

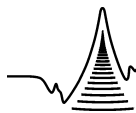


FAKULTA MECHATRONIKY A MEZIOBOROVÝCH INŽENÝRSKÝCH STUDIÍ

Určení kritérií k optimalizaci báze pro snížení dimenze problému modelování transportu mnohasložkových vodných roztoků

Vedoucí učitel projektu: Doc. Ing. Jan Šembera, Ph.D.

Vypracoval: Bc. Lukáš Zedek



Prohlášení

Byl(a) jsem seznámen(a) s tím, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 o právu autorském, zejména § 60 (školní dílo).

Beru na vědomí, že TUL má právo na uzavření licenční smlouvy o užití mé diplomové práce a prohlašuji, že souhlasím s případným užitím mé diplomové práce (prodej, zapůjčení apod.).

Jsem si vědom(a) toho, že užít své diplomové práce či poskytnout licenci k jejímu využití mohu jen se souhlasem TUL, která má právo ode mne požadovat přiměřený příspěvek na úhradu nákladů, vynaložených univerzitou na vytvoření díla (až do jejich skutečné výše).

Diplomovou práci jsem vypracoval(a) samostatně s použitím uvedené literatury a na základě konzultací s vedoucím diplomové práce a konzultantem.

Datum:

Podpis:

Abstrakt

Tato práce si kladla za cíl nalézt, systematizovat a zobecnit postupy použitelné pro nalezení optimální matematické interpretace informace o koncentracích látek rozpuštěných ve vodných roztocích. Systematičnost a obecnost hledaného postupu měly zajistit jeho širší použitelnost v porovnání s obvykle heuristicky postavenými metodami, založenými na přesné znalosti problému na jednom konkrétním místě. Optimálnost zmíněné matematické interpretace byla hodnocena primárně z hlediska místa v paměti informací obsazeného se zřetelem braným na přesnost uchované informace. Správný poměr mezi velikostí a přesností zpracovávané informace je důležitý z hlediska rychlosti práce s rozsáhlými soubory dat.

Předchozí odstavec znamená v jednoduchosti následující věc. V práci jsme se přestali na *roztok X* dívat jako na vodu, která obsahuje na jeden litr 30 mg *látky A*, 50 mg *látky B*, ... až 35 mg *látky Q*. Místo toho jsme na roztok pohlédli takovým způsobem, který říká: „V *roztoku X* naměříme přibližně stejné koncentrace, jako ve směsi tvořené 50 ml známého *roztoku 1* a 37 ml známého *roztoku 2*.“

Znamená to, že s určitou chybou je možné popsat *roztok X* dvěma čísly místo původních čísel v počtu Q . Přičemž předpokládáme že $Q > 2$.

Abstract

The main goal of this work was to find a general way for an optimal matematic description of chemical contamination, which is contained in underground water nearby the hill Ralsko. Generality was required to make the searched way usable for other aplications and sites. Optimal description should not take to much memory and it should store enough information. The right ralation betwen size and stored information can dramatically speed up the work with huge sets of data.

Simply said. We stopped to regard the *solution X* as a water containing 30 mg/l of the *substance A*, 50 mg/l of the *substance B*, ... and 35 mg/l of the *substance Q*. We started thinking so that: „Concentrations mesured in the *solution X* are similar to concentrations in the mixture containing 50 ml of *solution 1* with 37 ml of *solution 2*.“

That means, we need only 2 numbers to describe the *solution X* containing Q substances. We suppose $Q > 2$.

Obsah

Zadání.....	2
Prohlášení.....	3
Abstrakt.....	4
Značení.....	6
1. Úvod.....	7
2. Používané matematické postupy.....	8
2.1 Data reprezentovaná vektory chemických analýz, redukce dimenze.....	8
2.2 Analýza hlavních komponent.....	11
2.3 Metoda nejmenších čtverců, QR-rozklad.....	14
3. Lineární vektorový podprostor.....	15
4. Afinní vektorový podprostor.....	21
4.1 Výpočet posunu počátku.....	27
5. Výsledky použití nalezeného algoritmu.....	30
5.1 Práce s maticí 638x6.....	30
5.1.1 Vypočtené bázevé vektory.....	34
5.2 Práce s maticí 2515x6.....	36
5.3 Práce s maticí 90x22.....	37
5.4 Souhrnné stanovisko k výsledkům.....	40
6. Kritéria optimálnosti báze.....	40
7. Odlišnosti od existujících postupů.....	42
7.1 Srovnání chyb obou představených přístupů	49
8. Závěr.....	54
Literatura.....	55

Značení

Na tomto místě jsou uvedeny symboly a pojmy, které jsou pro zjednodušení popisu představovaných postupů v této práci hojně používány. Většina zavedených značek identifikuje vektorové prostory, které se vzájemně liší buď svými dimenzemi, rozdílnými bázemi a nebo obojím. Pro přehlednost jsou zmíněné značky umístěny v následující tabulce.

$V = [\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n]_\lambda$ $\vec{a} \in V, \vec{a} = [a_1, a_2, \dots, a_n]^T$ $a_1 \dots$ všechny látky $a_2 \dots$ síranový aniont $a_3 \dots$ amonný kationt $a_4 \dots$ hlinitý kationt $a_5 \dots$ kationty železa $a_6 \dots$ fluoridový aniont	V je orthonormální vektorový prostor, ve kterém jsou zobrazena poskytnutá data o chemických analýzách. Prostor V má dimenzi n (5, 6, nebo 22). Jednotlivé složky vektoru (např. vektoru $\vec{a}$) v prostoru V odpovídají koncentracím chemických látek ve vektorem reprezentovaném roztoku. Koncentrace mají fyzikální rozměr mg/l . Pro dále často používanou dimenzi prostoru V rovnou 6-ti odpovídají složky vektoru $\vec{a}$ chemickým látkám zapsaným v levém sloupci této tabulky. Symboly $\vec{v}_i$ označují jednotkové vektory standardní báze prostoru V s jedničkami na i -té pozici vektoru.
$\vec{m}_i \in V$	$\vec{m}_i$ je vektor obsahující data z i -té provedené chemické analýzy.
$B = [\vec{b}_1, \vec{b}_2, \dots, \vec{b}_n]$ ${}_i B = [\vec{b}_1, \vec{b}_2, \dots, \vec{b}_i], i \leq n$ ${}_i V = [\vec{b}_1, \vec{b}_2, \dots, \vec{b}_i]_\lambda, i \leq n$	B je matice sestavená ze sloupcových vektorů optimální orthonormální báze prostoru V . Tyto vektory jsou nalezeny za použití Analýzy hlavních komponent. Matice ${}_i B$ obsahuje prvních i vektorů z matice B a její lineární obal je i -rozměrným podprostorem ${}_i V$ prostoru V .
Hlavní podprostor prostoru V určený bází $[\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n]$	Pojem hlavní podprostor dimenze k prostoru V s bází $[\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n]$ byl definován, pro potřeby této práce, jako každý podprostor dimenze k prostoru V generovaný libovolnou k -ticí vektorů z báze prostoru V .
$\vec{p}_j = (p_{j1} \ p_{j2} \dots \ p_{jn})^T$	$\vec{p}_j$ jsou vektory ležící v průniku podprostoru ${}_i V$ a vybraného hlavního podprostoru prostoru V určeného bází $[\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n]$. Mezi vektory $\vec{p}_j$ je hledána nová báze redukovaného prostoru ${}_i V$.
${}_P \vec{p}_j$	Toto označení reprezentuje vektor $\vec{p}_j$ jako prvek prostoru ${}_i V$ se souřadnicemi vyjádřenými v bázi ${}_i B$. Má i složek.

Poznámka k obrázkům

Obrázky obsažené v této práci mají za úkol dokreslit na konkrétních, jednoduchých příkladech čtenáři představu o popisované problematice. Jedná se výhradně o náčrtky, které vzhledem k technickým možnostem a znalostem autora mohou obsahovat drobné faktické a také zejména geometrické nepřesnosti.

1. Úvod

Tato práce navázala na projekty [1] a [2], které jsou uvedené v seznamu literatury. Projekt [1] se zabýval redukcí dimenze vektorového prostoru, který obsahuje vektorově interpretované informace o chemických analýzách koncentrací kontaminantů ve vodných roztocích. Úkolem projektu [2] bylo potvrzení hypotéz, které se týkaly vybraných metod používaných v souvislosti se zmíněnou redukcí dimenze ve firmě DIAMO s.p. Samotná diplomová práce si kladla za cíl nalézt postupy použitelné k určení „vhodné“ (optimální) báze známého prostoru s redukovanou dimenzí (prostoru určeného jinou „nevyhovující“ bází). Požadavky na „vhodnost“ báze vycházejí z potřeby interpretovat výsledné bázevé vektory jako roztoky určitých chemických látek. Složky jednotlivých bázevých vektorů představují koncentrace zmíněných látek.

Z této interpretace přirozeně plynou dva základní požadavky:

1. Složky bázevých vektorů nemohou být záporné.
2. Poskytnutá data o chemických analýzách by mělo být možné aproximovat jako lineární kombinace vektorů zvolené „vhodné“ báze, přičemž v oněch lineárních kombinacích mohou vystupovat pouze nezáporné koeficienty.

Oba dva z uvedených požadavků plynou z jednoduchých představ. Požadavek 1) říká asi tolik, že roztok je popsán nezápornými koncentracemi chemických látek v něm přítomných a nikoli určením toho, co v něm chybí. Požadavek 2) má podobný význam, který je posunut o úroveň „výše“. Předpokládá totiž, že výsledný roztok může vzniknout smícháním několika roztoků bázevých a že naopak nikdy nevznikne vyjmutím určitého množství látek v původním roztoku obsažených. Tedy nevznikne odebráním roztoku.

Na cíle této práce se je možné dívat také následujícím způsobem. Mějme n -rozměrný vektorový prostor V . Vektorový prostor V nechť obsahuje vektory, které nesou informace o určitém druhu prováděných chemických analýz. Úkolem této práce bylo nalezení postupu vedoucího k určení „nejlepší“ báze známého k -rozměrného podprostoru kV prostoru V . Hledaná báze podprostoru kV musela splňovat výše stanovené požadavky.

K hledání zmíněné báze posloužily algebraické a statistické metody implementované do jazyka GNU Octave. GNU Octave je vyšší programovací jazyk určený k provádění numerických výpočtů a je freeware-ovou obdobou jazyka Matlab.

Praktický výsledek diplomové práce představuje spustitelný skript napsaný v jazyce GNU Octave. Tento skript poskytuje možnost nalézt zmíněnou bázi v krátkém čase a s minimálními požadavky na chemické znalosti jeho uživatele. Dvěma podmínkami pro úspěšný průběh celého výpočtu jsou vyhovující formát vstupních dat a zadaná dimenze redukovaného prostoru kV .

2. Používané matematické postupy

V praktické části této diplomové práce byly hojně využívány zejména algebraické a statistické postupy, které jsou představeny v této kapitole. První podkapitola je věnována vysvětlení vektorové interpretace chemických analýz koncentrací chemických složek zkoumaného roztoku. Poté je zde zmíněna Analýza hlavních komponent, která byla z hlediska této práce naprosto nepostradatelná. Dále je tu krátce představen QR-rozklad, který byl při výpočtech často používán k určení psoudoinverzních matic a k řešení přeuročených soustav metodou nejmenších čtverců.

2.1 Data reprezentovaná vektory chemických analýz, redukce dimenze

Data zpracovávaná v rámci této diplomové práce byla poskytnutá firmou DIAMO s.p. Z hlediska formy měla tato data podobu tabulek, jejichž řádky odpovídaly vždy právě jedné chemické analýze koncentrací látek rozpuštěných v podzemní vodě na lokalitě blízce Stráži pod Ralskem. Každá sada dat o chemické analýze se vztahovala

k jednomu konkrétnímu vrtu a k času kdy byla tato analýza prováděna. Sloupce zmíněných tabulek s daty by bylo možno nadepsat vzorci chemických látek, jejichž koncentrace jsou v těchto zaneseny.

Látek, jejichž koncentrace byly v podzemních vodách proměřovány, bylo celkem 22 a koncentrace šesti nejdůležitějších byla analyzována ve všech případech. Nepracovali jsme se všemi daty, ale zvlášť pouze s množinami dat, které měly změřené koncentrace všech šesti (resp. 22, resp. 5) chemikálií. Jinými slovy, z tabulek s daty jsme vybrali buď pouze prvních 6 vždy zaplněných sloupců nebo všechny zcela zaplněné řádky.

Chemické analýzy zkoumající koncentrace látek rozpuštěných v podzemních vodách jsou nákladné. Cena analýzy koncentrace jedné látky se pohybuje v řádu tisíců korun. Z toho důvodu je před započítáním analýzy na konkrétním vrtu vhodné zvážit, které látky je nutné „hlídat“. Základní analýza koncentrací chemických látek se provádí vždy a pro každé místo (vrt) a čas. Tato základní analýza poskytuje informace o celkové koncentraci všech rozpuštěných látek a jednotlivě o koncentracích pěti chemikálií, které je nutné sledovat kvůli jejich případným negativním dopadům na zdraví člověka nebo na životní prostředí.

Uspořádání poskytnutých dat do zcela zaplněných tabulek umožnilo analýzami získaná data matematicky interpretovat a tím je připravilo pro využití výpočetní techniky při jejich zpracování. V tabulkách uspořádané podmnožiny poskytnutých dat je možno vnímat jako matice tvořené řádkovými vektory těchto dat. Jednotlivé složky vektorů zkoumaných dat jsou koncentracemi přítomných a sledovaných chemických látek.

Obrázek 1 ukazuje princip přechodu od báze tvořené šesti původními vektory k bázi se 3 vektory. Snížení počtu bázevých vektorů je spojeno se snížením dimenze vektorů chemických analýz koncentrací (dimenze = počet sloupců matice).

Jinými slovy. Vektor $\vec{m}_i$ má v původní podobě šest složek. Tyto složky jsou násobky vektorů standardní báze $\vec{v}_i, i \in \{1, \dots, 6\}$, které mají vždy právě jednu nenulovou složku na své i -té pozici. Základní princip redukce dimenze vektoru $\vec{m}_i$ spočívá ve vyjádření jeho souřadnic jakožto násobků jiných vektorů, kterých je méně a které

mohou mít všechny složky různé od nuly. Rozklad dat z chemických analýz do jiné báze je přirozeně zatížen chybou a bázi je nutné volit takovým způsobem aby byla tato chyba co nejmenší. Velikost chyby může být přirozeně kvantifikována například normou rozdílu původního vektoru s analyzovanými daty a aproximativního k -rozměrného vektoru zobrazeného zpět do šesti-rozměrného (n -rozměrného) prostoru.

Pokud bychom se chtěli zaměřit na přesnou aproximaci určité složky analyzovaných dat, potom by bylo vhodné jednotlivé složky vektorů před zpracováním přeskálovat takovým způsobem, aby koncentrace chemikálie, která nás zajímá, měla z hlediska určování k -rozměrné aproximace, větší váhu.

Byly učiněny pokusy dokládající vliv škálování na změnu směru bázových vektorů hledaných týmiž postupy jako v případě neškálovaných dat a bylo potvrzeno, že „správné“ škálování sníží chybu jím zvýrazněné složky. Tyto pokusy byly nejprve prováděny pouze s jednou složkou a ukázaly, že na úkor přesnosti škálované složky se sníží přesnost složek ponechaných bez úpravy. Viz Kapitola 7.1 .



Obrázek 1: Princip redukce dimenze

2.2 Analýza hlavních komponent

Analýza hlavních komponent je standardní statistický algoritmus sloužící k nalezení a odstranění existujících lineárních funkčních závislostí mezi různými složkami známého souboru dat, chemických analýz, vektorů, atp. Po odstranění takovýchto funkčních závislostí je možné v datech transformovaných do nového orthonormálního souřadného systému vysledovat rozdílnou důležitost směrů jednotlivých báзовých vektorů. Tyto směry odpovídají směrům maximalizujícím směrodatné odchylky naměřených dat v původním prostoru a ve všech jeho podprostorech vznikajících postupným odstraňováním dříve nalezených směrů báзовých vektorů. Různá „důležitost“ jednotlivých báзовých vektorů je základem pro redukci dimenze zkoumaného prostoru.

Postup analýzy hlavních komponent je následující:

1. Data, na kterých si přejeme analýzu provést je třeba uspořádat do matice. V našem případě představovaly sloupce matice koncentrace chemických složek zkoumaných roztoků a každý řádek odpovídal vždy jedné provedené chemické analýze.
2. Data uspořádaná v matici mohou být při standardním postupu dále upravena odečtením průměrů všech naměřených koncentrací jednotlivých chemických složek od příslušných komponent každého ze známých vektorů chemických analýz. Jinými slovy, provedeme-li n chemických analýz koncentrací šesti chemických látek, pak je úkolem tohoto kroku nejprve výpočet šesti-složkového vektoru průměrné koncentrace každé jedné látky a poté následuje odečtení vypočteného vektoru po složkách od prvního až n -tého řádku zaznamenávajícího chemické analýzy. Tento postup se zdál nejdříve nevhodný, protože nevedl k nalezení lineárního vektorového podprostoru, ve kterém by spolu s každými dvěma vektory ležela také jejich libovolná lineární kombinace. Z tohoto důvodu jsme nejprve používali upravenou „Analýzu hlavních komponent bez odečtení průměru“. Výsledkem aplikace „Analýzy hlavních komponent s odečtení průměru“ by bylo nalezení podprostoru afinního. Práce s afinním podprostorem byla později také provedena a měla charakter zobecnění postupů práce s daty v lineárním

podprostoru. Části textu, které jsou výše vymezené ovozovkami, představují pojmy na které je dále v práci odkazováno. Dopad odečtení průměru je zachycen na Obrázku 2.

3. Z dat získaných použitím předchozích dvou kroků byly vypočteny dvě kovarianční matice příslušející po řadě původní matici a matici s odečtenými průměry.

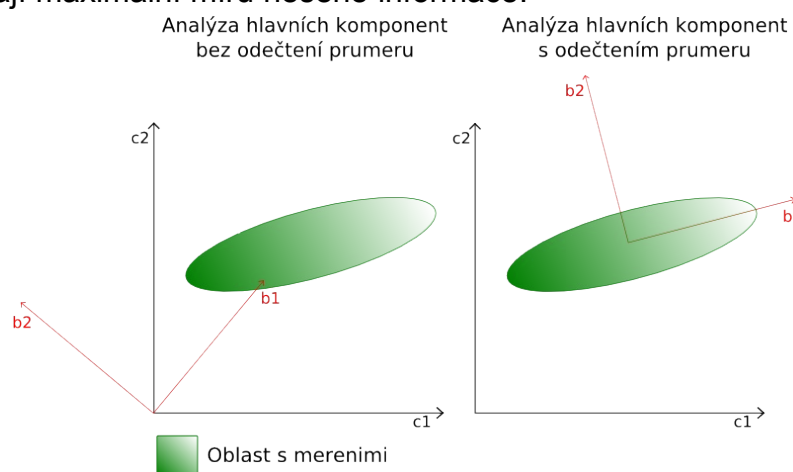
Kovarianční maticí je čtvercová matice vzniklá násobením transponované matice upravených poskytnutých dat s maticí netransponovanou (viz Obrázek 3). Výsledek zmíněného násobení matic je standardně vydělený konstantou $(n-1)$. V této práci jsme vzhledem k povaze další manipulace s kovariančními maticemi neprováděli zmíněné dělení. Naším cílem bylo totiž nalezení směrů určených vlastními vektory kovariančních matic (viz dále). Vydělení matic reálnými čísly by se podepsalo pouze na velikosti vlastních čísel matice a nikoli na směru jejich vlastních vektorů. Viz vzorce.

$$\begin{aligned} C \cdot \vec{x} &= \lambda \vec{x} \\ (k \cdot C) \cdot \vec{x} &= k \cdot \lambda \vec{x} \end{aligned}$$

Je patrné, že jak matice C matice $(k \cdot C)$ mají stejné vlastní vektory $\vec{x}$.

4. U každé kovarianční matice jsme vypočetli její vlastní čísla $\lambda_1, \dots, \lambda_n$ a jednotkové vlastní vektory $\vec{b}_1, \vec{b}_2, \dots, \vec{b}_n$. Kovarianční matice, která je počítaná z matice s lineárně nezávislými sloupci, je symetrická a pozitivně definitní, takže všechna její vlastní čísla jsou kladná. S tímto vědomím jsme vlastní vektory uspořádali do řádků transformační matice v pořadí odpovídajícím klesající velikosti příslušných vlastních čísel. Nalezené vlastní vektory tvoří novou orthonormální bázi B zkoumaného prostoru V .

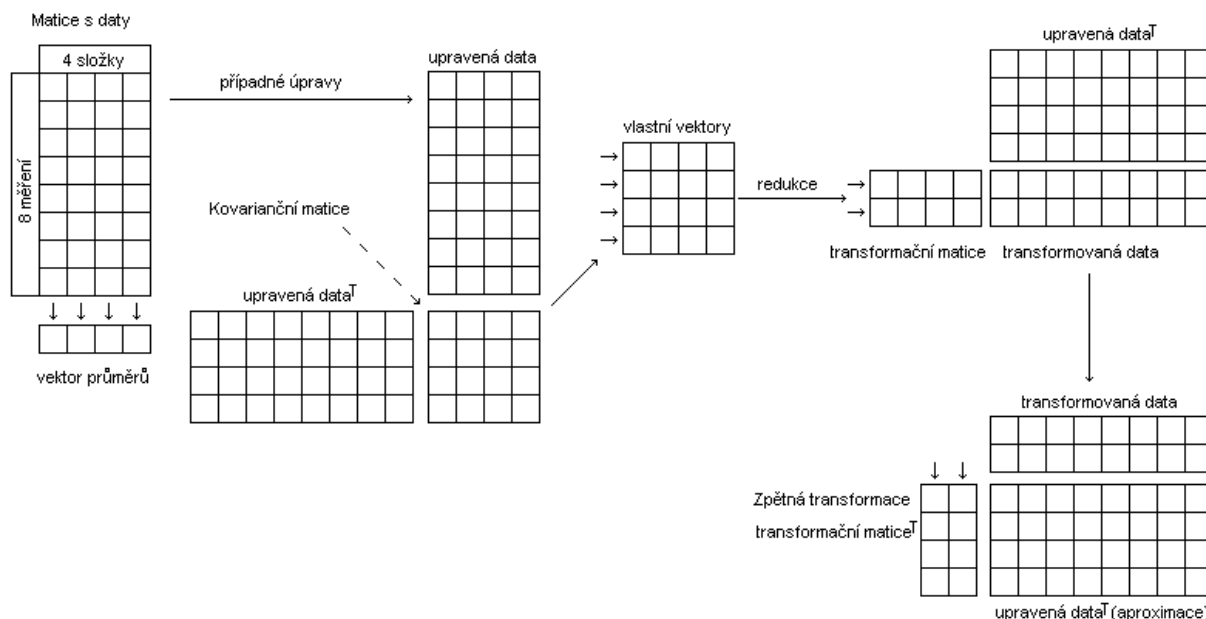
Obrázek 2 odhaluje vliv odečtení (resp. neodečtení) průměrných koncentrací od každého jednoho řádku analyzovaných dat ze zkoumané množiny na směr vektorů nové báze B prostoru V . Z Obrázku je zřejmé, že Analýza hlavních komponent s odečtením průměru „respektuje“ orientaci souboru dodaných dat v grafu a po redukci dimenze vektorů s daty odebráním několika jejich složek si redukované vektory uchovávají maximální míru nesené informace.



Obrázek 2: Vliv odečtení průměru

5. Dalším krokem byla redukce rozměru nově nalezeného prostoru. Samotné snížení dimenze prostoru jsme realizovali výběrem prvních k vlastních (bázových) vektorů odpovídajících k největším vlastním číslům kovarianční matice (tj. k prvních řádků transformační matice).
6. Po redukci transformační matice jsme provedli samotnou transformaci naměřených a upravených dat do redukovaného prostoru. Tento krok byl realizován násobením transpozice redukované transformační matice ${}_k B^T$ s transponovanou maticí obsahující upravená data.
7. Návrat do původního, neredukovaného prostoru, jsme později uskutečnili násobením pseudoinverze transformační matice s maticí transformovaných dat a případně ještě následným přičtením dříve vypočtených průměrů koncentrací jednotlivých složek. Použitá pseudoinverze orthogonální transformační matice se v tomto případě rovná její transpozici.

Maticové znázornění výše popsaných operací je přehledně zachyceno na následujícím obrázku.



Obrázek 3: Postup analýzy hlavních komponent

První z tabulek, které zachycuje Obrázek 3, představuje matici s osmi sadami dat o koncentracích čtyř chemických látek rozpuštěných v osmi proměřovaných roztocích. Počty roztoků a chemických složek byly zvoleny pouze kvůli vyhovující velikosti a názornosti obrázku. Tato čísla nemají žádnou souvislost se skutečně zpracovávanými sadami dat.

Z celé sady dat je vypočten řádkový vektor průměrných koncentrací, který může být volitelně odečten od všech ze sad zkoumaných dat. Upravenými daty jsou myšlena právě data s odečtenými průměry.

Uspořádání tabulek do tvaru „horizontálně obráceného L“ znázorňuje násobení matic.

'T' za jménem matice značí, že jde o transpozici matice, která je jménem určena.

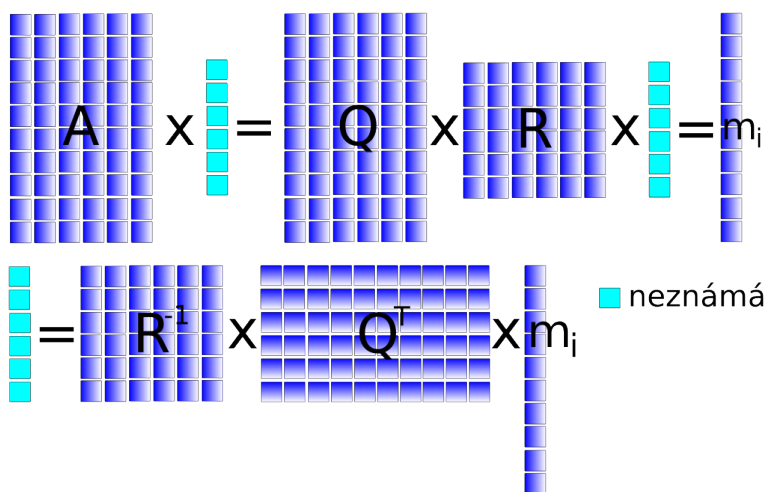
Operace pojmenovaná jako redukce spočívá na Obrázku v uchopení a dalším použití pouze prvních dvou ze čtyř předpokládaných vlastních vektorů spočtené čtvercové kovarianční matice.

2.3 Metoda nejmenších čtverců, QR-rozklad

QR-rozklad je algebraický postup k řešení přeuročených soustav lineárních rovnic za využití metody nejmenších čtverců (resp. takzvané pseudoinverze). Předpokládejme, že máme řešit přeuročenou soustavu lineárních rovnic danou maticí

A s rozměry $m \times n$, kde $m \geq n$. Dále předpokládejme, že soustava má nenulovou pravou stranu. Protože uvažovaná matice soustavy není čtvercová, nelze ji řešit s pomocí násobení soustavy inverzní maticí A^{-1} zleva. Místo toho je možné rozložit matici soustavy na součin orthonormální matice Q o rozměrech $m \times n$ a horní trojúhelníkové matice R s rozměry $n \times n$ (tj. $A = Q \cdot R$). Součin transpozice matice Q s maticí Q samotnou je jednotková matice (tedy $Q^T \cdot Q = I$). Čísla na diagonále matice R jsou kladná.

Volba použití QR-rozkladu vyplynula z požadavku na rychlost prováděných výpočtů. Popsaný postup je zakreslen na následujícím obrázku.



$A \cdot x = Q \cdot R \cdot x = m_i$

$x = R^{-1} \cdot Q^T \cdot m_i$

■ neznámá

Obrázek 4: Využití QR rozkladu pro pseudoinverzi

3. Lineární vektorový podprostor

K práci s lineárními vektorovými podprostory bylo přistoupeno na začátku vypracovávání praktické části této diplomové práce z důvodu předpokládané vyšší obtížnosti práce s afinními podprostory. Navíc existoval předpoklad, že nalezené postupy půjdou ve většině případů zobecnit tak aby byly použitelné i v afinních podprostorech a to proto, že lineární podprostor je speciálním případem podprostoru afinního. Tento předpoklad se potvrdil.

Nalezené postupy pro práci s lineárním podprostorem jsou následující:

- I) Samotnému hledání „vhodné“ báze i -rozměrného podprostoru ${}_iV$, který je

podmnožinou n -rozměrného prostoru V , předcházelo určení „optimální“ orthonormální báze B prostoru V , provedené za pomoci „Analýzy hlavních komponent bez odečtení průměru“. Získanou orthonormální bázi B (1) si je možné s určitým zjednodušením představit jako bázi původní, pootočenou okolo počátku.

$$B = \begin{bmatrix} b_{11} & b_{21} & b_{31} & b_{41} & b_{51} & b_{61} \\ b_{12} & b_{22} & b_{32} & b_{42} & b_{52} & b_{62} \\ b_{13} & b_{23} & b_{33} & b_{43} & b_{53} & b_{63} \\ b_{14} & b_{24} & b_{34} & b_{44} & b_{54} & b_{64} \\ b_{15} & b_{25} & b_{35} & b_{45} & b_{55} & b_{65} \\ b_{16} & b_{26} & b_{36} & b_{46} & b_{56} & b_{66} \end{bmatrix} \quad (1)$$

II) Z prvních i vektorů nově získané báze B byly vypočteny vektory $\vec{p}_j$, na které byly kladeny takové požadavky, aby bylo možno každý vektor $\vec{p}_j$ z alternativní báze podprostoru ${}_iB$ prohlásit za určitý roztok analyzovaných chemických látek (vektor $\vec{p}_j$ měl mít nezáporné složky) a aby šlo do báze tvořené vybranou k -ticí z vektorů $\vec{p}_j$ rozložit požadovaným způsobem (s nezápornými složkami rozkladů) maximální počet vektorů chemických analýz.

Zmíněné požadavky jsme přeformulovali do matematické podoby. Žádali jsme aby vektory $\vec{p}_j$ ležely zaprvé v některém z $(n-i+1)$ -rozměrných hlavních podprostorů prostoru V určeného standardní bází $[\vec{v}_1, \dots, \vec{v}_n]$ a zadruhé v podprostoru ${}_iV$ určeném vektory z matice ${}_iB$. Hlavním podprostorem dimenze $(n-i+1)$ prostoru V určeného bází $[\vec{v}_1, \dots, \vec{v}_n]$ je míněný libovolný $(n-i+1)$ -rozměrný podprostor prostoru V takový, který je definován $(n-i+1)$ -ticí vektorů z báze V .

Dále jsou složky takto určených vektorů $\vec{p}_j$ značeny následovně.

$$\vec{p}_j = (p_{j1} \ p_{j2} \ p_{j3} \ p_{j4} \ p_{j5} \ p_{j6})^T$$

Dimenze výsledných vektorů $\vec{p}_j$ odpovídá rozměru prostoru V (tj. n).

Realizace zmíněných výpočtů vektorů $\vec{p}_j$ bylo dosaženo za pomoci řešení soustav schematicky shodných s následujícím příkladem.

Příklad:

Výpočet vektorů $\vec{p}_j$ vychází z řešení soustav podobných té, kterou ukazuje vzorec (2). Popis soustavy je uveden za vzorcem (2).

$$\begin{bmatrix} \sum_i b_{1i} & \sum_i b_{2i} & \sum_i b_{3i} & \sum_i b_{4i} & \sum_i b_{5i} \\ b_{11} & b_{21} & b_{31} & b_{41} & b_{51} \\ b_{12} & b_{22} & b_{32} & b_{42} & b_{52} \\ b_{13} & b_{23} & b_{33} & b_{43} & b_{53} \\ b_{14} & b_{24} & b_{34} & b_{44} & b_{54} \end{bmatrix} \cdot \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad (2)$$

Pro osvětlení významu soustavy (2) uvažujme, že redukuje 6-ti rozměrný prostor o jednu dimenzi. Protože pozice vektoru v matici B odpovídá ve své podstatě relativní míře informace nesené příslušnou složkou průmětu korelovaných dat do prostoru touto maticí určeného, prováděli jsme z důvodu minimalizace celkové chyby zmíněnou redukci vždy prostřednictvím výběru prvních i (nyní $i = 5$) vektorů (resp. sloupců) z matice B .

V příkladu je rozměr redukovaného prostoru 5 a proto má soustava (2) právě 5 neznámých. Ponechme prozatím stranou význam, zdůrazněme že velmi **důležité**, první rovnice. Druhá až pátá rovnice jsou kombinacemi vždy čtyř ($i-1$) ze šesti (n) složek báзовých vektorů soustředěných v matici B a společně definují 4 (v obecném případě $i-1$) složky nových báзовých vektorů $\vec{p}_j$, které mají být rovny nule. V uvažovaném případě tak budou rovny nule souřadnice $p_{j1}, p_{j2}, p_{j3}, p_{j4}$.

Na pozicích druhé až páté (tj. i -té) rovnice se může vyskytnout $c = \binom{n}{i-1}$ různých kombinací složek vektorů z matice B a index j vektorů $\vec{p}_j$ tak může nabývat hodnot z intervalu $(1, \dots, c) \in \mathbb{N}$. Není to sice zcela zřejmé, ale při výpočtech se ukázalo, že počet vektorů $\vec{p}_j$ je ve většině případů výrazně vyšší než i (tj. dimenze podprostoru) a řada z těchto vektorů je tedy zbytných.

„Opomenutá“ první rovnice soustavy určuje, v kontrastu k popsaným čtyřem rovnicím,

ty souřadnice vektoru $\vec{p}_j$, které rovný nule být nemusí. Součet složek bázevých vektorů b_i je proveden z toho důvodu, aby nebylo nutné pro různě zvolené čtveřice nulových složek $\vec{p}_j$ přímo určovat také zbývající, obecně nenulovou dvojici. Vycházelo se při tom ze skutečnosti, že přičteme-li k nenulovému součtu dvou složek vektorů $\vec{b}_i$, $i \in (1, \dots, 5)$ lineární kombinace čtyř složek zbývajících, přičemž tyto lineární kombinace jsou všechny rovný nule, dostaneme tentýž výsledek jako kdybychom lineární kombinace nepřičítali. Tento postup výrazně zjednodušil programování celého problému. Vyjádřena vzorci, vypadá myšlenka konstrukce první rovnice soustavy (2) takto:

$$\sum_{i=1}^5 \sum_{j=1}^6 c_i \cdot b_{ij} = \sum_{i=1}^5 \sum_{j=5}^6 c_i \cdot b_{ij} = 1, \quad tj.: \\ \left[(b_{15} + b_{16}) \quad (b_{25} + b_{26}) \quad (b_{35} + b_{36}) \quad (b_{45} + b_{46}) \quad (b_{55} + b_{56}) \right] * \begin{bmatrix} c_1 & c_2 & c_3 & c_4 & c_5 \end{bmatrix}^T = 1 \quad (3)$$

Ve vzorci na prvním řádku se u sum s indexem j liší výchozí hodnoty od kterých je sumace prováděna. Z toho plyne následující skutečnost

$$\sum_{i=1}^5 c_i \cdot b_{ij} = 0, \quad \forall j \in (1, \dots, 4), \quad tj.: \\ \begin{bmatrix} b_{11} & b_{21} & b_{31} & b_{41} & b_{51} \\ b_{12} & b_{22} & b_{32} & b_{42} & b_{52} \\ b_{13} & b_{23} & b_{33} & b_{43} & b_{53} \\ b_{14} & b_{24} & b_{34} & b_{44} & b_{54} \end{bmatrix} \cdot \begin{bmatrix} c_1 & c_2 & c_3 & c_4 & c_5 \end{bmatrix}^T = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad (4)$$

Čímž se dostáváme zpět k soustavě (2) protože

$$\sum_{i=1}^5 \sum_{j=1}^4 c_i \cdot b_{ij} = 0 \Rightarrow \sum_{i=1}^5 \sum_{j=5}^6 c_i \cdot b_{ij} + \sum_{i=1}^5 \sum_{j=1}^4 c_i \cdot b_{ij} = \sum_{i=1}^5 \sum_{j=1}^6 c_i \cdot b_{ij} = \sum_{i=1}^5 \sum_{j=5}^6 c_i \cdot b_{ij} = 1 \\ tj.: \begin{bmatrix} \sum_i b_{i1} & \sum_i b_{i2} & \sum_i b_{i3} & \sum_i b_{i4} & \sum_i b_{i5} \\ b_{11} & b_{21} & b_{31} & b_{41} & b_{51} \\ b_{12} & b_{22} & b_{32} & b_{42} & b_{52} \\ b_{13} & b_{23} & b_{33} & b_{43} & b_{53} \\ b_{14} & b_{24} & b_{34} & b_{44} & b_{54} \end{bmatrix} \cdot \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad (5)$$

Jednička na pravé straně první rovnice je námi zvolená hodnota. Na její místo je možné dosadit libovolné kladné číslo. Velikost vybraného čísla ovlivní pouze

eukleidovské normy vektorů $\vec{p}_j$.

Samotný výpočet bázových vektorů $\vec{p}_j$ má tvar:

$$\vec{p}_j = {}_j\mathbf{B} * \vec{c}_j \quad (6)$$

Je tedy patrné, že $\vec{p}_j$ jsou lineárními kombinacemi bázových vektorů z báze ${}_j\mathbf{B}$ s koeficienty získanými řešením soustav (2).

III) Z vypočtených vektorů $\vec{p}_j$, jež jsou „kandidáty“ na vektory nové báze, byly vybrány ty, které měly všechny složky nezáporné. Tím došlo místy sice k významné avšak nikoli dostatečné redukci počtu nalezených vektorů. Pokud se nabízí otázka: „Proč nevyhovuje vektor se samými zápornými (resp. nekladnými) složkami?“, odpověď je nasnadě. Soustava (2) nám říká o vektoru $\vec{p}_j$ ve své podstatě následující:

$$\vec{p}_j = [p_{j1}, p_{j2}, p_{j3}, p_{j4}, p_{j5}, p_{j6}]^T$$

$$\sum_{i=1}^6 p_{ji} = 1, \quad p_{j1} = 0, \quad p_{j2} = 0, \quad p_{j3} = 0, \quad p_{j4} = 0 \quad (7)$$

Pokud by byly všechny složky vektoru $\vec{p}_j$ nekladné, jejich suma by nikdy nemohla být rovna jedné. Z toho plyne spor s první rovnicí soustavy (2).

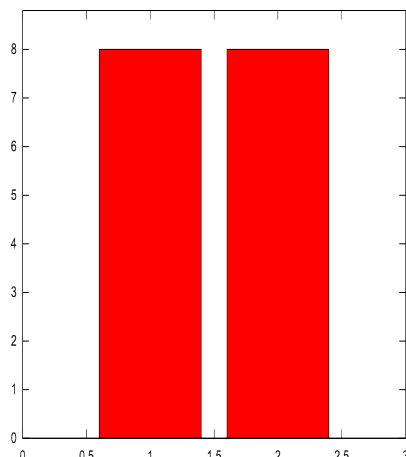
IV) Souřadnice vybraných vektorů $\vec{p}_j$ s výhradně nezápornými složkami byly transformovány zpět do prostoru s bází $\mathbf{B}$. Transformace měla charakter násobení matic. Matice $\mathbf{B}$ je orthonormální a její inverze je rovna její transpozici. Pseudoinverze submatice ${}_i\mathbf{B}$ byla také prováděna jako prostá její transpozice.

$${}_p\vec{p}_j = {}_i\mathbf{B}^T * \vec{p}_j, \quad \text{dimenze } [i \times 1] = [i \times n] * [n \times 1]$$

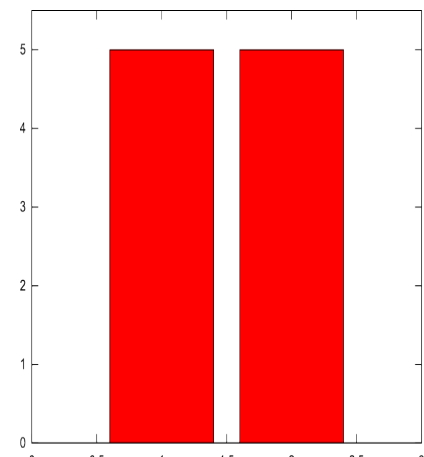
$${}_p\vec{m}_i = {}_i\mathbf{B}^T * \vec{m}_i, \quad \text{dimenze } [i \times 1] = [i \times n] * [n \times 1] \quad (8)$$

V) Souřadnice potenciálních bázových vektorů ${}_p\vec{p}_j$ byly vyjádřeny v souřadnicích sférických a tím promítnuty na kouli se středem v počátku. Tento krok byl přípravou pro zamýšlenou redukci počtu zbývajících vektorů ${}_p\vec{p}_j$ s využitím konstrukce konvexního obalu. Konvexní obal byl zkonstruován, ale k výraznému snížení počtu „použitelných“ bázových vektorů nedošlo.

V případech redukci prostoru o dimenzi $n=6$ jsme nepozorovali dokonce žádný úbytek. Dokladem této skutečnosti je několik následujících grafů. První ze sloupců zachycuje počet vektorů ${}_P\vec{p}_j$ zobrazených ve sférických souřadnicích a sloupec druhý nese informaci o počtu vektorů v konvexním obalu vektorů ${}_P\vec{p}_j$ také ve sférických souřadnicích.

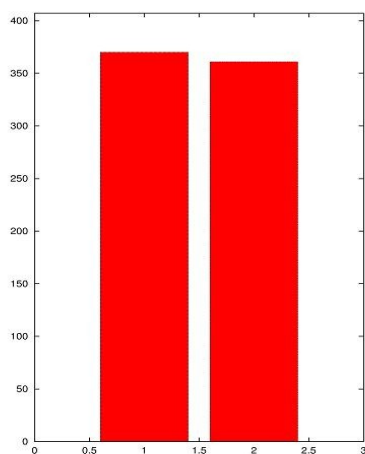


Obrázek 6: Redukce na 4 ze 6

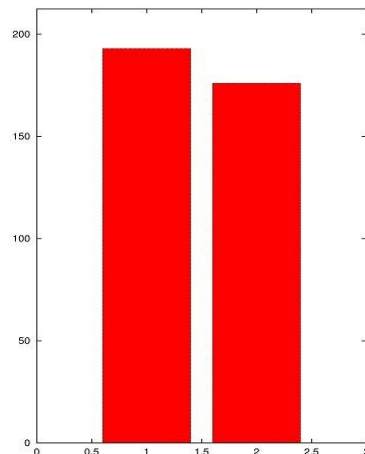


Obrázek 5: Redukce na 3 ze 6

Při redukci prostoru s dimenzí $n=22$ byl úbytek počtu vektorů ${}_P\vec{p}_j$ sice pozorován, ale byl o několik řádů nižší než jejich původní počet a neměl tedy praktického významu. Pro dokumentaci této skutečnosti jsou přiloženy následující grafy.



Obrázek 7: Redukce na 8 ze 22



Obrázek 8: Redukce na 7 ze 22

VI) Po neúspěšných pokusech s konvexním obalem jsme se rozhodli vybrat vhodnou bázi z množiny vektorů $\{\vec{p}_j\}$ za pomoci statistiky. Z m „použitelných“ vektorů $\vec{p}_j$ o dimenzi n jsme proto vybrali vždy právě k a zkoušeli jsme do takovéto báze rozložit sadu naměřených dat. Přitom jsme sledovali, kolik rozkladů zkoumaných sad dat (řádků) má všechny složky nezáporné. Logickým vyústěním tohoto kroku je volba báze s největším počtem rozkladů, které neobsahují záporné složky.

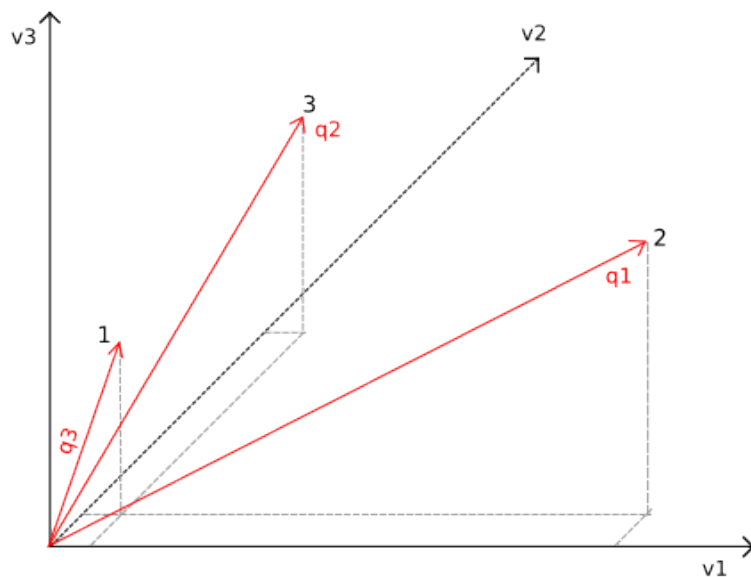
VII) Taktéž bylo prozkoumáno, zda je možné promítnout do více bází stejné dimenze všechny sady analyzovaných dat (řádky) tak, že všechny složky průmětů jsou nezáporné. S výjimkou případu redukce dimenze ze 6 na 4 k ničemu podobnému nedošlo.

Práce s lineárním podprostorem měla pouze charakter hledání postupů pro práci s podprostorem afinním. Z tohoto důvodu tu také nejsou uvedeny její konkrétní číselné výsledky. Vyjádření souřadnic sad analyzovaných dat $\vec{m}_i$ a potenciálních bázevých vektorů $\vec{p}_j$ jako násobků bázevých vektorů z matice ${}_k\mathbf{B}$ nebylo pro afinní vektorové podprostory použito. Stejně se ukázalo zbytečným použití konvexního obalu k redukci počtu vektorů $\vec{p}_j$.

4. Afinní vektorový podprostor

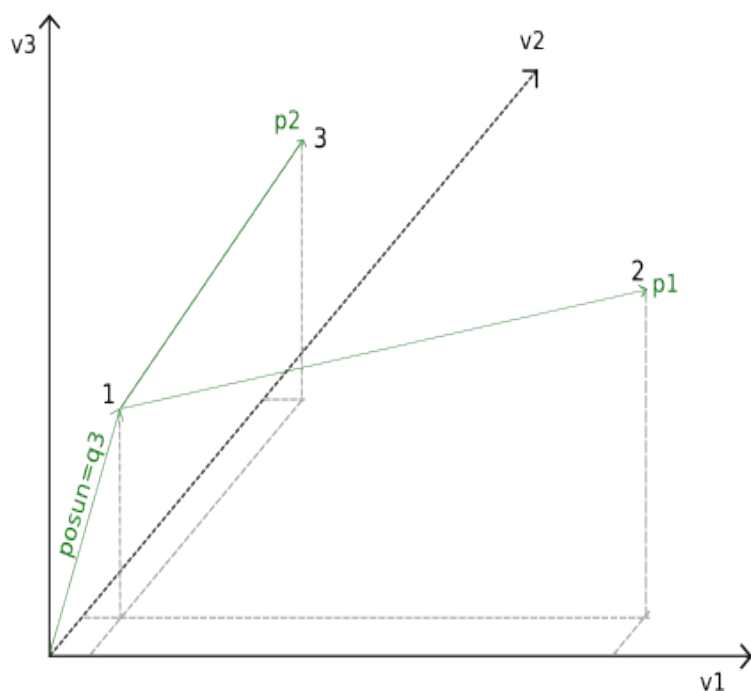
Po vyzkoušení práce s lineárními podprostory jsme přešli ke zobecnění nalezených postupů pro podprostory afinní. Afinní podprostor dimenze k je jednoznačně určený $(k+1)$ afinně nezávislými vektory přičemž je několik možností jak mohou být tyto vektory vzájemně orientovány.

Uvažujme například trojrozměrný prostor a jeho afinní podprostor s dimezí 2. Následující dva obrázky ukazují po řadě za prvé obvyklý a za druhé v diplomové práci použitý způsob uspořádání vektorů, které určují afinní podprostor (v příkladu jde o rovinu).



Obrázek 9: Obvyklé uspořádání

Afinní nadrovina, již zachycuje Obrázek 9, je určena obvyklým způsobem, tj. třemi vektory se společným počátkem a podmínkou, podle které souřadnice bodů v ní ležících jsou reálná čísla se součtem rovným jedné. Viz níže zapsané vzorce.



Obrázek 10: Použité uspořádání

V diplomové práci bylo k určování afinních podprostorů využito jiných vektorů než které je k tomuto účelu zvykem používat. Ukazuje je Obrázek 10. Vektorů byl sice

použít stejný počet, ale měly mírně odlišný význam. Koncový bod jednoho z vektorů totiž představoval společný počátek všech vektorů zbývajících. Vektor posunu počátku měl pro všechny rozklady sad analyzovaných dat stejnou velikost a proto nebylo třeba tuto informaci zanášet do každého jednoho rozkladu, ale bylo možné ji uchovat separátně. To znamená, že vektory aproximativních rozkladů do afinního podprostoru nemusely mít $(k+1)$ složek, ale vystačili jsme s k složkami. Zmenšení dimenze rozkladů vedlo ke zrychlení výpočtů a k úspoře paměti.

Číslice 1, 2 a 3 označují na obou obrázcích v této kapitole tytéž body, které leží ve zkoumaném afinním podprostoru určeném oběma zobrazenými způsoby.

Následuje matematický popis souvislosti mezi rozklady sad analyzovaných dat do obou typů graficky představených bází. J -tá sada dat $\vec{m}_j$ zobrazená do k -rozměrného afinního vektorového podprostoru ${}_kV$ prostoru V je popsána vzorcí:

$$\vec{m}_j = po\vec{sun} + \sum_{i=1}^k \alpha_i \cdot \vec{p}_i = \sum_{i=1}^{k+1} \beta_i \cdot \vec{q}_i, \quad \sum_{i=1}^{k+1} \beta_i = 1 \quad (9)$$

Vzorce (9) říkají, že k určení polohy bodu v k -rozměrném podprostoru ${}_kV$ je potřeba buď jeden vztažný bod s polohou definovanou vektorem $po\vec{sun}$ a k -tice lineárně nezávislých vektorů $\vec{p}_i, i \in \{1, \dots, k\}$ s počátky v bodě určeném vektorem $po\vec{sun}$ nebo $(k+1)$ libovolných lineárně nezávislých vektorů $\vec{q}_i$ takových, že koeficienty jejich násobků dají dohromady součet rovný jedné.

Další vzorec (10) ukazuje vztah mezi vektory $\vec{p}_i$ a $\vec{q}_i$.

$$\vec{q}_i = \vec{p}_i + po\vec{sun}, i \in \{1, \dots, k\}; \quad \vec{q}_{k+1} = po\vec{sun} \quad (10)$$

Poslední skupina vzorců, které posuzují vztah (resp. ekvivalenci) mezi oběma uvažovanými bázemi, porovnává velikosti násobků jednotlivých bázevých vektorů v rozkladech těchto dat do obou bází.

$$\begin{aligned} \vec{m}_j - po\vec{sun} &= \sum_{i=1}^k \alpha_i \cdot \vec{p}_i = \sum_{i=1}^k \beta_i \cdot \vec{q}_i + po\vec{sun} \cdot (\beta_{k+1} - 1) = \sum_{i=1}^k \beta_i \cdot \vec{q}_i + po\vec{sun} \cdot (1 - \sum_{i=1}^k \beta_i - 1) \\ \sum_{i=1}^k \beta_i \cdot \vec{q}_i + po\vec{sun} \cdot (1 - \sum_{i=1}^k \beta_i - 1) &= \sum_{i=1}^k \beta_i \cdot (\vec{p}_i + po\vec{sun}) - po\vec{sun} \cdot \sum_{i=1}^k \beta_i = \sum_{i=1}^k \beta_i \cdot \vec{p}_i \\ \alpha_i &= \beta_i, i \in \{1, \dots, k\} \end{aligned} \quad (11)$$

Ze vzorců je vidět, že navzdory zdánlivé výrazné odlišnosti obou sad vektorů, se shodují dokonce i koeficienty, které určují násobky báзовých vektorů ve složkách rozkladů jednotlivých sad dat a oba popisy jsou plně ekvivalentní. Důvod shodnosti koeficientů je ukrytý v charakteru vztahu mezi vektory $\vec{p}_i$ a $\vec{q}_i$, kde jedny jsou pouze speciální lineární kombinací druhých.

Hlavním úkolem této části práce bylo nalezení (resp. určení) vektoru posunu počátku báze tvořené vektory $\{\vec{p}_j\}$, $j=1, \dots, n$. Kritéria pro přijetí vektoru *posun* byla následující:

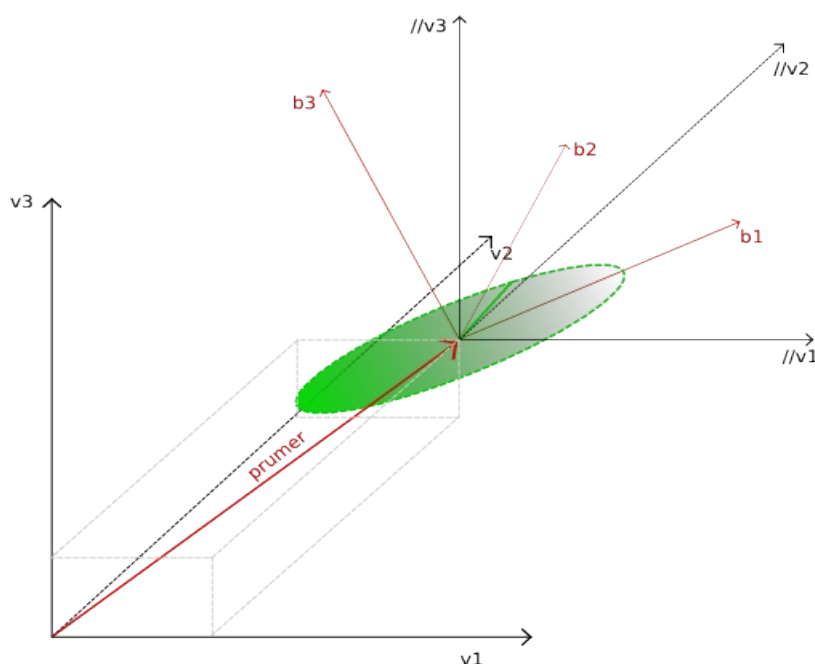
- 1)Vektory vzniklé rozdíly vektorů sad dat a vektoru posunu počátku měly být rozložitelné do báze $\{\vec{p}_j\}$, $j=1, \dots, n$ tak, aby všechny složky rozkladů byly nezáporné. Důvodem pro tento požadavek byla snaha interpretovat jednotlivé sady nám dodaných dat, které představují roztoky chemických látek, jako směsi báзовých roztoků $\{\vec{p}_j\}$, $j=1, \dots, n$ a to za předpokladu, že při uvažovaném mísení báзовých roztoků nedochází k chemickým reakcím, ale pouze k jejich „slévání“.
- 2)Vektor posunu počátku měl mít výhradně nezáporné složky. Vysvětlení je podobné jako v předchozím bodě. Vektor posunu počátku doplňuje bázi. Jde ve své podstatě o roztok, který je obsažen v každém analyzovaném roztoku. Jak bylo řečeno již dříve, roztok je v pojetí této práce představován vektorem nezáporných koncentrací rozpuštěných chemických látek.

Ukázalo se, že oba dva požadavky nebylo s poskytnutými soubory dat o analyzovaných roztocích možné splnit. Při zachování druhého, silnějšího požadavku jsme museli upustit od úmyslu rozložit do zvolené báze požadovaným způsobem všechna data. Druhý požadavek je silnější z toho důvodu, že jím reprezentovaný roztok je součástí všech analyzovaných roztoků. Nezáporné složky rozkladů zůstaly při tomto omezení, v nejnepríznivějším případě, například cca. 92%-ům dat ze souboru 638 chemických analýz roztoků s šesti složkami. Viz dále.

Postup určení „nejlepšího“ vektoru posunu počátku (tj. posunu předcházejícímu rozkladu do navrhované báze) využíval části dříve navržených algoritmů doplněné o výpočet vektoru *posun*.

Kroky použitého postupu jsou následující:

- Načtení naměřených dat
- Provedení „Analýzy hlavních komponent s odečtením průměru“, která poskytuje jako výsledek vlastní čísla a vlastní vektory kovarianční matice příslušející matici dat s upravenými daty a vrací taktéž data transformovaná do systému s bází $B = [\vec{b}_1, \vec{b}_2, \dots, \vec{b}_n]$. Vztah standardní báze $[\vec{v}_1, \vec{v}_2, \dots, \vec{v}_n]$ s bází B ukazuje, pro $n = 3$, Obrázek 11.



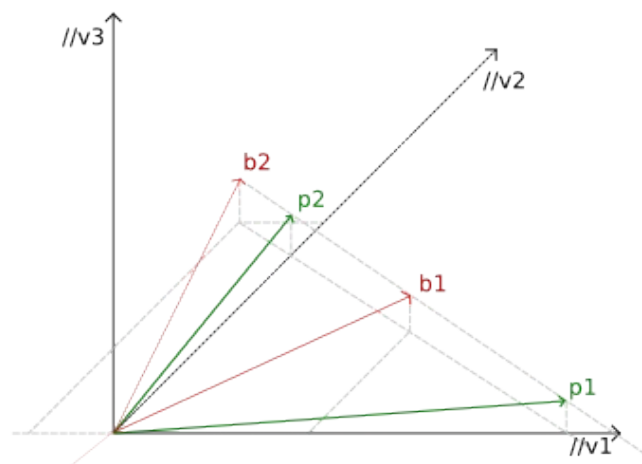
Obrázek 11: Báze získaná pomocí Analýzy hlavních komponent

V souřadném systému s osami $\vec{v}_i, i \in 1, 2, 3$ jsou zobrazena poskytnutá data o provedených chemických analýzách. Oblast ve které data leží je v obrázku ohraničena přerušovanou zelenou čarou. V dalším textu bude tato oblast místy označována jako „mrak dat“.

Červeně zakreslené vektory jsou spojeny s „Analýzou hlavních komponent s odečtením průměru“. Vektor směřující do středu mraku dat má souřadnice rovné průměrným hodnotám koncentrací chemických složek roztoků popsanych zpracovávaným souborem dat. Vektory $\vec{b}_i, i \in 1, 2, 3$ tvoří novou bázi prostoru V a jsou výstupem „Analýzy hlavních komponent s odečtením průměru“.

Vektory báze $\parallel \vec{v}_i, i \in 1, 2, 3$ jsou na Obrázku pouze pro lepší představu orientace vektorů $\vec{b}_i, i \in 1, 2, 3$ a zdvojená lomítka v jejich názvech naznačují rovnoběžnost s vektory $\vec{v}_i, i \in 1, 2, 3$.

- Z báze B byly vybírány k -tice vektorů tak, že jich vždy bylo vzato právě prvních k (tj. báze byla zredukována na ${}_k B = [\vec{b}_1, \dots, \vec{b}_k]$). Na dalším Obrázku jde o redukci dimenze ze tří na dva odebráním vektoru $\vec{b}_3$ a k -tice tak existuje pouze jedna.



Obrázek 12: Průnik podprostoru s hlavními nadrovinami

- Byly vypočteny vektory, které leží v průnicích ${}_k V$ a hlavních podprostorů dimenze $(n-k+1)$ prostoru V s počátkem posunutým o vektor *průmer* (viz Obrázek 11). Pro lepší představu tohoto kroku je zde Obrázek 12. Zdvojená lomítka v označení některých os na předchozích dvou obrázcích značí rovnoběžnost s osami určenými zbytkem lomítky uvedeného jména.

Obrázek 12 ukazuje podprostor ${}_k V$, který je kvůli v příkladu prováděné redukci dimenze prostoru V ze tří na dva určený dvojicí vektorů $[\vec{b}_1, \vec{b}_2]$.

Vektor $\vec{p}_1$ leží v průniku rovin určených vektorovými součiny $\vec{b}_1 \times \vec{b}_2$ a $\parallel \vec{v}_1 \times \parallel \vec{v}_3$. Vektor $\vec{p}_2$ je průnikem roviny $\vec{b}_1 \times \vec{b}_2$ s rovinou $\parallel \vec{v}_2 \times \parallel \vec{v}_3$. Roviny určené vektorovými součiny dvojic vektorů $\parallel \vec{v}_i, i \in 1, 2, 3$ jsou hlavními podprostory dimenze 2 prostoru V s počátkem posunutým o vektor *průmer*. Vektory $\vec{b}_i, i \in \{1, \dots, k\}$

definují podprostor ${}_kV$, který stejně tak může definovat jiná k -tice (pokud existuje) lineárně nezávislých vektorů v něm ležících, jako například námi nalezené a použité vektory $[\vec{p}_1, \dots, \vec{p}_k]$. Pro jednoznačné určení podprostoru ${}_kV$ je zvolenou k -tici samozřejmě nutné doplnit ještě vektorem *posun* (tj. jedním pevným bodem). Viz dále.

- Z nalezených vektorů $\vec{p}_j, j > k$ ležících na popsanych průnicích byly dále vybrány pouze ty s výhradně nezápornými složkami.
- Poté následoval algoritmus určující vektor *posun* a spolu s ním definitivně také bázi, která k vektoru *posun* „patří“. Tento algoritmus popisuje kapitola 4.1 Výpočet posunu počátku.

4.1 Výpočet posunu počátku

Funkci k výpočtu posunu počátku jsou jako argumenty poskytnuta data o chemických analýzách, potenciální bazové vektory $\vec{p}_j, j \in \{1, \dots, m\}$ a dimenze zamýšlené redukce (k). Navrací se jednak vektor *posun* a pak také identifikátory bazových vektorů $\vec{p}_i$, kterých je právě k a které tvoří novou bázi.

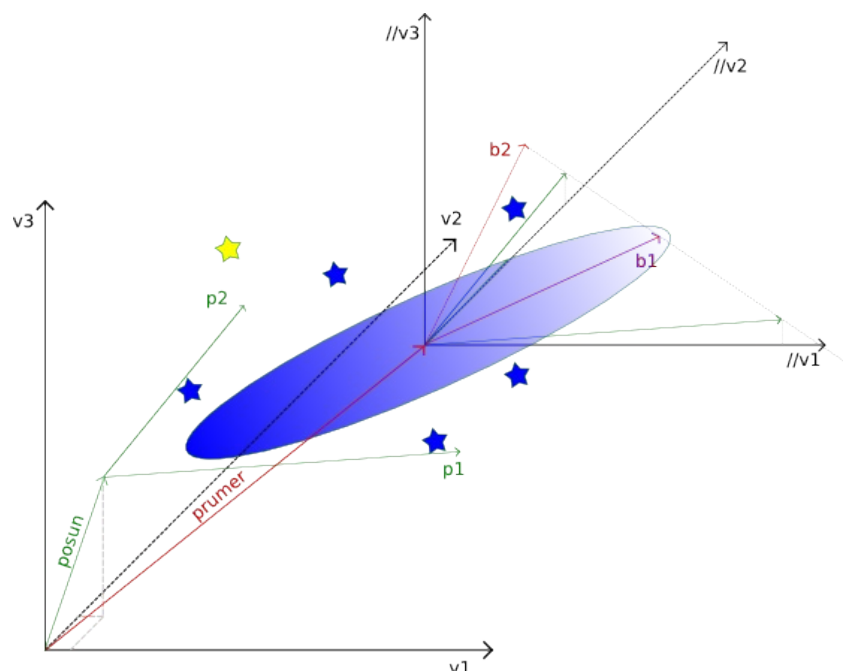
Výběr bazových vektorů $\vec{p}_i, i \in \{1, \dots, k\}$ probíhá následovně: Z m potenciálních bazových vektorů $\vec{p}_j, j \in \{1, \dots, m\}$ jsou sestaveny k -tice. Do každé z takto získaných k -tic jsou postupně rozložena data o chemických analýzách od kterých jsou předtím odečteny průměrné koncentrace. Z rozkladů jsou určeny vektory *minima*. Tyto jsou transformovány z podprostoru ${}_kV$ do prostoru $V = [\vec{v}_1, \dots, \vec{v}_n]_\lambda$ (v dalším textu je předpokládáno, že $\vec{minima} \in V$) a je prověřena nezápornost složek součtu takovéto transformace *minim* s vektorem průměrných koncentrací. Pokud jsou všechny složky vektoru součtu *minim* a průměrů nezáporné, práce s vybranou k -ticí končí zjištěním a případným uložením počtu řádků v matici s daty o chemických analýzách (tj. zjištěním počtu rozložitelných sad dat). K uložení dojde ve chvíli, kdy je počet rozložitelných sad dat vyšší než pro kteroukoli z ostatních doposud prověřovaných k -tic (viz dále). Povahu vektoru *posun* ukazuje Obrázek 13.

Modře vyznačená oblast spolu s hvězdičkami téže barvy představuje sady dat

(reprezentované řádky), která se podařilo rozložit do báze $[\vec{p}_1, \vec{p}_2]$ doplněné vektorem *pošun* tak, že složky rozkladů jsou nezáporné. Konečná báze spolu s vektorem *pošun* je na Obrázku zakreslena zelenou barvou. Žlutá hvězdička na obrázku symbolizuje data, která není možné do nové báze transformovat tak, aby byly všechny složky vektoru *pošun* nezáporné. Je také třeba zdůraznit, že Obrázek ukazuje ideální případ počtu vektorů $\vec{p}_j$, a to z toho důvodu, že z nich lze vytvořit pouze jednu *k*-tici (tj. dvojici, $m=k$). Připomeňme si, že obvykle je

$$\binom{m}{k} \gg 1$$

k-tic je velké množství a je z nich nutné vybrat tu „nejvhodnější“.



Obrázek 13: Nejde rozložit všechna dodaná data

V případě, že součet vektorů *minima* a *prumer* má záporné složky, potom jsou identifikovány ty řádky matice s daty, které tuto situaci vyvolávají. Nalezené řádky jsou dále zkoumány a je zjištěno, které složky jejich rozkladů se vyskytují v rozkladu vektoru *minima*. To znamená, že jsou vytvořeny jakési „uspořádané dvojice“, které říkají: „První složka rozkladu vektoru *minima* patří k rozkladu sady dat na řádku s číslem *i*. Druhá složka rozkladu vektoru *minima* patří k rozkladu sady dat na řádku číslo *j*. *K*-tá složka rozkladu vektoru *minima* patří k rozkladu sady dat na

řádku y ."

Následně je zjištěno, jak velké násobky α_i vektorů $\vec{p}_i$ je nutné přičíst k vektoru $\vec{minima}$, aby měl takový součet pouze nezáporné složky. Rozepsaný po složkách ukazují tento postup vzorce (12).

$$\begin{aligned} minima_j + \alpha_{ij} \cdot p_{ij} &\geq 0, \quad j \in \{1, \dots, n\} \\ \alpha_i &= \max(\alpha_{ij}) \end{aligned} \quad (12)$$

Nalezení takovýchto násobků α_i vektorů $\vec{p}_i$ umožní zjistit, kolik rozkladů sad zkoumaných dat do báze $\vec{p}_i, i \in \{1, \dots, k\}$ leží v „pásech“ ohraničených dvěma $(k-1)$ -rozměrnými podprostory. Zmíněné podprostory vznikají redukcemi prostoru ${}_k V$ odstraněním jednoho z vektorů báze $\vec{p}_i, i \in \{1, \dots, k\}$. Jeden z této dvojice podprostorů vždy obsahuje rozklad vektoru $\vec{minima}$, zatímco ve druhém pokaždé leží rozklad vektoru $\vec{minima} + \alpha_i \cdot \vec{p}_i$.

Počet rozkladů sad poskytnutých dat, které leží v jednotlivých pásech, je rozhodujícím kritériem při volbě jedné sady dat, která bude z „mraku dat“ vyčleněna a odstraněna. Pro připomenutí zopakujme, že rozklad každé ze sady poskytnutých dat, která má být odstraněna, musí obsahovat složku, která je zároveň na stejné pozici v rozkladu vektoru $\vec{minima}$. Ze sad zpracovávaných dat (koncentrací), které se svými rozklady podílí na rozkladu $\vec{minima}$, je k odstranění vybrána ta, která nabývá minima ve směru vektoru s nanejvýš rozklady ležícími v jemu příslušejícím pásu.

Ve chvíli, kdy jsou prověřeny všechny k -tice vytvořené z m poskytnutých vektorů $\vec{p}_j, j \in \{1, \dots, m\}$, je do hlavního programu navrácen vektor *pošun* spolu s bází z vektorů $\vec{p}_i, i \in \{1, \dots, k\}$. Zvolená báze se vyznačuje největším počtem proveditelných a „vyhovujících“ rozkladů. Vektor *pošun* je pro každou k -tici z $\vec{p}_j$ totožný s prvním nalezeným vektorem $\vec{minima}$ takovým, který má všechny složky nezáporné.

5. Výsledky použití nalezeného algoritmu

Pátá kapitola obsahuje konkrétní a okomentované číselné výsledky představených postupů aplikovaných na tři různé množiny zpracovávaných dat. Jde po řadě o množiny dat zanesené do matic o rozměrech 638x6, 2515x6 a 90x22. První a poslední ze zmíněných množin byly získány v průběhu jednoho roku provádění chemických analýz koncentrací rozpuštěných látek. Data zanesená do matice s rozměrem 2515x6 jsou záznamem chemických analýz v trvání pěti let.

5.1 Práce s maticí 638x6

Následující tabulka ukazuje porovnání počtů rozložitelných sad poskytnutých dat do konkrétních „nejlepších“ nalezenýchází tvořených vždy K vektory, kde každému K od dvou do šesti náleží jeden řádek.

K	Počet1	Procenta1	Počet2	Procenta2
2	547	85,7%	587	92,0%
3	552	86,5%	587	92,0%
4	586	91,8%	587	92,0%
5	621	97,3%	636	99,7%
6	638	100,0%	638	100,0%

Tabulka 1: Počty úspěšných rozkladů

Druhý a třetí sloupec ukazují počet a procentuální zastoupení vektorů rozdílů sad dat $\vec{m}_i$ a $pošun$ takových, které mají jenom nezáporné složky. Tyto dva sloupce jsou sice uvedeny pouze proto, aby bylo skutečné výsledky s čím porovnat. Druhý a třetí sloupec vlastně nesou informaci o tom, kolika z poskytnutých sad dat by zůstaly nezáporné složky, pokud by došlo k pouhému přesunutí počátku souřadné soustavy prostoru $V = [\vec{v}_1, \dots, \vec{v}_n]_\lambda$ o vektor $pošun$.

O skutečném výsledku nás informují čtvrtý a pátý sloupec. Tato část tabulky totiž říká, kolik z rozdílů $\vec{m}_i - pošun$ lze rozložit na součet nezáporných násobků bázevých vektorů $\vec{p}_i, i \in \{1, \dots, k\}$.

V tabulce jsou zřetelné dva trendy. Zdá se, že počet rozložitelných sad dat neklesá s růstem dimenze redukovaného prostoru a dokonce roste s upuštěním od požadavku vzájemné orthogonálnosti bázevých vektorů (tj. roste s rozkladem do nalezené báze). První z těchto jevů by mohl znamenat třeba tolik, že s klesající dimenzí redukovaného

prostoru roste norma vektroú *pošun* , protože se ale jedná pouze o nepodloženou domněnku a záhy dojde k popření tohoto „prvního jevu“ nebudeme se jím zabývat. Zajímavější pro zamyšlení je jev druhý. Vysvětlení toho, že po rozkladu do báze, jejíž vektory nejsou nutně orthogonální je situace příznivější, nemusí být ale nijak složité. Při provádění rozkladů (spojených s redukcí dimenze) bylo využito řešení přeúřčených soustav rovnic pomocí metody nejmenších čtverců a z toho plyne, že výsledky jsou zatíženy chybou a tedy nepřesné (minimálně nepřesné). V důsledku zanesení chyby je část záporných složek pravděpodobně odstraněna.

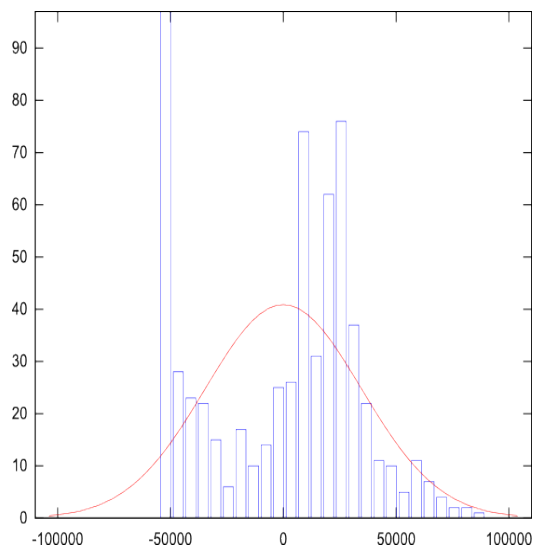
Na druhou stranu si je ale také nutné uvědomit, že pokud by měl vektor *pošun* takové souřadnice, aby všechna data ležela v prvním oktantu (resp. $2^n - tantu$) posunutě soustavy souřadnic, potom by **žádné** „zanesení chyby“ nemohlo vyvolat zvýšení množství rozložitelných dat.

Z dále uváděných výsledků vyplyne, že i druhý pozorovaný jev je s nejvyšší pravděpodobností vyvolán pouze speciálními (blíže neurčenými) vlastnostmi báze *B* určené „Analýzou hlavních komponent s odečtením průměru“ aplikovanou na množinu dat 638x6.

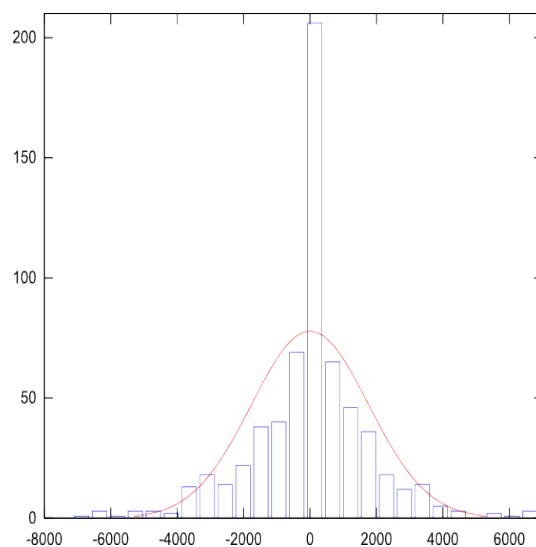
Krátce shrnuto, nezápornost složek porováváme u dvou různých sad dat ze kterých jedna sada jsou poskytnutá data o chemických analýzách koncentrací chemických látek v roztocích a sada druhá jsou poskytnutá data cíleně zatížená chybou.

Dále je z tabulky možné usoudit, že pokud se nebudeme ptát na chyby aproximací sad dat s redukovanou dimenzí, potom je nejvhodnější používat redukci na dimenzi 2, protože redukcí ze 6-ti na 5 se mnoho paměti počítače neušetří a ostatní prováděné redukce poskytují v nejlepším případě stejné kvantitativní výsledky.

Chyba prováděných aproximativních rozkladů (tj. kvalita výsledků), která byla vyvolána redukcemi dimenze vektorů sad dat, byla v rámci této práce posuzována pouze okrajově, zato však dvěma způsoby. Ptát se totiž můžeme buď na chybu celého vektoru nebo na chyby jednotlivých složek. Viz Kapitola 7.1 .

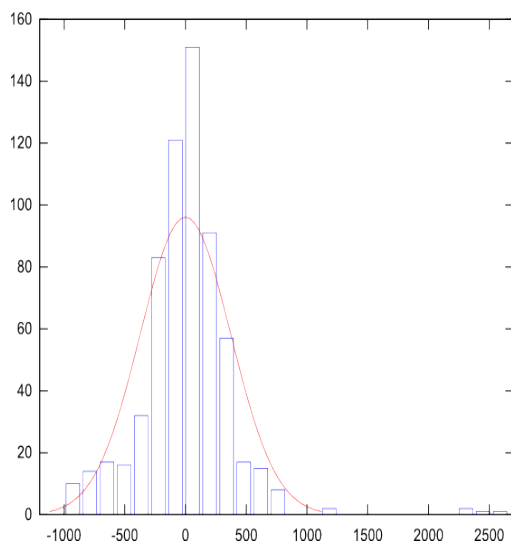


Obrázek 14: První složka

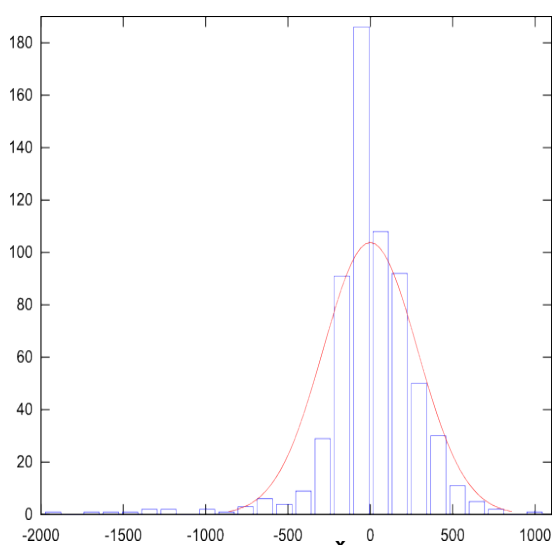


Obrázek 15: Druhá složka

Předpokládejme, že výsledné pravděpodobnosti zapsané v tabulce nemohou nabývat libovolně malých, nezáporných hodnot. Pokud jsou souřadnice dat o chemických analýzách vyjádřeny jako násobky vektorů ${}_k\mathbf{B}=[\vec{b}_1, \vec{b}_2, \dots, \vec{b}_k]$, potom mají složky těchto vektorů přibližně normální rozdělení. Toto zjištění dokládají Obrázky 14-17. Rozdělení páté a šesté složky zde není zobrazeno. Svým charakterem se obě rozdělení podobají rozdělení pravděpodobnosti výskytu hodnot druhé až čtvrté složky.



Obrázek 16: Třetí složka



Obrázek 17: Čtvrtá složka

Jak je vidět, výjimku v charakteru rozložení pravděpodobnosti výskytu hodnot tvoří

pouze první ze složek. Je tomu tak proto, že první složka odpovídá násobkům prvního vektoru, který byl nalezen „Analýzou hlavních komponent s odečtením průměru“, a z tohoto důvodu nese tato složka největší množství informace. V tomto smyslu výkladu, obsahují další složky čím dál tím větší podíl „šumu“ ku užitečné informaci. Výjimku, kterou představuje první složka, nebudeme pro jednoduchost dále uvažovat. Pokud tedy přistoupíme k tomuto zjednodušení a složky vektorů sad dat prohlásíme, z hlediska popisu pravděpodobnosti nabývání hodnot jednotlivých složek, za nezávislé, můžeme spočítat pravděpodobnost výskytu roztoku (vektoru), ve kterém jsou „díličí pravděpodobnosti“ rovny $X\%$, jako součin takovýchto „díličích pravděpodobností“. Zvolíme-li si pro všech šest složek $P(X) = 0,975$ (nepokryje pouze 2,5% případů), dostaneme

$$P(\mathbf{B}) = \prod_{i=1}^6 P(b_i) = 0.859 \quad (13),$$

což je méně než minimální pravděpodobnost zapsaná v tabulce a z tohoto hlediska je tedy tato hodnota vyhovující.

5.1.1 Vypočtené bázové vektory

Na tomto místě jsou zapsány konkrétní výsledky, tj. nalezené vektory *posun* a báze $\vec{p}_i, i \in \{1, \dots, k\}$ získané při různých redukcích 6-ti rozměrného prostoru, který obsahoval data získaná během jednoletého období provádění chemických analýz koncentrací kontaminantů v podzemní vodě. U redukce na dimenzi 3 jsou pro srovnání zapsány také posun a báze používané v podniku DIAMO s. p. První z vektorů reprezentuje vždy určený posun. Zbýlé sloupce jsou vektory samotné báze.

Při zmenšení rozměru prostoru ze 6-ti na 2 byly určeny vektory:

Posun2	P1	P2
1229,757	0,000	0,914
688,334	0,949	0,000
1,901	0,001	0,013
176,791	0,039	0,057
70,946	0,011	0,013
1,841	0,000	0,003

Tabulka 2: Posun a báze,
dimenze prostoru 2

Redukce na dimenzi 3 měla výsledek:

Posun3	P1	P2	P3
1162,918	0,000	0,921	0,891
688,283	0,955	0,000	0,000
0,185	0,000	0,012	0,016
177,537	0,044	0,063	0,035
59,142	0,000	0,000	0,058
2,351	0,001	0,004	0,000

*Tabulka 3: Posun a báze,
dimenze prostoru 3*

Báze poskytnutá ing. Wasserbauerem je takováto (dimenze 3):

W0	W1	W2	W3
216,000	75838,800	121705,000	113803,000
15,000	43314,600	77639,200	76371,200
0,030	886,721	1222,190	3208,140
0,250	5494,920	10736,200	10095,900
0,460	4483,170	327,274	888,792
0,050	182,600	512,400	415,000

Tabulka 4: báze ing. Wasserbauera

Sloupec W0 je posunem. Ostatní z vektorů se stanou vektory báze (ve smyslu postupu použitého v této práci) afinního podprostoru po odečtení W0 po složkách od každého z nich. Vektory W1 totiž určují afinní podprostor výše popsaným „obvyklým způsobem“.

Po úpravě odečtením W0 vypadá báze následovně.

PosunW	P1	P2	P3
216,000	75622,800	121489,000	113587,000
15,000	43299,600	77624,200	76356,200
0,030	886,691	1222,160	3208,110
0,250	5494,670	10735,950	10095,650
0,460	4482,710	326,814	888,332
0,050	182,550	512,350	414,950

Tabulka 5: upravená báze

ing. Wasserbauera

Podrostor o dimenzi 4 je definován vektory posunu a báze:

Posun4	P1	P2	P3	P4
1064,498	0,000	0,773	0,979	0,923
692,017	0,955	0,000	0,000	0,000
1,411	0,000	0,000	0,017	0,018
96,406	0,045	0,223	0,000	0,000
17,142	0,000	0,000	0,000	0,059
4,457	0,001	0,004	0,004	0,000

*Tabulka 6: Posun a báze,
dimenze prostoru 4*

Redukce o jednu dimenzi, tj. ze 6-ti na 5 skončila výsledkem:

Posun5	P1	P2	P3	P4	P5
69,660	0,000	0,000	0,995	0,932	0,866
13,800	0,000	0,999	0,000	0,000	0,000
0,024	0,000	0,000	0,000	0,000	0,134
0,248	0,998	0,000	0,000	0,000	0,000
0,053	0,000	0,000	0,000	0,068	0,000
0,200	0,002	0,001	0,005	0,000	0,000

*Tabulka 7: Posun a báze, dimenze
prostoru 5*

Bez redukce je nejvhodnějším posunem vektor *ninim* a báze je tvořena lineárně nezávislými, orthogonálními vektory, které mají vždy právě jednu nenulovou složku.

Posun6	P1	P2	P3	P4	P5	P6
121,000	0,000	0,000	0,000	0,000	0,000	1,000
13,800	0,000	0,000	0,000	0,000	1,000	0,000
0,030	0,000	0,000	0,000	1,000	0,000	0,000
0,250	0,000	0,000	1,000	0,000	0,000	0,000
0,030	0,000	1,000	0,000	0,000	0,000	0,000
0,050	1,000	0,000	0,000	0,000	0,000	0,000

Tabulka 8: Posun a báze, dimenze prostoru 6

Při porovnání vektoru posunu počátku neredukované báze s prvním vektorem báze ing. Wasserbauera je zjevné, že vektor tvořený minimálními koncentracemi jednotlivých složek je blízký (alespoň v řádech) cenomanské vodě, kterou právě první vektor ing. Wasserbauera představuje.

5.2 Práce s maticí 2515x6

Vzhledem ke skutečnosti, že k „pětiletým datům“ nebyla dodaná ani báze využívaná v DIAMO s.p. ani tamtéž nalezené rozklady, nejsou v této kapitole zapsány nalezené posuny a bazové vektory, protože by tyto nebylo s čím porovnat. Uvádíme zde tabulku

s počty úspěšně provedených rozkladů.

K	Počet1	Procenta1	Počet2	Procenta2
2	2223	88,4%	2515	100,0%
3	2368	94,2%	2499	99,4%
4	2512	99,9%	2420	96,2%
5	2513	99,9%	2482	98,7%
6	2515	100,0%	2515	100,0%

Tabulka 9: Počty úspěšných rozkladů

Tabulka jasně vyvrací platnost obou trendů popsanych v předchozí podkapitole. Počet rozložitelných sad dat nemusí při použití navržených postupů růst spolu s dimenzí redukovaného prostoru a nemusí dokonce růst ani v případě přechodu od ortogonální souřadné soustavy k obecněji definované souřadné soustavě.

Co se týče volby dimenze redukovaného prostoru, je vidět, že jednoznačným favoritem je dimenze 2. Do podprostoru této dimenze jdou totiž rozložit všechna analyzovaná data. Zároveň je ale před redukcí vždy nutné kvantifikovat míru zachování nesené informace. Z principu analýzy hlavních komponent totiž vyplývá, že čím menší je dimenze redukovaného prostoru, tím méně užitečné informace v upravených datech zbývá.

5.3 Práce s maticí 90x22

Práce s daty o dimenzi 22 dala podnět k modifikaci Analýzy hlavních komponent s odečtením průměru. Jednalo se o to, že po odečtení průměru a vypočtení vlastních vektorů příslušné kovarianční matice měl první z těchto vektorů několik záporných složek. Tato skutečnost vyvolávala chyby v navržených algoritmech a taktéž znemožňovala nalezení básových vektorů splňujících stanovené požadavky.

Výsledky práce s daty o dimenzi 22 jsou pravděpodobně zatíženy velkou chybou, protože analýz koncentrací všech dvaadvaceti chemických látek je v poměru k jejich počtu malé množství. Pro porovnání následuje několik podílů:

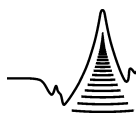
$$\frac{\text{počet měření}}{\text{dimenze prostoru}} = \frac{90}{22} = 4,091; \quad \frac{638}{6} = 106,330; \quad \frac{2515}{6} = 419,170$$

Řešení výše nastíněné situace se zápornými složkami vektoru $\vec{b}_1$ spočívalo v odečtení jiného vektoru než byl vektor průměrných koncentrací. Odečítaný vektor

byl násobkem vektoru průměrných koncentrací a byl hledán za pomoci metody půlení intervalu tak, aby ležel vektoru průměrů co možná nejbližší a jeho eukleidovská norma byla menší než norma vektoru průměrů. „Blížkost“ obou vektorů je požadována proto, aby byla po případné redukci nalezené nové báze zachována maximální míra nesené informace. Odečítání jiného vektoru než je vektor průměrných koncentrací (jeho násobku) je v popsaném případě pozvolným přechodem od podprostoru afinního k podprostoru lineárnímu.

Jako potvrzení přechodu od afinního k lineárnímu podprostoru je dále vypsáno několik sloupcových vektorů uspořádaných do tabulky. První ze sloupců, které ukazuje Tabulka 10 obsahuje pouze pořadová čísla přiřazená vždy jedné konkrétní chemikálii, jejíž koncentrace byla analyzována. Ve druhém sloupci je zaznamenaný vektor $\vec{od\acute{e}cet}$, který byl v Analýze hlavních komponent odečtený namísto vektoru průměrných koncentrací. Třetí sloupec představuje $(k+1)$. báze vektoru afinního podprostoru. Tomuto vektoru říkáme $\vec{posun22}$. Čtvrtý sloupec zachycuje vektor minimálních analýzami zjištěných koncentrací jednotlivých v podzemní vodě rozpuštěných látek. A konečně pátý sloupec popisované tabulky nese doplňkovou informaci o tom, zda je konkrétní jedna složka vektoru $\vec{posun22}_k$ větší nebo menší než příslušná složka vektoru minimálních koncentrací.

Přechod k lineárnímu podprostoru je patrný na složkách vektoru $\vec{od\acute{e}cet}$, které jsou ve většině případů (výjimka je 11.složka) menší než složky vektoru minimálních koncentrací. Pro lepší představu plynoucí z podstaty hledání vektoru odečítaného v Analýze hlavních komponent poznamenejme, že $\vec{od\acute{e}cet}$ je přesně padesátinou vektoru průměrných koncentrací.



K	Odečet	Posun	Minima	Indikátor
1	1380,76	10997,8	44760	1
2	841,98	6890,18	25450	1
3	19,73	155,59	477	1
4	122,42	1009,24	3930	1
5	29,23	254,15	631	1
6	4,92	37,84	152,6	1
7	1,42	10,92	45,68	1
8	0,44	3,73	15,12	1
9	0,02	0,15	0,47	1
10	5,95	52,12	177,2	1
11	11,19	111,05	1,58	-1
12	5,24	48,37	33,2	-1
13	0,26	2,18	10,04	1
14	50,69	170,79	337	1
15	1,16	10,03	35,14	1
16	0,44	4,1	6,72	1
17	0,75	6,79	12,14	1
18	0,36	2,41	2,2	-1
19	1,22	10,7	28,36	1
20	0,25	2,14	8,29	1
21	0,13	1,07	1,3	1
22	0,31	2,66	9,4	1

Tabulka 10: Porovnání odečtu s minimem a průměrem

Tabulka 11 ukazuje počty rozložitelných sad dat a jejich procentuální zastoupení v celkem devadesáti sadách dat s dvaceti dvěma složkami.

K	Počet1	Procenta1	Počet2	Procenta2
2	72	80,0%	84	93,3%
3	77	85,6%	87	96,7%
4	85	94,4%	72	80,0%
5	86	95,6%	78	86,7%

Tabulka 11: Počet úspěšných rozkladů

Podobně jako výsledky práce s „pětiletými daty“ o dimenzi šest, tak i „jednoroční data“ o dimenzi dvacet dva ukazují skutečnost, že při využitých postupech nemá maximální počet rozložitelných sad dat zřetelnou souvislost s dimenzí redukovaného prostoru a zanesení chyby nemusí nutně znamenat kvantitativní zlepšení výsledku.

Z prověřovaných dimenzí afinního podprostoru se zdá být nejlepší volbou redukce na dimenzi 3. Přičemž zdůrazníme, že dimenze původního prostoru byla 22 a její zmenšení tak činí plných 19D.

5.4 Souhrnné stanovisko k výsledkům

Z výše zapsaných výsledků je jasně patrné, že navržené postupy aplikované na různé sady dat vedou k sadám výsledků s různými vlastnostmi. Na místě je otázka: „Jaké musí soubor dat splňovat požadavky aby bylo pro redukci dimenze vektorové reprezentace těchto dat možno využít nalezené postupy?“. Z porovnání zkušeností získaných redukcemi šesti- a dvaadvacetirozměrného prostoru je možno usoudit, že dat v poměru k dimenzi redukovaného prostoru by mělo být co možná nejvíce. Čím více bude analyzovaných dat, tím spíše se projeví korelace mezi jejich složkami a to povede k přesnějšímu určení vektorů nové báze B prostoru V . V ideálním případě má vektor $\vec{b}_1$ pouze nezáporné složky a pokud tomu tak není, potom je nutné Analýzu hlavních komponent upravit takovým způsobem, aby se nezápornosti složek docílilo. Úpravu jsme prováděli prostřednictvím odečtení jiného vektoru než *průměr*. Je ale nutné mít na paměti, že podobnými modifikacemi se zhoršuje přesnost aproximací jednotlivých sad měření.

6. Kritéria optimálnosti báze

Protože jako bázi vektorového prostoru dimenze k je možné vybrat jakoukoli k -tici lineárně nezávislých vektorů, bylo na začátku práce nutné stanovit vlastnosti, které má splňovat „přijatelná“ báze. V průběhu vypracovávání diplomové práce byly tyto požadavky zpřesňovány a rozšiřovány takovým způsobem, aby nás dovedly k výběru jedné, „optimální“ báze. Na „optimálnost“ báze je nutno hledět s určitou rezervou. S velkou pravděpodobností nejde o nejlepší možnou, ale pouze o nejlepší nalezenou $(k+1)$ -tici báзовých vektorů.

Společným požadavkem kladeným na všechny vektory byla nezápornost jejich složek. Nejlepší $(k+1)$ -tice byla vybrána na základě největšího počtu dodaných dat, která byla do ní požadovaným způsobem rozložitelná. Požadavek na nezápornost složek báзовých vektorů vyplynul ze záměru interpretovat proměřené roztoky jako směsi **roztoků** báзовých.

Následuje přehled dalších kritérií uvedených v pořadí, které odpovídá jejich průběžnému přidávání.

1. $(k+1)$ bázových vektorů afinního podprostoru (v případě lineárního podprostoru k) má mít takový směr, aby v průmětech dat do podprostoru jimi určeného byla zachována maximální míra celkové, původně nesené informace. To znamená, že aproximativní vektor nemá ležet daleko od původního vektoru odpovídajícího zvolené sadě chemických analýz. Splnění tohoto požadavku vyplývá z použití analýzy hlavních komponent.
2. Bázové vektory $\vec{p}_1, \dots, \vec{p}_k$ zvoleného afinního podprostoru původního n -rozměrného prostoru V mají mít pouze nezáporné složky. Způsob nalezení takových vektorů je popsán v Kapitole 4.
3. Případný $(k+1)$. bázový vektor „fixuje polohu“ k -rozměrného podprostoru v původním n -rozměrném prostoru. Z tohoto hlediska má podobný význam jako v případě přímky bod, kterým přímka se známou směrnici prochází. Na tento vektor je kladen takový požadavek, aby ležel v prvním kvadrantu (oktantu, resp. $2^n - tantu$) původního prostoru (tj. aby měl všechny složky nezáporné).
4. Vybraná k -tice bázových vektorů $\vec{p}_1, \dots, \vec{p}_k$ se vyznačuje tím, že při jejím „vhodném“ umístění, určeném příslušným $(k+1)$. vektorem *pošun*, do původního n -rozměrného prostoru, lze do ní provést nejvíce rozkladů poskytnutých sad dat takových, že v těchto rozkladech figurují pouze a jenom nezáporné násobky v bázi obsažených vektorů. „Vhodné“ umístění má pozici $(k+1)$. bázového vektoru.

Krom kvantitativních měřítek pro výběr báze je, jak bylo už zmíněno, přirozeně žádoucí sledovat také kvalitu výsledku. To se projevilo při srovnávání výsledků našeho postupu s výsledky dosaženými v DIAMu s. p. (viz Kapitola 7). Během většiny práce jsme předpokládali, že kvalita výsledku je zajištěna analýzou hlavních komponent. Ukázalo se však, že PCA preferuje uchování informace o nejsilněji zastoupených složkách na úkor informace o složkách přítomných v roztoku v marginálních množstvích. Tento problém je možno řešit škálováním dodaných dat, předcházejícím jejich dalšímu zpracování. Po přeškálování není nutné nalezená kritéria výběru bázových vektorů měnit.

7. Odlišnosti od existujících postupů

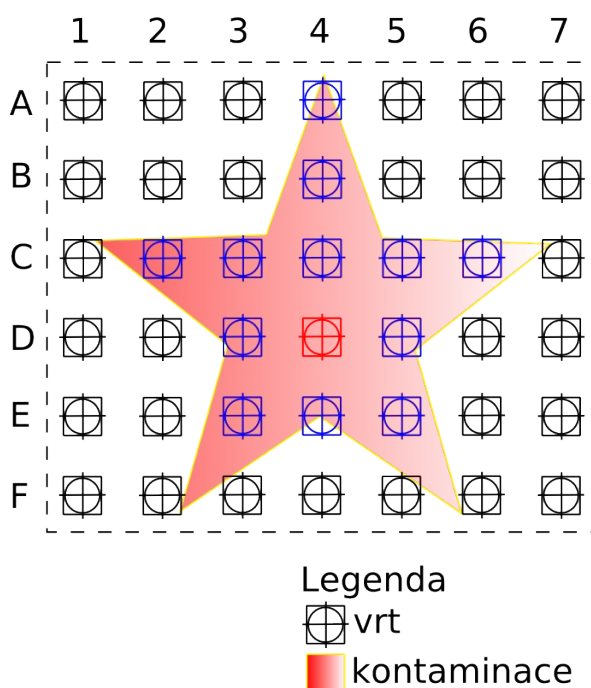
Jediným nám známým postupem (krom námi narhovaného) k redukci dimenze vektorového prostoru obsahujícího informace o provedných chemických analýzách je metoda pro toto využívaná ve firmě DIAMO s.p. ing. Wasserbauerem. Vzhledem k odlišnému přístupu k problému se v této práci představený a v DIAMu s.p. používaný postup od sebe navzájem poněkud liší. Zatímco v této práci je k určení báze využito algebraických výpočtů, v podniku DIAMO s.p. je při redukci dimenze prostoru z 22 na 2 (resp. na 3) využit vektor reprezentující cenomanskou vodu, dále 2 vektory, které představují vtláčené technologické roztoky a nakonec vektor takový, který je dopočtený tím způsobem aby součet koeficientů rozkladů analyzovaných dat byl roven jedné.

Prvním krokem postupu ing. Wasserbauera (v DIAMu s.p.) je prohlášení roztoků s malým množstvím rozpuštěných látek za cenomanskou vodu. Malé množství je v tomto případě stanoveno na 2500 mg/l. Pro srovnání uveďme limit pro pitnou vodu, který činí 250 mg/l. Ve vtláčených technologických roztocích je koncentrace rozpuštěných látek blízka 40000 mg/l.

Nahrazení slabých roztoků vodou má význam z hlediska dopočítávaného vektoru (v našem případě všech). Po takovéto záměně se v matici s daty může vyskytnout cenomanská voda na „velkém“ množství míst (řádků). Jestliže všechny řádky představující vodu vyjmem a zapomeneme, znamená to, že uchováme pouze informaci o oblasti do které se kontaminant rozšířil. Jde-li nám o popis rozložení chemických látek v horninovém prostředí, potom zmenšení zkoumané oblasti znamená obvykle kvalitativní zlepšení tohoto popisu.

Příklad:

Příklad osvětluje možné důvody případného vyřazování analýz roztoků s malými koncentracemi všech rozpuštěných látek. Mějme výsledky chemických analýz koncentrací látek rozpuštěných v podzemní vodě na hypotetické lokalitě Valsko. Zkoumaná oblast v lokalitě Valsko nechť má půdorys obdélníku, který ukazuje Obrázek 18. Ve zkoumané oblasti ať jsou ekvidistantně rozloženy vrty mající souřadnice A1 až F7. Vrtem se souřadnicí D3 byl v minulosti do podloží vtlačén určitý technologický roztok.



Obrázek 18: Kontaminovaná oblast

Vtláčený technologický roztok se za dobu provádění chemických analýz rozšířil do určité oblasti. Tvar kontaminované oblasti, zvolený čistě pro příklad, na vloženém obrázku ukazuje, že kontaminace se v závislosti na vlastnostech horniny v podloží nemusí šířit všemi směry stejně rychle. Kontaminovaná oblast tedy nemusí bezpodmínečně mít tvar kruhu nebo elipsy.

Jestliže tedy chceme popsat co nejpřesněji rozložení kontaminantu v oblasti, ve které se vyskytuje, potom není účelné pracovat s informacemi z vrtů v místech, kam kontaminant nedoputoval v dostatečném množství. Vrt, ve kterých nebyla významná přítomnost kontaminantu zjištěna, jsou na obrázku nakresleny černou barvou. Naproti tomu informace, se kterými si přejeme pracovat, pochází z vrtů nakreslených modrou,

případně červenou barvou.

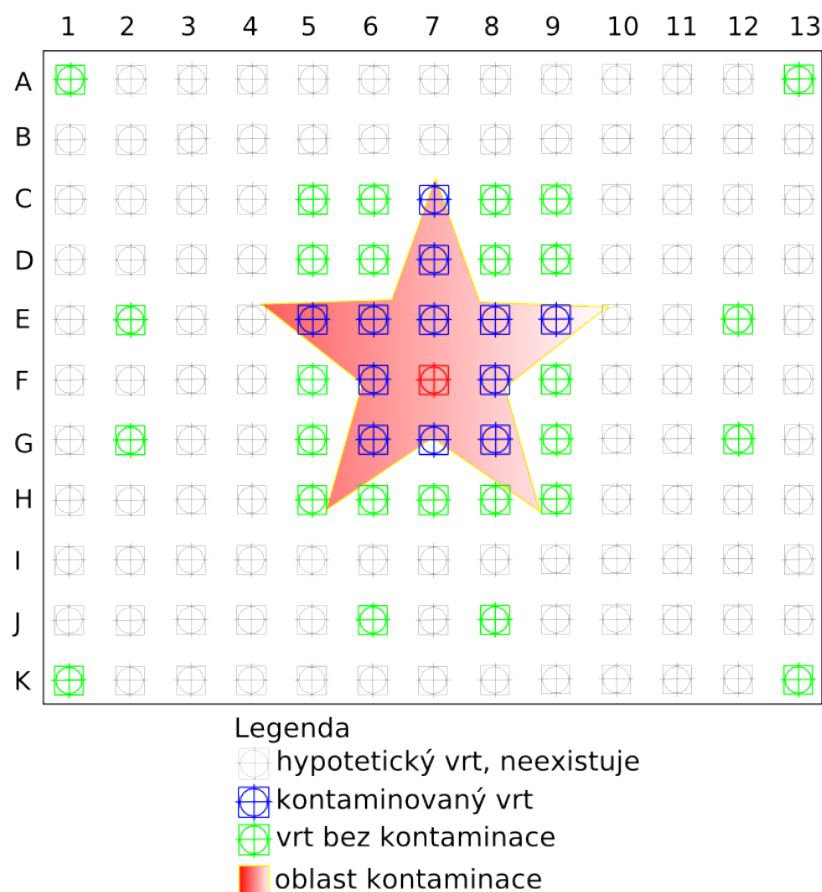
- - konec příkladu - - -

Pokud jsou naše priority jiné než přesné vymezení a popis kontaminované oblasti, například popis rozložení chemikálií na celém vrty „proděravěném“ území, potom se může zdát, že odstranění informace o vodách je nesprávný postup. Dále je ukázáno, že i v takovém případě má tento krok své opodstatnění.

Významnou skutečností, kterou musíme brát v úvahu při snaze o co nejpresnější popis zkoumané lokality, je fakt, že vrty nejsou s nejvyšší pravděpodobností rozmístěny ekvidistantně. Situaci s rozložením vrtů ukazuje Obrázek 19. Největší počet vrtů je přirozeně v místech, kde byla expertními odhady předpovězena a průběžně měřeními potvrzována největší míra kontaminace podzemní vody chemickými látkami. Takové území je v relativním měřítku k velikosti zkoumané lokality pravděpodobně velmi malé. Vzdálenost mezi krajními vrty v kontaminované oblasti a vrty na hranicích celé zkoumané lokality může být značně vyšší než průměrná vzdálenost vrtů na samotném kontaminovaném území. Odstranění vod ze sady dodaných dat vychází z předpokladu, že v podzemí je od místa poklesu koncentrace rozpuštěných látek pod stanovenou mez až k hranici prozkoumané oblasti přítomna plus-mínus čistá cenomanská voda a uchovávat tuto informaci na velkém množství míst není třeba. Na druhou stranu, pokud chceme popsat rozlehlé území potom není možné informaci o vodě „zahodit“ a vzhledem k relativní velikosti části lokality, kde se voda podle předpokladů nachází, je nutné přisoudit informaci o vodě patřičnou váhu.

Za předpokladu, že bychom z množiny poskytnutých dat neodstraňovali sady odpovídající vodě blízkým roztokům a měření by byla na celé lokalitě prováděna v ekvidistantně rozmístěných vrtech, potom by se váha informace o vodě sama přirozeně projevila v důsledku majoritního zastoupení vodu zachycujících analýz. Protože ale vrty ekvidistantně rozmístěny nejsou, proto je nutné přisoudit vodě váhu uměle. V limitním případě, kdy by se počet analýz vody z ekvidistantně rozložených vrtů blížil 100% tak by průměrné koncentrace chemických látek rozpuštěných v podzemní vodě byly blízké koncentracím látek v cenomanské vodě. S průměrnými

koncentracemi pracuje Analýza hlavních komponent (viz příslušná kapitola). To znamená, že přisouzení velké váhy cenomanské vodě je možno provést úpravou „Analýzy hlavních komponent s odečtením průměru“ na „Analýzu hlavních komponent s odečtením cenomanské vody“.



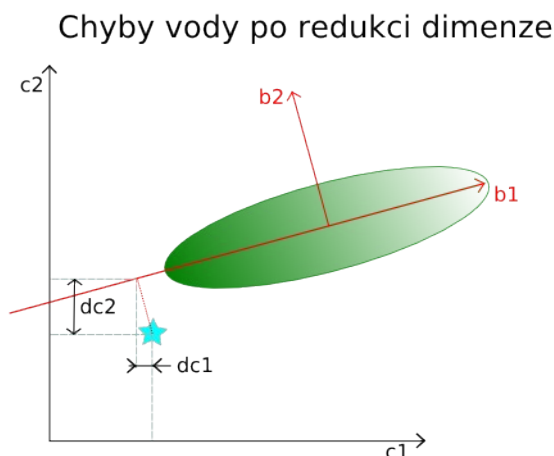
Obrázek 19: Důvod důležitosti vody

Ing. Wasserbauer postupoval při vážení významu vody jinak. Nevyužíval totiž analýzy hlavních komponent a namísto toho do báze podprostoru rovnou zahrnul vektor koncentrací chemických látek rozpuštěných v cenomanské vodě.

Pro připomenutí, naším postupem, stejně jako postupem ing. Wasserbauera se předchází zanesení výrazné chyby do aproximací roztoků blízkých cenomanské vodě. Spolu se zpřesněním aproximací koncentrací ve vodě zmenší přesnost aproximací všech ostatních proměřených roztoků. Požadavek na přesné přiblížení koncentrací v cenomanské vodě plyne, krom už napsaných důvodů, také ze snadné úvahy, která je představena v následujícím příkladu.

Příklad:

Mějme sadu dodaných chemických analýz koncentrací látek rozpuštěných v podzemní vodě. Obrázek 20 ukazuje zelenou barvou vyznačenou oblast (oblast v grafu, nikoliv území) s analyzovanými daty ke kterým náleží navíc také analýza cenomanské vody zachycená na obrázku v podobě modré hvězdičky.



Obrázek 20: Chyba vody

Analýza hlavních komponent s odečtením průměru vede k nalezení báze

$$\mathbf{B} = \{\vec{b}_1, \vec{b}_2\}$$

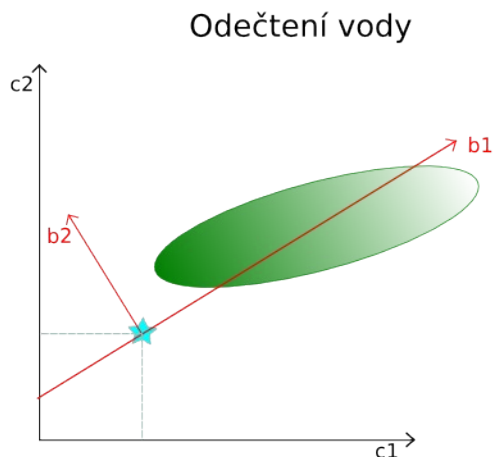
. Po vyjádření souřadnic analyzovaných dat v bázi $\mathbf{B}$ a po následné redukcí dimenze prostoru „zapomenutím“ složek souřadnic příslušejících vektoru $\vec{b}_2$ bude do vektoru analýzy cenomanské vody zanesena chyba. Může dojít například k tomu, že chyba koncentrace jedné z látek (c_1) je přijatelná, zatímco chyba koncentrace jiné látky (c_2) je například vlivem její zdravotní závadnosti neakceptovatelná. Zjednodušeně řečeno, z analyzované pitné vody může redukce **z pohledu hygienika** udělat účinný prostředek pro hubení hlodavců.

Jestliže tedy chceme aby byl určitý roztok vyjádřen naprosto přesně, potom je možným řešením odečíst v analýze hlavních komponent namísto vektoru průměrných koncentrací vektor představující vybraný roztok. Pro případ, kdy je vybraným roztokem cenomanská voda vyvolá její odečtení změnu směru vektorů báze

$$\mathbf{B} = \{\vec{b}_1, \vec{b}_2\}$$

. Tato změna bude mít charakter podobný tomu, který ukazuje Obrázek 21. Je patrné, že v tomto případě voda a jí blízké roztoky nebudou redukcí dimenze žádnou výraznou chybou zatíženy. Stejně dobře je, ale vidět skutečnost, že vektory báze $\mathbf{B} = \{\vec{b}_1, \vec{b}_2\}$ nemají směry, které by maximalizovaly směrodatné odchylky

v ostatních dodaných datech (viz PCA) a z toho důvodu nebudou minimalizovány chyby aproximací vektorů příslušných ostatním roztokům.

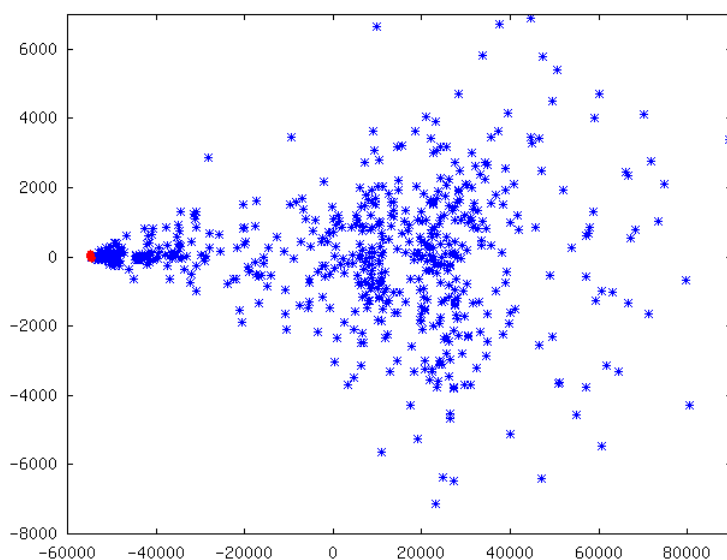


Obrázek 21: Modifikace PCA

--- konec příkladu ---

To, že uvedený příklad představující chybu cenomanské vody není fikcí odtrženou od reality, je doloženo několika následujícími obrázky. Tyto grafy zachycují data z analýz roztoků vyjádřená v podobě 2D aproximací původních vektorů se šesti složkami.

Barvy jednotlivých bodů v grafu signalizují, jestli chyba aproximace překročila 25%. Přičemž relativní chyba aproximace analýzy j -tého roztoku byla počítána podle vzorce:

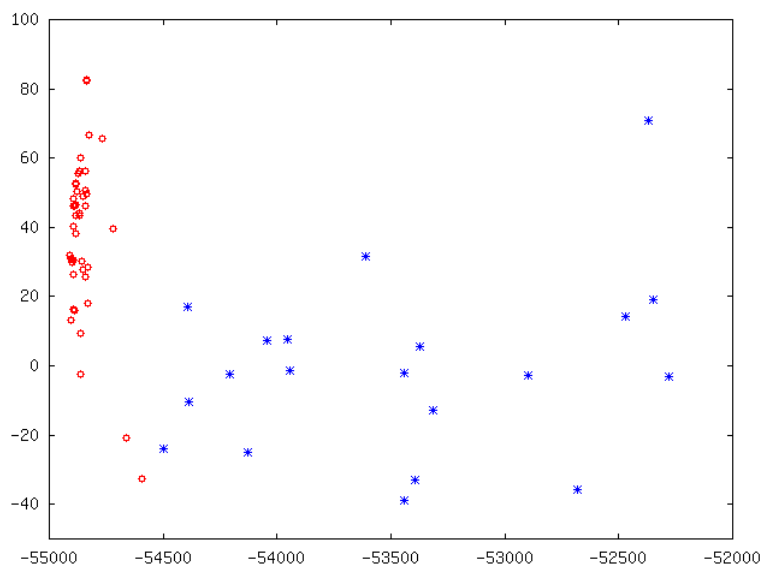


Obrázek 22: Koncentrování chyb v grafu

$$chybaR_j = \frac{\sqrt{\sum_{i=1}^n ({}_j m_i - {}_z m_i)^2}}{\sqrt{\sum_{i=1}^n {}_j m_i^2}}, \text{ kde } {}_z m_i = {}_2 B * \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}, \quad {}_2 B = \begin{bmatrix} b_{11} & b_{21} \\ b_{12} & b_{22} \\ b_{13} & b_{23} \\ b_{14} & b_{24} \\ b_{15} & b_{25} \\ b_{16} & b_{26} \end{bmatrix} \quad (14)$$

Čitatel zlomku pro výpočet relativní chyby představuje normu rozdílu j -té analýzy a její aproximace a jmenovatel je normou vektoru provedené analýzy roztoku.

Detail na oblast (ne prostorová oblast, ale oblast v grafu) roztoků blízkých vodě ukazuje Obrázek 23. Je na něm vidět, že roztoky aproximované s chybou vyšší než stanovená hranice se opravdu koncentrují na jednom místě a tato oblast se zdá být disjunktní s oblastí ve které leží roztoky s menší chybou aproximace.

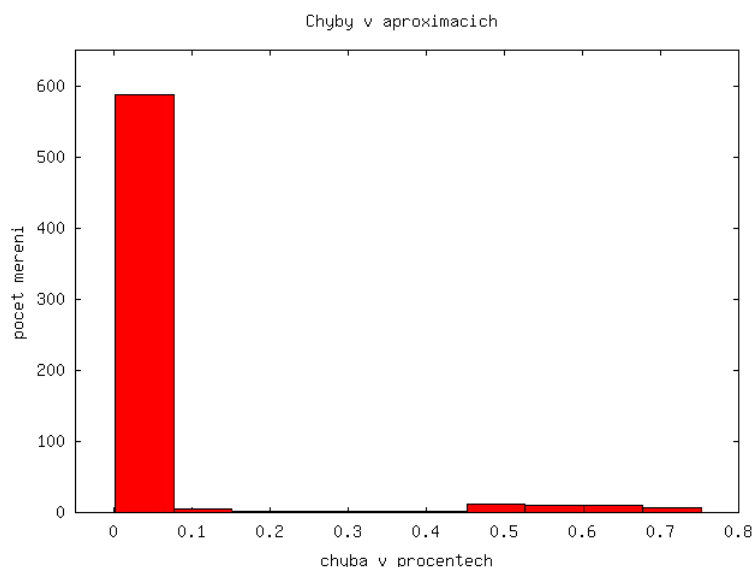


Obrázek 23: Rozhraní dané chybou 25%

Důvod pro sledování relativní chyby aproximace pramení z faktu, že absolutní chyby by vzhledem k velkému rozsahu norem vektorů koncentrací analyzovaných roztoků nebylo možné vůbec srovnávat. Srovnání by nebylo proveditelné z toho důvodu, že zanedbatelná procentuální chyba aproximace koncentrovaného roztoku může být v absolutních číslech větší než norma vektoru představujícího roztok blízký cenomanské vodě.

Volba meze relativní chyby (25%) pro barevné rozlišení bodů v grafu byla učiněna

na základě posouzení histogramu četnosti výskytů relativních chyb aproximací zkoumané sady dat.



Obrázek 24: Četnost výskytů relativních chyb

Po vyzkoušení navrženého postupu s odečtením vody a s redukcí souboru měření se povedlo do příslušných bází všech dimenzí rozložit všechna zbylá data. Rozměr bázových vektorů jsme na doporučení ing. Wasserbauera zmenšili ze šesti na pět. Jinými slovy, odebrali jsme z vektorů chemických analýz složku s informací o koncentraci rozpuštěného fluoru. V případě matice s rozměrem 638x5 (jednoletá data) došlo vlivem odstranění vody k redukcí rozměru na 570x5 zatímco data v matici o velikosti 2515x5 (pětiletá data) se podařilo rozložit všechna. Pětiletá data se podařilo rozložit všechna z toho důvodu, že jsme z jejich souboru již dříve vody odstranili. Původní počet všech analýz provedených v průběhu pěti let byl 3055.

7.1 Srovnání chyb obou představených přístupů

U námi navržené metody aproximace dat (s odečtením průměrných koncentrací, případně se škálováním) jsme posoudili střední chyby aproximací jednotlivých chemických složek roztoků. Tyto chyby byly vyjádřeny v procentech způsobem, který ukazuje vzorec (15).

$$k_j[\%] = \frac{1}{\bar{x}_j} \cdot \sqrt{\frac{\sum_{i=1}^n (x_{ij} - x_{ij}^{apr})^2}{n}} \quad (15)$$

Kde j značí složku, jejíž chybu posuzujeme, a i probíhá přes všechna měření, která k posouzení takové chyby využíváme. Označení $\bar{x}_j$ má význam střední hodnoty počítané z koncentrací j -té složky.

Ing. Wasserbauer nám dal k dispozici jím vypočtené hodnoty, které získal zkoumáním souboru pětiletých dat. Následující tabulka obsahuje pro porovnání jak hodnoty vypočtené ing. Wasserbauerem, tak chyby určené na stejné sadě dat námi.

	RL	SO4-a	NH4	Al	Fe-c
Err_Wass	5,9%	7,5%	13,8%	8,2%	3,7%
Err_Zed	0,04%	0,04%	21,05%	7,20%	34,28%

Tabulka 12: Procentuální chyby apromace

Je vidět, že v nejméně zastoupených složkách jsme dosáhli několikařádově lepší přesnosti, ale stalo se tak na úkor přesnosti aproximace složek, které jsou v roztoku přítomny v marginálních množstvích. Citelné zhoršení je patrné u železa (poslední sloupec). Tento negativní výsledek nás dovedl ke dvěma novým postupům.

První se týkal přeškálování původních dat takovým způsobem, aby se informace o marginálně zastoupených chemikáliích spolu s redukcí neztrácely. Pokud totiž vynásobíme vybrané složky vektorů správnými čísly, projeví se to zřetelnou změnou směru vektorů báze $B = \{\vec{b}_1, \dots, \vec{b}_n\}$ a to s sebou nese právě zpřesnění aproximace vybraných složek.

Škálování jsme prakticky realizovali násobením matice s daty škálovací maticí se speciálně zvolenými diagonálními prvky. Diagonála škálovací matice obsahuje převrácené hodnoty velikostí středních hodnot těch složek vektorů dodaných dat, které mají být příslušným číslem přenásobeny. Význam naší volby velikosti diagonálních prvků spočívá v jistém „zrovnoprávnění“ koncentrací všech proměřovaných složek roztoků. Po přeškálování je střední hodnota koncentrace každé složky rovna jedné. Když jsme data přenásobili škálovací maticí

$$skala = \begin{bmatrix} 2.13e-005 & 0.00e+000 & 0.00e+000 & 0.00e+000 & 0.00e+000 \\ 0.00e+000 & 3.51e-005 & 0.00e+000 & 0.00e+000 & 0.00e+000 \\ 0.00e+000 & 0.00e+000 & 1.49e-003 & 0.00e+000 & 0.00e+000 \\ 0.00e+000 & 0.00e+000 & 0.00e+000 & 2.41e-004 & 0.00e+000 \\ 0.00e+000 & 0.00e+000 & 0.00e+000 & 0.00e+000 & 9.57e-004 \end{bmatrix}$$

bázové vektory se změnily následujícím způsobem.

B1	B2	B3	B4	B5
0,85	-0,52	0,01	-0,06	0,02
0,52	0,86	0,01	-0,04	0
0,01	-0,01	0,07	-0,04	-1
0,07	0	-0,41	0,91	-0,06
0,02	0	0,91	0,41	0,04
B1sk	B2sk	B3sk	B4sk	B5sk
0,44	0,21	-0,24	-0,01	-0,84
0,44	0,21	-0,3	-0,73	0,38
0,46	0,23	0,85	0,03	0,05
0,43	0,26	-0,35	0,69	0,39
0,45	-0,89	-0,01	0,03	0,02

Tabulka 13: Vliv škálování na směr bázových vektorů

Pokud by se na diagonále škálovací matice nacházelo pouze jedno číslo různé od jedné a toto by nabývalo výrazně řádově vyšší hodnoty (testováno na 10^8 až 10^{12} pro železo, tj. pátá složka), potom by se to projevilo tím, že z vektoru B1 by se stal vektor jednotkový s jedničkou na místě složky škálované velkým číslem.

Druhý nově použitý postup se týkal posouzení míry uchování nesených informací o jednotlivých chemických složkách v aproximativních vektorech. Namísto doposud k tomuto používané procentuální chyby jsme tentokrát využili koeficientů korelace mezi dodanými daty a jejich aproximacemi.

Nejprve jsme vypočetli matici korelačních koeficientů jednotlivých dvojic složek dat s neredukovanou dimenzí (dodaných dat). Každý jeden z prvků této matice vypovídá o tom, nakolik hodnota i -té složky vektoru, který představuje analýzu konkrétního roztoku, závisí na hodnotě jeho j -té složky. Tuto závislost je možné vnímat tak, že s odebráním jedné ze dvou silně korelovaných složek (korelační koeficient blízký jedné) se ztratí minimum informace.

1,00	0,99	0,95	0,99	0,87
0,99	1,00	0,94	0,99	0,87
0,95	0,94	1,00	0,94	0,84
0,99	0,99	0,94	1,00	0,85
0,87	0,87	0,84	0,85	1,00

Tabulka 14: Korelace složek

Pozornost si zaslouží poslední sloupec, který ukazuje Tabulka 14. Z důvodu symetrie uvedené matice (tabulky), nese tutéž informaci také poslední řádek. Čísla v posledním řádku se vyznačují tím, že v porovnání se všemi ostatními jsou zřetelně

menší. Jestliže jsou korelační koeficienty jedné ze složek výrazně menší než pro všechny složky ostatní, projeví se to výrazně vyšší chybou aproximace, provedené za pomoci redukce dimenze (při použití PCA). Toto dokládá Tabulka 12. Bez úpravy dat škálováním, byly s největší chybou opravdu aproximovány nejméně s ostatními korelované složky (železo, amonné kationty).

Příklad:

Představme si zaznamenané koncentrace dvou chemických látek. Tyto koncentrace necht' jsou zobrazeny v grafu. Body v grafu necht' je možné s přijatelnou přesností nahradit přímkou. Složky souřadnic takto „nahraditelných“ bodů jsou silně korelované (koeficient korelace je blízký jedné) a pokud budeme schopni nalézt předpis přímky, kterou tyto body obklopují, potom nám postačí uchovat informaci pouze o jedné složce (je jedno které).

--- konec příkladu ---

V případě posuzování korelace mezi složkami vektorů a příslušnými složkami jejich aproximací nám koeficienty říkají, jak výrazné je sepejetí mezi skutečnou a aproximativní hodnotou.

Vztah původních, neškálovaných dat s jejich aproximacemi je zaznamenán v následující tabulce.

Vše	SO4-a	NH4	Al	Fe
1,0000	1,0000	0,9504	0,9932	0,8686

Tabulka 15: Korelace neškálovaných dat

Za povšimnutí stojí v porovnání s ostatními malá míra korelace koncentrace a aproximace koncentrace páté složky (železa).

Přeškálování dat výše uvedenou maticí změní tabulku s korelačními koeficienty následujícím způsobem.

Vše	SO4-a	NH4	Al	Fe
0,9950	0,9931	0,9720	0,9916	0,9999

Tabulka 16: Korelace škálovaných dat

Je naprosto zřejmé, že přeškálování matice s daty výše uvedeným způsobem vedlo zejména ke zdůraznění „důležitosti“ informace o rozpuštěném železe a mírně také o amonném kationtu. Pohledem na ostatní sloupce tabulky zjistíme, že zbylé složky

vektorů a jejich aproximací škálování nijak výrazně nedekorelovalo.

Pro úplnost zbývalo provést posouzení středních kvadratických chyb aproximací přeškálovaných dat. Tyto byly vypočteny s použitím vzorce (15). Vypočtené hodnoty ukazuje Tabulka 17.

	RL	SO ₄ -a	NH ₄	AL	FEC
Err_Wass	5,9%	7,5%	13,8%	8,2%	3,7%
Err_Zed	6,0%	7,4%	15,8%	8,1%	0,3%

Tabulka 17: Procentuální chyby škálovaných dat

Oproti Tabulce 12 došlo ke snížení chyb aproximací koncentrací amonných kationtů a železa. Naopak relativní chyba síranů se o dva řády zhoršila. Vzhledem k tomu, že hodnoty koncentrací síranů se pohybují v řádu 10^4 mg/l, není možno považovat tuto relativně (procentuálně) stále malou chybu za zanedbatelnou.

Konkrétními dopady námi provedeného škálování se ovšem není možné nikterak vážně zabývat, protože škálovací matice byla zkonstruována pouze na základě úsudku a nikoliv s využitím promyšlené metodiky.

8. Závěr

Teoretická část diplomové práce nabízí seznámení s problematikou počítačové interpretace dat získaných chemickými analýzami koncentrací jednotlivých složek vodných roztoků. Představené metody se liší svou přesností a délkou trvání výpočtů. Při zpracovávání dat je tak nutné brát do úvahy jejich množství, čas, který je na výpočet k dispozici a akceptovatelnou chybu aproximace.

Při vyhotovování praktické části této diplomové práce byl identifikován postup, který umožňuje nalézt v relativně krátkém čase bázi vektorového podprostoru stanovené dimenze. Oproti postupu použitému k témuž ve firmě DIAMO s.p. mají zde navržené metody několik výhod. První z výhod plyne z použití metod vektorové algebry, díky kterému není navržený postup natolik úzce svázán s chemickým významem poskytnutých dat. Tato volnější vazba přímo vybízí k využití v práci s naměřenými daty získanými i v jiném oboru lidské činnosti než v chemii. Druhou výhodou je skutečnost, že lze nalézt báze pro různou zamýšlenou míru redukce (obecně od dimenze 2 do $n-1$). Na základě rozkladů dat do těchto bází lze posoudit jak velká redukce dimenze je s ohledem na kombinaci zachování maximální míry informace a dosažení co nejjednodušší práce s vektory redukované dimenze nejvhodnější.

Otázka, která nebyla v práci do detailu dořešena, se týká chyb aproximací koncentrací jednotlivých složek aproximovaných roztoků.

Literatura

- [1] Lukáš Zedek: Vyhledání a aplikace algoritmů informatiky vhodných pro snížení dimenze problému modelování transportu mnohasložkových vodných roztoků, 2007, ročníková práce, NTI, FM TUL
- [2] Lukáš Zedek: Identifikace postupů ing. Wasserbauera, 2007, projekt, NTI, FM TUL
- [3] www.wikipedia.org
- [4] Chemické analýzy cenomanských roztoků v lokalitě Stráž p. Ralskem, DIAMO,s.p., osobní sdělení