OBECNÁ **METODA HLAVNÍCH KOMPONENT**
PC (Principal Component Method)

Metoda hlavních komponent je postup vyvinutý původně jako statistická technika v z psychometrii a sociometrii, kde se uplatňuje jako jedna z metod užitých k výpočtům parametrů *modelu faktorové analýzy*. Jejím hlavním konceptem je transformace proměnných (v lineárním regresním modelu jde o vysvětlující proměnné obsažené v matici plánu X) takovým způsobem, aby v nich obsažená informace byla „rozdělena" jiným způsobem než v původních proměnných. Je tím míněna jednak vzájemná „nezávislost" těchto transformovaných proměnných, jednak „ostřejší" rozdělení informace v nich obsažené (oproti původním proměnným).

V průběhu nasazení metody se konstruují tzv. *hlavní komponenty*, což jsou lineární kombinace vysvětlujících proměnných. Nejprve se sestrojí první hlavní komponenta jako taková lineární kombinace vysvětlujících proměnných, která vysvětluje nejvíce variability závisle proměnné. Poté se (ze „zbytku" informace obsažené v matici X) sestrojí druhá hlavní komponenta, která z tohoto zbytku vysvětlí další část variability, poté třetí hlavní komponenta, až konečně poslední **k-tá hlavní komponenta.**

Hlavní komponenty jsou *ortogonální*, tzn. jsou *vzájemně* *naprosto* ***lineárně nezávislé*. Lze je navíc znormovat tak, aby byly** dokonce ***ortonormální*, tzn. aby měly jedničkovou normu. Matematicky to lze uskutečnit tak, že jsou hlavní komponenty vzaty jako normované hlavní (charakteristické) vektory příslušné momentové matici $X'X$. *Obvykle se tato matice sestavuje z normovaných vektorů vysvětlujících proměnných (nebo aspoň z matice X, ze které je vynechán první sloupec, protože vektor konstant nemající žádnou variabilitu k vysvětlení rozptylu závisle proměnné přispět nijak nemůže.)***

Převodem na *k* hlavních komponent získáme v modelu „jinak rozdělenou", (a to „nerovnoměrněji") **informaci obsaženou ve vysvětlujících proměnných. První hlavní komponenta bude** obvykle **obsahovat větší část „modelové informace", než kterýkoliv sloupec matice X uvažovaný samostatně. Představu o rozdělení této informace poskytne nejlépe rozložení hlavních (charakteristických) čísel matice X.** Obrazně řečeno, **podíl informačního obsahu 1. a 2. hlavní komponenty bude dán podílem** příslušných **dvou (**hodnotami **největších) charakteristických čísel,** analogicky tomu bude u dalších hlavních komponent**.**

Výhody nasazení metody hlavních komponent do lineárního regresního modelu jsou:

a) **Omezením se na relativně malý počet *r* hlavních komponent (cca 2-4) dosáhneme úspornějšího provedení kvantitativní analýzy regresní rovnice. Sníží se tak počet „vysvětlujících proměnných", zcela se odstraní riziko multikolinearity** (nové proměnné jsou totiž vzájemně ortogonální)

b) Hlavní komponenty jsou typické vždy individuálním přínosem k vysvětlení závisle proměnné, tzn. informace v hlavních komponentách se vzájemně „nemísí".

Nevýhodou nasazení metody hlavních komponent v lineárním regresním modelu je :

- Interpretace obsahu čistě matematicky zkonstruovaných hlavních komponent je obvykle problematická a často nelze „obsahově pojmenovat" ani první hlavní komponentu. Toto určitě **znesnadňuje analytické zkoumání v rámci přijaté modelové specifikace, ale nemusí to vadit při predikčním uplatnění této metody.**

Metoda hlavních komponent se v ekonometrii uplatňuje ve více oblastech, zejména

a) při odstranění problému multikolinearity
b) pokud potřebujeme snížit počet vysvětlujících proměnných modelu, tj. když $K \geq T$.
c) při aplikaci techniky instrumentálních proměnných, a to jak v jednorovnicovém modelu, tak v simultánní soustavě regresních rovnic.

Formálně zapsáno : $Z_{[T,k]} = X_{[T,k]}.A_{[k,k]}$

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & .... & x_{1k} \\ x_{21} & x_{22} & x_{23} & .... & x_{2k} \\ x_{31} & x_{32} & x_{33} & ..... & x_{3k} \\ ..... & ..... & ..... & ..... & ..... \\ x_{T1} & x_{T2} & x_{T3} & ..... & x_{Tk} \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} & .... & a_{1k} \\ a_{21} & a_{22} & a_{23} & ..... & a_{2k} \\ a_{31} & a_{32} & a_{33} & ..... & a_{3k} \\ ..... & ..... & ..... & ..... & \\ a_{k1} & a_{k2} & a_{k3} & ..... & a_{kk} \end{pmatrix} = \begin{pmatrix} z_{11} & z_{12} & z_{13} & ..... & z_{1k} \\ z_{21} & z_{22} & z_{23} & ..... & z_{2k} \\ z_{31} & z_{32} & z_{33} & ..... & z_{3k} \\ ..... & ..... & ..... & ..... & ..... \\ z_{T1} & z_{T2} & z_{T3} & ..... & z_{TK} \end{pmatrix}$$

přičemž platí $X'X = A.D.A'$, **kde**

A **je matice, jejíž *sloupce* tvoří vlastní (charakteristické) vektory** $a_{.j}$ **, pro něž platí**

$A'.A = A.A' = I_k$ v důsledku ortonormality vlastních vektorů

D **je diagonální matice, jejíž diagonální prvky tvoří charakteristická čísla (kořeny) matice** $X'X$ **. Obvykle se tato charakteristická čísla uspořádávají sestupně, aby největší bylo první z nich (v levém horním rohu).**

$$\begin{pmatrix} \xi_{11} & \xi_{12} & \xi_{13} & ... & \xi_{1k} \\ \xi_{21} & \xi_{22} & \xi_{23} & .. & \xi_{2k} \\ \xi_{31} & \xi_{32} & \xi_{33} & .. & \xi_{3k} \\ ..... & .... & ..... & .. & ..... \\ \xi_{k1} & \xi_{k2} & \xi_{k3} & ... & \xi_{kk} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & .. & a_{1k} \\ a_{21} & a_{22} & a_{23} & .. & a_{2k} \\ a_{31} & a_{32} & a_{33} & .. & a_{3k} \\ ..... & ..... & ..... & ... & \\ a_{k1} & a_{k2} & a_{k3} & ... & a_{kk} \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 & 0 & ... & 0 \\ 0 & \lambda_2 & 0 & .. & 0 \\ 0 & 0 & \lambda_3 & .. & 0 \\ ... & .... & .... & .. & ... \\ 0 & 0 & 0 & ... & \lambda_k \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} & a_{31} & .. & a_{k1} \\ a_{12} & a_{22} & a_{32} & .. & a_{k2} \\ a_{13} & a_{23} & a_{33} & .. & a_{k3} \\ ... & .... & ... & .. & ... \\ a_{1k} & a_{2k} & a_{3k} & .. & a_{kk} \end{pmatrix}$$

Jako ξ_{ij} **jsme označili prvky momentové matice** $X'X$ **. Pro charakteristická čísla na diagonále D platí uspořádání**

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq .... \geq \lambda_k$$

Původní lineární regresní model $y = X\beta + \varepsilon$ **lze takto po transformaci proměnných přepsat do tvaru**

$$y = (XA).(A'\beta) + \varepsilon = Z.\delta + \varepsilon$$ **, v němž**

$Z = X.A$ **je matice s po sloupcích uloženými hlavními komponentami**
$\delta' = A'\beta$ **je nový (** transformovaný **) vektor parametrů**

Přitom pro součin matice hlavních komponent $Z'Z$ **platí :**

$$Z'Z = A'X'XA = A'ADA'A = D$$

Jednotlivé hlavní komponenty jsou obsaženy ve *sloupcích* matice Z **. Každý *sloupec* matice** Z **může být vyjádřen jako ortogonální lineární kombinace** (obecně všech) **sloupců matice** X **, kde** $j-tý$ **vlastní vektor** z_j **určuje váhy v této lineární kombinaci, tj.** $z_j = X.a_j$ **,**

přičemž jednotlivá a_j **jsou tvořena prvky charakteristického vektoru.**

Může se přirozeně stát, že některá z charakteristických čísel budou nulová. Tak tomu bude tehdy, jestliže některé sloupce matice X **budou lineárně závislé a jak matice** X **, tak momentová matice** $X'X$ **budou singulární.**

Nalezení hlavních komponent se zpravidla provede nasazením vhodné procedury v matematických softwarových prostředcích – obvykle tato procedura nese v sobě název *eig* nebo *eigen (vlastní čísla jsou anglicky eigenvalues a vlastní vektory eigenvectors).*

V MATLABu je syntax této procedury **[V,D] = eig (X) , kde**
Vstupním polem je matice X , u níž se vlastní čísla a vektory vyčíslují
V je výstupní matice vlastních vektorů (s uložením po sloupcích)
D je diagonální matice vlastních čísel.

Vlastní čísla na diagonále matice D poskytují informaci o tom, jak je „rozložena“ variabilita proměnných obsažených v matici X. Obvykle k dostatečnému vystižení variability X postačí vzetí 2-4 hlavních komponent, které často obsahují až 98% variability obsažené v matici X.