INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Geocomputation

Principal component analysis (PCA)

Mgr. Michaela Tučková

Letní škola Geocomputation

29. června 2011

Obsah přednášky

Co je PCA?

- jedna z nejdůležitějších úloh vícerozměrné statistické analýzy, jejíž tvůrcem je Karl Pearson (1901)
- konkrétně se PCA řadí do třídy úloh, které redukují dimensi mnohorozměrných dat ⇒ snahou je získat jednodušší a lépe čitelný soubor tak,aby v něm zůstalo zachováno co největší množství informace z původního datového souboru (tj. redukcí bychom měli ztratit co nejméně informace)
- výhodou této metody je její použitelnost na jekýkoliv typ rozdělení pravděpodobnosti, tj. neklade důraz na rozdělení pravděpodobnosti náhodných proměnných

⇓

snaha o vyjádření proměnlivosti v datech za pomoci menšího počtu veličin **hlavních komponent**

Oblasti využití

Příkladem využití PCA je práce s družicovými snímky, kdy často dochází k vysokemu stupni pozitivni korelace mezi různými pasmy. Pokud je odrazivost nějaké oblasti v jednom pásmu vysoká, pak bude vysoká i v jiném pásmu.

⇓

Tedy obrázek snímaný v sedmi pásmech v sobě nese mnohem méně informace než je sedminásobná informační hodnota jednoho pásma.

⇓

? Otázkou nyní je, zda by nebylo možné využít pro charakteristiku odrazivosti zemského povrchu méně vlnových pásem ?

⇓

Na tuto otázku je možné nalézt odpověď právě pomocí metody PCA.

Kovariance, kovarianční matice

- **Kovariance** je míra měřená vždy mezi dvěmi náhodnými proměnnými X_i a X_j

$$\mathrm{Cov}(X_i, X_j) = \sqrt{\frac{\sum_{i=1}^{l}\sum_{j=1}^{l}(x_i - \bar{x}_i)(x_j - \bar{x}_j)}{l-1}}$$

- Vlastnosti **kovarianční matice** $\mathbf{\Sigma}$

- čtvercová
- symetrická $\rightarrow \mathrm{cov}(X_i, X_j) = \mathrm{cov}(X_j, X_i)$
- positivně semidefinitní $\rightarrow \mathbf{h}'\mathbf{\Sigma}\mathbf{h} \geq 0$ pro jakýkoliv vektor $\mathbf{h} \neq 0$
$\rightarrow$ všechna vlastní čísla matice $\mathbf{\Sigma} \geq 0$

$$\mathbf{\Sigma} = \begin{pmatrix} \mathrm{cov}(X_1,X_1) & \mathrm{cov}(X_1,X_2) & \dots & \mathrm{cov}(X_1,X_n) \\ \mathrm{cov}(X_2,X_1) & \mathrm{cov}(X_2,X_2) & \dots & \mathrm{cov}(X_2,X_n) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{cov}(X_{n-1},X_1) & \mathrm{cov}(X_{n-1},X_2) & \dots & \mathrm{cov}(X_{n-1},X_n) \\ \mathrm{cov}(X_n,X_1) & \mathrm{cov}(X_n,X_2) & \dots & \mathrm{cov}(X_n,X_n) \end{pmatrix}$$

Spektrální rozklad kovarianční matice

Hlavní nástroj metody PCA je právě spektrální rozklad kovarianční matice $\mathbf{\Sigma}$.

$$\begin{aligned} \mathbf{\Sigma} &= \mathbf{P}\mathbf{\Lambda}\mathbf{P}' \quad \text{a platí} \quad \mathbf{P}\mathbf{P}' = \mathbf{I} \\ &= \sum_{j=1}^{n} \Lambda_j \mathbf{p}_j \mathbf{p}_j' \end{aligned}$$

kde

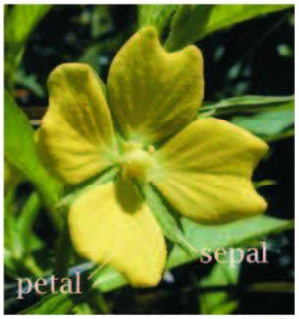
Λ_j ... jsou vlastní čísla matice $\mathbf{\Sigma}$
$\mathbf{p}_j$ ... jsou ortonormální vlastní vektory matice $\mathbf{\Sigma}$
$\mathbf{p}_1, \dots, \mathbf{p}_n$ ... je báze v R^n
$\mathbf{P}$ ... je ortogonální matice

Poznámka

Pro vlastní čísla platí, že jsou uspořádaná dle velikosti:

$$\Lambda_1 \geq \Lambda_2 \geq \Lambda_3 \geq \dots \geq \Lambda_n \geq 0$$

Příklad



Příklad

```
library(MASS)
data=data(iris)
iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
```


Spektrální rozklad

Poznámka

Ortonormální vektory jsou takové, které splňují dvě vlastnosti:

- *jsou navzájem na sebe kolmé* $\rightarrow (\mathbf{p}_i \perp \mathbf{p}_j) = 0$
- *mají jednotkovou délku* $\rightarrow \|\mathbf{p}_i\| = \|\mathbf{p}_j\| = 1$

Poznámka

Báze vektorového prostoru je množina lineárně nezávislých vektorů $\rightarrow$ *lineární kombinace vektorů splňuje podmínku:*

$$\sum_{i=1}^{n} a_i \mathbf{p}_i \neq 0$$

pro alespoň jeden koeficient $a_i \neq 0$.
$\rightarrow$ *Žádný z vektorů* $\mathbf{p}_i$ *není možné vyjádřit jako lineární kombinaci zbývajících vektorů.*

Postup

Mějme n - dimenzionální náhodný vektor (tj. n náhodných proměnných)

$$\mathbf{X} = (X_1, \ldots, X_n).$$

Úlohou PCA je nahradit tento vektor menším počtem latentních (skrytých) proměnných

$$\mathbf{Y} = (Y_1, \ldots, Y_k)$$

kdy $k < n$, které by co nejpřesněji popisovali původní soubor ve smyslu zachování varianční struktury.

$\Downarrow$

Budeme hledat novou náhodnou veličinu, která vznikne lineární transformací vektoru

$$\mathbf{X} = (X_1, \ldots, X_n) \quad \rightarrow \quad \mathbf{Y} = \mathbf{DX},$$

kde $\mathbf{D}$ je nějaká matice rozměru $(k \times n)$.

Postup

Ukazuje se, že nejvhodnějším kandidátem pro matici $\mathbf{D}$ je právě ortonormální matice vlastních vektorů $\mathbf{P}$, kterou získáme ze spektrálního rozkladu kovarianční matice $\mathbf{\Sigma}$ náhodného vektoru $\mathbf{X} = (X_1, \ldots, X_n)$. Lineární transformací tedy vytvoříme nový náhodný vektor

$$\mathbf{Y} = \mathbf{PX}.$$

Definice

Náhodnou veličinu

$$Y_j = \mathbf{p}_j' \mathbf{X}, \quad kde \quad j = 1, \ldots, n$$

nazveme j*-tou hlavní komponentou vektoru* $\mathbf{X}$.

Proces vytváření tohoto nového souboru probíhá postupně, tj. $Y_1, Y_2, \ldots$ Díky tomu bude dosaženo efektu postupného vyčerpání maxima zbývající variability náhodné proměnně $\mathbf{X} = (X_1, \ldots, X_n)$.

Konstrukce hlavních komponent

Hledáme takový vektor $\mathbf{d} = (d_1, \ldots, d_n)$ reálných čísel, který spňuje podmínku

$$\mathbf{d}'\mathbf{d} = 1$$

a pro který má náhodná veličina $\mathbf{d}'\mathbf{X}$ získaná lineární transformací původního náhodného vektoru $\mathbf{X}$ největší rozptyl.
Vzhledem k tomu, že platí

$$\mathrm{Var}(\mathbf{d}'\mathbf{X}) = \mathbf{d}'\mathbf{\Sigma}\mathbf{d}$$

tak se tedy snažíme o maximalizaci výrazu

$$\mathbf{d}'\mathbf{\Sigma}\mathbf{d}.$$

To nastane právě tehdy, když

$$\mathbf{d} = \mathbf{p}$$

a potom tato **maximální hodnota** odpovídá právě číslu Λ_1.
Tímto postupem tedy získáme **první hlavní komponentu**

$$Y_1 = \mathbf{p}_1'\mathbf{X}.$$

Konstrukce hlavních komponent

Výše uvedený postup budeme znovu opakovat, abychom nalezli druhou hlavní komponentu Y_2. Budeme tedy opět hledat vektor $\mathbf{d} \in R^n$, který bude splňovat podmínku

$$\mathbf{d}'\mathbf{d} = 1.$$

Nyní ovšem k tomuto požadavku přibývá další podmínka kladená na vektor $\mathbf{d}$ a to požadavek nekorelovanosti s již existující náhodnou veličinou Y_1. Této podmínky je dosaženo právě tehdy když

$$\mathbf{d}'\mathbf{p}_1 = 0.$$

Tato situace nastane opět tekdy, když

$$\mathbf{d} = \mathbf{p}_2$$

kdy získáme **druhou hlavní komponentu**

$$Y_2 = \mathbf{p}_2'\mathbf{X}$$

a k tomu odpovídající **druhou největší hodnotu** odpovídající právě číslu Λ_2. Takto pokračujeme dále dokud nenajdeme všechny hlavní komponenty.

Vlastnosti hlavních komponent

- variance j-té hlavní komponenty je přímo úměrná j-tému vlastnímu číslu

$$\mathrm{Var}(Y_j) = \mathrm{Var}(\mathbf{X})\mathbf{p}_j = \mathbf{p}_j'\mathrm{Var}(\mathbf{X})\mathbf{p}_j = \mathbf{p}_j'\mathbf{\Sigma}\mathbf{p}_j = \mathbf{p}_j'\sum_{i=1}^{n}\Lambda_i\mathbf{p}_i\mathbf{p}_i'\mathbf{p}_j = \Lambda_j$$

- libovolné dvě hlavní komponenty i, j jsou vzájemně nekorelované

$$\mathrm{Cov}(Y_i, Y_j) = \mathrm{Cov}(\mathbf{p}_i'\mathbf{X}, \mathbf{p}_j'\mathbf{X}) = \mathbf{p}_i'\mathrm{Var}(\mathbf{X})\mathbf{p}_j = \mathbf{p}_i'\sum_{k=1}^{p}\Lambda_k\mathbf{p}_k\mathbf{p}_k'\mathbf{p}_j = 0$$

- jestliže je hodnost kovariační matice menší než její rozměr, tj.

$$rank(\mathbf{\Sigma}) = k,$$

kdy $k < n \longrightarrow$ posledních $(n-k)$ vlastních čísel je rovno nule

$\downarrow$

posledních $(n-k)$ hlavních komponent jsou skoro jisté konstanty

Příklad

```
> a=iris[, c(1:4)]
> a=scale(iris[, c(1:4)],center=TRUE)
> cov=cov(a)
> cov
             Sepal.Length Sepal.Width Petal.Length Petal.Width
Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
> eigen(cov)
$values
[1] 2.91849782 0.91403047 0.14675688 0.02071484
$vectors
           [,1]        [,2]       [,3]       [,4]
[1,]  0.5210659 -0.37741762  0.7195664  0.2612863
[2,] -0.2693474 -0.92329566 -0.2443818 -0.1235096
[3,]  0.5804131 -0.02449161 -0.1421264 -0.8014492
[4,]  0.5648565 -0.06694199 -0.6342727  0.5235971
```


Vlastnosti hlavních komponent

Pokud nastane vlastnost 3, tedy $rank(\mathbf{\Sigma}) = k$, přičemž $k < n$, pak platí

$$\sum_{j=1}^{n}\mathrm{Var}(X_j) = \sum_{j=1}^{k}\mathrm{Var}(Y_j).$$

Nyní nastává otázka, kolik hlavních komponent bychom měli vytvořit, abychom dostatečně popsali variabilitu původního datového souboru?

Definice

Míra variability v $\mathbf{X} = (X_1, \ldots, X_n)$ *je vyjádřena jako*

$$\mathrm{Tr}(\mathbf{\Sigma}) = \sum_{j=1}^{n}\Lambda_j, \quad kde \quad j = 1, \ldots, n.$$

Kolik hlavních komponent je třeba?

Při stanovení počtu komponent zohledňujeme tato hlediska:

- snažíme se zmenšit dimenzi původního datového souboru $\mathbf{x} \rightarrow$ přiznivý je malý počet hlavních komponent
- snažíme se zachovat informaci o variabilitě původního datového souboru $\mathbf{x} \rightarrow$ potřebujeme dostatečné množství hlavních komponent

Hodnotíme současné naplnění těchto cílů.

$\Downarrow$

Definice

Relativní příspěvek i-té hlavní komponenty R_i do celkové variability původního datového souboru, tj. náhodného vektoru $\mathbf{X}$ *je možné vyjádřit jako*

$$R_i = \frac{\Lambda_i}{\sum_{j=1}^{n}\Lambda_j}, \quad kde \quad j = 1, \ldots, n.$$

Obvykle se uvažuje několik největších hlavních komponent tak, aby jejich příspěvek vyčerpal např. 80% celkové variability.

Příklad

```
eigen=eigen(cov)$values
norm=sum(eigen)
norm
[1] 4

> eigen[1]/norm
[1] 0.7296245
> eigen[2]/norm
[1] 0.2285076
> eigen[3]/norm
[1] 0.03668922
> eigen[4]/norm
[1] 0.005178709
```

Problém jednotek

Analýza hlavních komponent je shodná, pokud jsou všechny složky náhodného vektoru $\mathbf{x} = (X_1, \ldots, X_n)$ měřeny ve stejných jednotkách. Problém nastává v situaci, kdy jsou složky vektoru $\mathbf{x}$ měřeny v různých jednotkách.

Změnou měřítka by se mohly podstatně změnit hodnoty hlavních komponent, které potom nemusí vysvětlovat postupně největší část celkové variance a teorie výběrových vlastností je složitější !!!

V případě takové různorodosti datového souboru $\mathbf{x}$ můžeme provést normování původního datového souboru ve tvaru:

$$U_i = \frac{X_i - \mathrm{E}(X_i)}{\sqrt{\mathrm{Var}(X_i)}}, \quad \text{kde} \quad i = 1, \ldots, n.$$

Analýzu hlavních komponent pak provedeme s korelační maticí namísto matice kovarianční.

Geometrická interpretace

Hlavní komponenty je možné si představit jako shluk k bodů v n rozměrném prostoru. Konstrukcí hlavních komponent jakoby rotujeme souřadnicovou soustavou tak, aby nové osy procházely směry největší variability bodů prostoru. Na několika prvních osách se takto zachytí maximum informace o prostorové struktuře dat.

⇓

Vícerozměrná data se snažíme zobrazit na:

- přímku - 1 komponenta
- do roviny - 2 komponenty
- do prostoru - 3 komponenty, ...

KONEC

DĚKUJI

VÁM

ZA

VAŠI

POZORNOST