, ale 2. neuron, neboli 2. vlastní vektor, pravděpodobně správně nezkonverguje, dokud se 1. neuron alespoň částečně nepřiblíží ke správné hodnotě 1. vlastního vektoru.

Výhodou použití metody GHA pro analýzu hlavních komponent je to, že se nemusíme zdržovat výpočtem korelační matice $\mathbf{R}$ a všech jejích vlastních vektorů. Pro malé rozměry vstupního vektoru $\mathbf{x}$ je výhodnější použít jiné metody, které počítají vše a přesto jsou rychlejší, jako je například Jacobiho metoda, o které pojednávám v následující kapitole. Od určitého rozměru vektoru $\mathbf{x}$ však začne být výhodnější použití GHA, který bude opakovaně zpracovávat jednotlivé exempláře vektoru $\mathbf{x}$ v náhodném pořadí a upravovat zpočátku náhodné hodnoty m vlastních vektorů tak, aby konvergovaly k vlastním vektorům pro m nejvyšších vlastních čísel. Podmínkou výhodnosti dále je, že m je podstatně menší než rozměr $\mathbf{x}$. V porovnání s Jacobiho metodou metoda GHA při realizaci na počítači potřebuje podstatně méně paměti.

3.4 Jacobiho metoda nalezení vlastních čísel a vlastních vektorů symetrické matice [2]

Autorem této metody z roku 1846 je německý matematik, který významně ovlivnil mnoho odvětví matematiky, Karl Gustav Jacob (Jacobi) (1804 – 1851). [4]

Cílem je nalézt vlastní čísla a vlastní vektory reálné symetrické matice $\mathbf{A}$ řádu n . Za tím účelem se matice $\mathbf{A}$ bude iterativně násobit zprava a zleva maticí $\mathbf{P}_{pq}$.

$$\mathbf{A}' = \mathbf{P}_{pq}^T \cdot \mathbf{A} \cdot \mathbf{P}_{pq} \quad (3.4.1)$$

$\mathbf{P}_{pq}$ je ortogonální matice. Všechny její prvky na hlavní diagonále jsou rovny číslu 1 kromě prvku na p -tém řádku a sloupku a prvku na q -tém řádku a sloupku. Tyto dva prvky jsou rovny kosinu úhlu φ . Všechny její prvky mimo hlavní diagonálu jsou rovny číslu 0 kromě prvku na p -tém řádku a q -tém sloupku, to je sinus úhlu φ , a kromě prvku na q -tém řádku a p -tém sloupku, což je minus sinus úhlu φ . Platí, že $p < q$. Pro zjednodušení zápisu dalších výpočtů zavedu označení s pro sinus φ , c pro kosinus φ a t pro tangens φ . Platí tedy, že $s^2 + c^2 = 1$ a $t = s/c$. Příkladem matice $\mathbf{P}_{pq}$ může být:

$$\mathbf{P}_{24} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & c & 0 & s \\ 0 & 0 & 1 & 0 \\ 0 & -s & 0 & c \end{bmatrix}$$

Provedeme-li výpočet (3.4.1), dostaneme symetrickou matici $\mathbf{A}'$, která se od původní matice $\mathbf{A}$ liší v p -tém a q -tém řádku a sloupku. Vzhledem k ortogonalitě matice $\mathbf{P}$ platí, že $\mathbf{P}^T = \mathbf{P}^{-1}$. Rovnici (3.4.1) tedy můžeme chápat jako změnu báze matice $\mathbf{A}$ podle vzorce $\mathbf{P}^{-1} \mathbf{A} \mathbf{P}$. Matice $\mathbf{A}$ tak přejde do nového souřadného systému, který je vůči původnímu otočen o úhel φ . Matice $\mathbf{A}'$ je podobná matici $\mathbf{A}$, má tedy stejná vlastní čísla.

Pro prvky mimo hlavní diagonálu v p -tém řádku a sloupku platí rovnice

$$a'_{ip} = ca_{ip} - sa_{iq}, \quad i = 1, 2, \dots, n, \quad i \neq p, \quad i \neq q. \quad (3.4.2)$$

Pro prvky mimo hlavní diagonálu v q -tém řádku a sloupku platí rovnice

$$a'_{iq} = ca_{iq} + sa_{ip}, \quad i = 1, 2, \dots, n, \quad i \neq p, \quad i \neq q. \quad (3.4.3)$$

Pro diagonální prvek v p -tém řádku a sloupku platí rovnice

$$a'_{pp} = c^2 a_{pp} + s^2 a_{qq} - 2sca_{pq} \quad (3.4.4)$$

Pro diagonální prvek v q -tém řádku a sloupku platí rovnice

$$a'_{qq} = s^2 a_{pp} + c^2 a_{qq} + 2sca_{pq} \quad (3.4.5)$$

Pro prvky mimo hlavní diagonálu na p -tém řádku a q -tém sloupku a na q -tém řádku a p -tém sloupku platí rovnice

$$a'_{pq} = a'_{qp} = (c^2 - s^2) a_{pq} + sc (a_{pp} - a_{qq}) \quad (3.4.6)$$

Zvolíme vždy takový úhel φ , pro který bude prvek a'_{pq} roven nule. Při dalším násobení výsledné matice $\mathbf{A}'$ zprava a zleva jinou maticí $\mathbf{P}_{pq}$ podle vztahu (3.4.1) se předchozí nuly zase odstraní, ale suma čtverců hodnot prvků mimo hlavní diagonálu se bude stále zmenšovat, jak dále dokáží. Položím-li $a'_{pq} = 0$, dostanu z rovnice (3.4.6) vztah

$$\theta = \cot 2\varphi = \frac{c^2 - s^2}{2sc} = \frac{a_{qq} - a_{pp}}{2a_{pq}}, \quad (3.4.7)$$

kde jsem využila známých goniometrických vztahů

$$\begin{aligned} \sin 2\varphi &= 2sc, \\ \cos 2\varphi &= c^2 - s^2, \\ \cotg 2\varphi &= (c^2 - s^2) / 2sc. \end{aligned}$$

Číslo θ dám do vztahu s tangentem úhlu $\varphi = t = s/c$:

$$\theta = \frac{c^2}{2sc} - \frac{s^2}{2sc} = \frac{c}{2s} - \frac{s}{2c} = \frac{1}{2t} - \frac{t}{2} \quad / \cdot 2t$$

$$2t\theta = 1 - t^2$$

$$t^2 + 2t\theta - 1 = 0 \quad (3.4.8)$$

Kořen kvadratické rovnice (3.4.8) s nižší absolutní hodnotou koresponduje s rotačním úhlem nižším než $\pi/4$. Tato volba vede k nejstabilnější redukci prvků mimo hlavní diagonálu. Řeším-li rovnici (3.4.8), dostanu

$$\frac{-2\theta \pm \sqrt{4\theta^2 + 4}}{2} = \frac{-2\theta \pm 2\sqrt{\theta^2 + 1}}{2} = -\theta \pm \sqrt{\theta^2 + 1} = t_{1,2}.$$

Abych vyjádřila kořen této rovnice s nižší absolutní hodnotou, budu tento výsledek dále upravovat:

$$t_{1,2} = \frac{-\theta \pm \sqrt{\theta^2 + 1}}{1} \cdot \frac{-\theta \mp \sqrt{\theta^2 + 1}}{-\theta \mp \sqrt{\theta^2 + 1}} = \frac{\theta^2 - \theta^2 - 1}{-\theta \mp \sqrt{\theta^2 + 1}} = \frac{1}{\theta \pm \sqrt{\theta^2 + 1}}.$$

Podle této rovnice je t s nižší absolutní hodnotou rovno

$$t = \frac{\text{sign}(\theta)}{|\theta| + \sqrt{\theta^2 + 1}} = \text{sign}(\theta) \cdot (\sqrt{\theta^2 + 1} - |\theta|), \quad (3.4.9)$$

kde $\text{sign}(\theta)$ znamená číslo 1 se znaménkem patřícím číslu θ , viz (3.1.15.).

V průběhu výpočtů absolutní hodnota θ roste, jak prvky mimo hlavní diagonálu a_{pq} klesají. Když θ naroste tak, že θ^2 by v počítači přeteklo, použijeme limitu rovnice (3.4.9) pro $\theta \rightarrow \infty$:

$$t = \frac{1}{2\theta}$$

Pomocí známých goniometrických vztahů vyjádřím pomocí t hodnoty c a s :

$$c = \frac{1}{\sqrt{t^2 + 1}} \quad (3.4.10)$$

$$s = tc \quad (3.4.11)$$

Dále vyjádřím rovnice (3.4.2), (3.4.3), (3.4.4) a (3.4.5) pomocí předchozí hodnoty a její korekce:

$$a'_{ip} = a_{ip} - s \left(a_{iq} - \frac{c}{s} a_{ip} + \frac{1}{s} a_{ip} \right) = a_{ip} - s \left(a_{iq} + \frac{1-c}{s} a_{ip} \right) = a_{ip} - s(a_{iq} + \tau a_{ip}), \quad (3.4.12)$$

$$a'_{iq} = a_{iq} + s \left(a_{ip} + \frac{c}{s} a_{iq} - \frac{1}{s} a_{iq} \right) = a_{iq} + s \left(a_{ip} - \frac{1-c}{s} a_{iq} \right) = a_{iq} + s(a_{ip} + \tau a_{iq}), \quad (3.4.13)$$

kde

$$\tau = \frac{1-c}{s} \cdot \frac{1+c}{1+c} = \frac{1-c^2}{s+sc} = \frac{s^2}{s+sc} = \frac{s}{1+c} = \operatorname{tg} \frac{\varphi}{2},$$

neboť platí goniometrické vztahy

$$\cos^2 \frac{\varphi}{2} + \sin^2 \frac{\varphi}{2} = 1,$$

$$\cos^2 \frac{\varphi}{2} - \sin^2 \frac{\varphi}{2} = c,$$

které po vzájemné kombinaci dají

$$2 \cos^2 \frac{\varphi}{2} = 1 + c,$$

$$2 \sin^2 \frac{\varphi}{2} = 1 - c,$$

$$\operatorname{tg} \frac{\varphi}{2} = \sqrt{\frac{2 \sin^2 \frac{\varphi}{2}}{2 \cos^2 \frac{\varphi}{2}}} = \sqrt{\frac{1-c}{1+c} \cdot \frac{1+c}{1+c}} = \sqrt{\frac{1-c^2}{(1+c)^2}} = \sqrt{\frac{s^2}{(1+c)^2}} = \frac{s}{1+c}.$$

Pro zjednodušení rovnic (3.4.4) a (3.4.5) použijí rovnici (3.4.6), ($a'_{pq} = 0$).

$$0 = (c^2 - s^2) a_{pq} + sc (a_{pp} - a_{qq})$$

$$sc (a_{qq} - a_{pp}) = (c^2 - s^2) a_{pq}$$

$$a_{qq} - a_{pp} = \left(\frac{c}{s} - \frac{s}{c} \right) a_{pq}$$

Z rovnice (3.4.4) vyjádřím a_{qq}

$$a_{qq} = \frac{1}{s^2} a'_{pp} - \frac{c^2}{s^2} a_{pp} + \frac{2sc}{s^2} a_{pq}$$

a sestavím rovnici z toho, čemu se rovná a_{qq} v rovnicích (3.4.4) a (3.4.6).

$$\frac{1}{s^2} a'_{pp} - \frac{c^2}{s^2} a_{pp} + \frac{2sc}{s^2} a_{pq} = a_{pp} + \left(\frac{c}{s} - \frac{s}{c} \right) a_{pq} \quad / \cdot s^2$$

$$a'_{pp} - c^2 a_{pp} + 2sca_{pq} = s^2 a_{pp} + \left(sc - \frac{s^3}{c} \right) a_{pq}$$

$$a'_{pp} = (s^2 + c^2) a_{pp} - \left(2sc - sc + \frac{s^3}{c} \right) a_{pq}$$

$$a'_{pp} = a_{pp} - t a_{pq} \quad (3.4.14)$$

Z rovnice (3.4.5) vyjádřím a_{pp}

$$a_{pp} = \frac{1}{s^2} a'_{qq} - \frac{c^2}{s^2} a_{qq} - \frac{2sc}{s^2} a_{pq}$$

a sestavím rovnici z toho, čemu se rovná a_{pp} v rovnicích (3.4.5) a (3.4.6).

$$\frac{1}{s^2} a'_{qq} - \frac{c^2}{s^2} a_{qq} - \frac{2sc}{s^2} a_{pq} = a_{qq} - \left(\frac{c}{s} - \frac{s}{c} \right) a_{pq} \quad / \cdot s^2$$

$$a'_{qq} - c^2 a_{qq} - 2sca_{pq} = s^2 a_{qq} - \left(sc - \frac{s^3}{c} \right) a_{pq}$$

$$a'_{qq} = (s^2 + c^2) a_{qq} + \left(2sc - sc + \frac{s^3}{c} \right) a_{pq}$$

$$a'_{qq} = a_{qq} - t a_{pq} \quad (3.4.15)$$

Neboť

$$2sc - sc + \frac{s^3}{c} = sc + \frac{s^3}{c} = \frac{s}{c} (c^2 + s^2) = t.$$

A konečně z rovnice (3.4.6) dostanu díky volbě vhodného úhlu φ

$$a'_{pq} = 0. \quad (3.4.16)$$

Takovéto vyjádření nových prvků matice $\mathbf{A}'$ v p -tém a q -tém řádku a sloupku minimalizuje zaokrouhlovací chyby. Vzhledem k symetrii matice $\mathbf{A}$ podle hlavní diagonály doporučuji při počítačovém zpracování postupně aktualizovat prvky pouze nad diagonálou

nebo pouze pod diagonálou a aktualizované prvky nejdříve „odkládat“ na druhou stranu matice, dokud se nevypočtou všechny nové prvky pro danou jednu transformaci $\mathbf{P}_{pq}$.

Nyní dokáží konvergenci Jacobiho metody do stavu, kdy mimodiagonální prvky se budou blížit nule a na hlavní diagonále zbydou vlastní čísla.

Vzhledem k tomu, že matice $\mathbf{P}_{pq}$ je ortogonální, bude suma čtverců všech prvků matice $\mathbf{A}'$ z rovnice (3.4.1) stejná jako suma čtverců matice $\mathbf{A}$. Násobení ortogonální maticí, jak zprava, tak i zleva, zachovává délku vektoru, tj. řádku nebo sloupku matice $\mathbf{A}$. Tato délka je rovna odmocnině ze sumy čtverců prvků původního vektoru.

Suma čtverců mimodiagonálních prvků matice $\mathbf{A}'$ je však nižší o $2a_{pq}^2$, neboť podle rovnic (3.4.2) a (3.4.3) je čtverec změněných prvků na řádku a sloupku p a q , když $p \neq q$

$$a'_{ip}{}^2 = (ca_{ip} - sa_{iq})^2 = c^2 a_{ip}^2 - 2sca_{ip}a_{iq} + s^2 a_{iq}^2,$$

$$a'_{iq}{}^2 = (ca_{iq} + sa_{ip})^2 = c^2 a_{iq}^2 + 2sca_{ip}a_{iq} + s^2 a_{ip}^2.$$

Čtverce prvků a'_{ip} a a'_{iq} můžeme sečíst, protože je jich v matici $\mathbf{A}'$ stejný počet $2(n-1)$:

$$a'_{ip}{}^2 + a'_{iq}{}^2 = a_{ip}^2 (s^2 + c^2) + a_{iq}^2 (s^2 + c^2) = a_{ip}^2 + a_{iq}^2.$$

Suma čtverců prvků a_{ip} a a_{iq} , kde $p \neq q$, $i \neq p$, $i \neq q$ se tedy transformací (3.4.1) nemění. Zbývajících mimodiagonálních prvků jsou a_{pq} , a_{qp} , které se transformací (3.4.1) vynulují, viz rovnice (3.4.6) a (3.4.16), jejich čtverce se tedy transformací (3.4.1) sníží na nulu. Suma čtverců mimodiagonálních prvků matice $\mathbf{A}'$ tedy při každém provedení rovnice (3.4.1) klesne o $2a_{pq}^2$, což vede k postupnému přibližování všech mimodiagonálních prvků k nule. Tím jsem dokázala konvergenci Jacobiho metody.

Suma prvků na hlavní diagonále, takzvaná *stopa matice*, se ortogonální transformací (3.4.1) nemění, avšak suma čtverců těchto prvků roste při každé iteraci o $2a_{pq}^2$, protože

jinak by nebyla splněna podmínka zachování sumy čtverců *všech* prvků matice A' . V konečném stavu je suma čtverců prvků na hlavní diagonále zvýšena oproti původní hodnotě v počáteční matici A o sumu čtverců počátečních mimodiagonálních prvků.

Rozepíšu-li iterační proces podle rovnice (3.4.1), dostanu

$$A'_1 = P_1^T \cdot A_1 \cdot P_1,$$

$$A'_2 = P_2^T \cdot A'_1 \cdot P_2 = P_2^T \cdot P_1^T \cdot A_1 \cdot P_1 \cdot P_2,$$

$$A'_3 = P_3^T \cdot A'_2 \cdot P_3 = P_3^T \cdot P_2^T \cdot P_1^T \cdot A_1 \cdot P_1 \cdot P_2 \cdot P_3$$

a tak dále.

Proces spěje do stavu, kdy A' se blíží diagonální matici Λ s vlastními čísly na hlavní diagonále.

$$\Lambda = V^T \cdot A \cdot V, \quad (3.4.17)$$

kde

$$V = P_1 \cdot P_2 \cdot P_3 \cdot \dots \cdot P_k, \quad k \rightarrow \infty. \quad (3.4.18)$$

Jelikož matice V vznikla vynásobením ortogonálních matic, je také ortogonální. Platí $V^T = V^{-1}$ a tedy

$$V\Lambda = AV.$$

Sloupce matice V jsou tedy normalizované (jednotkové) vlastní vektory a na hlavní diagonále matice Λ jsou vlastní čísla matice A . Při výpočtu může být matice V počítána iterativně podle vzorce

$$\mathbf{V}' = \mathbf{V} \cdot \mathbf{P}_{pq}, \quad (3.4.19)$$

přičemž počáteční matice $\mathbf{V}$ je jednotková.

Při operaci (3.4.19) se v matici $\mathbf{V}$ změní pouze p -tý a q -tý sloupek. Vzhledem k tomu, že matice $\mathbf{P}_{pq}$ je ortogonální, řádky a sloupky matice $\mathbf{V}$ zůstávají jednotkové. Pro prvky v matici $\mathbf{V}$ platí rovnice:

$$v'_{ij} = v_{ij}, \quad j \neq p, \quad j \neq q, \quad (3.4.20)$$

$$v'_{ip} = cv_{ip} - sv_{iq} = v_{ip} - s \left(v_{iq} - v_{ip} \frac{c}{s} + v_{ip} \frac{1}{s} \right) = v_{ip} - s \left(v_{iq} + v_{ip} \frac{1-c}{s} \right) = v_{ip} - s(v_{iq} + \tau v_{ip}) \quad (3.4.21)$$

$$v'_{iq} = sv_{ip} + cv_{iq} = v_{iq} + s \left(v_{ip} + v_{iq} \frac{c}{s} - v_{iq} \frac{1}{s} \right) = v_{iq} + s \left(v_{ip} - v_{iq} \frac{1-c}{s} \right) = v_{iq} + s(v_{ip} - \tau v_{iq}) \quad (3.4.22)$$

Shrnutí

Jacobiho algoritmus konverguje do stavu, kdy na hlavní diagonále zpracovávané symetrické matice zůstanou její vlastní čísla, ostatní prvky jsou rovny nule a součin všech transformačních matic činí matici, jejímiž sloupky jsou jednotkové vlastní vektory zpracovávané matice.

3.5 Jackknife [5]

Jackknife je neparametrická metoda odhadu rozptylů umožňující stanovit také odpovídající intervaly spolehlivosti. Její výhodou oproti parametrickým metodám je snadnost jejího popisu a nezávislost na typu rozdělení dat. Je počítačově zaměřená a hodí se v kombinaci s robustními odhady nebo při určování rozptylu střední hodnoty transformovaných náhodných proměnných.

Vedle této techniky se často uvádí i metoda *Bootstrap*, která se používá ke stejným účelům. Metoda *Bootstrap* je založena na náhodném výběru a tak při každém provedení vyjde trochu jinak. Při zpracování svých dat jsem ji neužívala.

Metody *Jackknife* a *Bootstrap* byly poprvé publikovány v roce 1981 Bradley Efronem.

Robustní odhady parametrů polohy a rozptýlení dat se užívají, pokud data pocházejí z jiného než normálního rozdělení nebo jsou-li v datech vybočující hodnoty. Mezi jejich zástupce patří *medián* jako 50% kvantil, kdy polovina prvků leží pod a polovina nad jeho hodnotou, a *uřezaný průměr* jako aritmetický průměr z dat, ze kterých byl předtím odstraněn určitý počet nejvyšších a nejnižších hodnot.

Metoda *Jackknife* využívá takzvaných pseudohodnot y_i definovaných vztahem

$$y_i = n\hat{\Theta} - (n-1)\hat{\Theta}_{(i)}, \quad (3.5.1)$$

kde $\hat{\Theta}_{(i)}$ jsou odhady získané ze všech prvků výběru kromě i -tého. Pro střední hodnotu $\bar{\Theta}_J$ pak platí

$$\bar{\Theta}_J = \frac{1}{n} \sum_{i=1}^n y_i = n\hat{\Theta} - \frac{n-1}{n} \sum_{i=1}^n \hat{\Theta}_{(i)}. \quad (3.5.2)$$

Pro střední a větší výběry lze pseudohodnoty y_i považovat za přibližně normálně rozdělené. Náhodná veličina

$$T_J = \frac{\bar{\Theta}_J - \Theta}{\hat{\sigma}_J} \quad (3.5.3)$$

má normované normální rozdělení. Rozptyl Jackknife odhadu $\hat{\sigma}_J^2$ se počítá z

$$\hat{\sigma}_J^2 = \frac{1}{n(n-1)} \sum_{i=1}^n (y_i - \bar{\Theta}_J)^2. \quad (3.5.4)$$

Neparametrický odhad $\hat{\sigma}_J^2$ umožňuje stanovit rozptyly libovolných odhadů parametru Θ .

3.6 Kombinace metody Jackknife s metodou PCA

V této práci jsou odhadovaným parametrem Θ z předchozí kapitoly souřadnice polohy vzoru na rozptylovém grafu získaném metodou PCA.

Podle metody Jackknife se ze vstupní tabulky vyřadí vždy jeden vzor (řádek) a vypočtou se souřadnice všech ostatních vzorů metodou PCA. Pro každý vzor tak dostaneme $n - 1$ horizontálních a vertikálních souřadnic na rozptylovém grafu, kde n je počet všech vzorů ve vstupní tabulce.

Na tyto souřadnice se potom aplikuje vzorec (3.5.1). Parametr $\hat{\Theta}$ bude souřadnice vypočtená z celé tabulky a $\hat{\Theta}_{(i)}$ bude ta samá souřadnice tak, jak vyjde, když ve vstupní tabulce chybí i -tý řádek. Pro každý vzor tak vznikne $n - 1$ dvojic hodnot y_i , které budu v dalším textu nazývat *pseudosouřadnice*. Z nich budu následně konstruovat konfidenční elipsoidy určující interval spolehlivosti polohy vzoru na rozptylovém grafu.

Metoda Jackknife se dá použít také tím způsobem, že místo vzorů (řádků) se vynechávají vlastnosti (sloupky vstupní tabulky). Výsledný pohled na podobnost vzorů je potom jiný, středy konfidenčních elipsoidů jsou od poloh vzorů více odchýleny.

V této práci uvádím pouze výsledky vzniklé vynecháváním vzorů.

Na tomto místě může vzniknout otázka, proč nejsou bodové odhady ve středu elipsoidů. Metoda Jackknife zaručuje nestrannost odhadu poloh obrazů vzorů. Podle [5] je odhad nestranný, když pro daný rozsah výběru n je jeho střední hodnota rovna skutečné hodnotě parametru Θ celého souboru. Vzory pocházejí z p -rozměrného prostoru a Jackknife to zohledňuje tím, že střední hodnoty Jackknife poloh, tj. pseudosouřadnic, lépe vystihují vzájemné vzdálenosti mezi vzory i v tom směru, který je metodou PCA redukován. To znamená, že původní bodový odhad polohy obrazu je více vychýlen od skutečné hodnoty než korigovaný bodový odhad (střed konfidenčního elipsoidu). Skutečnou hodnotu polohy bodu však v aplikacích neznáme kvůli prostorové redukci, a proto nemůžeme uvedenou vlastnost Jackknife odhadů ověřit.

3.7 Konstrukce konfidenčních elipsoidů

Nejprve znormuji $n - 1$ dvojic pseudosouřadnic patřících stejnému vzoru tak, že od nich odečtu aritmetický průměr $\bar{x}$, $\bar{y}$ jejich odpovídajících složek. Ze znormovaných pseudosouřadnic této množiny bodů potom vypočtu kovarianční matici $(n - 1)\mathbf{C}$.

$$(n - 1)\mathbf{C} = \begin{bmatrix} \sum_{i=1}^{n-1} x_i^2 & \sum_{i=1}^{n-1} x_i y_i \\ \sum_{i=1}^{n-1} x_i y_i & \sum_{i=1}^{n-1} y_i^2 \end{bmatrix}, \quad (3.7.1)$$

kde x_i je normovaná horizontální souřadnice a y_i normovaná vertikální souřadnice.

Normování znamená, že pro každý vzor platí $\sum_{i=1}^{n-1} x_i = \sum_{i=1}^{n-1} y_i = 0$. Tuto matici vydělím hodnotou $(n-1)$ a potom invertuji, abych dostala matici $\mathbf{C}^{-1}$.

Konfidenční elipsoid jednoho vzoru má potom rovnici

$$(x - \bar{x}, y - \bar{y})\mathbf{C}^{-1}\begin{pmatrix} x - \bar{x} \\ y - \bar{y} \end{pmatrix} = kvantil^2, \quad (3.7.2)$$

kde $\bar{x}$, $\bar{y}$ jsou aritmetické průměry původních (nenormovaných) pseudosouřadnic patřících jednomu vzoru a *kvantil* je číslo určující velikost konfidenčního elipsoidu. Platí, že čím je *kvantil* vyšší, tím je větší i konfidenční elipsoid.

Konfidenční elipsoid má střed o souřadnicích $\bar{x}$, $\bar{y}$. Jeho hlavní poloosa je orientována ve směru nejvyššího rozptylu pseudosouřadnic a jeho vedlejší poloosa leží ve směru nejnižšího rozptylu pseudosouřadnic patřících danému vzoru. Délka poloos konfidenčního elipsoidu je rovna součinu *kvantilu* a směrodatné odchylky pseudosouřadnic na dané poloose. Odtud plyne zdůvodnění termínu *kvantil*, protože ten se ve statistické teorii násobí se směrodatnou odchylkou.

Hodnotě *kvantilu* nedávám přesný statistický význam. Volím jej pro každou úlohu dle svého úsudku. V rámci jedné úlohy je *kvantil* pro všechny konfidenční elipsoidy stejný.

4. Aplikace zjištěných poznatků na ekonomický výzkum

Metodu analýzy popsanou v kapitolách 2 a 3 jsem použila pro zpracování třech úloh, o jejichž obsahu jsem se již krátce zmínila na začátku kapitoly 3.1. Při volbě těchto úloh jsem kladla důraz na jejich vzájemnou odlišnost, co se týká povahy dat, aby tak mohla být otestována i všestrannost navrhované metody analýzy.

1. úloha „Profese“, viz kapitola 4.2.1, je demonstrací takzvaného *měkkého problému*. To znamená, že její data pocházejí z výzkumu subjektivního vnímání určité třídy objektů, kterou jsou zde profese. Subjektivita znamená, že data nemusí být přesná a mohou se lišit, budou-li je generovat různí lidé.

2. úloha „Měny“, viz kapitola 4.2.2, obsahuje ekonomická *data růstového charakteru*.

3. úloha „Podniky“, viz kapitola 4.2.3, zastupuje problém zpracování typických *tvrdých dat*. Zde jsou jimi výstupy účetnictví podniků strojírenského sektoru. Tvrdá data jsou většinou charakterizována jako údaje neměnné, objektivní a měřitelné.

V následujícím textu hojně používám termín „*analýza*“, který má v této části práce svůj specifický význam. *Analýza* zde znamená použití souboru metod popisovaných ve 2. a 3. kapitole. Jmenovitě to je zobrazení dat do dvourozměrného prostoru metodou PCA, doplnění intervalových odhadů k bodově zobrazeným datům, a zařazení ekonomických objektů do kategorií podle toho, zda se jejich intervalové odhady překrývají.

Jak jsem se již zmínila v závěru kapitoly 2, analýzu zvolených ekonomických dat jsem prováděla pomocí vlastních programů vytvořených v jazyce Turbo Pascal. Vstupy a výstupy mých programů zobrazuji v přílohách. Výstupy jsou tvořeny grafy, které vysvětlují v následující kapitole.

4.1 Výstupy analýzy

Pro lepší orientaci v kapitole 4.2 se nejprve zaměřím na vysvětlení formy výsledků mé analýzy, jejímž výstupem jsou 4 grafy.

4.1.1 Sloupcové grafy 1. a 2. pohledu

Způsobem popsaným v kapitole 3.2 o PCA vznikne ze vstupní tabulky dat takzvaná korelační matice. Ta má určitá vlastní čísla a jim náležející vlastní vektory.

1. pohled jsou výškami sloupců vyjádřené hodnoty složek vlastního vektoru patřícího nejvyššímu vlastnímu číslu korelační matice. **2. pohled** je to samé pro 2. nejvyšší vlastní číslo. Tyto sloupcové grafy objasňují, které vlastnosti se zasloužily o umístění vzoru v horizontálním (vodorovném) směru daném osou X nebo ve vertikálním (svislém) směru daném osou Y na rozptylových grafech v dalších přílohách. Polohy obrazů vzorů na rozptylovém grafu jsou totiž váženými součty jejich vlastností ze vstupní tabulky, kde vahami jsou složky 1. a 2. vlastního vektoru.

Například v příloze 3 je profese 18 (manuální dělník) umístěna extrémně nalevo. Na grafu 1. pohledu jsou nejvyšší ty sloupce patřící vlastnostem, které má profese 18 nejméně rozvinuté, konkrétně 3, 4, 5, 6, 7, 8, 9 a 11, jejich popis viz příloha 1.

Profese 36 (elitní voják) je umístěna extrémně nahoře. Vysvětlení podá graf 2. pohledu. Tam jsou nejvyšší váhy dány vlastnostem 1, 2, 4 a 12, a ty má profese 36 skutečně relativně rozvinuté.

4.1.2 Sloupcové grafy vlastních čísel

Všechna vlastní čísla korelační matice jsou zde srovnána sestupně podle velikosti a vyjádřena výškou sloupců. První dva sloupce patří 1. a 2. nejvyššímu vlastnímu číslu, jimž náležející vlastní vektory tvoří osy rozptylových grafů. V kapitole 3.2 o PCA ukazují, že vlastní číslo se rovná rozptylu dat v daném směru, viz vzorec (3.2.23). Poměr plochy prvních dvou sloupců k ploše všech sloupců vyjadřuje podíl zachované informace na celé informaci obsažené ve vstupních datech.

Při zběžném pohledu na přílohu 2 a přílohu 7 můžeme tedy odhadnout, že metoda PCA pracovala úspěšněji v úloze o měnách než v úloze o profesích. Numericky se tento

poměr rovná $(\lambda_1 + \lambda_2) / \sum_{i=1}^p \lambda_i$, kde λ_1 je nejvyšší vlastní číslo, λ_2 je 2. nejvyšší vlastní číslo a

$\sum_{i=1}^p \lambda_i$ je díky normalizaci korelační matice rovno p , tedy jejímu rozměru neboli počtu vlastností u každého vzoru.

4.1.3 Rozptylové grafy

Body na nich zobrazují umístění vzorů po prostorové redukci metodou PCA. Každému vzoru odpovídá bodový odhad jeho polohy na grafu znázorněný bodem a intervalový odhad polohy znázorněný konfidenčním elipsoidem. Ten je určen podle kovarianční matice z bodů patřících stejnému vzoru vzniklých metodou Jackknife. Kvantil konfidenčních elipsoidů určující jejich velikost je zvolen subjektivně a je pro všechny elipsoidy na grafu stejný.

4.1.4 Seznam minimálních kvantilů a mapa podobnosti objektů

U každé dvojice vzorů je uveden minimální kvantil, který musí mít obě elipsy těchto vzorů, aby se začaly dotýkat. Čím je jeho hodnota vyšší, tím méně jsou si vzory podobné.

Kvantily nižší než subjektivně zvolená mez jsou zvýrazněny barvou a na rozptylovém grafu, který za seznamem následuje, je ve stejné barvě nakreslena čára spojující tyto dva body. Vzniká tak mapa podobnosti vzorů umožňující kategorizaci ekonomických objektů.

4.2 Ekonomické úlohy

Navrhovaný způsob analýzy ekonomických dat demonstruji na třech již zmíněných ekonomických úlohách.

4.2.1 Profese

Tuto úlohu pojmám jako testovací příklad s úzkou vazbou na personalistiku.

Vstupní tabulka dat je v příloze 1. Jejím obsahem je mé subjektivní hodnocení mnou zvolených parametrů určitých profesí na škále 0 až 100 bodů. Moje analýza zobrazila vzory profesí na rozptylový graf v příloze 3. Polohy obrazů jednotlivých profesí na tomto grafu vysvětlují sloupcové grafy z přílohy 2.

Na **grafu 1. pohledu** vidíme, že extrémní hodnoty zde nabývají parametry profesí č.:

- 3 - Požadavky na analytické myšlení,
- 5 - Požadavky na smysl pro etiku a morální bezúhonnost,
- 6 - Kreativita,
- 7 - Samostatnost,

9 - Profesionální hrdost,

11 - Osobní zaujetí (Zda se prací zabýváme i ve volném čase.)

Shodou okolností jsou všechny tyto vysoce rozvinuté vlastnosti kladné. Tyto aspekty profesí se dají zobecnit jako nároky na intelekt a vysoké osobní nasazení.

Profese umístěné v rozptylovém grafu na pravé straně by měly mít tuto stránku podstatně vyvinutější než profese na opačné straně. Mezi profese na pravé straně grafu patří:

22 - Přírodovědec (exaktní vědy),

4 - Novinář, Recenzent,

25 - Duchovní,

5 - Politik,

6 - Psycholog, Psychiatr,

8 - Chirurg,

11 - Veterinář,

21 - Designér informačních systémů.

Můžeme usoudit, že k řádnému výkonu těchto profesí je skutečně třeba vlastností, které zdůraznil graf 1. pohledu.

Na levé straně grafu se nacházejí profese:

18 - Dělník manuální,

13 - Uklízečka,

29 - Účetní na nejnižší úrovni,

O těchto profesích je známo, že nekladou vysoké nároky na duševní schopnosti typu nonverbální inteligence a jejich vykonavatelé většinou po odpracování své směny pracovní problémy pouštějí z hlavy. To jsou vlastnosti, které lze považovat za protikladné k těm, které zdůraznil graf 1. pohledu.

Na **grafu 2. pohledu** vidíme, že extrémní hodnoty zde nabývají parametry profesí č.:

1 - Fyzická náročnost,

2 - Požadavky na manuální zručnost,

4 - Požadavky na sociální inteligenci (umění jednat s lidmi a odhadnout je),

12 - Riziko onemocnění nebo úrazu.

Shodou okolností jsou všechny tyto zdůrazněné vlastnosti opět kladné. Vyjmenované aspekty profesí můžeme shrnout jako nároky profese na tělesnou konstrukci zaměstnance a umění vycházet s lidmi.

Profese umístěné v horní polovině rozptylového grafu by měly mít tyto stránky vyvinutější než profese umístěné dole. Profese, které se umístily nejvíce nahoře jsou:

36 - Voják elitní,

12 - Zdravotní sestra,

9 - Masér,

33 - Řezník,

7 - Lékař (obvodní), Zubař,

35 - Policista, Kriminalista,

Domnívám se, že tyto profese skutečně vyžadují to, co zvýraznil 2. pohled.

V dolní části grafu se nacházejí profese:

20 - Programátor,

15 - Inženýr, Konstruktor,

14 - Architekt,

21 - Designér informačních systémů

29 - Účetní na nejnižší úrovni

Tyto profese většinou nevyžadují fyzickou sílu a obratnost a některé z nich lze vykonávat v odloučenosti od lidí, což jsou aspekty protikladné k těm, které zdůraznil graf 2. pohledu.

Konfidenční elipsoidy na rozptylovém grafu v příloze 3 ukazují, jaké profese by se při zvoleném kvantilu mohly považovat za podobné. Příloha 5 obsahuje **mapu podobnosti objektů**. Zde jsou podobné profese spojeny barevnými čarami. Profese tak byly zařazeny do skupin neboli kategorií. Záleží jen na výši kvantilu, budou-li se skupiny profesí rozšiřovat, nebo ztrácet své členy.

Jelikož rozložení profesí na rozptylovém grafu odpovídá obecnému názoru na ně, považuji výsledky této testovací úlohy za uspokojivé. Analýza zde byla úspěšná.

Podobná úloha by se dala prakticky využít pro kategorizaci profesí využitelnou na úřadech práce, kde by se profese patřící do stejné kategorie prezentovaly uchazečům o zaměstnání společně a byly by u nich uvedeny jejich typické nároky na osobnost pracovníka. Kategorie profesí sestavené podle vhodných parametrů by se daly také využít v pojišťovnictví, kde je potřeba vytipovat určitá rizika a ta mohou být více profesím společná.

4.2.2 Měny

Cílem úlohy je zmapovat vzájemné ovlivňování kurzů vybraných světových měn od roku 1992 do roku 1997. Výběr měn je dán údaji dostupnými ve Statistické ročence ČR [6]. Sledované časové období se snaží respektovat požadavek na konzistenci dat, aby zde nebyly srovnávány údaje z podstatně odlišných hospodářských epoch. Vstupní tabulka dat je v příloze 6. Do analýzy vstupuje 5 sloupců pod nadpisem „Logaritmované podíly“.

Podíly kurzů v jednotlivých letech logaritmuji, protože přijímám hypotézu o tom, že ekonomický růst je podobného charakteru jako růst biologický a vyvíjí se tedy exponenciálně. Logaritmováním se tento typ vývoje převede na lineární a ten je snáze analyzovatelný. U mé úlohy o měnách je logaritmování užitečné především proto, že při jeho užití nezáleží na měně, ke které se všechny kurzy vztahují. Přepočet kurzů na jinou základní měnu a zlogaritmování jejich meziročních podílů způsobí, že každý sloupec nové tabulky je oproti stejnému sloupci tabulky původní zvýšen či snížen o nějakou konstantu. Tento rozdíl se potom zruší při normování dat popsaném v kapitole 3.2 o PCA.

Moje analýza zobrazila měny na rozptylový graf v příloze 8. Polohy obrazů jednotlivých měn na tomto grafu vysvětlují sloupcové grafy z přílohy 7.

Na **grafu 1. pohledu** vidíme, že záporných hodnot zde nabývají meziroční pohyby 93/92, 94/93, 95/94 a kladné jsou váhy pohybů 96/95, 97/96. Kurzy měn umístěných v pravé části rozptylového grafu (ITL, CAD, GBP, USD, SEK) tedy v rozmezí let 1992 až 1994 relativně klesaly, zatímco relativně rostly mezi roky 1995 až 1997. Kurz JPY v levé části rozptylového grafu měl opačný vývoj.

Na **grafu 2. pohledu** vidíme, že záporných hodnot zde nabývají meziroční pohyby 95/94, 96/95 a kladné jsou váhy pohybů 93/92, 94/93, 97/96. Kurzy měn umístěných v horní části rozptylového grafu (JPY, USD, GBP) tedy v rozmezí let 1992 až 1993 relativně rostly, v letech 1994 a 1995 klesaly, a mezi roky 1996 a 1997 opět relativně rostly. Kurzy měn SEK, ESP, FIM se v tomto období vyvíjely opačně.

Tabulka v příloze 6 to do velké míry potvrzuje, zvláště u extrémně umístěných měn. Při výkladu poloh obrazů měn na rozptylovém grafu je třeba brát v úvahu, že čísla vstupní tabulky nabývají kladných i záporných hodnot.

Konfidenční elipsoidy na rozptylovém grafu v příloze 8 ukazují, jaké měny by při zvoleném kvantilu mohly být považovány za podobné ve svém vývoji za sledované období. Příloha 9 obsahuje seznam všech minimálních kvantilů a **mapu podobnosti objektů**. Zde je vidět významná skupinka měn BEF, NLG, DEM, ATS, DKK, o které se můžeme domnívat, že odráží snahu o koordinaci měnové politiky, nebo alespoň to, že tyto měny byly vystaveny srovnatelným hospodářským vlivům. Fakt, že se jedná o evropské měny potvrzuje oprávněnost vzniku takového shluku. Maximálně nepodobné jsou si měny JPY a SEK (kvantil 2,64). Nejpodobnější jsou si v tomto období měny CZK a FRF (kvantil 0,02).

Domnívám se, že způsob rozložení měn na rozptylovém grafu dává ekonomický smysl v návaznosti na hospodaření jednotlivých států a systémy měnových kurzů existující v popisovaném období, a proto považuji výsledky této úlohy rovněž za uspokojivé.

Výsledky úloh stejného druhu by mohly být využity ke krátkodobému předvídání vývoje kurzů měn podle vývoje kurzů jiných měn, protože měny na opačných stranách grafu se vyvíjejí většinou opačně a měny s nízkými minimálními kvantily konfidenčních elipsoidů mají převážně podobný vývoj.

4.2.3 Podniky

Cílem úlohy je zkoumání vzájemných podobností a rozdílů mezi podniky ze strojírenského oboru co se týká jejich majetkové struktury a několika růstových ukazatelů. Podniky jsem vybrala náhodně, údaje z jejich účetní závěrky jsou z Obchodního věstníku [7]. Vzhledem k velkému objemu dat uvádím pouze sekundární data v příloze 12, která byla použita jako vstup mého programu. Jejich správnost lze ověřit nahlédnutím do Obchodního věstníku a využitím informací v mých přílohách 10 a 11.

Moje analýza zobrazila podniky na rozptylový graf v příloze 14. Polohy obrazů jednotlivých podniků na tomto grafu vysvětlují sloupcové grafy z přílohy 13. Orientaci v rozptylovém grafu dále umožňuje příloha 15, v níž uvádím souřadnice všech podniků.

Na **grafu 1. pohledu** vidíme, že vysokých kladných hodnot zde nabývají váhy č.:

- 2 - Hmotný investiční majetek / Aktiva v roce 1997,
 - 8 - Základní jmění / Pasiva v roce 1997,
 - 18 - Hmotný investiční majetek / Aktiva v roce 1996,
 - 24 - Základní jmění / Pasiva v roce 1996,
- a vysokých záporných hodnot váhy č.:
- 14 - Závazky / Pasiva v roce 1997,
 - 30 - Závazky / Pasiva v roce 1996.

Na **grafu 2. pohledu** vidíme, že vysokých kladných hodnot zde nabývají váhy č.:

- 12 - Hospodářský výsledek účetního období / Pasiva v roce 1997,
- 28 - Hospodářský výsledek účetního období / Pasiva v roce 1996,

- 36 - Tržby za prodané vlastní výrobky, služby a zboží / Aktiva v roce 1997,
37 - Přidaná hodnota / Aktiva v roce 1997,
38 - Osobní náklady / Aktiva v roce 1997.

Údaje zdůrazněné 1. pohledem lze zobecnit jako kapitálovou náročnost nebo sílu. Údaje zdůrazněné 2. pohledem shrnují spíše hospodářskou výkonnost podniků. Obojí relativně k výši celkových aktiv. Analýza nejvíce zvýrazňuje ty rysy, ve kterých se podniky v daném výběru nejvíce navzájem liší.

Znamená to, že podniky umístěné na pravé straně rozptylového grafu v příloze 14 by měly být kapitálově silné, co se týká podílu hmotného investičního majetku a základního jmění k celkovým aktivům. Nejvíce napravo je zde podnik č. 64 - Zbrojovka Brno a. s.

Podniky umístěné na levé straně by měly mít nízkou kapitálovou náročnost. Nejvíce nalevo je podnik č. 44 - Stabilní hasící zařízení, spol. s r. o.

Podniky v horní části grafu by měly vynikat v hospodářských výsledcích. Měl by to být především podnik č. 19 - Chemopetrol Chemstroj, s. r. o.

Podniky v dolní části grafu mají pravděpodobně hospodářské problémy. Nejvíce dole je podnik č. 38 - Přerovské strojírný a. s.

V následující tabulce jsou o těchto podnicích údaje z účetní závěrky za rok 1997 v tis. Kč, které nejlépe vystihují daný problém.

Podnik	Aktiva	HIM	Z. jmění	Závazky	HV	Tržby	PH	Osobní n.
64	875 571	544 098	768 041	129 205	-82 382	432 256	224 850	221 482
44	108 888	3 915	200	76 638	12 316	185 322	63 579	30 335
19	16 958	2 803	5 550	7 262	376	70 620	32 462	31 072
38	3 216 734	1 012 717	1 729 704	770 410	-1 416 859	416 362	126 374	43 439

Vysvětlivky:

HIM - Hmotný investiční majetek

Z. jmění – základní jmění

HV – hospodářský výsledek účetního období

PH – přidaná hodnota

Osobní n. – osobní náklady

Pramen: [7]

Podnik č. 64 má skutečně vysokou výši kapitálu (základní jmění). Podnik č. 44 doměnkou potvrzuje díky nízkému základnímu jmění. Podnik č. 19 není příliš hospodářsky výkonný v absolutní výši, ale vzhledem k výši svých aktiv ano. Podnik č. 38 má dle svého hospodářského výsledku hluboký hospodářský propad.

Domnívám se, že vzhledem k zadaným údajům přinesla analýza podniků uspokojivé výsledky. Z rozložení podniků na rozptylovém grafu lze činit hypotézy o tom, jak tyto podniky naplňují výše identifikované zobecněné rysy vstupních parametrů, to jest relativní kapitálovou sílu či náročnost a hospodářskou výkonnost. To znamená, že zde došlo k užitečné kompresi informací obsažených v rozsáhlé vstupní tabulce, ze které na první pohled není nic zřejmé.

Výsledky úloh stejného druhu by mohly být využity ke kontrole majetkové přiměřenosti, kterou provádějí auditoři. Pokud by ve vstupní tabulce byl takový soubor podniků, aby většina z nich byla normální tj. měla přiměřený majetek a hospodářský výkon, mohla by vybočující pozorování na rozptylovém grafu svědčit o nezákonných nebo alespoň neobvyklých hospodářských operacích s majetkem.

Závěr

Cílem mé práce bylo dosáhnout výsledků srovnatelných s vícekritériálním (paralelním) úsudkem člověka – specialisty v oblasti ekonomie.

Podařilo se mi zpracovat tři různé teorie GHA, PCA a Jackknife. Klasickou metodu PCA jsem porovnávala s neuronovou sítí samoorganizující se podle GHA. Zjistila jsem, že obě dvě metody poskytují stejné hodnoty vah sítě a stejné obrazy ekonomických objektů na dvourozměrném grafu. Zavedením metody Jackknife jsem sledovala citlivost sítě na vynechávání objektů. Tak vznikla metoda sestavování intervalových odhadů, která umožňuje analyzovat vztahy mezi ekonomickými objekty.

Celou v této práci popsanou metodu analýzy jsem aplikovala na tři příklady z oblasti ekonomie. V úlohách jsem kladla důraz zejména na rozmanitost možných zadání ekonomického problému.

1. úloha je z oblasti personalistiky. S využitím měkkých dat o nárocích na lidské vlastnosti řeší kategorizaci profesí. Přestože jsem data zvolila dle svého osobního názoru, domnívám se, že vzájemné podobnosti profesí získané neuronovou sítí nejsou v rozporu s obecnou lidskou představou.

2. úloha ukazuje možnosti použití navrhované metody v oblasti analýzy časových řad v ekonomii. Cílem bylo zjistit vzájemné podobnosti významných světových měn z časového trendu jejich vývoje. Při řešení úlohy jsem využila poznatků z ekonomie a průměrné roční hodnoty kurzů měn jsem převedla na růstové ukazatele. Analýzou jsem zjistila významnou podobnost ve vývoji třech skupin měn za sledované období 1992 až 1997: (BEF, NLG, DEM, ATS, DKK), (CZK, FRF) a (GBP, CAD).

3. úloha zpracovává tvrdá data z oficiálních účetních údajů strojírenských podniků České republiky. Zde se podařilo nalézt dva hlavní pohledy na soubor podniků, pro něž

existuje ekonomická interpretace, a to kapitálová síla či náročnost a hospodářská výkonnost. To jsou faktory, které daný soubor popisují nejvýstižněji, neboť v nich existuje mezi popisovanými podniky největší rozdílnost.

V práci se podařilo prokázat použitelnost PCA a GHA pro analýzu vícerozměrných dat z oblasti ekonomie.

Na tomto místě bych si dovolila krátké zamyšlení nad možnostmi a omezeními neuronového řešení nejen ekonomických úloh. Výstupy analýzy popsané v této práci je třeba posuzovat vždy vzhledem ke vstupním údajům, stejně jako se to dělá i u tradičnějších přístupů k řešení. Čím kvalitnější budou vstupní informace, tím více můžeme očekávat od výsledků analýzy. Z toho důvodu navrhuji, aby při profesionálním provádění navrhované analýzy byli u sběru dat experti na analyzovanou oblast. Ti by byli schopni posoudit, jaké vstupní údaje jsou pro získání požadovaných výsledků nejdůležitější. Zároveň by se na přípravě dat měli podílet i lidé nespecializující se na daný obor, aby se tak zabránilo úzkému jednostrannému pohledu specialisty.

Výstupy neuronové sítě prokazují tak dobré výsledky při zobecňování mnohovýznamových dat, že to může svádět k přehnaným požadavkům. Mám zde na mysli například to, že požadovaným výstupem touto prací popisované neuronové sítě nemohou být měkké, subjektivní informace typu „podnik je dobrý nebo špatný“, když veškerým vstupem byly tvrdé údaje z účetnictví, které samozřejmě neříkají nic o podnikatelských schopnostech lidí, kteří se svěřeným majetkem hospodaří.

Pro dosažení měkkého výstupu z tvrdých dat by bylo pravděpodobně vhodnější použít složitější způsob zpracování přiblížený v kapitole 1.2. I zde je však nutné měkký výstup kvantifikovat na nějaké škále, například „dobrý“ = 100, „špatný“ = 0, protože i tato data musí do sítě vstoupit jako kvantitativní informace pro její učení. Neuronovým sítím tak asi navždy zůstanou nedostupné nekvantifikovatelné kategorie jako je krása, láska nebo smysl života.

Seznam literatury

1. HAYKIN SIMON: Neural Networks, A Comprehensive Foundation, Macmillan College Publishing Company, Inc., New York, 1994.
2. PRESS, W. H., TEUKOLSKI, S. A., VETTERLING, W. T., FLANNERY, B. P.: Numerical Recipes in C, Cambridge University Press, Cambridge, 1992.
3. BEALE, R., JACKSON, T.: Neural Computing: An Introduction, IOP Publishing Ltd, Bristol, Philadelphia, 1992.
4. GROlier INCORPORATED, Encyclopedia Americana - Deluxe Library Edition, Danbury, Connecticut 061816, USA, 1992.
5. MELOUN MILAN, MILITKÝ JIŘÍ, Statistické zpracování experimentálních dat, PLUS spol. s r. o., 1. vydání, Praha, 1994.
6. ČESKÝ STATISTICKÝ ÚŘAD: Statistická ročenka České republiky, 1. vydání, Praha 1993 – 1998, 6 svazků.
7. Obchodní věstník, Praha 1998 - 1999.
8. SYNEK MILOSLAV A KOLEKTIV: Manažerská ekonomika, Grada Publishing, spol. s r. o., 1. vydání, Praha, 1996.
9. PAO, Y. H.: Adaptive Pattern Recognition & Neural Networks, Addison Wesley, 1989.
10. McCLELLAND, J. L. & RUMELHART, D. E.: Parallel Distributed Processing, Volume 1, MIT Bradford Press, 1986.
11. CRICK, FRANCIS: Věda hledá duši, Mladá fronta, Edice Kolumbus, 1. vydání, Praha, 1997.
12. KOHONEN, T.: Self Organisation and Associative Memory, third edition, Springer-Verlag, 1990.
13. HOPFIELD, J. J.: Neural Networks and Physical Systems with Emergent Collective Computational Abilities, in Proc. Natl. Acad. Sci. USA., Volume 81, pages 3088-3092, 1984.
14. ANDĚL, JIŘÍ: Matematická statistika, SNTL/ALFA, 2. vydání, Praha, 1985.

Seznam příloh

	Počet stran
1. Profese – Vstupní tabulka dat	1
2. Profese – Sloupcové grafy 1. a 2. pohledu a vlastních čísel	1
3. Profese – Rozptylový graf	1
4. Profese – Seznam minimálních kvantilů	1
5. Profese – Mapa podobnosti objektů	1
6. Měny – Vstupní tabulka dat	1
7. Měny – Sloupcové grafy 1. a 2. pohledu a vlastních čísel	1
8. Měny – Rozptylový graf	1
9. Měny – Seznam minimálních kvantilů a mapa podobnosti objektů	1
10. Podniky – Popis řádků vstupní tabulky	1
11. Podniky – Popis sloupků vstupní tabulky	1
12. Podniky – Vstupní tabulka dat	4
13. Podniky – Sloupcové grafy 1. a 2. pohledu a vlastních čísel	1
14. Podniky – Rozptylový graf a vybrané konfidenční elipsoidy	1
15. Podniky – Souřadnice objektů na rozptylovém grafu	1

Vstupní tabulka dat

Profese

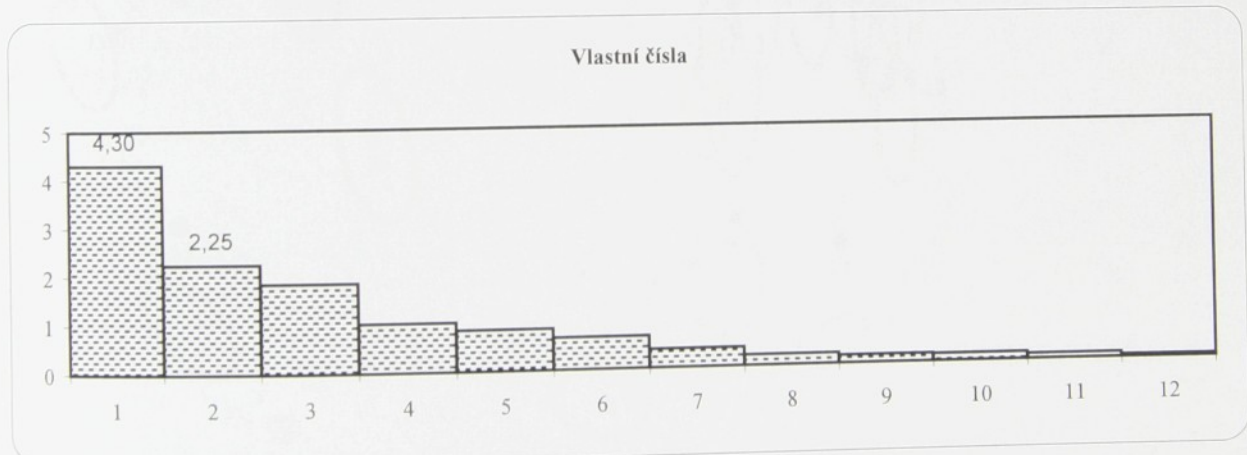
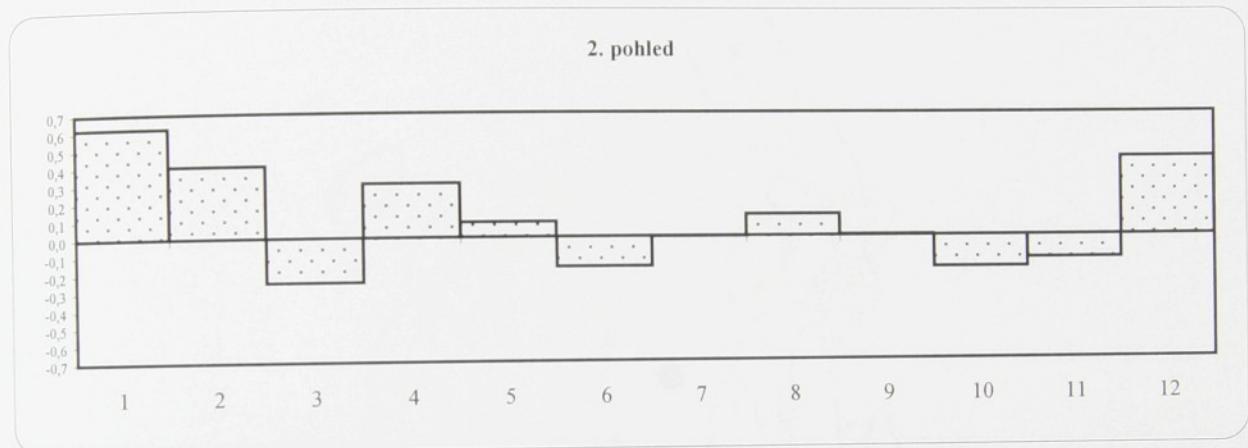
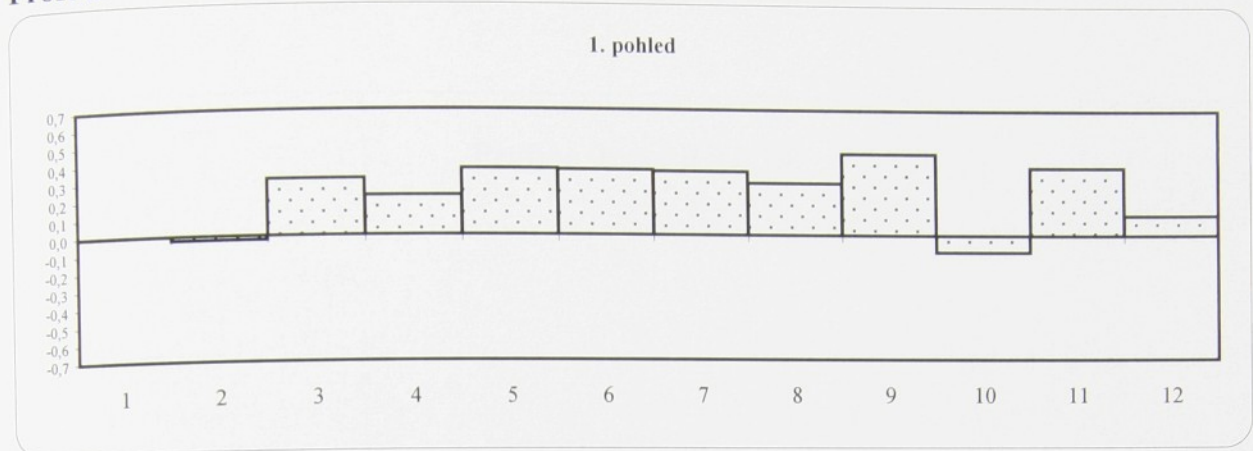
	1	2	3	4	5	6	7	8	9	10	11	12	
1	40	100	0	50	50	100	100	100	100	100	90	5	Malíř (umělec)
2	80	50	0	100	80	100	50	100	100	80	70	30	Herec
3	80	60	0	90	80	100	90	100	100	80	70	30	Hudebník, Zpěvák, D.J.
4	40	10	80	100	100	80	100	90	100	100	70	40	Novinář, Recenzent
5	40	0	80	100	100	90	100	50	100	80	70	40	Politik
6	30	0	80	100	100	50	100	80	95	100	70	50	Psycholog, Psychiatr
7	70	90	70	100	100	50	100	90	95	95	20	40	Lékař (obvodní), Zubař
8	80	100	80	40	100	80	100	90	100	95	70	40	Chirurg
9	90	90	0	100	50	0	100	100	80	100	0	5	Masér
10	50	90	0	80	50	50	50	80	70	100	0	10	Kadeřník
11	80	100	80	50	100	80	100	90	90	80	70	40	Veterinář
12	80	90	30	80	100	50	50	80	70	100	0	40	Zdravotní sestra
13	50	10	0	10	50	0	20	30	10	100	0	30	Uklízečka
14	10	50	90	50	50	95	95	80	90	90	60	5	Architekt
15	10	50	100	0	50	90	95	80	90	90	50	5	Inženýr, Konstruktor
16	50	70	70	90	70	50	100	50	60	90	0	10	Místr na dílně
17	40	10	90	90	90	80	100	40	90	50	60	30	Manažer střední a vyšší
18	80	50	10	10	10	10	10	10	10	90	0	40	Dělník manuální
19	70	80	40	20	50	60	40	60	60	90	0	30	Řemeslník, Opravář
20	5	10	100	0	50	90	80	70	80	90	20	10	Programátor
21	5	10	100	70	90	95	95	80	90	90	70	10	Designér informačních systémů
22	50	50	100	0	100	100	100	100	100	90	100	30	Přírodovědec (exaktní vědy)
23	40	60	95	70	100	80	100	40	80	100	20	30	Učitel exaktních věd
24	40	0	50	70	100	80	100	50	80	100	20	30	Učitel společenských věd
25	40	0	50	100	100	50	100	100	100	100	100	30	Duchovní
26	20	10	40	60	90	0	80	10	40	95	0	10	Bankovní úředník za přepážkou
27	40	40	10	100	50	80	100	50	50	50	5	30	Prodejní agent (např. pojišťovák)
28	40	0	80	100	100	50	100	50	90	60	20	30	Právník
29	5	10	20	10	50	10	40	20	20	90	0	10	Účetní na nejnižší úrovni
30	30	10	80	70	100	50	100	40	90	90	20	30	Auditor, Daňový poradce
31	40	60	0	90	50	0	100	80	50	95	0	40	Řidič - taxikář
32	60	70	0	50	50	0	100	70	60	95	0	40	Řidič nákladáku
33	80	70	0	50	50	0	80	50	50	80	0	30	Řezník
34	40	30	0	50	50	0	80	50	40	95	0	20	Zelinář - prodavač
35	90	50	95	100	100	50	100	60	80	70	70	50	Policista, Kriminalista
36	100	80	75	90	100	50	100	50	80	50	0	70	Voják elitní

Hodnocené parametry profesí

- 1 Fyzická náročnost
- 2 Požadavky na manuální zručnost
- 3 Požadavky na analytické myšlení
- 4 Požadavky na sociální inteligenci (umění jednat s lidmi a odhadnout je)
- 5 Požadavky na smysl pro etiku a morální bezúhonnost
- 6 Kreativita
- 7 Samostatnost
- 8 Pocit uspokojení z práce
- 9 Profesionální hrdost
- 10 Pocit správnosti volby (Opakem by byla míra nesouhlasu s některými praktikami v povolání.)
- 11 Osobní zaujetí (Zda se prací zabýváme i ve volném čase.)
- 12 Riziko onemocnění nebo úrazu

Sloupcové grafy 1. a 2. pohledu a vlastních čísel

Profese



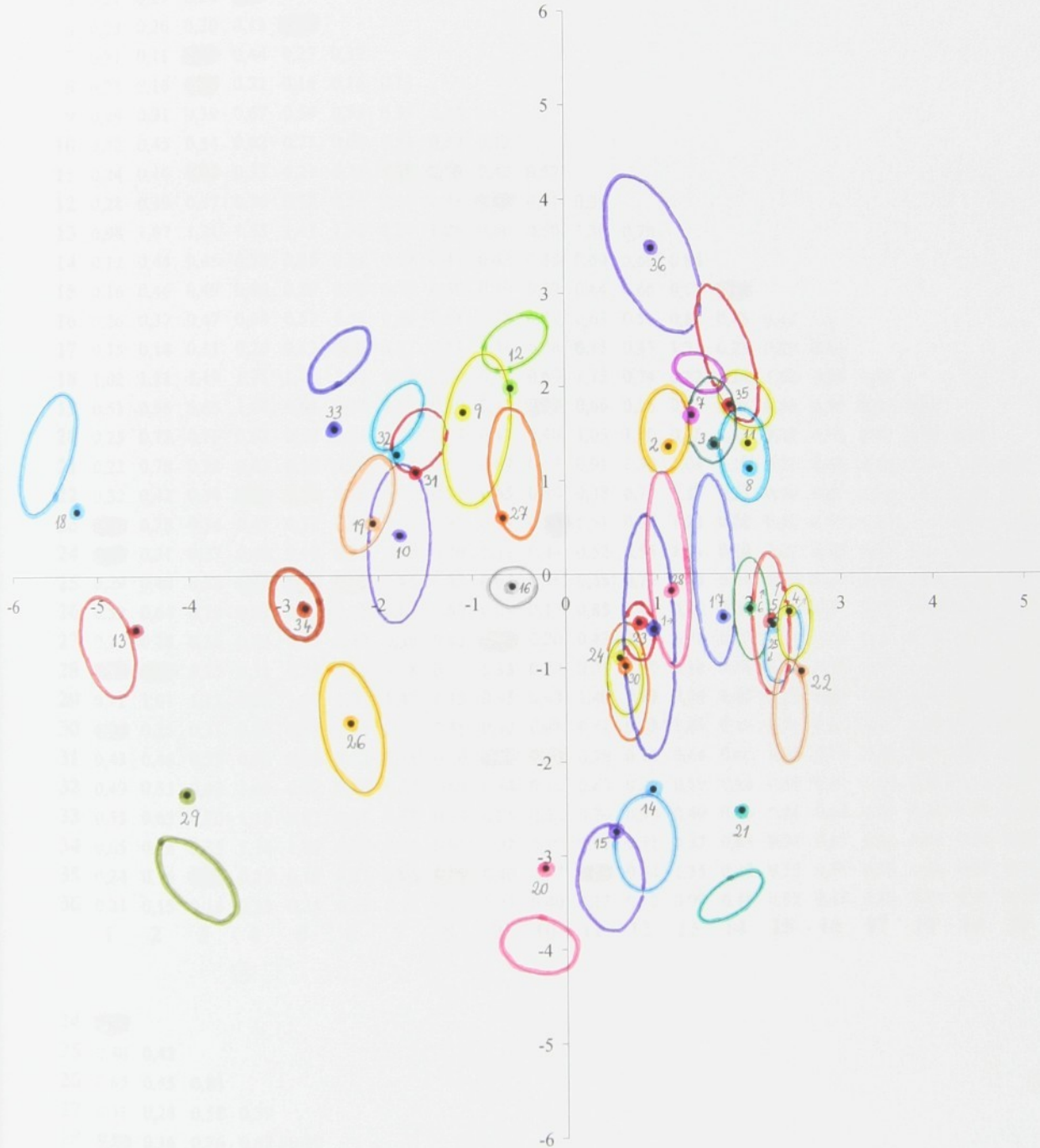
	1. pohled	2. pohled
1	-0,001948020	0,621031793
2	-0,023929287	0,407298702
3	0,315468936	-0,247894721
4	0,219038019	0,304980319
5	0,368816694	0,084076166
6	0,363760296	-0,167860152
7	0,355434429	-0,002249429
8	0,289378466	0,121328234
9	0,456535337	0,004592633
10	-0,094218068	-0,179459423
11	0,379660818	-0,130069971
12	0,109820579	0,441413735

Vlastní čísla
4,30206529
2,24551964
1,85052295
1,01507231
0,85203576
0,66840071
0,38497612
0,22000009
0,16321804
0,14381963
0,10186929
0,05250017

Zachovaná informace
55 %

Rozptylový graf

Profese (kvantil = 0,10)



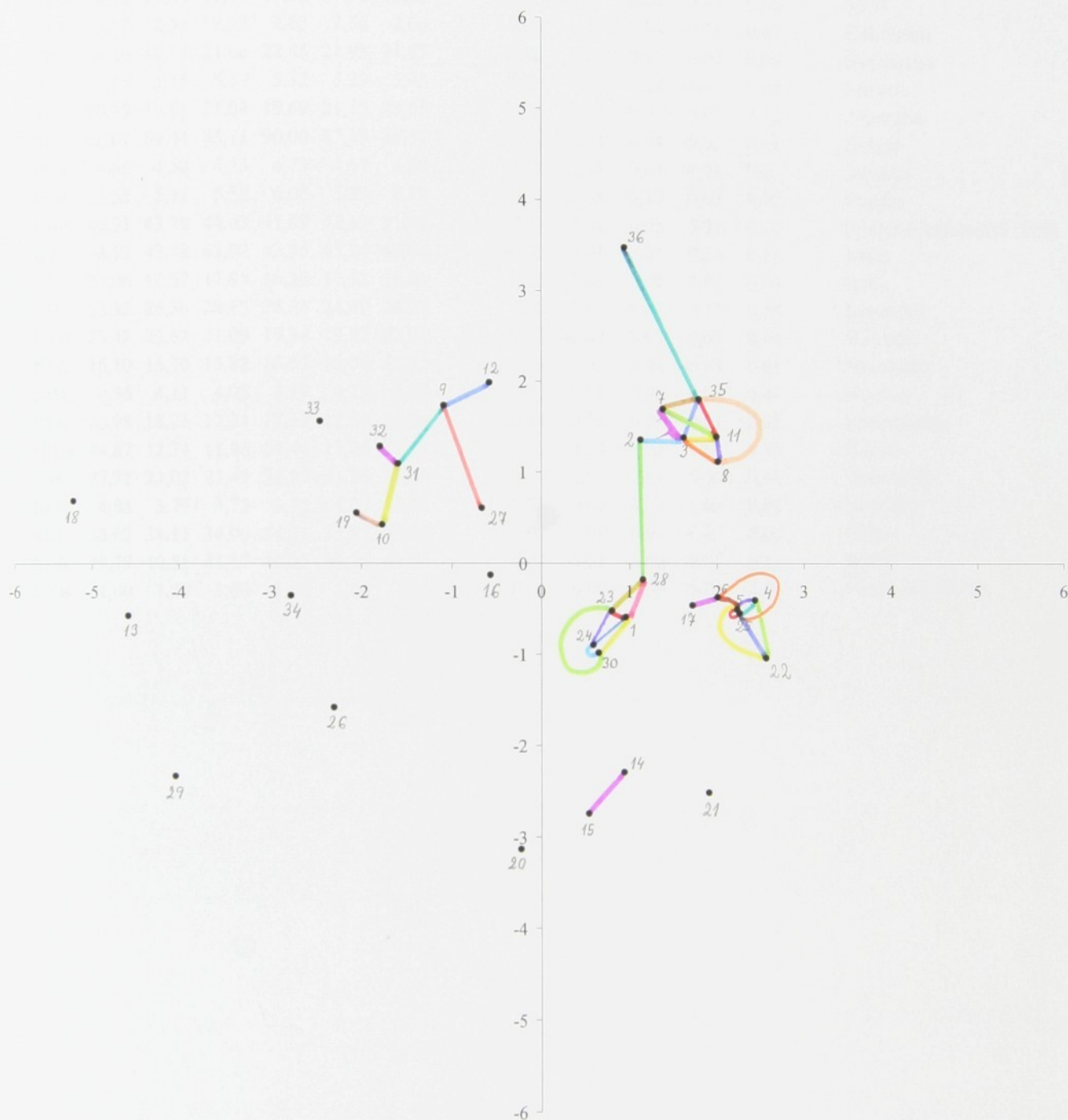
Seznam minimálních kvantilů

Profese

[illegible]

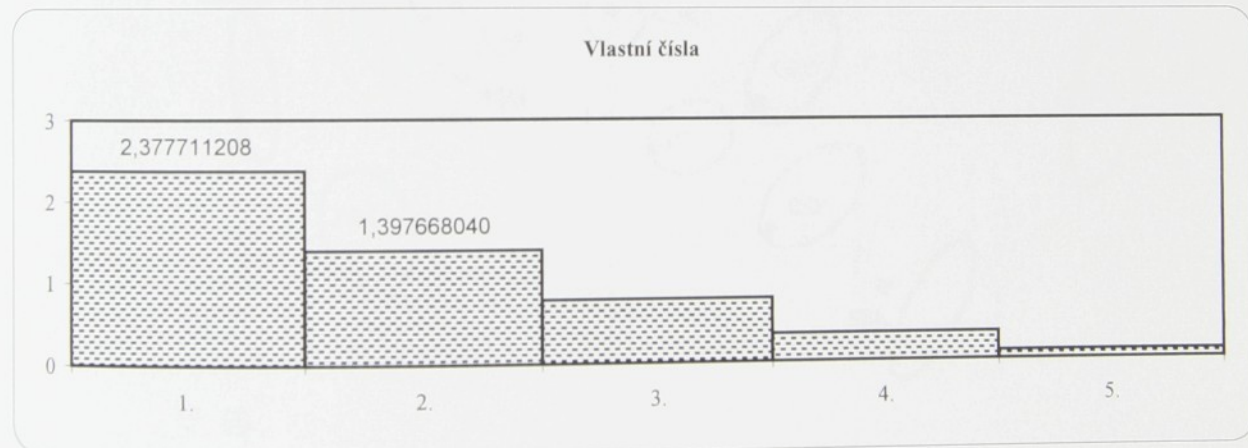
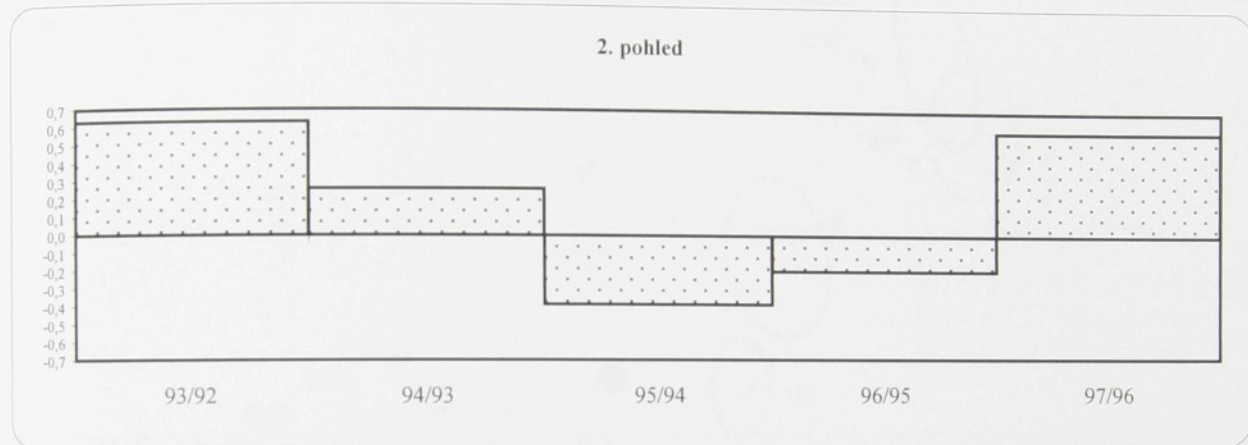
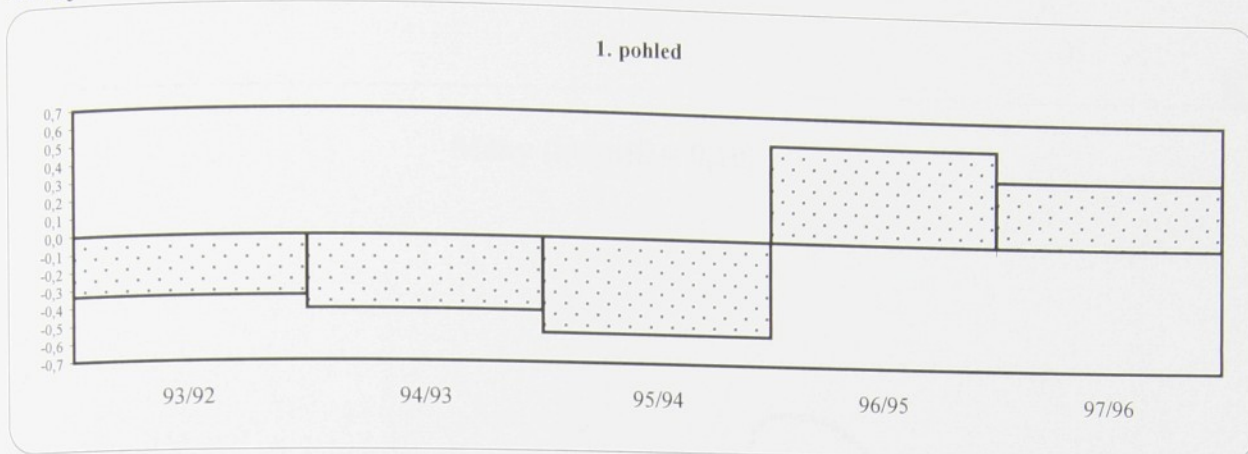
Mapa podobnosti objektů

Profese (kvantil = 0,10)



Sloupcové grafy 1. a 2. pohledu a vlastních čísel

Měny



	1. pohled	2. pohled
1. 93/92	-0,336643861	0,631009375
2. 94/93	-0,410851520	0,259113412
3. 95/94	-0,535896574	-0,388189376
4. 96/95	0,540400457	-0,199193582
5. 97/96	0,372362926	0,586786446

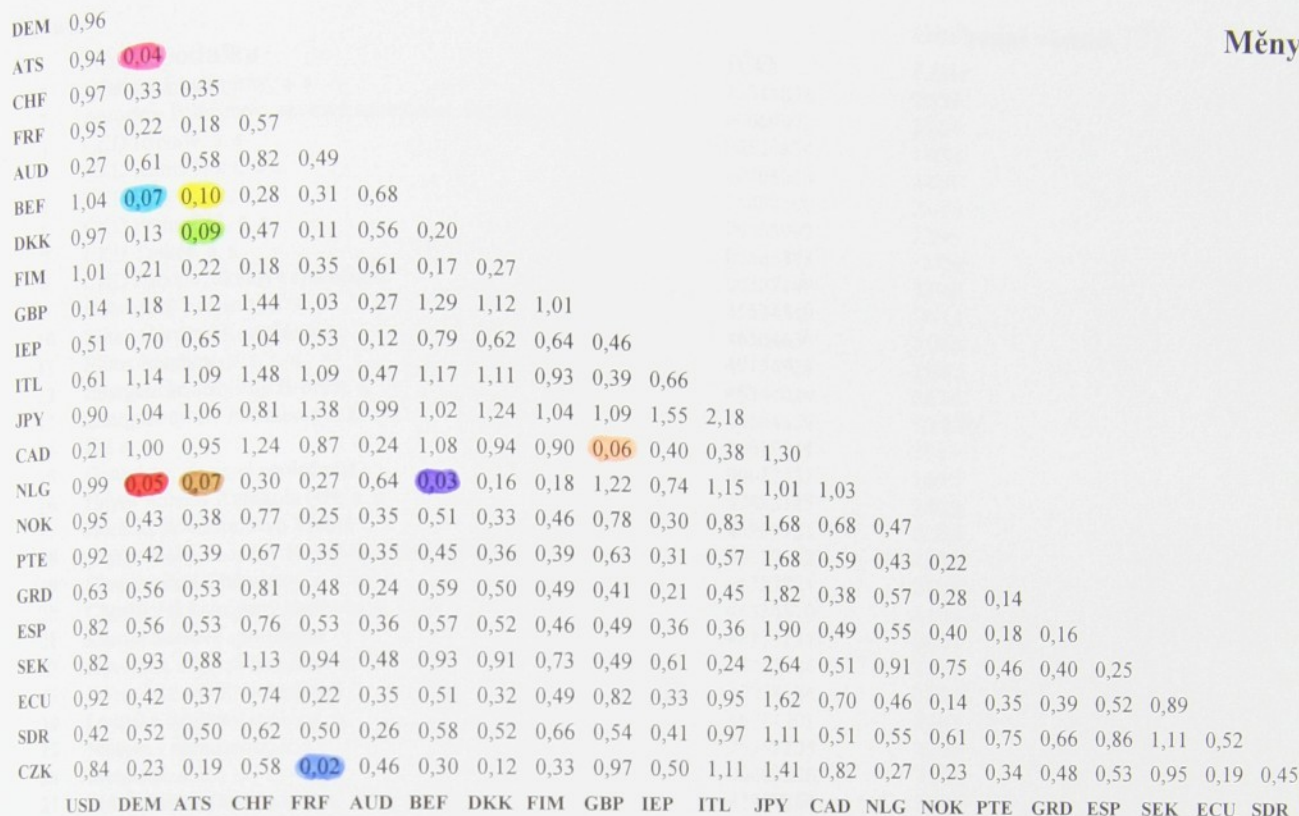
Vlastní čísla
2,377711208
1,397668040
0,782478437
0,339869689
0,102272626

Zachovaná informace
76 %

Měny (kvantil = 0,10)



Měny



Popis řádků vstupní tabulky podniků

Podnik

č. Název podniku

IČO

Obchodní věstník [7]

č./str

1	Adamovské strojírny, a. s.	46345833	29/78
2	Agrostroj Pelhřimov, akciová společnost, Pelhřimov	00009971	28/24
3	ČKD Hronov, a. s.	00529834	14/15
4	ČKD Kutná Hora, a. s.	00508055	24/39
5	ČKD Pragoimex, a. s.	15888100	23/20
6	ČKD Praha DIZ, a. s.	00565997	22/43
7	ČKD Trakce, a. s.	00565393	15/10
8	ČKD Žandov, akciová společnost	00527149	22/40
9	Dobrušské strojírny, a. s.	45534349	26/43
10	Elitex Cervený Kostelec a. s.	46504630	25/46
11	Elitex Machinery, s. r. o.	49196928	25/52
12	Energetické strojírny Brno, a. s.	46347020	28/56
13	Energostrojírny Pardubice a. s.	46504656	5/12/99
14	Eta a. s.	00010341	25/59
15	Gumokov, akciová společnost	00012131	16/15
16	Hayes Wheels Autokola NH, a. s.	47673125	22/30
17	Horácké kovodružstvo Třebíč	46969721	21/26
18	Humpolecké strojírny Humpolec, akciová společnost	60827882	28/58
19	Chemopetrol Chemstroj, s. r. o.	61327727	27/19
20	Chotěbořské strojírny služby a. s.	45534519	21/18
21	Karosa akciová společnost	48171131	28/31
22	Kovona a. s. Lysá nad Labem	14801884	22/38
23	Kovoplast, výrobní družstvo	49974084	22/25
24	Lounské strojírny spol. s r. o.	46711201	27/16
25	Meopta - nářad'ovna, a. s.	25352105	22/28
26	Motorpal, a. s.	00009296	23/10
27	MS - Unikov Ostrava s. r. o.	47977264	27/41
28	MZD Dobrovické strojírny a. s.	46351370	29/17
29	Ondřejovická strojírna, spol. s r. o.	42766982	12/14
30	Opravy a kovovýroba Záchlumí a. s.	45358656	25/21
31	Praga Čáslav, a. s.	45147914	23/27
32	Praga Hrádek n. N. a. s.	46708251	24/39
33	Pragometal s. r. o.	14892600	5/18/99
34	Pražská strojírna, a. s.	60193298	29/19
35	První brněnská strojírna Holding, akciová společnost	49455737	29/57
36	První brněnská strojírna Třebíč, a. s.	60703952	27/24
37	První brněnská strojírna Velká Bíteš, a. s.	00176109	27/38
38	Přerovské strojírny a. s.	00008311	29/74
39	PSP Speciální strojírna a. s.	47677953	23/23
40	Roudnické strojírny a slévárny a. s.	49903187	29/42
41	Sázavan strojírny s. r. o.	48950874	18/20
42	Slovácké strojírny akciová společnost	00008702	30/70
43	Sokolovské strojírny	45359105	22/18
44	Stabilní hasící zařízení, spol. s r. o.	46507892	5/16/99
45	Strojírny a opravy a. s. Kaplice	60827467	24/22
46	Strojírny Bohdalice a. s.	45274029	27/40
47	Strojírny Poldi, s. r. o., Kladno	46358404	24/12
48	Strojírny Prostějov, a. s.	46900331	27/57
49	Strojírny Slaný s. r. o.	62957007	1/12/99
50	Strojtex a. s.	46504893	22/24
51	Škoda Elcar s. r. o.	61776297	29/72
52	Škoda Praha a. s.	00128201	21/6
53	Škoda, Hutě, Plzeň, s. r. o.	47718706	5/17/99
54	Škoda, Kovárny, Plzeň, s. r. o.	47718587	5/17/99
55	TOS Svitavy, a. s.	15034020	27/26
56	TOS Znojmo, akciová společnost	46347691	27/29
57	Toshulin, s. r. o.	49974785	24/13
58	Továrny textilních potřeb - ELITEX, a. s.	00010936	28/33
59	TRW DAS a. s.	48244716	27/13
60	Vagónka Česká Lípa akciová společnost	49902083	26/12
61	Válcovny plechu, a. s.	14613581	5/9/99
62	Vitkovice, a. s.	45193070	29/78
63	Východočeské plynárenské strojírny a. s.	60108819	25/45
64	Zbrojovka Brno a. s.	46346066	26/13
65	ZPS - Elektromontáže, a. s.	44118856	25/21
66	ZPS - Nástrojárna, a. s.	48531367	26/41
67	ZDAS, a. s.	46347160	28/62
68	Železniční opravy a strojírny spol. s r. o.	46507035	11/8

Popis sloupků vstupní tabulky podniků

1	Nehmotný investiční majetek / Aktiva	v roce 1997
2	Hmotný investiční majetek / Aktiva	v roce 1997
3	Finanční investice / Aktiva	v roce 1997
4	Zásoby / Aktiva	v roce 1997
5	Pohledávky / Aktiva	v roce 1997
6	Finanční majetek / Aktiva	v roce 1997
7	Ostatní aktiva / Aktiva	v roce 1997
8	Základní jmění / Pasiva	v roce 1997
9	Kapitálové fondy / Pasiva	v roce 1997
10	Fondy tvořené ze zisku / Pasiva	v roce 1997
11	Hospodářský výsledek minulých let / Pasiva	v roce 1997
12	Hospodářský výsledek účetního období / Pasiva	v roce 1997
13	Rezervy / Pasiva	v roce 1997
14	Závazky / Pasiva	v roce 1997
15	Bankovní úvěry a výpomoci / Pasiva	v roce 1997
16	Ostatní pasiva / Pasiva	v roce 1997
17	Nehmotný investiční majetek / Aktiva	v roce 1996
18	Hmotný investiční majetek / Aktiva	v roce 1996
19	Finanční investice / Aktiva	v roce 1996
20	Zásoby / Aktiva	v roce 1996
21	Pohledávky / Aktiva	v roce 1996
22	Finanční majetek / Aktiva	v roce 1996
23	Ostatní aktiva / Aktiva	v roce 1996
24	Základní jmění / Pasiva	v roce 1996
25	Kapitálové fondy / Pasiva	v roce 1996
26	Fondy tvořené ze zisku / Pasiva	v roce 1996
27	Hospodářský výsledek minulých let / Pasiva	v roce 1996
28	Hospodářský výsledek účetního období / Pasiva	v roce 1996
29	Rezervy / Pasiva	v roce 1996
30	Závazky / Pasiva	v roce 1996
31	Bankovní úvěry a výpomoci / Pasiva	v roce 1996
32	Ostatní pasiva / Pasiva	v roce 1996
33	ln (Tržby v roce 1997 / Tržby v roce 1996)	
34	ln (Přidaná hodnota 97 / Přidaná hodnota 96)	
35	ln (Osobní náklady 97 / Osobní náklady 96)	
36	Tržby za prodané vlastní výrobky, služby a zboží / Aktiva	v roce 1997
37	Přidaná hodnota / Aktiva	v roce 1997
38	Osobní náklady / Aktiva	v roce 1997

Vstupní tabulka dat

Podniky

1. část

	1	2	3	4	5	6	7	8	9	10
1	0,002311	0,290118	0,141151	0,237378	0,249175	0,023980	0,055888	0,355083	0,015049	0,028
2	0,005103	0,328486	0,013531	0,391478	0,170498	0,058573	0,032331	0,565947	0,000000	0,066
3	0,006792	0,463889	0,000347	0,286235	0,150105	0,082789	0,009843	0,596090	0,000000	0,102
4	0,016993	0,458277	0,081901	0,243389	0,142468	0,026445	0,030526	0,514761	0,000000	0,041
5	0,000640	0,002187	0,022062	0,371740	0,113528	0,462704	0,027139	0,030041	0,000000	0,049
6	0,001356	0,087285	0,000979	0,583632	0,179673	0,110216	0,036859	0,074528	0,000000	0,010
7	0,002230	0,097687	0,041075	0,217753	0,640610	0,000222	0,000424	0,412370	0,000000	0,035
8	0,005226	0,332912	0,013718	0,488666	0,121133	0,034965	0,003381	0,807853	0,000000	0,000
9	0,005720	0,373043	0,000000	0,514527	0,091952	0,004874	0,009885	0,362064	0,000338	0,305
10	0,059215	0,326337	0,000219	0,349481	0,207600	0,053664	0,003483	0,278394	0,000000	0,013
11	0,000959	0,456282	0,000000	0,379164	0,122758	0,019936	0,020901	0,004447	0,000000	0,006
12	0,001118	0,590644	0,007910	0,183217	0,181840	0,021539	0,013731	0,678913	0,001060	0,085
13	0,000750	0,761566	0,000000	0,156989	0,068091	0,007873	0,004731	0,641205	0,000000	0,069
14	0,000512	0,396464	0,011349	0,217466	0,299024	0,061898	0,013287	0,403318	0,060621	0,075
15	0,022586	0,639964	0,006792	0,164838	0,115577	0,031864	0,018378	0,185032	0,312647	0,054
16	0,006365	0,708488	0,000000	0,085866	0,054125	0,097897	0,047260	0,324947	0,000099	0,000
17	0,000000	0,355849	0,000287	0,194847	0,246260	0,202126	0,000630	0,078208	0,000401	0,555
18	0,025314	0,559214	0,010992	0,227176	0,158080	0,006349	0,012874	0,394130	0,000116	0,000
19	0,011263	0,165291	0,002359	0,100543	0,498172	0,204328	0,018045	0,327279	0,000000	0,017
20	0,002197	0,632722	0,002680	0,135223	0,156311	0,046617	0,024251	0,675560	0,009200	0,084
21	0,011083	0,231093	0,000211	0,272222	0,267904	0,215448	0,002038	0,450551	0,051875	0,026
22	0,002107	0,489852	0,000349	0,196092	0,082818	0,226058	0,002724	0,685969	0,005421	0,079
23	0,001226	0,492678	0,003381	0,238299	0,110771	0,130212	0,023435	0,451134	0,107707	0,047
24	0,000000	0,320272	0,001473	0,283090	0,348736	0,034413	0,012017	0,141409	0,008856	0,014
25	0,005175	0,749426	0,000000	0,082711	0,146931	0,015176	0,000581	0,872169	0,000000	0,000
26	0,001010	0,399224	0,032782	0,134943	0,239162	0,164441	0,028437	0,434007	0,002134	0,006
27	0,001394	0,425710	0,120634	0,125145	0,221002	0,036050	0,070065	0,358443	0,000000	0,000
28	0,000000	0,424615	0,001005	0,380691	0,175533	0,016679	0,001477	0,669286	0,000000	0,113
29	0,007173	0,336568	0,000000	0,475494	0,169653	0,000443	0,010669	0,057386	0,000000	0,005
30	0,000000	0,610405	0,009406	0,109487	0,166598	0,094555	0,009549	0,651097	0,004451	0,081
31	0,000961	0,234031	0,000000	0,392118	0,299316	0,064935	0,008640	0,422203	0,000000	0,006
32	0,081370	0,319681	0,000000	0,223050	0,296956	0,057621	0,021321	0,587200	0,000000	0,079
33	0,000798	0,642460	0,000000	0,073611	0,192066	0,046826	0,044238	0,006607	0,000000	0,020
34	0,001636	0,570154	0,000000	0,189553	0,121591	0,114265	0,002801	0,581333	0,000036	0,006
35	0,001871	0,028798	0,418635	0,027969	0,514284	0,005905	0,002539	0,028909	0,000000	0,001
36	0,009642	0,411077	0,001718	0,211094	0,311916	0,007868	0,046685	0,318384	0,000000	0,036
37	0,069547	0,324531	0,131564	0,223241	0,142773	0,067163	0,041181	0,456693	0,000000	0,071
38	0,026315	0,314828	0,384491	0,037483	0,228970	0,000592	0,007321	0,537721	0,000010	0,100
39	0,000000	0,438555	0,000000	0,263579	0,209146	0,020056	0,068664	0,806949	0,000000	0,000
40	0,000277	0,354562	0,020452	0,079103	0,507751	0,034246	0,003609	0,673958	0,000000	0,013
41	0,001164	0,180267	0,002695	0,414086	0,282228	0,117348	0,002211	0,157004	0,095354	0,034
42	0,005181	0,472251	0,000000	0,191688	0,165216	0,150196	0,015467	0,679410	0,014198	0,000
43	0,001700	0,392664	0,173040	0,200645	0,185229	0,042646	0,004075	0,708967	0,000000	0,073
44	0,000689	0,035954	0,000000	0,478014	0,256980	0,201997	0,026367	0,001837	0,000000	0,002
45	0,000000	0,558283	0,030868	0,182249	0,169902	0,058697	0,000000	1,007181	-0,067373	0,090
46	0,003353	0,358925	0,000000	0,255177	0,250822	0,129265	0,002458	0,581439	0,000000	0,015
47	0,000505	0,364678	0,037938	0,334893	0,212437	0,047371	0,002179	0,454838	0,000000	0,050
48	0,000000	0,536183	0,002236	0,124921	0,264000	0,069001	0,003659	0,758075	0,000112	0,000
49	0,000000	0,185570	0,494557	0,077703	0,155748	0,086243	0,000178	0,001485	0,000000	0,000
50	0,002121	0,551272	0,001337	0,216687	0,186340	0,036942	0,005300	0,643191	0,000000	0,122
51	0,002161	0,233392	0,000000	0,108593	0,642584	0,013131	0,000139	0,869882	0,009724	0,027
52	0,000722	0,020415	0,012286	0,387792	0,501646	0,069326	0,007812	0,027245	0,000039	0,000
53	0,012731	0,400912	0,000000	0,339002	0,204575	0,004647	0,038132	0,269292	0,000000	0,010
54	0,001685	0,586420	0,000000	0,207391	0,164896	0,008114	0,031494	0,331965	0,000000	0,000
55	0,002371	0,429948	0,002702	0,353183	0,180826	0,025086	0,005884	0,683494	0,000000	0,000
56	0,001099	0,301551	0,063755	0,458602	0,124166	0,044402	0,006426	0,565222	0,000000	0,000
57	0,003322	0,242919	0,003816	0,429797	0,274214	0,026239	0,019693	0,000508	0,000000	0,000
58	0,004332	0,484014	0,004870	0,173469	0,325168	0,005642	0,002506	0,712159	0,000000	0,030
59	0,002559	0,453234	0,000000	0,212539	0,283710	0,042828	0,005130	0,406478	0,000000	0,130
60	0,016248	0,448051	0,000275	0,160976	0,283258	0,075173	0,016019	0,262857	0,000000	0,000
61	0,001196	0,339715	0,041272	0,202396	0,370216	0,034792	0,010413	0,400995	0,000548	0,040
62	0,007991	0,545989	0,043086	0,242671	0,125092	0,023342	0,011830	0,480237	0,025989	0,070
63	0,001098	0,418931	0,012248	0,292597	0,123168	0,076988	0,074970	0,577230	0,000000	0,170
64	0,014039	0,621421	0,014566	0,190606	0,121402	0,010820	0,000893	0,877189	0,000442	0,080
65	0,035498	0,260823	0,000000	0,543175	0,148792	0,031445	0,013188	0,369826	0,000000	0,000
66	0,041849	0,436352	0,001553	0,047078	0,428536	0,031445	0,013188	0,535380	0,000000	0,020
67	0,006290	0,386146	0,012731	0,223510	0,344957	0,008508	0,017859	0,398953	0,000798	0,050
68	0,050220	0,201430	0,001273	0,216223	0,521818	0,001179	0,007857	0,000936	0,000000	0,000

Vstupní tabulka dat

Podniky

2. část

	11	12	13	14	15	16	17	18	19	20
1	0,000000	0,008262	0,018131	0,279643	0,280652	0,014837	0,002163	0,284043	0,147548	0,237
2	0,009689	0,003159	0,000000	0,079535	0,262920	0,012609	0,005963	0,356687	0,012392	0,409
3	0,009183	0,022974	0,008538	0,146950	0,111662	0,002525	0,007336	0,426013	0,000362	0,304
4	-0,097602	0,010671	0,032103	0,222390	0,267908	0,008361	0,008345	0,489752	0,089861	0,238
5	-0,025595	0,014384	0,000000	0,578699	0,301744	0,051163	0,000518	0,005097	0,028507	0,276
6	0,034293	0,019339	0,000107	0,808831	0,016685	0,035270	0,001251	0,085179	0,000962	0,518
7	0,000000	-0,064741	0,000000	0,186814	0,426272	0,003403	0,002583	0,287879	0,054774	0,244
8	-0,133199	-0,041339	0,000000	0,211873	0,136379	0,018434	0,001966	0,387581	0,010809	0,447
9	0,000000	0,022407	0,008058	0,097523	0,187554	0,016932	0,007448	0,312167	0,000000	0,403
10	-0,137282	0,024633	0,061349	0,409746	0,287175	0,062753	0,104508	0,331468	0,000285	0,399
11	0,044064	0,000667	0,053015	0,274340	0,596555	0,020558	0,003742	0,489489	0,000000	0,328
12	0,004136	-0,085085	0,015078	0,152855	0,146683	0,001284	0,002091	0,595859	0,007166	0,183
13	0,002264	-0,016151	0,000000	0,202870	0,099928	0,000577	0,001079	0,663001	0,036463	0,177
14	-0,058743	0,017958	0,014389	0,182244	0,279961	0,024901	0,001092	0,423040	0,016287	0,253
15	0,064237	-0,016639	0,012753	0,155519	0,225064	0,006732	0,023725	0,631285	0,006367	0,166
16	-0,102331	-0,012099	0,078930	0,267380	0,395291	0,047734	0,008214	0,774233	0,000000	0,077
17	-0,002837	0,091362	0,100017	0,159884	0,008970	0,008970	0,000000	0,404843	0,000317	0,259
18	-0,039237	-0,015997	0,000000	0,264641	0,373223	0,023123	0,015291	0,567416	0,008603	0,218
19	0,154794	0,022172	0,042576	0,428234	0,000000	0,007843	0,013780	0,196385	0,002564	0,089
20	-0,074377	-0,143878	0,061292	0,131704	0,252144	0,004082	0,002108	0,622880	0,016323	0,129
21	0,060130	0,071516	0,105464	0,228913	0,000000	0,005520	0,002058	0,245286	0,000243	0,287
22	0,106117	0,026843	0,014470	0,077648	0,000000	0,003872	0,002863	0,506761	0,000682	0,211
23	0,000000	0,130359	0,000000	0,253555	0,000000	0,009319	0,002205	0,545441	0,000216	0,203
24	0,032548	0,055182	0,041884	0,282495	0,386528	0,036649	0,000006	0,319105	0,001790	0,395
25	-0,010379	-0,019973	0,050063	0,101143	0,000000	0,006977	0,000000	0,831758	0,000000	0,088
26	0,000000	0,058664	0,000000	0,130936	0,358492	0,009128	0,001325	0,448428	0,001584	0,183
27	-0,207752	-0,049203	0,002297	0,244086	0,608542	0,043551	0,002262	0,378180	0,154261	0,152
28	-0,175553	-0,014074	0,000000	0,245881	0,159510	0,001498	0,000000	0,434965	0,000983	0,407
29	0,466703	0,120901	0,000000	0,220727	0,128388	0,000157	0,008517	0,350464	0,000000	0,411
30	0,031053	0,004761	0,065793	0,078486	0,081513	0,001410	0,000000	0,550914	0,106277	0,097
31	-0,011710	0,000232	0,037732	0,471407	0,013410	0,060116	0,000787	0,360330	0,000000	0,251
32	-0,012146	0,001613	0,017536	0,159245	0,159181	0,007720	0,093030	0,324113	0,000000	0,288
33	0,055387	0,021858	0,084815	0,480097	0,156307	0,174476	0,000365	0,681387	0,000000	0,038
34	0,000000	0,005048	0,093746	0,215845	0,082842	0,014772	0,000463	0,717581	0,000000	0,134
35	0,002170	0,017837	0,000000	0,785348	0,048996	0,115279	0,000796	0,002227	0,210016	0,031
36	0,048705	-0,061656	0,001073	0,424098	0,196641	0,036278	0,008633	0,361381	0,002377	0,262
37	0,073128	0,015035	0,000143	0,126557	0,252635	0,004294	0,006027	0,370699	0,042194	0,321
38	0,000000	-0,440465	0,002893	0,239501	0,477444	0,082126	0,059492	0,285420	0,389147	0,039
39	-0,092080	-0,223319	0,000000	0,494073	0,000000	0,014378	0,000318	0,519537	0,000000	0,207
40	0,000000	-0,067724	0,003022	0,306020	0,067333	0,003697	0,000349	0,425203	0,017467	0,331
41	0,259280	0,181052	0,080229	0,102746	0,075571	0,013986	0,002705	0,222611	0,002917	0,471
42	-0,091885	0,004752	0,014267	0,258522	0,105350	0,015385	0,002149	0,496717	0,059759	0,240
43	0,010495	0,008731	0,001516	0,080774	0,112723	0,003225	0,004269	0,407921	0,175739	0,204
44	0,123374	0,113107	0,027129	0,703824	0,000000	0,028102	0,000592	0,075112	0,000000	0,544
45	-0,074651	-0,028057	0,000000	0,055887	0,016246	0,000406	0,000000	0,556646	0,000000	0,198
46	-0,009806	0,048040	0,000000	0,228137	0,133046	0,003807	0,004912	0,546193	0,000000	0,279
47	0,000000	0,010157	0,024015	0,165989	0,290823	0,004123	0,000194	0,300797	0,057398	0,292
48	-0,221110	0,008568	0,000000	0,114971	0,331639	0,005651	0,000000	0,535417	0,002217	0,142
49	-0,026525	0,056525	0,036386	0,928965	0,000000	0,003163	0,000000	0,000000	0,000000	0,409
50	0,003910	-0,004593	0,000000	0,104986	0,117892	0,012247	0,034685	0,531670	0,001706	0,214
51	-0,459942	0,015707	0,000000	0,523658	0,000000	0,013851	0,002987	0,227369	0,000000	0,242
52	0,000742	0,000798	0,003871	0,894725	0,017273	0,051514	0,000595	0,021794	0,013240	0,344
53	-0,216100	-0,206163	0,034871	0,583894	0,506710	0,014450	0,011440	0,389858	0,000000	0,353
54	-0,000478	-0,044355	0,027057	0,134347	0,501584	0,041255	0,003175	0,547674	0,000000	0,268
55	-0,075985	-0,048386	0,010695	0,125647	0,295140	0,009194	0,001216	0,418345	0,002460	0,365
56	0,000000	-0,017755	0,002165	0,164413	0,278576	0,002108	0,000337	0,315136	0,108019	0,372
57	0,037899	0,030835	0,021381	0,447329	0,457757	0,003287	0,001975	0,303321	0,006483	0,416
58	-0,231057	-0,099906	0,000000	0,350301	0,212501	0,019448	0,003382	0,504975	0,001340	0,161
59	0,000000	0,153669	0,006664	0,151544	0,089795	0,055112	0,003219	0,548037	0,000000	0,207
60	-0,152994	-0,078867	0,022643	0,904122	0,027356	0,014620	0,000703	0,479632	0,000403	0,156
61	-0,025641	0,000464	0,000000	0,298497	0,252406	0,030588	0,001107	0,366314	0,013140	0,244
62	-0,068391	0,002736	0,013068	0,271742	0,172862	0,030622	0,007009	0,564081	0,030935	0,251
63	0,000000	0,054835	0,000647	0,132344	0,054575	0,004372	0,001022	0,412457	0,000000	0,279
64	-0,243001	-0,094089	0,034263	0,147567	0,186912	0,004956	0,026515	0,610583	0,009857	0,179
65	0,000000	0,000586	0,000000	0,586064	0,036983	0,000565	0,058700	0,361694	0,000000	0,266
66	0,000000	0,065834	0,000089	0,185705	0,175712	0,008298	0,088010	0,515829	0,000413	0,051
67	0,068549	0,014178	0,000000	0,176059	0,278777	0,011726	0,004038	0,423330	0,015373	0,211
68	0,207895	-0,044530	0,107264	0,356997	0,335207	0,029448	0,012349	0,205872	0,000409	0,351

Vstupní tabulka dat

Podniky

3. část

	21	22	23	24	25	26	27	28	29	30
1	0,265163	0,023117	0,040693	0,386537	0,016375	0,043909	0,015985	-0,028476	0,003677	0,283
2	0,172891	0,026317	0,015950	0,597157	0,000000	0,068475	0,010223	0,003142	0,000000	0,070
3	0,230735	0,029187	0,001641	0,622017	0,000000	0,105280	0,000000	0,021732	0,000000	0,133
4	0,126258	0,037004	0,010058	0,508830	0,000000	0,040328	-0,107796	0,012125	0,005333	0,244
5	0,102547	0,566887	0,019811	0,042835	0,000000	0,070673	-0,067979	0,031484	0,000000	0,80
6	0,322664	0,068358	0,003350	0,072753	0,000000	0,009702	0,019769	0,018665	0,002547	0,818
7	0,398265	0,009160	0,002470	0,315477	0,000000	0,032010	0,026635	-0,030925	0,002600	0,361
8	0,120110	0,026940	0,005425	0,760260	0,000000	0,001554	0,000743	-0,126845	0,000000	0,230
9	0,241941	0,023899	0,010996	0,295861	0,000000	0,051346	0,147828	0,051074	0,009316	0,155
10	0,137949	0,025819	0,000329	0,246127	0,000091	0,011699	-0,032944	-0,088426	0,157484	0,188
11	0,138705	0,021456	0,017937	0,004113	0,000041	0,000000	0,000229	0,046349	0,032348	0,322
12	0,181277	0,015093	0,015322	0,677791	0,001058	0,074152	-0,080840	0,098779	0,003949	0,095
13	0,105510	0,014082	0,001984	0,623628	0,000000	0,066064	0,017145	-0,005943	0,000000	0,195
14	0,236705	0,056934	0,012015	0,426593	0,064118	0,080425	-0,063026	0,000893	0,012351	0,192
15	0,151841	0,013759	0,006776	0,173410	0,293019	0,049806	0,055264	0,009468	0,026953	0,184
16	0,065886	0,072124	0,002353	0,361562	0,000110	0,000053	-0,086065	-0,027797	0,025000	0,194
17	0,142513	0,190123	0,002631	0,094713	0,000444	0,652529	-0,055788	0,052650	0,038196	0,192
18	0,184161	0,002287	0,003246	0,416377	0,000123	0,000000	0,000000	-0,041373	0,000000	0,266
19	0,538008	0,148571	0,011473	0,355724	0,000000	0,024740	0,127227	0,081079	0,053903	0,357
20	0,172148	0,030312	0,026252	0,536777	0,000000	0,125073	0,002399	-0,119452	0,099090	0,100
21	0,310532	0,151733	0,002956	0,507717	0,058457	0,026850	0,022164	0,050322	0,113014	0,196
22	0,109183	0,166221	0,003084	0,670147	0,005296	0,076252	0,097889	0,031142	0,037661	0,078
23	0,108673	0,135782	0,003826	0,500584	0,110186	0,053483	-0,066886	0,108803	0,000000	0,284
24	0,241602	0,037369	0,004450	0,086417	0,009277	0,008836	0,025175	0,087111	0,138702	0,296
25	0,062860	0,017295	0,000000	0,877937	0,000000	0,000000	0,000000	-0,010447	0,006702	0,122
26	0,259267	0,103900	0,002079	0,536988	0,000539	0,029317	0,029458	-0,047028	0,008560	0,088
27	0,207860	0,069707	0,034823	0,326572	0,000000	0,000033	-0,010913	-0,178367	0,017242	0,200
28	0,098702	0,055251	0,002549	0,654725	0,000000	0,113854	-0,143577	-0,028157	0,000000	0,223
29	0,156567	0,061687	0,011621	0,003980	0,000000	0,000478	0,346776	0,180791	0,023693	0,230
30	0,162566	0,072727	0,009978	0,616850	0,004217	0,079677	0,043516	-0,012001	0,141654	0,102
31	0,325106	0,057039	0,004974	0,603297	0,000000	0,009446	-0,011644	-0,005089	0,023534	0,228
32	0,232937	0,040074	0,021784	0,522194	0,000000	0,078016	0,000000	-0,011807	0,034648	0,138
33	0,133915	0,026429	0,119831	0,006734	0,000000	0,020846	0,037877	0,018546	0,057629	0,574
34	0,118645	0,025811	0,002953	0,722328	0,000045	0,006253	0,000000	0,005610	0,023192	0,165
35	0,751494	0,001306	0,002358	0,004778	0,000000	0,001075	0,020520	0,001203	0,000116	0,891
36	0,263389	0,042807	0,058965	0,323174	0,000000	0,037025	0,054407	-0,004967	0,024612	0,436
37	0,175823	0,053416	0,030603	0,424117	0,000000	0,065316	0,057721	0,012529	0,000132	0,205
38	0,192507	0,025995	0,008433	0,402864	0,027809	0,066407	0,000000	0,009155	0,001966	0,099
39	0,154071	0,024681	0,094029	0,794775	0,000000	0,000000	-0,078712	-0,011979	0,000000	0,271
40	0,118151	0,032115	0,075394	0,575613	0,000000	0,076296	0,001297	-0,065897	0,000517	0,226
41	0,128194	0,156914	0,015493	0,169943	0,101852	0,038110	0,107390	0,188552	0,090707	0,210
42	0,116999	0,076602	0,007121	0,701439	0,014659	0,027518	0,000000	-0,121736	0,005655	0,210
43	0,183922	0,019568	0,004571	0,720625	0,000000	0,074093	0,000000	0,013545	0,001568	0,051
44	0,169929	0,161245	0,048398	0,002041	0,000000	0,002551	0,084194	0,062827	0,016673	0,786
45	0,178221	0,065830	0,000900	0,945795	-0,063267	0,084854	-0,053381	-0,016721	0,000000	0,062
46	0,125940	0,042954	0,000408	0,710290	0,000000	0,019176	0,000000	-0,011979	0,000000	0,217
47	0,345988	0,000769	0,001997	0,576828	0,000000	0,046913	-0,040144	0,007347	0,014903	0,165
48	0,309218	0,001431	0,009422	0,751650	0,000111	0,002076	-0,193589	-0,025647	0,000000	0,152
49	0,406617	0,181527	0,002459	0,005345	0,000000	0,000000	0,000000	-0,095467	0,001978	1,072
50	0,192828	0,020420	0,003703	0,605550	0,000000	0,114193	0,000000	0,006171	0,000000	0,104
51	0,495952	0,031677	0,000000	0,822726	0,009196	0,011292	-0,435009	0,017633	0,000000	0,574
52	0,556599	0,054127	0,009122	0,027936	0,000040	0,003811	0,001916	0,000152	0,004563	0,900
53	0,219055	0,019787	0,005914	0,268541	0,000000	0,014046	-0,138801	-0,076696	0,000001	0,538
54	0,152560	0,021024	0,007057	0,380174	0,000000	0,010540	-0,003632	0,003247	0,009333	0,134
55	0,169685	0,040866	0,002409	0,627403	0,000000	0,064179	0,018991	-0,152697	0,011171	0,128
56	0,164889	0,034742	0,004412	0,533344	0,000000	0,056908	-0,047608	-0,004813	0,000000	0,189
57	0,189390	0,067728	0,014887	0,000671	0,000000	0,000023	0,000000	0,056751	0,005054	0,348
58	0,321400	0,006104	0,001080	0,783769	0,000000	0,040231	-0,012088	-0,242205	0,020864	0,176
59	0,231373	0,001432	0,008246	0,499961	0,000000	0,053007	0,000000	0,115175	0,000000	0,160
60	0,150718	0,187972	0,024078	0,387602	0,000000	0,001130	-0,078610	-0,145666	0,021674	0,481
61	0,345460	0,025172	0,004241	0,441021	0,000600	0,045208	-0,049634	0,022825	0,002180	0,281
62	0,110458	0,031198	0,005291	0,500986	0,028106	0,073807	0,000000	-0,071346	0,014304	0,261
63	0,125745	0,084145	0,096759	0,577175	0,000000	0,097762	0,038389	0,042513	0,000000	0,160
64	0,144733	0,011973	0,016832	0,821131	0,000414	0,080996	-0,182673	-0,044800	0,034123	0,117
65	0,257731	0,053643	0,001591	0,551398	0,000000	0,007215	0,000000	0,033887	0,000000	0,248
66	0,257687	0,084001	0,003244	0,619937	0,000000	0,029282	0,000000	0,057489	0,005269	0,197
67	0,329053	0,013211	0,000589	0,430776	0,000775	0,054119	0,075762	0,000764	0,000169	0,165
68	0,412626	0,001721	0,011615	0,001003	0,000000	0,005888	0,161665	0,064092	0,100092	0,380

Vstupní tabulka dat

Podniky

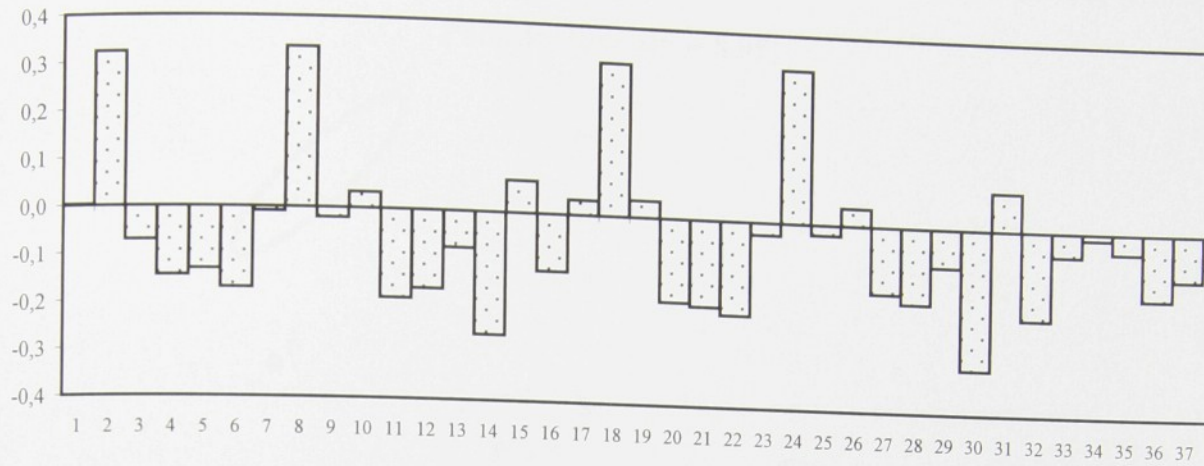
4. část

	31	32	33	34	35	36	37	38
1	0,274428	0,003983	0,134692	0,235804	0,197885	0,528162	0,250137	0,149198
2	0,240805	0,009512	0,007030	-0,019938	0,036619	0,748560	0,306571	0,229133
3	0,115261	0,002651	0,094702	0,126293	0,095835	0,909943	0,432362	0,287173
4	0,287021	0,009474	0,090912	0,128008	0,057738	0,935991	0,467304	0,315244
5	0,000000	0,119977	-0,770344	-1,367192	-0,407504	0,230562	0,015678	0,042196
6	0,018662	0,038950	0,032002	-0,031377	0,083727	0,869261	0,124509	0,086656
7	0,286467	0,006082	-0,605134	-1,744852	-2,631189	0,283816	0,028567	0,011967
8	0,133391	0,000250	-0,161794	-0,264056	0,009278	0,523351	0,231807	0,239409
9	0,250295	0,039071	0,190407	-0,075362	0,081749	1,273455	0,387143	0,254779
10	0,481586	0,036191	0,189880	0,540035	0,091605	1,492339	0,378851	0,215614
11	0,581687	0,012960	0,048742	0,146070	0,141008	1,428081	0,594986	0,409867
12	0,127205	0,002277	-0,004329	-0,017259	-0,002571	0,985893	0,399976	0,335785
13	0,102097	0,001736	-0,104536	-0,024005	-0,017857	0,505226	0,147211	0,123600
14	0,260947	0,024753	0,025643	0,052076	0,012486	1,251182	0,377047	0,225711
15	0,201784	0,005620	0,041568	-0,010503	0,021416	0,924927	0,352906	0,232504
16	0,492446	0,040634	0,132839	0,211442	0,053089	0,411588	0,135136	0,058907
17	0,014866	0,009763	0,283794	0,351366	0,119722	1,432223	0,757580	0,453115
18	0,343943	0,014289	0,004081	0,200199	0,088841	0,570451	0,236283	0,180996
19	0,000000	0,000320	0,284590	0,128869	0,151681	4,164406	1,914259	1,832292
20	0,237051	0,018247	-0,105303	-0,137997	-0,040838	0,956135	0,441013	0,350352
21	0,001429	0,023109	0,048372	0,088981	0,152109	1,363200	0,307002	0,120456
22	0,000000	0,002989	-0,088054	-0,197399	-0,013478	0,710883	0,216196	0,155533
23	0,000000	0,009706	0,184919	0,207894	0,189956	0,710437	0,542168	0,421761
24	0,290952	0,057458	-0,008233	0,063492	0,149044	1,381658	0,563122	0,470366
25	0,000000	0,003102	0,907651	0,889149	0,915019	0,846004	0,549467	0,433933
26	0,339307	0,013884	-0,036148	-0,138894	0,047768	0,685776	0,454539	0,277856
27	0,617792	0,027565	0,091723	0,459497	0,066459	1,757005	0,271218	0,186308
28	0,178036	0,001284	-0,105269	-0,031908	-0,003626	0,765300	0,330304	0,254129
29	0,214036	-0,000133	-0,119483	-0,141367	0,123193	1,301615	0,605186	0,401257
30	0,022065	0,001630	0,050598	0,040095	0,082187	1,169832	0,511929	0,444571
31	0,146664	0,005367	0,735360	0,306377	0,370021	1,139211	0,251342	0,184341
32	0,219832	0,018701	0,231179	0,225406	0,261310	0,788647	0,382377	0,220972
33	0,262078	0,022165	0,014069	0,182876	0,254781	1,107389	0,337664	0,184083
34	0,051467	0,025682	0,032482	0,586166	0,295231	0,698062	0,361429	0,239133
35	0,067051	0,013415	0,816479	0,515397	0,739314	0,421158	0,151910	0,052155
36	0,123335	0,005691	-0,036703	-0,295312	-0,018663	0,913443	0,201079	0,219027
37	0,230149	0,004176	0,043537	-0,141616	-0,052796	0,738960	0,227731	0,214250
38	0,372901	0,019418	-0,326251	-0,665415	-0,504127	0,129436	0,039286	0,013504
39	0,000000	0,024869	-0,365315	-0,438062	-0,048079	1,078773	0,424939	0,534992
40	0,182627	0,002651	-1,002371	-1,292471	-1,297706	0,404263	0,087231	0,082120
41	0,070809	0,022362	0,069732	0,109224	0,120407	1,373134	0,730620	0,410710
42	0,156204	0,005639	0,010449	-0,202054	0,018640	0,765849	0,269453	0,227857
43	0,138223	0,000132	0,264734	0,043718	0,030152	0,617188	0,220066	0,176030
44	0,000000	0,045071	0,161990	0,293302	0,165461	1,701951	0,583894	0,278589
45	0,039666	0,000122	-0,238288	-0,038846	0,013975	0,809368	0,308433	0,303202
46	0,064643	0,000816	0,224270	0,113832	0,088629	0,792249	0,292289	0,287774
47	0,221810	0,006676	-0,033466	-0,064343	0,121085	1,236584	0,498512	0,373868
48	0,308200	0,004988	0,068570	0,411021	0,202423	0,730928	0,350361	0,201941
49	0,016036	0,000000	0,043108	0,414577	0,107687	0,753271	0,356126	0,295457
50	0,163049	0,006624	-0,035868	-0,054214	-0,002239	1,267867	0,533844	0,408238
51	0,000000	0,000000	-0,397765	-0,133119	0,190785	0,762009	0,394371	0,269544
52	0,006743	0,054323	0,134354	0,031887	0,106676	0,477624	0,034957	0,027772
53	0,363166	0,031412	0,081738	-0,132890	0,023397	1,547554	0,308753	0,241757
54	0,445616	0,019933	0,016133	-0,145306	-0,027661	0,831959	0,122058	0,094188
55	0,295835	0,006782	0,012423	0,085760	-0,058537	0,718023	0,273724	0,231444
56	0,271180	0,001679	-0,137423	-0,084536	0,059190	0,692913	0,298156	0,203323
57	0,584216	0,004521	0,653811	0,640003	0,846852	1,047025	0,422882	0,246737
58	0,228992	0,003797	0,079883	-0,405492	0,076402	0,558933	0,092090	0,130688
59	0,133721	0,037190	0,218654	0,247824	0,125029	1,914846	0,631676	0,206980
60	0,299535	0,033184	0,755044	0,681857	0,141967	1,004085	0,215970	0,165872
61	0,246829	0,009057	0,137060	-0,113118	0,115878	1,083367	0,168855	0,116174
62	0,174737	0,018073	0,032175	0,194548	0,016889	0,648698	0,211099	0,121840
63	0,077413	0,006212	0,132337	0,210719	0,125178	1,150823	0,390076	0,223071
64	0,170545	0,002769	-0,007116	0,025394	-0,012821	0,493685	0,256804	0,252957
65	0,157542	0,000969	0,231579	-0,049481	0,080405	0,699685	0,181643	0,136017
66	0,090449	0,000165	0,110496	0,168964	0,158376	0,926671	0,608869	0,376729
67	0,261365	0,007087	0,192502	0,129285	0,035409	0,755975	0,284981	0,190249
68	0,285778	0,000939	0,015149	-0,009610	0,069144	1,184346	0,444620	0,343469

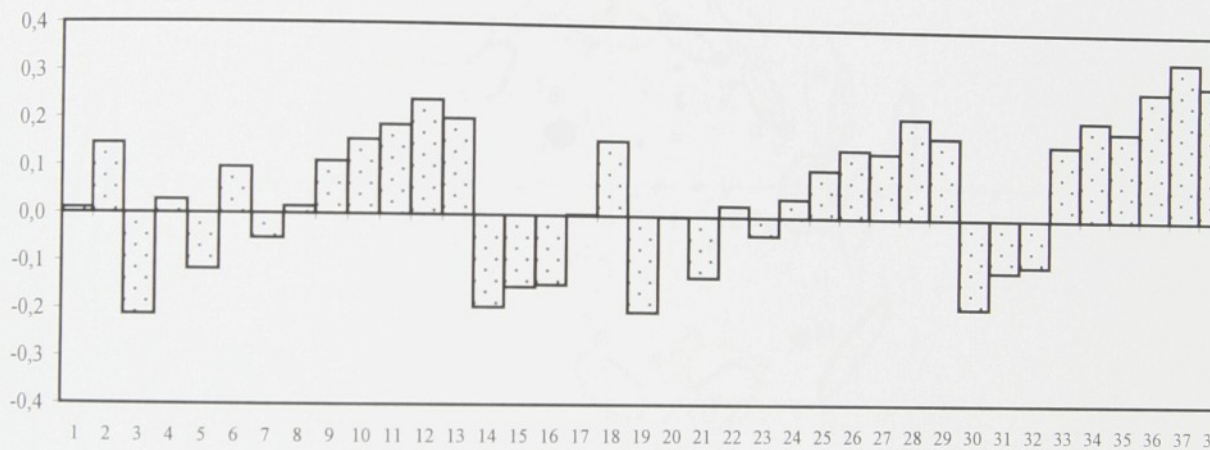
Sloupcové grafy 1. a 2. pohledu a vlastních čísel

Podniky

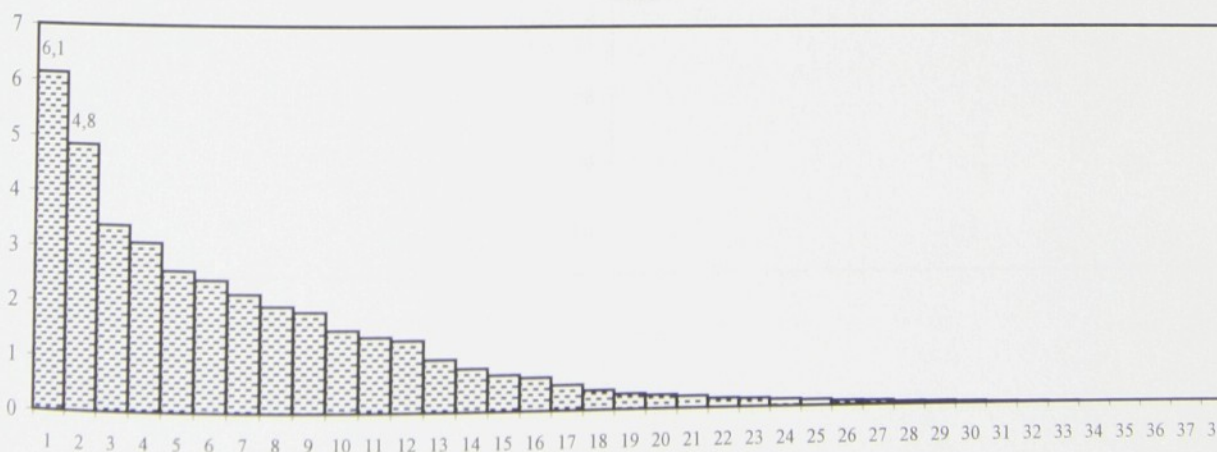
1. pohled



2. pohled



Vlastní čísla



Zachovaná informace
29 %

Rozptylový graf a vybrané konfidenční elipsoidy



Souřadnice objektů na rozptylovém grafu

Podniky

	Bodový odhad		Středý elipsoidů	
	X	Y	X	Y
1	-0,138651	-1,562297	0,140893	-1,915468
2	0,804107	0,055855	0,520704	0,198436
3	1,062701	1,334721	0,533622	1,784583
4	1,351648	-0,007890	1,361946	0,057640
5	-6,135921	-3,924554	-6,027020	-3,669555
6	-4,902406	-2,049165	-5,273508	-3,423302
7	0,066812	-5,269830	0,929075	-6,138738
8	1,673880	-1,130713	2,017200	-1,158564
9	-0,822104	1,308961	-1,876922	1,653824
10	-1,078156	0,497516	-1,166738	0,381133
11	-0,946426	0,806862	-1,502190	0,202750
12	2,482970	1,103720	2,298115	1,805407
13	3,223218	0,299461	3,681209	1,087190
14	0,192197	0,323322	0,200355	0,173998
15	0,868492	2,477776	-0,428066	4,049439
16	2,204757	-0,447224	3,044769	-0,040796
17	-1,630994	5,345129	-1,405514	5,679892
18	1,619826	-0,836297	2,142723	-0,747312
19	-4,989574	6,559604	-4,299395	9,382689
20	2,736127	0,873726	2,391521	1,535374
21	-2,281168	2,266051	-2,271299	2,024724
22	1,063130	1,746549	0,903860	2,126669
23	-0,002487	2,599393	-0,015829	2,352250
24	-2,807137	1,144781	-3,657701	0,687018
25	3,592040	3,465188	3,549584	4,595195
26	0,703338	-0,039759	0,888969	-0,235310
27	1,048228	-2,492528	1,770997	-2,792035
28	1,808537	-0,106851	1,497707	0,145968
29	-3,170393	2,778101	-4,615220	2,644188
30	1,580641	2,746607	0,946212	3,380240
31	-1,093544	0,284153	-0,920706	0,103715
32	0,472333	0,596740	0,105740	0,751133
33	-1,364330	-0,069306	-1,646319	-0,755074
34	1,731098	1,911040	1,988753	2,587825
35	-4,821410	-4,513474	-3,135267	-5,060201
36	-0,647574	-1,200675	-0,749647	-1,517386
37	0,299111	-0,390878	0,008729	-0,324877
38	2,265761	-5,911627	2,861808	-5,819188
39	1,675204	-1,065197	1,515934	-1,211473
40	1,647190	-3,208868	1,576497	-3,329719
41	-4,172661	4,184539	-4,609914	3,944545
42	1,749290	-0,322878	2,351629	-0,339570
43	1,962903	-0,380452	2,187845	-0,092414
44	-6,464558	0,618309	-7,607111	-0,013742
45	3,461523	0,468505	3,284195	1,264937
46	1,013300	0,711187	1,183207	0,843599
47	0,203182	0,361144	-0,209324	0,252806
48	3,035318	-0,485118	3,930661	-0,326745
49	-4,929752	-2,571693	-4,931661	-4,788997
50	1,895110	1,526517	1,206881	2,164733
51	0,738265	-2,594541	1,783425	-3,154568
52	-5,486410	-3,566455	-5,147262	-5,714869
53	-0,116920	-2,256560	0,044211	-2,833321
54	1,567476	-0,907110	2,054670	-0,768530
55	1,922663	-0,590310	2,143618	-0,522009
56	0,737892	-0,900641	0,701768	-1,015108
57	-2,290335	0,374715	-2,453982	-0,320638
58	3,002027	-2,233782	3,946224	-2,364530
59	-0,301465	2,142374	-1,006811	2,581855
60	-0,637879	-1,090345	-0,077704	-1,613773
61	-0,083089	-1,108459	0,099820	-1,322788
62	1,550685	-0,202179	1,876022	0,095873
63	0,546704	1,421180	-0,223586	1,571438
64	4,122358	-0,085856	4,631884	0,585445
65	-0,425151	-0,487403	-0,370246	-0,620839
66	1,088911	1,193543	0,996496	1,235965
67	0,411892	-0,158499	0,391125	-0,089413
68	-3,442348	0,642142	-4,061661	0,108378