Univerzita Pardubice
Chemicko-technologická fakulta
Katedra analytické chemie

12. licenční studium PYTHAGORAS
Statistické zpracování dat

3.1 Matematické principy vícerozměrných metod statistické analýzy

Semestrální práce
2010

RNDr. Markéta Vaňková, Ph.D.
Endokrinologický ústav
Národní 8, 116 94 Praha 1

Otázka 1

Najděte vlastní (charakteristická) čísla a vlastní vektory, determinant, stopu a odmocninu od této matice

1 2 3
2 8 2
3 2 10

Řešení:

matematický software Matlab

Odmocnina matice A

A=[1 2 3; 2 8 2; 3 2 10]

>> sqrtm(A)

ans =

0.3214 + 0.3439i 0.5699 - 0.0625i 0.8379 - 0.0894i
0.5699 - 0.0625i 2.7586 + 0.0114i 0.2636 + 0.0162i
0.8379 - 0.0894i 0.2636 + 0.0162i 3.0393 + 0.0232i

Toto je řešení komplexní matice velikosti 3x3.

odmocnina je A^(1/2)= suma_{k=1}^3 sqrt(lambda_k) U_kU_k^T

kde lambda_k, k=1,...3 jsou vlastní čísla
U_k, k=1,...3 jsou odpovídající normované vlastní vektory (sloupce matice U níže) "^T" je operátor transpozice

Odmocnina bude opět symetrická matice, zde ovšem Hermitovsky symetrická.
Bude splňovat vlastnost A^(1/2)A^(1/2)=A.

Stopa matice se většinou značí tr(A) a determinant jako det(A).
tr(A)=19, det(A)=-12

Stopa je součct prvků na diagonále tj. 19, determinant je -12,

Vlastní čísla jsou řešením kubické rovnice
jejich numerické řešení je

-0.1433
6.7657
12.3776

Vlastní vektory lze též snadno dopočítat.

```
A=[1 2 3; 2 8 2; 3 2 10]

[E L]=eig(A)

E =
   -0.9532   -0.0143    0.3020
    0.1733   -0.8444    0.5069
    0.2478    0.5355    0.8074


L =
   -0.1433         0         0
         0    6.7657         0
         0         0   12.3776
```

E obsahuje vlastní vektory a L vlastní čísla na diagonále.

Otázka 2

Pro typická data z pracoviště, minimálně 4 rozměrná, určete projekci do prvních dvou komponent, dvojný graf a diskutujte jeho význam.

Data

Soubor 332 žen, sledované parametry:

věk

Rohrer index – poměr porodní hmotnosti a délky

BMI – body mass index; poměr tělesné hmotnosti v kg na tělesnou výšku v m^2

poměr sval/tuk – poměr hmotnosti svaloviny v kg na hmotnost podkožního tuku v kg

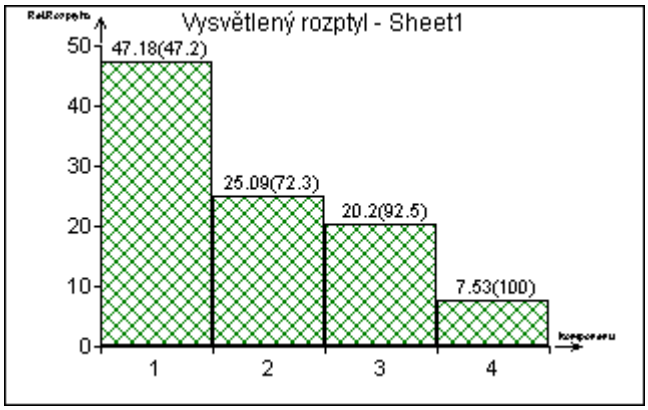
soubor v excelu: uloha_3.1b.xls

program: QCExpert 3.1

Původní data neměla normální rozdělení, byla proto transformována. Byl využit postup založený na kombinaci jedno- a vícerozměrné průzkumové analýzy dat a parametrické transformace s použitím funkcí „hledání řešení“ a „řešitel“ v tabulkovém procesoru Excel, dále došlo k identifikaci vícerozměrných outliers a jejich následnému vyloučení (autor automatizovaného postupu ing. M. Hill).

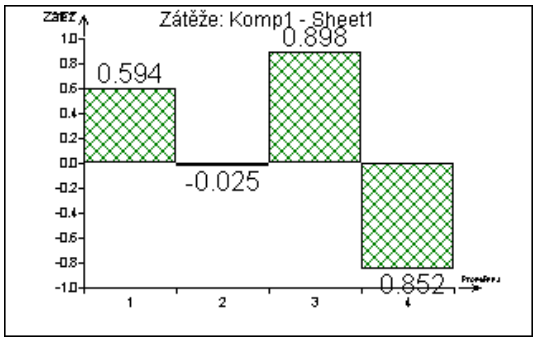
Grafický výstup

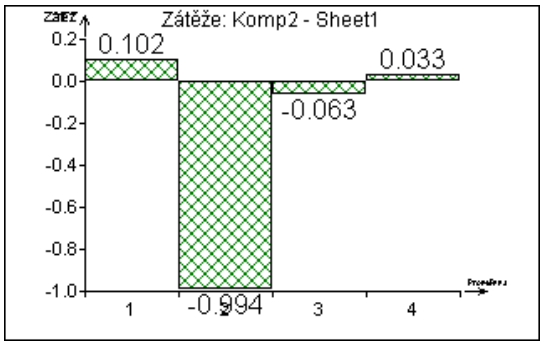
Graf vysvětleného rozptylu, relativní variabilita vysvětlená jednotlivými komponentami. Komponenty jsou vždy seřazeny tak, že první komponenta vysvětluje největší díl variability, poslední komponenta nejmenší díl. Na ose x jsou pořadová čísla komponenty, na ose y jsou procenta rozptylu vztažená na celkový rozptyl. Čísla nad sloupci udávají rozptyl a kumulativní rozptyl v procentech.

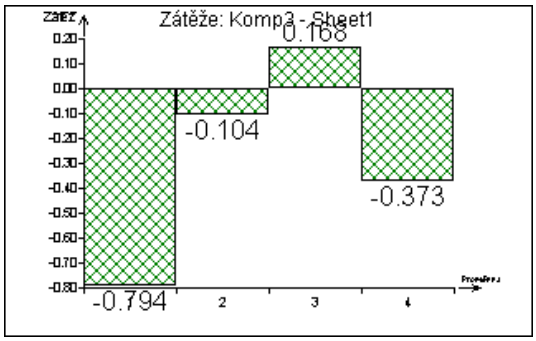


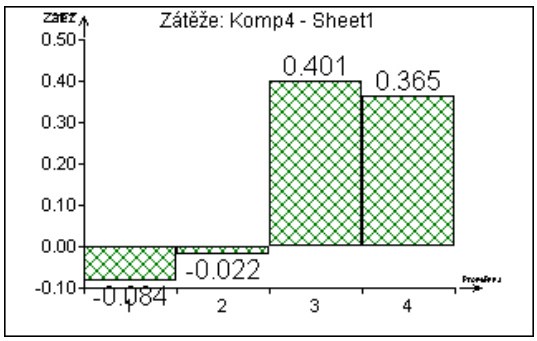
První dvě komponenty vysvětlují 72,3 % variability rozptylu.

Grafické vyjádření zátěží pro jednotlivé komponenty. Zátěže udávají strukturu jednotlivých komponent. Někdy vystihují komponenty různé rysy dat určené jednou nebo několika proměnnými. Tyto proměnné pak mají v dané komponentě výrazně vyšší absolutní hodnotu zátěže.









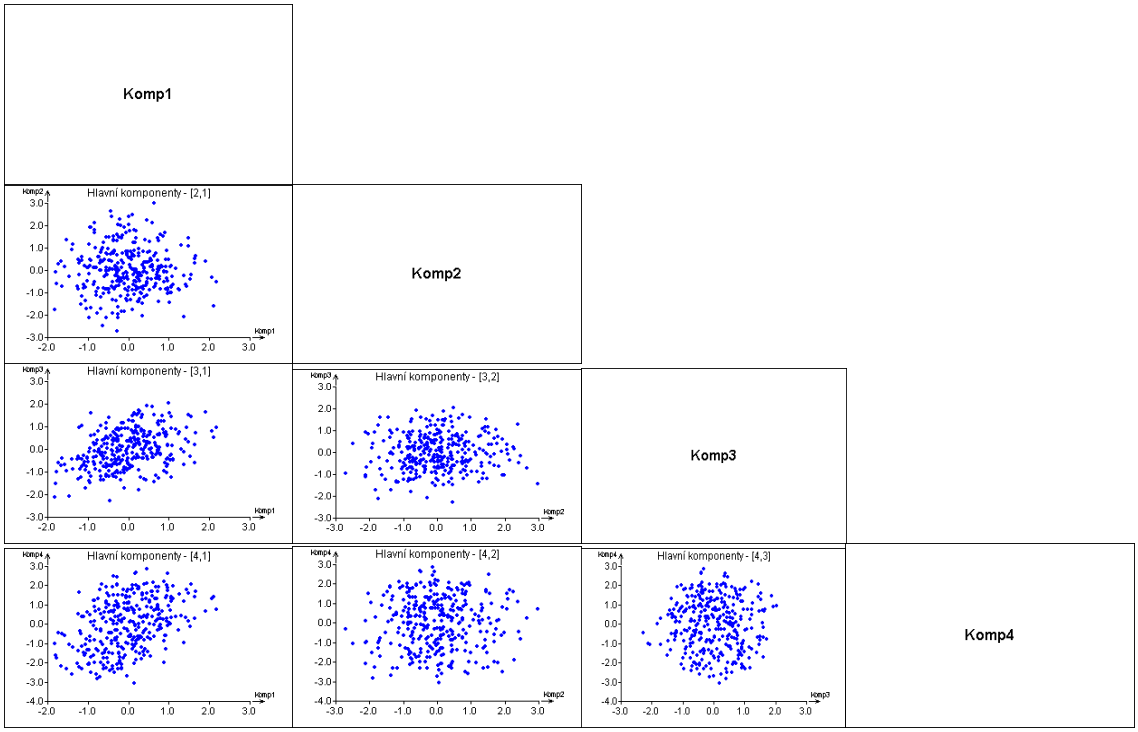
V první komponentě mají nejvyšší absolutní hodnotu zátěže parametry BMI a poměr sval/tuk, nejmenší pak Rohrerův index.

Ve druhé komponentě má parametr Rohrerův index obrovskou převahu nad zbylými parametry.

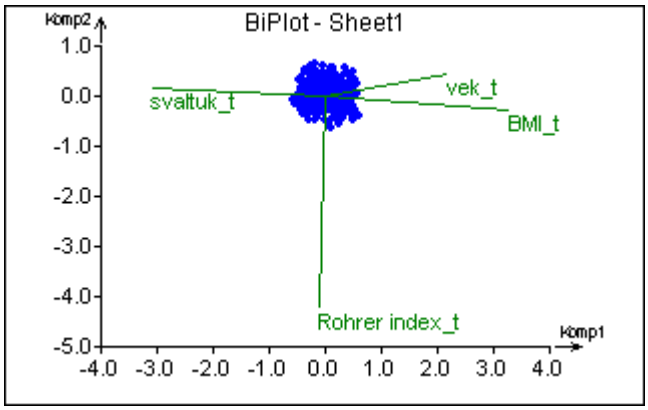
Ve třetí komponentě má nejvyšší absolutní hodnotu parametr věk.

Ve čtvrté komponentě mají nejvyšší absolutní hodnotu parametry BMI a poměr sval/tuk.

Grafy hlavních komponent jsou rozptylové grafy pro všechny kombinace komponent. Mohou sloužit k posouzení homogenity dat.



Biplot je projekce vícerozměrných dat do plochy. Body reprezentují řádky, paprsky odpovídají sloupcům. Při interpretaci grafu se vychází z toho, že aproximace původních dat úměrná vektorovému součinu jednotlivých bodů a úseček (bod reprezentuje konec vektoru s počátkem v bodě (0,0)). Blízké vektory řádků (body) nebo sloupců (paprsky) budou zřejmě vzájemně korelované. Vektory řádků, ležící ve směru některého vektoru sloupce budou mít v tomto sloupci vyšší, resp. nižší hodnoty. Znaménko (smysl vektoru) při tom nehraje roli.



Největší korelace je mezi BMI a poměrem sval/tuk, což se v grafu promítá tím, že oba parametry leží na společné přímce a každá z nich směřuje na opačnou stranu. Další významné korelace jsou mezi věkem a BMI a věkem a poměrem sval/tuk. Rohrerův index se zdá být nezávislým parametrem.

Závěr

První dvě komponenty popisují 72,3 % variability, spolu se třetí komponentou pak 92,5 % variability. Do první komponenty nejvíce prispívají antropometrické parametry jako je BMI a poměr sval/tuk a také věk, který významně koreluje s těmito parametry. Do druhé komponenty nejvíce přispívá Rohrerův index, který vypovídá o porodní velikosti. Porodní velikost (malá nebo naopak velká) je často dávána do souvislosti s nejvyšší dosaženou hmotností v dospělosti a dále také s rizikem rozvoje některých metabolických onemocnění jako je např. diabetes 2 typu.