4.5.2 Faktorová analýza (FA)

Podobně jako metoda hlavních komponent patří také faktorová analýza mezi metody redukce počtu původních proměnných. Ve faktorové analýze předpokládáme, že každou vstupující proměnnou můžeme vyjádřit jako lineární kombinaci nevelkého počtu společných skrytých faktorů a jediného chybového faktoru. Na rozdíl od komponentní analýzy se při faktorové analýze snažíme vysvětlit závislost proměnných. K nevýhodám metody patří zejména nutnost zadat *počet společných faktorů* ještě před prováděním vlastní analýzy.

Předpokládáme-li, že $\boldsymbol{x}^{\mathrm{T}} = (x_1, x_2, \dots, x_m)^{\mathrm{T}}$ je jeden objekt pozorovaných proměnných s korelační $\boldsymbol{R}$ nebo kovarianční maticí $\boldsymbol{C}$, potom můžeme všechny objekty $\boldsymbol{X}$, rozměru $n \times m$ zapsat jako *model faktorové analýzy (FA)* ve tvaru

$$x_1 = l_{11}f_1 + l_{12}f_2 + \dots\ l_{1m}f_m + \varepsilon_1\ ,$$

$$x_2 = l_{21}f_1 + l_{22}f_2 + \dots\ l_{2m}f_m + \varepsilon_2\ ,$$

$$\dots\dots\dots\dots$$

$$x_n = l_{n1}f_1 + l_{n2}f_2 + \dots\ l_{nm}f_m + \varepsilon_n\ ,$$

kde $f_1, f_2, \dots, f_m$ jsou *faktory*, které vyvolávají korelace mezi proměnnými a $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ jsou *chybové faktory*, které přispívají k rozptylu jednotlivých proměnných. Koeficienty l_{ik} nazýváme *faktorové zátěže i-tého objektu u k-tého* společného faktoru f_k a představují prvky matice faktorových zátěží. Model můžeme přepsat v maticové podobě jako

$$\boldsymbol{x} = \boldsymbol{L}\,\boldsymbol{f} + \varepsilon.$$

Pro ortogonální faktorový model lze kovarianční matici vektoru vstupujících proměnných čili sloupců zdrojové matice napsat ve formě tzv. *základní faktorové věty* ve tvaru

$$\boldsymbol{C} = \boldsymbol{L}\,\boldsymbol{L}^{\mathrm{T}} + \boldsymbol{\Gamma}^2,$$

kde $\boldsymbol{L}$ je *matice faktorových zátěží*, dále $\boldsymbol{L}\,\boldsymbol{L}^{\mathrm{T}}$ představuje vlastně kovarianční matici vektoru $\boldsymbol{L}\,\boldsymbol{f}$ a konečně $\boldsymbol{\Gamma}^2$ je *matice jedinečností*.

Matice jedinečností $\boldsymbol{\Gamma}^2$ čili *kovarianční matice chybových faktorů* je maticí diagonální, protože předpokládáme nekorelované chyby. Uvědomíme-li si dále, že diagonální prvky matice $\boldsymbol{\Gamma}^2$ představují rozptyly jednotlivých sloupců zdrojové matice, lze psát

$$\boldsymbol{S}^2 = \boldsymbol{H}^2 + \boldsymbol{\Gamma}^2\ ,$$

kde $\boldsymbol{S}^2$ je diagonální matice rozptylů faktorů. Proměnlivost každého faktoru, vyjádřenou sloupcem zdrojové matice, můžeme rozdělit na součet dvou složek: *komunalitu* $\boldsymbol{H}^2$, která představuje proměnlivost společnou všem faktorům a *jedinečnost* $\boldsymbol{\Gamma}^2$, která představuje část proměnlivosti čili rozptylu, nevysvětlenou faktory.

Komunalita $\boldsymbol{H}$ vyjadřuje míru proměnlivosti a je vahou, s jakou jednotlivé faktory přispívají do rozptylu odpovídající proměnné. Čtverec komunality je suma faktorových zátěží faktorů.

Jedinečnost $\boldsymbol{\Gamma}^2$ bývá dále rozdělována na část *specifity* $\boldsymbol{\Gamma}^2_s$ a část *nespolehlivosti* $\boldsymbol{\Gamma}^2_n$.

Specifita představuje tu část proměnlivosti, kterou nelze vysvětlit ani chybou experimentu, ani společnými faktory, zatímco *nespolehlivost* představuje experi-mentální chybu při měření faktorů.

Uvedený způsob rozkladu proměnlivosti představuje základní hledisko pro klasifikaci metod faktorové analýzy. Metoda hlavních komponent je zvláštním případem faktorové analýzy, kdy je model definován tvarem $\boldsymbol{S}^2 = \boldsymbol{H}^2$ a předpokládá se, že prostřednictvím hlavních komponent lze proměnlivost zdrojové matice beze zbytku reprodukovat. Jde tedy o vhodnou ortogonální transformaci, která zachovává všechnu původní proměnlivost, a to beze zbytku. Hovoříme pak z hlediska faktorové analýzy o *úplné komponentní analýze*. Jestliže při reprodukci pomocí hlavních komponent reprodukujeme pouze podstatnou část proměnlivosti (ale ne všechnu), jedná se ve faktorové analýze o *neúplnou komponentní analýzu.*

Pro odhad parametrů faktorového modelu se často užívá metody hlavních komponent, která je aplikována na redukovanou kovarianční matici. Předpokládá se, že jsou známy nějaké počáteční odhady chybových rozptylů, které jsou odečítány od diagonálních prvků výběrové korelační matice $\boldsymbol{R}$. Takto upravenou kovariační matici rozkládáme na součin matic $\boldsymbol{L}_1\,\boldsymbol{L}_1^{\mathrm{T}}$, kde $\boldsymbol{L}_1$ představuje výchozí matici odhadů faktorových zátěží. Postup pokračuje iterativně a po několika krocích konverguje ke konečné matici odhadů faktorových zátěží. Pokud neznáme výchozí odhady chybových rozptylů (resp. komunalit), je možné je určit speciálním postupem.

Porovnání FA a PCA:

1. Obě metody nemá cenu použít, když jsou původní proměnné x_j^{T}, $j = 1, ..., m$, nekorelované. FA pak nemá co objasnit a PCA povede k hlavním komponentám totožným s původními proměnnými.

2. FA postuluje model pro data, PCA nikoliv.

3. FA se pokouší objasnit kovariance a korelace původních proměnných pomocí několika málo společných faktorů. PCA objasňuje pouze rozptyl původních promĕn-ných.

4. PCA: když zvýšíme počet použitých proměnných A o 1 na A+1, původní komponenty se nezmění. FA: když přidáme další faktor, ostatní faktory se podstatně změní.

5. PCA: výpočet je přímočarý, jednoduchý. FA: výpočet faktorového skóre je daleko komplexnější a byla pro něj navržena řada postupů.

6. Obyčejně není žádný vztah mezi hlavními komponentami PC a korelační maticí $\boldsymbol{R}$ anebo kovarianční maticí $\boldsymbol{C}$.