

LICENČNÍ STUDIUM STATISTICKÉ ZPRACOVÁNÍ DAT PŘI MANAGEMENTU JAKOSTI

PŘEDMĚT:

3.2 Metody s latentními proměnnými a klasifikační metody

Vypracoval: Ing. Adolf Goebel, Ph.D.

V Přerově, 17. listopadu 2000

3.2.1 Otázka 1

Vypočtete algoritmem NIPALS první latentní proměnnou z matice A:

$$\begin{bmatrix} 1 & 1 & 6 \\ 2 & 2 & 4 \\ 3 & 0 & 2 \end{bmatrix}$$

Matici před zpracováním standardizujte.

Prvním krokem je výpočet vektoru aritmetických průměrů $\bar{x}^T$ a vektoru směrodatných odchylek s^T .

$$\bar{x}^T = [(1+2+3)/3 \quad (1+2+0)/3 \quad (6+4+2)/3] = [2 \quad 1 \quad 4]$$

$$s^T = [s_1 \quad s_2 \quad s_3] = [1 \quad 1 \quad 2]$$

Provedením standardizace se získá matice A:

$$\begin{bmatrix} -1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & -1 & -1 \end{bmatrix}$$

Variabilita všech sloupců je stejná, proto je možné vzít za odhad hlavní komponenty libovolný sloupec, např. první:

$$t_1^T = [-1 \quad 0 \quad 1]$$

Po dosazení do vzorce

$$p_1^T = (t_1^T t_1)^{-1} t_1^T A$$

a znormování podle vztahu

$$p_1^N = (p_1^T p_1)^{-1/2} p_1$$

se získá počáteční odhad vektoru vyjadřujícího míru příspěvků odhadu hlavní komponenty t_1 .
Následujícím dosazením do vztahu:

$$t_1^T = (p_1^T p_1)^{-1} p_1^T A^T$$

se získá odhad hlavní komponenty t_1 . Opakováním postupu se získá stabilní rozklad vektorů t_1 a p_1 . Jako konvergenční kritérium se použije vztah:

$$d = (t_{\text{nové}} - t_{\text{staré}})^T (t_{\text{nové}} - t_{\text{staré}}) (t_{\text{nové}}^T t_{\text{nové}})^{-1}$$

Aby se mohl ukončit výpočet, je potřeba dosáhnout $d/N < 10^{-10}$.

K tomu, aby se toho dosáhlo, je třeba devíti kroků. Získá se následující stabilní rozklad vektorů:

$$\begin{aligned} p_1^T &= [0,62797 & -0,45969 & -0,62797] \\ t_1^T &= [-1,25594 & -0,45969 & 1,71562] \end{aligned}$$

3.2.2 Otázka 2

S použitím vhodných kritérií určete nezbytný počet latentních proměnných, bylo-li z dat určeno: PRESS(0) = S(0) = 100; PRESS(1) = 20; S(1) = 10; PRESS(2) = 3,5; S(2) = 3,4; PRESS(3) = 3,45; S(3) = 3,39.

Nezbytný počet latentních proměnných se určí s pomocí Woldova kritéria. To je založeno na poměru:

$$\text{PRESS}(P) / S_R(P - 1)$$

Je-li hodnota kritéria větší než 0,95, je zařazení další (P+1) latentní proměnné nevhodné.

Po dosazení dostaneme:

$$\begin{aligned} P = 1: & \quad \text{PRESS}(1) / S_R(0) = 20 / 100 = 0,2 \\ P = 2: & \quad \text{PRESS}(2) / S_R(1) = 3,5 / 10 = 0,35 \\ P = 3: & \quad \text{PRESS}(3) / S_R(2) = 3,45 / 3,4 = 1,01 \end{aligned}$$

Pro $P = 3$ je již hodnota vyšší než 0,95, čtvrtá latentní proměnná již není významná.

3.2.3 Otázka 3

Odhadněte hodnotu chybějícího prvku A [2,2], jestliže výpočtem z nekompletní matice byly určeny vektory p: 0,541 0,423 0,514 0,514 t: -1,340 -0,735 2,076

Prvky zdrojové matice odpovídající p-té latentní proměnné je možné získat ze vzorce:

$$A_p^{\text{pred}} = t_p p_p^T$$

Pro odhad hodnoty prvku A[2,2] takto platí:

$$A[2,2] = t[2] p[2] = -0,311$$

3.2.4 Otázka 4

Výpočtem metodou PCA byly určeny vektory:

$$p_1^T: \quad 0,012 \quad 0,458 \quad -0,352 \quad 0,987$$

$$p_2^T: \quad 0,926 \quad -0,238 \quad 0,872 \quad -0,115$$

Vyberte reprezentativní sloupce, které nejlépe charakterizují celou matici.

Míra příspěvku příslušné latentní proměnné k popisu variability sloupců zdrojové matice je představována vektory. Vektor p_1^T ukazuje, že pro zdrojovou matici je reprezentativním sloupцем sloupec čtvrtý $p_1^T [4] = 0,987$. V případě vektoru p_2^T je reprezentativním sloupцем sloupec první $p_2^T [1] = 0,926$ a sloupec třetí $p_2^T [3] = 0,872$.

3.2.5 Otázka 5

Vysvětlete, proč vysvětlená variabilita je při výpočtu metodou FA vždy nižší, než při výpočtu metodou PCA.

Metoda PCA je z hlediska faktorové analýzy považována za úplnou komponentní analýzu. Pomocí hlavních komponent je možné přesně reprodukovat variabilitu zdrojové matice. Metoda FA je z tohoto pohledu neúplnou komponentní analýzou, protože připouští existenci matice jedinečnosti, která zahrnuje část variability, která se nedá vysvětlit společnými faktory.

3.2.6 Otázka 6

Výpočtem metodou kanonických korelací bylo zjištěno:

$$0,297 X_1 + 0,298 X_2 + 0,050 X_3 + 0,256 X_4 = 0,493 Y_1 - 0,213 Y_2 \quad r_1 = 0,830$$

$$0,006 X_1 - 0,115 X_2 + 0,950 X_3 + 0,056 X_4 = 0,493 Y_1 + 0,213 Y_2 \quad r_2 = 0,512$$

Vypočtete skupinový korelační koeficient a interpretujte výsledky.

Skupinový korelační koeficient se vypočte ze vztahu:

$$R = [1 - (1 - r_1^2) (1 - r_2^2)]^{1/2}$$

$$R = [1 - (1 - 0,830^2) (1 - 0,512^2)]^{1/2} = 0,878$$

$$R^2 = 0,770, \text{ což znamená, že výsledek popisuje } 77 \% \text{ variability dat}$$

Parametr X_3 má vzhledem k ostatním parametrům zanedbatelný vliv na součet parametrů Y_1 a Y_2 , ostatní parametry mají vliv zhruba stejný. Na rozdíl parametrů Y_1 a Y_2 má parametr X_3 výrazný vliv, zatímco vliv ostatních parametrů je nevýrazný.

3.2.7 Otázka 7

Uveďte nějaký konkrétní případ vhodný pro zpracování metodou PLS.

Metoda projekce latentních struktur PLS (partial least square) je zobecněným postupem pro získání popisu vztahu mezi závislým náhodným vektorem y a vysvětlujícím náhodným vektorem x prostřednictvím latentních proměnných. Tímto postupem je možné opakováním kroků až do konvergence získat vektor latentních proměnných a vektor zátěží a nakonec i z residuální matice i matici latentních proměnných a zátěží. Možným příkladem z praxe, na který by se metoda dala aplikovat je např. následující tabulka, která se popisuje úpravu povrchu titanové běloby organickými povrchovými úpravami a vyhodnocení výsledného stavu pomocí smáčivosti různě polárními sloučeninami a měřením kontaktního úhlu:

Činidlo	spotřeba oleje	spotřeba vody	spotřeba DOP	kontaktní úhel
polyethylenglykol 200	*****	*****	*****	*****
polyethylenglykol 400	*****	*****	*****	*****
polyethylenglykol 600	*****	*****	*****	*****
trimethylolpropan	*****	*****	*****	*****
polyol TP 15	*****	*****	*****	*****
polyol TP 32	*****	*****	*****	*****
α - olefiny Neraten	*****	*****	*****	*****

3.2.8 Otázka 8

Jeden objekt je charakterizován metrickými znaky (2,10), druhý (3,8), třetí (4,9), čtvrtý (10,4) a pátý (11,5). Vypočítejte matici vzdálenosti v Euklidově metrice a dokumentujte výpočet shlukování některou z používaných metod. Výsledky interpretujte graficky.

Matice vzdálenosti znaků má tvar:

```

1      0,00
2      2,24  0,00
3      2,24  1,41  0,00
4      10,00 9,75  7,81  0,00
5      10,30 8,54  8,06  1,41  0,00

```

Nejmenší vzdálenost mají prvky 3-2 a 5-4. Z nich je možné vytvořit první shluky, spočítat těžiště nových shluků a spočítat těžiště nových shluků a opět vypočítat matici vzdáleností:

```

1      0,00
2-3    2,12  0,00
4-5    10,12 8,06  0,00

```

Nejmenší vzdálenost má shluk 2-3 a znak 1. Je tedy možné tyto prvky spojit do shluku a spočítat těžiště nového shluku a opět vypočítat matici vzdáleností:

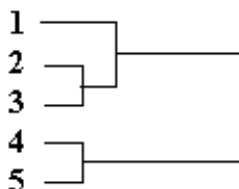
```

1-2-3  0,00

```

4-5 8,75 0,00

Tento proces je možné shrnout do následujícího dendogramu:



3.2.9 Otázka 9

Organické barvivo je vyráběno dvoustupňovou syntézou, v prvním stupni reakcí dvou komponent o standardním složení, v druhém stupni se k produktu ze stupně prvního přidává další komponenta rovněž o standardním složení. Poměr komponent má vliv na výtěžek a kvalitu výsledného produktu. Kontrola výtěžku kvality meziprojektu je prováděna dvěma metodami, u produktu je vyhodnocován celkový výtěžek na výchozí suroviny a prováděno 8 různých finančně a časově náročných zkoušek kvality. K dispozici jsou údaje z výroby za poslední dva roky, jedna šarže barviva se vyrobí za jednu směnu v třísměnném provozu. Navrhněte možné chemometrické postupy, na jejichž základě by bylo možno výrobu barviva racionalizovat.

Z údajů o četnosti výroby je zřejmé, že by mělo být k dispozici velké množství dat. Prvním krokem by mělo být zjištění, jaké veličiny se sledují a archivují. V uvedeném popisu chybí například jakákoli zmínka o teplotě a individuální době trvání operací. Ukončují se operace na základě uplynutí příslušné doby trvání nebo na základě testu? Pokud na základě testů, pak je nutno vědět, jak dlouho trvá testování a zda reakce až do zjištění výsledku neprobíhá významnou rychlostí. Může se totiž snadno stát, že výsledky testů toho moc neříkají o stavu ve chvíli reálného ukončení procesu. Bylo by vhodné několikrát laboratorními testy ověřit i konečné složení ukončené várky. Pomocí testů párových hodnot výsledků před ukončením reakce (kontrola meziprojektu) a po ukončení reakce (testy vyhovují požadavkům na várku a následuje další zpracování). Jsou-li rozdíly významné, pak by se do dalšího sledování mělo zařadit i zaznamenání času odběru a doby ukončení reakce.

V případě, že se várka zpracovává okamžitě, nemají testy smysl a mělo by se uvažovat spíše o archivaci meziprojektu pro případné vysvětlení anomálií kvality hotového produktu. Reakci ve vzorku by bylo třeba zastavit, což by u organických reakcí mohlo být realizovatelné silným ochlazením. Obě metody analýzy by se měly vyšetřit z hlediska jejich korelace a možnosti jednu z metod vypustit, aspoň pro rutinní analýzu (pokud se neukáže vliv v rámci latentní proměnné na výslednou kvalitu). Pokud nejsou záznamy pořízeny automaticky, stálo by zato alespoň orientačně prověřit kvalitu dat (autokorelace pro jednotlivé obsluhy, laborantky atd.). Obecně by se mělo provést obvyklé vyšetření kvality dat.

Existující data je možné zpracovat vícerozměrnou statistickou analýzou. Použít se mohou metody s latentními proměnnými. Všechna dostupná data se sestaví do zdrojové matice. Matice se zřejmě bude muset centrovat a standardizovat. Použije se vhodný počítačový software a metoda, např. metoda hlavních komponent. Cílem bude získat matici latentních proměnných

(score) a matici zátěží (loadings). Pokud chybí nějaká hodnota ve zdrojové matici, je možné se pokusit o rekonstrukci. Ze získaného výsledku metody se může posoudit, který nejmenší počet proměnných popisuje maximum variability zdrojové matice bez zahrnutí experimentální chyby. K získání dalších poznatků je možné pro latentní proměnné vypočítat modelovací síly a komunality (bývají obvykle standardním výstupem komerčních programů) a posoudit, jak který sloupec přispívá k vytvoření které latentní proměnné. Pro interpretaci výsledků z hlediska chemického a fyzikálního významu je nutno provést např. rotaci pro matici zátěží a např. cílené testování pro matici latentních proměnných. Další možností interpretace vektorů matice latentních proměnných a matice zátěží může být regresní a korelační analýza vztahu těchto vektorů s veličinami s jednoznačným fyzikálně chemickým významem.

Pro hodnocení se může rovněž použít metody faktorové analýzy s výpočtem matic faktorových zátěží jedinečnosti a komunalit. Metoda může opět pomoci redukovat počet proměnných a odhadnout nepřímo měřitelné veličiny. Pokud ve zdrojové matici chybí hodně údajů, měla by se použít metoda konjugovaných odchylek, která je na to méně citlivá a umožňuje odhad latentních proměnných ve statisticky validním modelu. Dobrou použitelnou metodou k objasnění vztahů a významů prvků je rovněž kanonická analýza nebo projekce latentních struktur.

Takový chemometrický rozbor bude nepochybně schopen odpovědět na otázky:

- S které analýzy jsou málo vypovídající nebo zbytečné?*
- S jaký výsledek analýz je potřebný pro potvrzení budoucích nejlepších vlastností maziprodktu i výsledného produktu?*
- S jaký poměr komponent a podmínky v prvním i druhém stupni technologie vedou k tomuto žádoucímu stavu?*