Faktorska analiza

Med metodami za pregledovanje podatkov smo omenili metodo glavnih komponent. Cilj te metode je določiti manjše število linearnih kombinacij merjenih spremenljivk tako, da z njimi pojasnimo kar se da velik del celotne razpršenosti (variance) podatkov. Faktorska analiza, ki je podobna metoda za redukcijo podatkov, se v osnovi razlikuje od metode glavnih komponent. V primeru faktorske analize gre za študij povezav med spremenljivkami, tako da poizkušamo najti novo množico spremenljivk (manj kot merjenih spremenljivk), ki predstavljajo to, kar je skupnega opazovanim spremenljivkam. Faktorska analiza poizkuša poenostaviti kompleksnost povezav med množico opazovanih spremenljivk z razkritjem skupnih razsežnosti ali faktorjev, ki omogočajo vpogled v osnovno strukturo podatkov.

V (tržnem) raziskovanju je velikokrat tako, da pojmov, ki so ključni, ne moremo neposredno meriti (npr. družbeni položaj ljudi, ekonomsko razvitost držav, zadovoljstvo z delom). Ponavadi jih merimo posredno z indikatorji tistega, kar naj bi merili. Zberemo torej nekaj direktno merljivih spremenljivk, ki so indikatorji pojma (konstrukta), ki ga želimo meriti, in nato poizkušamo razkriti ali so povezave med izbranimi opazovanimi spremenljivkami pojasnljive s predpostavljeno nemerljivo spremenljivko, ali pa je morda potrebno postaviti kompleksnejšo strukturo povezanosti. Ponavadi imenujemo merljive spremenljivke manifestne, nemerljive pa latentne spremenljivke. V takih študijah je najpogosteje uporabljena ena od metod faktorske analize. **Cilj teh metod je ugotoviti ali so zveze med opazovanimi spremenljivkami (kovariance ali korelacije) pojasnljive z manjšim številom posredno opazovanih spremenljivk ali faktorjev.**

Oče faktorske analize je Spearman (1904), ki je obravnaval skupni uspeh učencev na osnovi ocen treh predmetov:

- X_1 – klasične vede
- X_2 – francoski jezik
- X_3 – angleški jezik

Koeficienti korelacije teh treh predmetov na množici učencev so predstavljeni v tabeli.

	X_1	X_2	X_3
X_1	1.		
X_2	0.83	1.	
X_3	0.78	0.67	1.

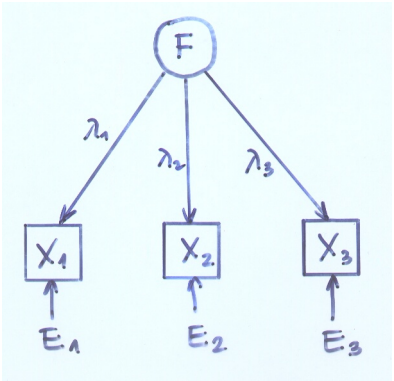
Spearman je predpostavljal, da en faktor ustrezno pojasnjuje dobljene korelacije. Faktorski model je zapisal takole:

$$X_1 = \lambda_1 F + E_1$$

$$X_2 = \lambda_2 F + E_2$$

$$X_3 = \lambda_3 F + E_3$$

kjer so λ_i faktorske uteži in E_i predstavljajo specifične faktorje. V tem primeru splošni faktor F pomeni splošno učenčevo uspešnost. Specifični faktorji E_i bodo imeli majhne variance, če so opazovane spremenljivke blizu faktorju F.

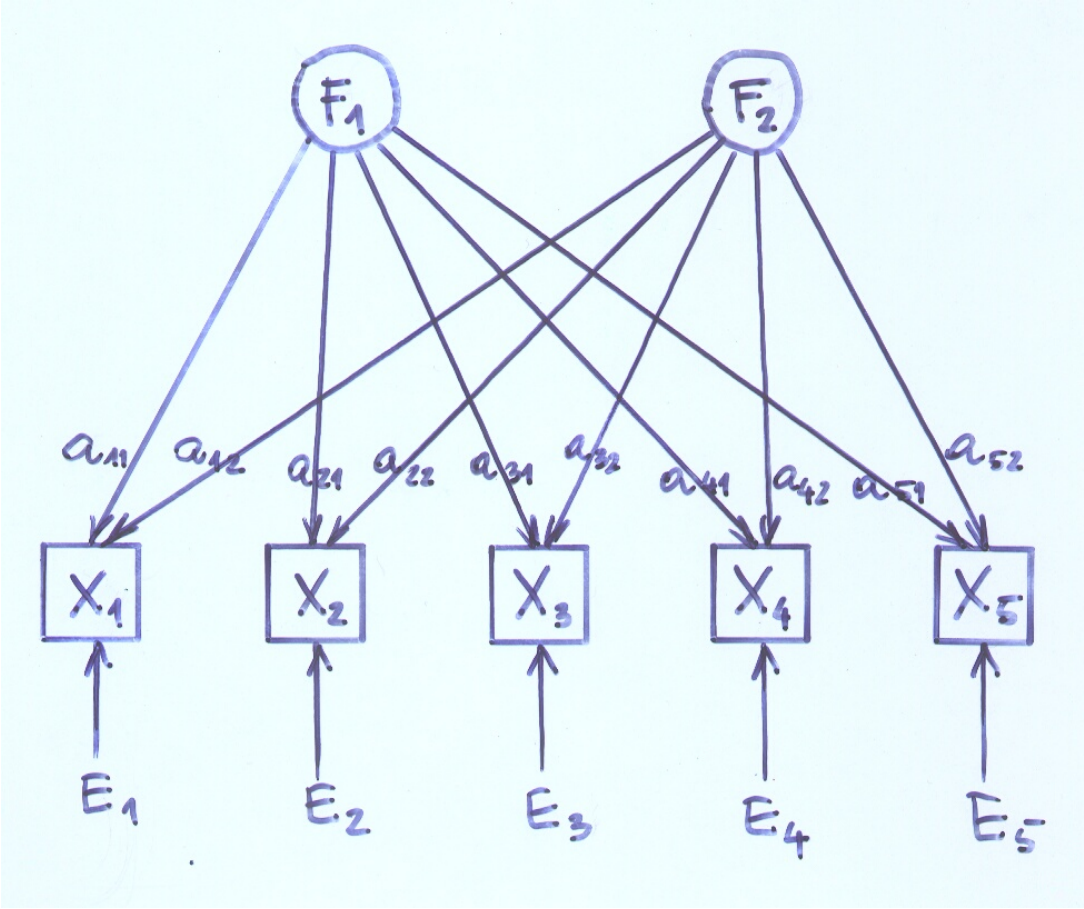


Splošni faktorski model

Dane naj bodo spremenljivke X_i (i=1,...,m), F_r (r=1,...,k) in E_i (i=1,...,m). Osnova faktorskega modela je domneva, da med spremenljivkami X_i, F_r in E_i velja zveza:

$$X_i = \sum_{r=1}^{k} a_{ir} F_r + E_i, \qquad i = 1, ..., m$$

kjer je $k < m$. X_i so merjene spremenljivke, F_r so v tem modelu skupni faktorji, E_i je specifični faktor, ki vpliva samo na X_i, a_{ir} pa je faktorska utež, ki kaže vpliv faktorja F_r na X_i.



Zapišimo faktorski model najprej v matrični obliki. Spremenljivke X_i zapišimo v matriki $X_{n\times m}$ (matrika podatkov)

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}$$

faktorje F_r v matriki $F_{n\times k}$ (matrika faktorjev)

$$F = \begin{bmatrix} f_{11} & f_{12} & \dots & f_{1k} \\ f_{21} & f_{22} & \dots & f_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \dots & f_{nk} \end{bmatrix}$$

faktorske uteži a_{ir} v matriki $A_{m\times k}$ (matrika uteži)

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mk} \end{bmatrix}$$

in specifična faktorje v matriki $E_{n\times m}$ (matrika specifičnih faktorjev)

$$E = \begin{bmatrix} e_{11} & e_{12} & \dots & e_{1m} \\ e_{21} & e_{22} & \dots & e_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ e_{n1} & e_{n2} & \dots & e_{nm} \end{bmatrix}$$

Splošni faktorski model potem lahko zapišemo v matrični obliki

$$X = FA' + E$$

Običajne predpostavke splošnega faktorskega modela so:

1. Specifični faktorji so pravokotni med seboj
   ($\text{cov}(E_i, E_j) = 0$, če je $i \neq j$);
2. Vsak specifični faktor E_i je pravokoten na vsak skupni faktor F_j ($\text{cov}(E_i, F_j) = 0$ za vsak i in j);
3. Skupni faktorji so pravokotni med seboj
   ($\text{cov}(F_i, F_j) = 0$, če je $i \neq j$);
4. Spremenljivke X_i, F_i in E_i naj bodo centrirane
   ($E(X_i) = E(F_i) = E(E_i) = 0$).

Posledice teh predpostavk so:

- Zaradi četrte predpostavke velja
$$\Sigma = \frac{1}{n} X'X$$
To je variančno - kovariančna matrika ali v primeru standardiziranih spremenljivk korelacijska matrika.

- Ker velja prva predpostavka je
$$\frac{1}{n} E'E = \Psi$$
kjer je Ψ diagonalna matrika z variancami specifičnih faktorjev na diagonali.

- Zaradi druge predpostavke je $E'F = 0$.

- Iz tretje predpostavke sledi
$$\frac{1}{n} F'F = I$$

Iz splošnega faktorskega modela in na osnovi omenjenih predpostavk lahko izpeljemo naslednjo **faktorsko enačbo**

$$\Sigma = AA' + \Psi$$

Preverimo:

$$\Sigma = \frac{1}{n}X'X = \frac{1}{n}(FA' + E)'(FA' + E) =$$

$$= \frac{1}{n}(AF' + E')(FA' + E) =$$

$$= \frac{1}{n}(AF'FA' + E'FA' + AF'E + E'E) = AA' + \Psi$$

V zgornji faktorski enačbi je Σ v splošnem matrika varianc in kovarianc, če pa so merjene spremenljivke standardizirane, je Σ korelacijska matrika. Najpogosteje velja slednje.

Če primerjavo na levi in desni strani faktorske enačbe diagonalne elemente, dobimo naslednje enačbe

$$\sigma_i^2 = \sum_{j=1}^{k} a_{ij}^2 + \psi_{ii}$$

kar pomeni, da varianco merjene spremenljivke X_i razbijemo na del, ki je pojasnjen s skupnimi faktorji, in na specifično varianco. Delež variance, ki je pojasnjena s skupnimi faktorji, imenujemo komunaliteta in jo označujemo s h_i^2:

$$\sum_{j=1}^{k} a_{ij}^2 = h_i^2$$

Naloga je, da iz znanih elementov matrike varianc in kovarianc (ali korelacijske matrike) Σ izračunamo neznane parametre faktorskega modela, se pravi faktorske uteži A in specifične variance Ψ.

Najprej naslednje: iz k-faktorskega modela, ki je podan s faktorsko enačbo $X = FA' + E$, sledi, da je variančno–kovariančna matrika opazovanih spremenljivk $\Sigma = AA' + \Psi$. Velja tudi obratno: če je kovariančno matriko Σ možno dekomponirati v zgornjo obliko, potem k-faktorski model velja za opazovane spremenljivke X.

Pred računanjem parametrov faktorskega modela se moramo najprej vprašati, kako je z

- identifikabilnostjo faktorskega modela in
- enoličnostjo ocen parametrov.

Identifikabilnost

Število vseh parametrov faktorskega modela, ki jih moramo oceniti na osnovi faktorske enačbe, je $m \times k$ (faktorskih uteži) in m (specifičnih varianc). Parametre ocenjujemo na osnovi informacij v matriki varianc in kovarianc opazovanih spremenljivk, ki jih je $\frac{m(m+1)}{2}$. Torej imamo $\frac{m(m+1)}{2}$ enačb za $mk + m$ parametrov. Za identifikacijo modela je ponavadi potreben pogoj, da je več enačb kot ocenjevanih parametrov. Od tu naslednji pogoj:

$$mk + m \leq m(m+1)/2$$

oziroma

$$k \leq \frac{(m-1)}{2}$$

Zgornji pogoj je potreben pogoj za identifikacijo faktorskega modela, ni pa zadosten.

Enoličnost

Naša naloga je, da iz danih elementov variančno–kovariančne ali korelacijske matrike Σ izračunamo neznane elemente matrik A in Ψ. Vprašanje, ki se ob tem ponuja, je: pri kakšnih pogojih je mogoče za dano matriko Σ enolično določiti matriki A in Ψ, ki zadoščata enačbi

$$\Sigma = AA' + \Psi$$

Če je $k > 1$ in če obstaja enolično določena matrika Ψ, obstaja neskončno mnogo matrik A, ki zadoščajo zgornji enačbi. Poglejmo zakaj: Naj bo $M_{k\times k}$ neka ortonormalna matrika ($MM' = I$) in naj bo

$$A^* = AM$$

Potem je

$$A^*A^{*\prime} = (AM)(AM)' = AMM'A' = AA' = \Sigma - \Psi$$

To pomeni, da je tudi A^* lahko rešitev zgornje enačbe. Ortonormalnih matrik je neskončno mnogo, zato tudi obstaja neskončno mnogo matrik, ki zadoščajo zgornji enačbi. Za enolično določitev matrike A je zato potrebno dodati še kakšen pogoj.

Ocenjevanje faktorskega modela

Faktorski model ocenjujemo v dveh korakih:

1. ocena komunalitet (t.j. skupnega prostora) z eno od metod faktorske analize,
2. ocena faktorskih uteži z eno od rotacij.

Faktorska analiza ni končana, če ni narejena tudi ustrezna rotacija.

1. Metode faktorske analize

- metoda glavnih osi (PAF)
- metoda največjega verjetja (ML)
- 'image' faktorska analiza
- alfa faktorska analiza
- ...

Metoda glavnih osi

Splošno faktorsko enačbo

$$\Sigma = AA' + \Psi$$

lahko zapišemo takole

$$\Sigma - \Psi = AA'$$

Denimo, da so merjene spremenljivke standardizirane. Ker je matrika Ψ diagonalna matrika z variancami specifičnih faktorjev na diagonali, je leva stran enačbe korelacijska matrika s komunalitetami na diagonali.

V splošnem lahko komunalitete določimo šele tedaj, ko določimo skupne faktorje, ki pa jih lahko izračunamo iz popravljene korelacijske matrike $\Sigma - \Psi$. Ta začarani krog za ocene komunalitet in skupnih faktorjev je osnovna pomanjkljivost splošnega faktorskega modela. Metoda glavnih osi (PAF) rešuje problem faktorske analize iteracijsko. Najprej v diagonalo korelacijske matrike namestimo neke ocene komunalitet. Komunalitete lahko ocenimo na več načinov, npr. z največjim koeficientom korelacije v vrstici korelacijske matrike ali z multiplim koeficientom korelacije posamezne spremenljivke s preostalimi spremenljivkami. Nato določimo uteži skupnih faktorjev A tako, da izračunamo lastne vrednosti in lastne vektorje korelacijske matrike z ocenjenimi komunalitetami na diagonali, pri čemer (tako kot pri metodi glavnih komponent) predstavljajo lastne vrednosti variance skupnih faktorjev in lastni vektorji njihove uteži. Na osnovi izračunanih uteži lahko izračunamo komunalitete, ki jih ponovno vstavimo v diagonalo korelacijske matrike. Ponovno izračunamo lastne vrednosti in lastne vektorje na novo popravljene korelacijske matrike itd.

Na žalost ni dokazano, da ta postopek vedno skonvergira k pravi rešitvi, vendar ponavadi da dobre rezultate. Z metodo glavnih osi torej enolično določimo matriki A in Ψ, kar z drugimi besedami pomeni, da v prostoru merjenih spremenljivk zakoličimo skupni prostor, tako da je varianca prvega dobljenega skupnega faktorja največja, od vseh možnih skupnih faktorjev, pravokotnih na prvi faktor, je izbran drugi skupni faktor z največjo varianco itd.

Poznanih je več drugih metod faktorske analize za oceno matrik A in Ψ (npr. metoda največjega verjetja, image faktorska analiza, alfa faktorska analiza, kanonična faktorska analiza).

2. Rotacije

Ob reševanju splošne faktorske enačbe

$$\Sigma = AA' + \Psi$$

kjer je znana matrika Σ in neznani matriki A in Ψ, smo ugotovili, da matrike A ne moremo enolično oceniti. Zato različne metode faktorske analize z dodatnimi pogoji enolično poiščejo matriko A. Ko so z izbrano metodo poiskani skupni faktorji, lahko pozabimo na privzete dodatne pogoje. Če dobljene rešitve ne moremo dobro interpretirati, lahko dobljeno rešitev v skupnem prostoru (določenim z dobljenimi skupnimi faktorji) transformiramo, zarotiramo. To pomeni, da dobljeno matriko A pomnožimo z neko transformacijsko matriko M

$$A^* = AM$$

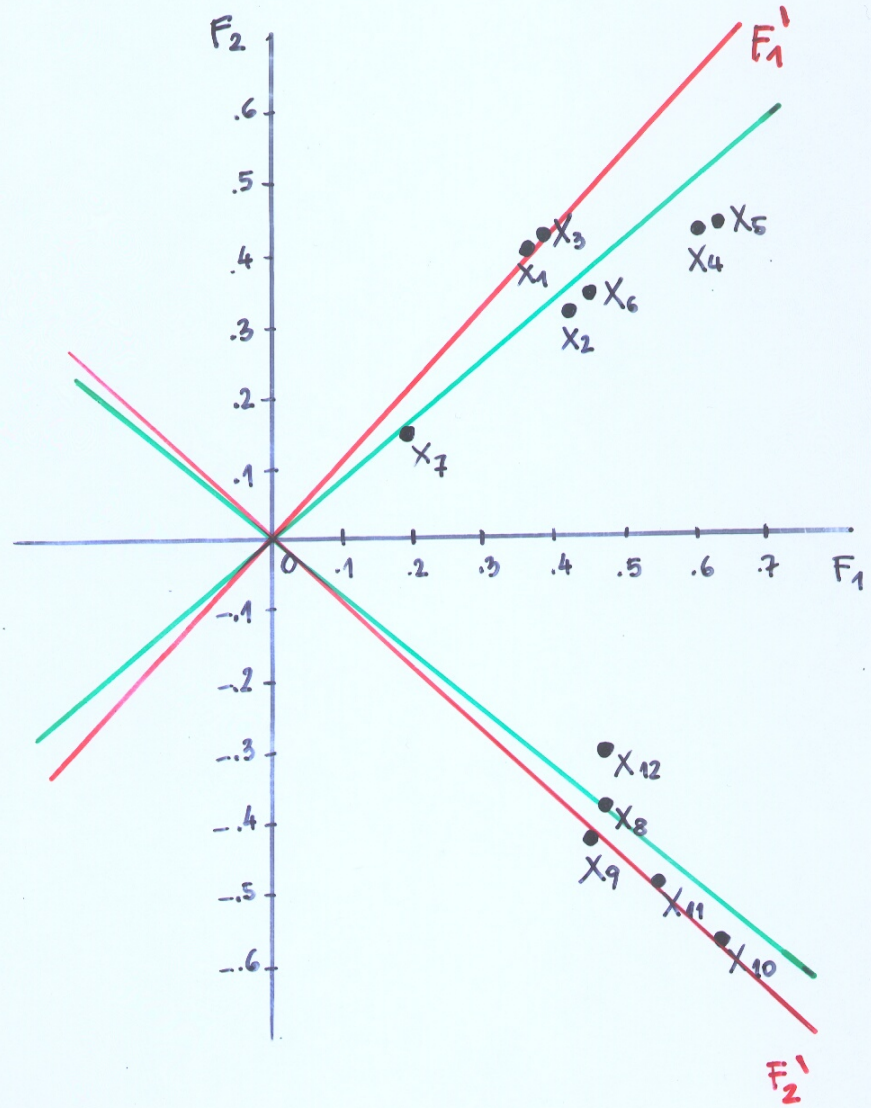
Rešitev A^* enako dobro reproducira originalne podatke kot prvotna rešitev A.

Za rotacijo se odločimo predvsem takrat, ko ne moremo smiselno interpretirati dobljene skupne faktorje. Npr., ko so projekcije iste spremenljivke precejšnje na več faktorjih, ali pa če so projekcije na prvem faktorju vseh spremenljivk precejšnje (splošen faktor).

Vzemimo primer dvanajstih dejavnikov, ki naj bi vplivali na poslovni uspeh malih podjetih v Sloveniji (J. Prašnikar (1994): Drobno gospodarstvo v Sloveniji). Z metodo glavnih osi smo dobili dva skupna faktorja, ki ju predstavljamo v naslednji tabeli:

	F_1	F_2	h^2
X_1 – PROD-MET	.38	.41	.31
X_2 – MARK-MET	.42	.32	.28
X_3 – PRODUKT	.39	.42	.33
X_4 – ODNOSI	.60	.43	.54
X_5 – USP-ZAP	.63	.44	.58
X_6 – USP-MAN	.46	.33	.32
X_7 – DRUZINA	.20	.15	.06
X_8 – GOSP-ZDR	.47	-.39	.37
X_9 – POL-ZVE	.46	-.44	.41
X_{10} – LOK-OBL	.63	-.59	.75
X_{11} – DRZAVA	.54	-.50	.55
X_{12} – PODJETJA	.48	-.30	.32
λ_i	2.82	2.00	
% p.v.	23.5	16.7	40.1

Prvi faktor je splošni faktor, drugi pa bipolarni. Za lažji vpogled v dobljeno faktorsko strukturo rešitev predstavimo v koordinatnem sistemu, kjer sta koordinatni osi oba dobljena skupna faktorja, točke pa spremenljivke:



- OSNOVNA FAKTORSKA REŠITEV
- PRAVOKOTNA ROTACIJA
- POŠEVNA ROTACIJA

Na sliki je vidno, da imamo dve izraziti skupini spremenljivk: X_1 do X_6 in X_8 do X_{12}. Spremenljivka X_7 (podpora družine) se ne uvršča v nobeno od omenjenih skupin. Tako izrazita razvrstitev spremenljivk ni razvidna iz tabele. Če zavrtimo koordinatni osi, tako da gredo kar se da skozi točke, ki predstavljajo spremenljivke, dobimo drugačne faktorske uteži:

	F_1	F_2
X_1 – PROD-MET	**.56**	-.03
X_2 – MARK-MET	**.52**	.06
X_3 – PRODUKT	**.57**	-.03
X_4 – ODNOSI	**.73**	.11
X_5 – USP-ZAP	**.75**	.13
X_6 – USP-MAN	**.56**	.09
X_7 – DRUZINA	.25	.03
X_8 – GOSP-ZDR	.06	**.60**
X_9 – POL-ZVE	.02	**.64**
X_{10} – LOK-OBL	.04	**.86**
X_{11} – DRZAVA	.03	**.74**
X_{12} – PODJETJA	.13	**.55**

Dobljene faktorske uteži so izrazitejše in rešitev je mogoče lažje razložiti. Prvi faktor ima izrazite uteži na izboljšavah produkcijskih in marketinških metodah ter samih produktov, dobrih odnosih med zaposlenimi in usposobljenosti zaposlenih in managementa. Drugi faktor pa na podpori gospodarskih združenj, lokalnih oblasti, države in drugih podjetih ter na zvezah v politiki. Podpora družine ni izrazita v nobenem od skupnih faktorjev. Prvi faktor torej obsega dejavnike za poslovni uspeh **znotraj podjetja**, drugi faktor pa obsega dejavnike v obliki podpore **izven podjetja**.

Bistvo rotiranja je, da dobimo teoretično pomembne faktorje in čim enostavnejšo faktorsko strukturo. Thurston je postavil nekaj **osnovnih načel** za iskanje take enostavne strukture:

1. vsaka vrstica v faktorski matriki A naj ima vsaj eno ničlo;
2. če je k skupnih faktorjev, naj ima vsak faktor v matriki vsak k ničel;
3. za vsak par faktorjev v matriki naj bo več spremenljivk, ki imajo močne uteži v enem stolpcu in majhne na ostalih;
4. za vsak par faktorjev v matriki naj ima velik del spremenljivk majhne uteži na obeh faktorjih (če je 4 ali več vseh faktorjev);
5. za vsak par faktorjev v matriki naj bo le majhen del spremenljivk z utežmi različnimi od 0 na obeh faktorjih (če je 4 ali več vseh faktorjev).

Postopki rotacij prevedejo Thurstonove enostavne kriterije na optimiziranje ustreznih kriterijskih funkcij, ki dajo rotirane faktorske rešitve.

Ločimo dve vrsti rotacij:

- pravokotne (rotirani faktorji so neodvisni med seboj) in
- poševne (rotirani faktorji so odvisni med seboj).

Pravokotne rotacije

Poznamo vsaj tri pravokotne rotacije:

- **QUARTIMAX** prevede problem na maksimizacijo četrtih potenc faktorskih uteži. Ta rotacijski postopek poenostavlja strukturo po vrsticah v faktorski matriki. Posledica je, da je običajno prvi dobljeni faktor splošni.
- **VARIMAX** maksimizira varianco kvadratov uteži v vsakem faktorju in s tem poenostavlja strukturo po stolpcih.
- **EQUIMAX** poenostavlja strukturo po vrsticah in stolpcih.

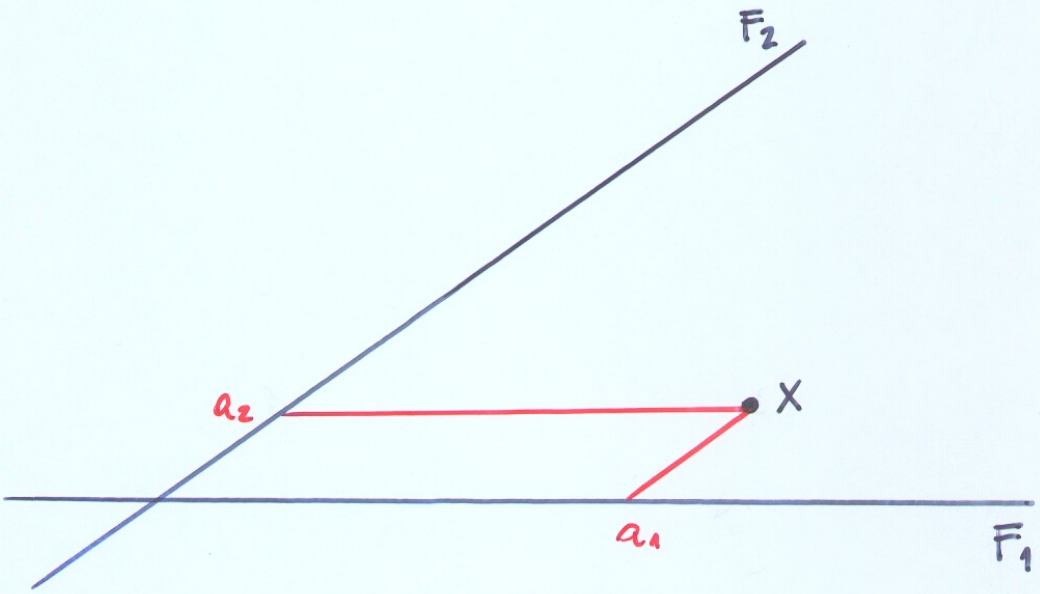
Poševne rotacije

Načela, ki so osnova posameznim postopkom poševne rotacije, so podobna kot pri pravokotnih rotacijah z razliko, da so v tem primeru rotirani faktorji korelirani med seboj. Poznanih je več postopkov poševne rotacije: OBLIMIN, OBLIMAX, QUARTIMIN, COVARMIN in BIQUARTIMIN. Nobeden od postopkov ni bistveno boljši od preostalih.

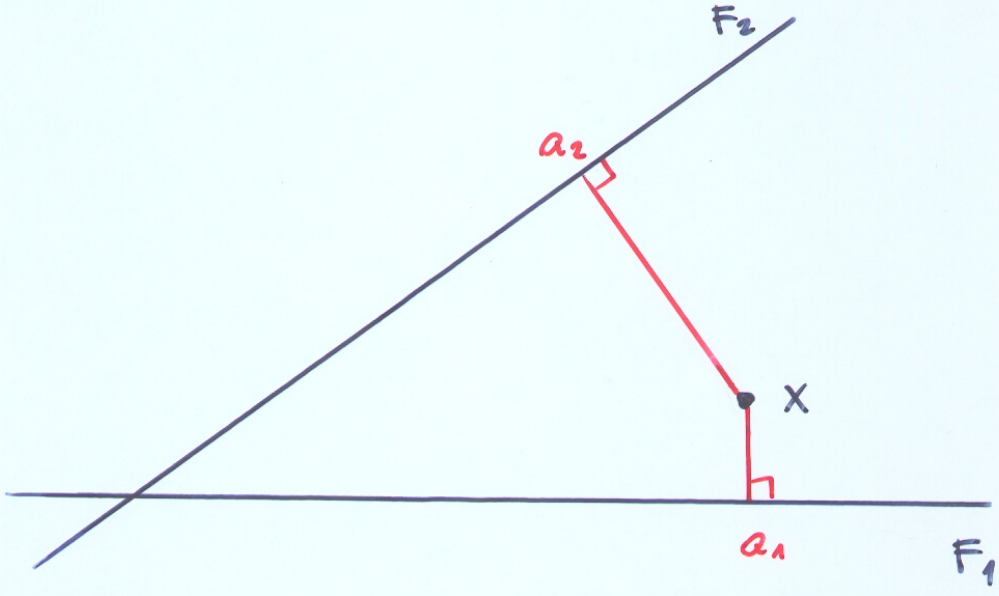
V primeru poševnih rotacij (grafično to pomeni, da kot med faktorjema, ki sta predstavljena s koordinatnima osema, ni pravi kot) lahko spremenljivke (točke v poševnem koordinatnem sistemu) projiciramo na poševne faktorje na dva načina:

- vzporedno, s čemer dobimo **'pattern' uteži** in
- pravokotno, s čemer dobimo **strukturne uteži**, ki so koeficienti korelacije med spremenljivko in faktorjem.

V primeru pravokotnih faktorjev so 'pattern' in strukturne uteži enake.



'PATTERN' UTEŽI



STRUKTURNE UTEŽI

Faktorske vrednosti (angl. factor scores)

Ocenjeno: A, Ψ (h_i^2)

Še ni ocenjeno: $F_{n\times k} = [f_{ij}]$

f_{ij} je vrednost j-tega faktorja na i-ti enoti.

$$F \neq \text{linearna kombinacija } X_i$$

$$\boxed{X_i = \text{linearna kombinacija } F_j + E_i}$$

Regresijska ocena faktorskih vrednosti

$$\boxed{\hat{F} = XB}$$

B je potrebno oceniti. Vzemimo, da so vse spremenljivke standardizirane. V vektorju B so standardizirani regresijski koeficienti.

Pomnožimo zgornjo enačbo z leve z $1/nX'$:

$$1/nX'\hat{F} = 1/nX'XB$$

$$1/nX'\hat{F} = \Sigma B$$

Levo je izraz enak strukturni matriki faktorskih uteži:

$$A = \Sigma B \Longrightarrow B = \Sigma^{-1}A$$

V matriki B so regresijski koeficienti (angl. factor score coefficients).

Ocena faktorskih vrednosti je torej

$$\hat{F} = X\Sigma^{-1}A$$