Kursus: Mitmemõõtmeline statistika

Seminar VIII: Tunnuste grupeerimine

Faktoranalüüs

Õppejõud: Katrin Niglas
PhD, dotsent
informaatika instituut

TALLINNA ÜLIKOOL

Tunnuste grupeerimine

Erinevad eesmärgid/olukorrad, miks **tunnuseid koondada**:

Testide korral (üksikküsimus, alamosa, kogu test):

- õigete/valede vastuste arvu leidmine
- punktide summeerimine

Ankeetide korral (sisuliselt haakuvate küsimuste plokk):

- teatud tüüpi vastuste arvu leidmine (nt "jah", vastamata,...)
- teatud küsimusteploki põhjal koondtunnuse arvutamine (nt keskmine hinnang, summaarne kulu, ...)
- **latentsete** e peidetud muutujate/tunnuste leidmine (nt hoiakud fenomeni eri aspektide suhtes, psühho-somaatilised seisundid, ...)

TALLINNA ÜLIKOOL

Tunnuste grupeerimine – näide "keelehoiakud"

Fail: Keelehoiakud.sav

Ankeetküsitlus koolinoorte keelehoiakute analüüsimiseks,
Keelekäitumise küsimused 5-palli skaalal (8 küsimust)
Keelehoiakute küsimused 4-palli skaalal (orig 35 küsimust)

Ül1: Leida keelehoiakute erinevad aspektid ja moodusta iga aspekti kohta koondtunnus/summamuutuja!

? Millised tunnused omavahel kokku panna?

Sisuline sobivus + statistiline sobivus (väljendub seoste tugevuses)

TALLINNA ÜLIKOOL

Tunnuste grupeerimine – latentsed muutujad

Latentsete e peidetud muutujate/tunnuste leidmine – sisuliselt otsitakse ühisosa omavaid algtunnuseid ja vajadusel moodustatakse omavahel sobivatest tunnustest koondtunnus.

Eelsamm: Uuri seoseid (seosekordajad, diagrammid, risttabelid, ...)

Alternatiiv1: Moodusta suurima korrelatsiooni tee

Alternatiiv2: Kasuta hierarhilist klasteranalüüsi

Alternatiiv3: Kasuta faktoranalüüsi

Järelsamm: Kontrolli tunnuste_grupi/koondtunnuse usaldusväärsust (Cronbach'i α)

TALLINNA ÜLIKOOL

Tunnuste grupeerimine – faktoranalüüs

Idee:

Eeldatakse, et algtunnused omavad ühisosa ja on seetõttu kirjeldatavad mingite ühiste muutujate e faktorite abil:

$$X_1 = l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + e_1$$
$$X_2 = l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + e_2$$
$$\ldots$$
$$X_k = l_{k1}F_1 + l_{k2}F_2 + \ldots + l_{km}F_m + e_k$$

X_i	algtunnused (stand.)	k	algtunnuste arv
F_j	ühised faktorid (stand.)	m	faktorite arv
l_{ij}	faktorite kordajad, mida nimetatakse **faktorkaaluks**		
e_i	algtunnuste **omapära**		

TALLINNA ÜLIKOOL

Faktoranalüüs - eeldused

Eeldused:

- arvtunnused (ja binaarsed e kahe väärtusega tunnused)
- olemas omavahel seotud tunnuste grupid
- lineaarsed seosed ja vastavad eeldused
- objektide arv soovitavalt üle 300
  (kirjeldava meetodina võib kasutada ka väiksemate valimite korral)
- objekte vähemalt 3X rohkem kui tunnuseid
- mitmemõõtmeline normaaljaotus
  (selle eelduse mittetäidetus mõjutab analüüsi tulemust üldjuhul vähe)

TALLINNA ÜLIKOOL

Faktoranalüüs – mille poole püüdleme!?

Idee:

Eeldatakse, et algtunnused omavad ühisosa ja on seetõttu kirjeldatavad mingite ühiste muutujate e faktorite abil:

$$X_1 = l_{11}F_1 + l_{12}F_2 + \ldots + l_{1m}F_m + e_1$$
$$X_2 = l_{21}F_1 + l_{22}F_2 + \ldots + l_{2m}F_m + e_2$$
$$\ldots$$
$$X_k = l_{k1}F_1 + l_{k2}F_2 + \ldots + l_{km}F_m + e_k$$

Ülesanne:

- Leida ühisosa kirjeldavad faktortunnused F_j
- Leida sobivad kordajad ehk faktorkaalud l_{ij}
- Leida **kommunaliteet**, ehk algtunnuse variatiivsuse see osa, mis on kirjeldatud faktorite poolt
  (kommunaliteet = tunnuse variatiivsus – omapära)

TALLINNA ÜLIKOOL

Faktoranalüüs - põhisammud

Põhisammud:

1. Standardiseerida algtunnused (SPSS'is tehakse vaikimisi)
2. Moodustada algtunnustest lineaarkombinatsioonid, mis annavad meile esialgsed faktorid
3. Visata välja need lineaarkombinatsioonid, mille variatiivsus on väike ja leida algtunnuste kommunaliteedid
4. Pöörata järelejäänud algfaktoreid nii, et iga algtunnus oleks võimalikult tugevalt seotud ainult ühega faktoritest
5. Arvutada välja faktortunnuste väärtused ja anda faktoritele sobivad nimed

TALLINNA ÜLIKOOL

Faktoranalüüs – esialgsete faktorite leidmine

2. Moodustada algtunnustest lineaarkombinatsioonid, mis annavad meile esialgsed faktorid => **Peakomponentide meetod:**

$$C_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1m}X_m$$
$$C_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2m}X_m$$
$$\ldots$$
$$C_k = a_{k1}X_1 + a_{k2}X_2 + \ldots + a_{km}X_m$$

Lineaarkombinatsioonide moodustamise tingimused:

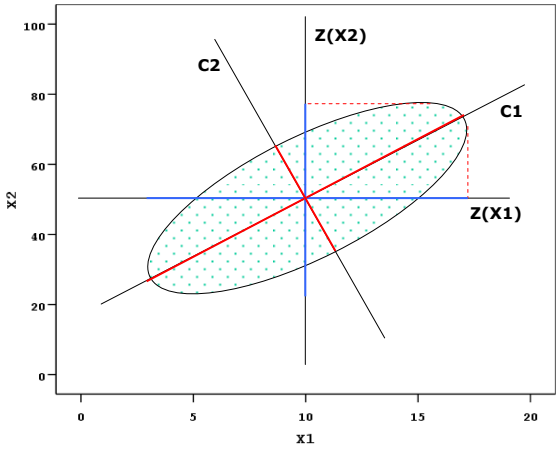
- Esimese peakomponendi dispersioon e **omaväärtus** (*eigenvalue*) on võimalikult suur; igal järgmisel väiksem, aga võimalikult suur
- Peakomponentide vahel ei ole korrelatsiooni

TALLINNA ÜLIKOOL

Faktoranalüüs - esialgsete faktorite leidmine

1. Standardiseerida algtunnused (SPSS'is tehakse vaikimisi)
2. Moodustada algtunnustest lineaarkombinatsioonid, mis annavad meile esialgsed faktorid => **Peakomponentide meetod:**

- Esimese peakomponendi dispersioon e **omaväärtus** (*eigenvalue*) on võimalikult suur; igal järgmisel väiksem, aga võimalikult suur
- Peakomponentide vahel ei ole korrelatsiooni



TALLINNA ÜLIKOOL

Faktoranalüüs – faktorite eraldamine

3. Visata välja need lineaarkombinatsioonid, mille variatiivsus on väike

$$C_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1m}X_m$$
$$C_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2m}X_m$$
$$\ldots$$

Esimesed mudelisse

$$\ldots$$
$$C_k = a_{k1}X_1 + a_{k2}X_2 + \ldots + a_{km}X_m$$

Viimased välja

Suure variatiivsusega peakomponentide eraldamise tingimus:

- peakomponendi dispersioon e **omaväärtus** (*eigenvalue*) on >1 ehk peakomponent kirjeldab rohkem, kui üksik algtunnus
- Mudelisse jäänud peakomponendid kokku kirjeldavad üle 60% algtunnuste variatiivsusest

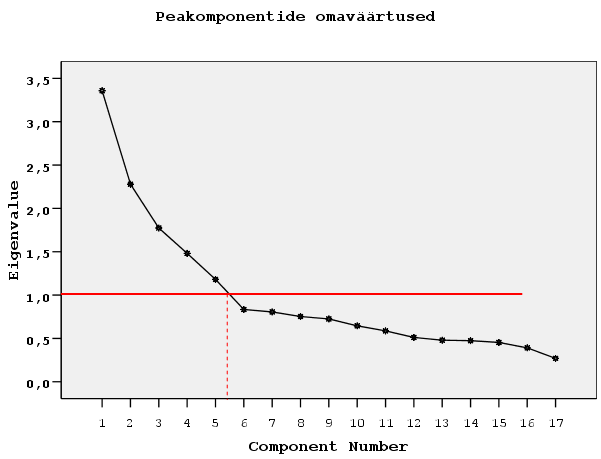
TALLINNA ÜLIKOOL

Total Variance Explained[a]

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,357	19,747	19,747	3,357	19,747	19,747
2	2,279	13,404	33,151	2,279	13,404	33,151
3	1,773	10,432	43,584	1,773	10,432	43,584
4	1,481	8,712	52,296	1,481	8,712	52,296
5	1,181	6,945	59,241	1,181	6,945	59,241
6	,834	4,904	64,145			
7	,806	4,739	68,884			
8	,753	4,429				
9	,725	4,263				
10	,645	3,796				
11	,587	3,452				
12	,511	3,009				
13	,480	2,823				
14	,474	2,788				
15	,453	2,668				
16	,391	2,302				
17	,270	1,586				

Extraction Method: Principal Component Analysi

a. Only cases for which Emakeel = eesti are u



TALLINNA ÜLIKOOL

Faktoranalüüs – teisendused ja omapärad

$$C_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1m}X_m$$
$$C_2 = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2m}X_m$$

Esimesed mudelisse

...

...

$$C_k = a_{k1}X_1 + a_{k2}X_2 + \ldots + a_{km}X_m$$

Viimased välja

Maatriks keeratakse ümber, ...

tehakse vajalikud matemaatilised teisendused ja ...

mudelist välja jäetud pea-komponentide poolt kirjeldatud algtunnuste variatiivsus loetakse algtunnuste omapäraks!

$$X_1 = l_{11}F_1 + l_{12}F_2 + \ldots + e_1$$
$$X_2 = l_{21}F_1 + l_{22}F_2 + \ldots + e_2$$
$$\ldots$$
$$X_k = l_{k1}F_1 + l_{k2}F_2 + \ldots + e_k$$

TALLINNA ÜLIKOOL

Faktoranalüüs - kommunaliteedid

Kommunaliteet on algtunnuse variatiivsuse see osa, mis on kirjeldatud faktorite poolt
(kommunaliteet = tunnuse variatiivsus – omapära)

NB! Madala kommunaliteediga (<0,3) tunnustel on teistega väike ühisosa ja seetõttu nad ei sobi mudelisse!

$$X_1 = l_{11}F_1 + l_{12}F_2 + \ldots + e_1$$
$$X_2 = l_{21}F_1 + l_{22}F_2 + \ldots + e_2$$
$$\ldots$$
$$X_k = l_{k1}F_1 + l_{k2}F_2 + \ldots + e_k$$

Communalities[a]

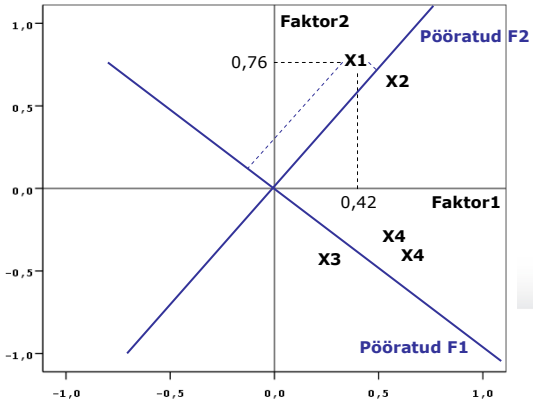
	Initial	Extraction
9 Kas eesti keele tunde peaks koolis rohkem olema?	1,000	,647
10 Kas inglise keele tunde peaks koolis rohkem olema?	1,000	,743
11 Kas osa õppeaineid võiks koolis olla inglise keeles?	1,000	,405
12 Kas Eestis peaks olema võimalus soovi korral õppida ingliskeelses gümnaasiumis?	1,000	,548
13 Kas Eestis peaks olema võimalus soovi korral omandada kõrgharidus inglise	1,000	,640

TALLINNA ÜLIKOOL

Faktoranalüüs – faktorite pööramine

NB! Iga peakomponent st esialgne faktor on võimalikult palju seotud kõigi algtunnustega => faktoreid ei saa hästi tõlgendada

4. Pöörata mudelisse jäänud esialgseid faktoreid nii, et iga algtunnus oleks võimalikult tugevalt seotud ainult ühega faktoritest



TALLINNA ÜLIKOOL

Component Matrix[a,b]

	Component				
	1	2	3	4	5
14 Kas Eesti riik peaks finantseerima eestikeelse keskhariduse kõrval ka ingliskeelset keskharidust?	,617	,546			,165
15 Kas Eesti riik peaks finantseerima eestikeelse kõrghariduse kõrval ka ingliskeelset kõrgharidust?	,608	,485			,148
13 Kas Eestis peaks olema võimalus soovi korral omandada kõrgharidus inglise keeles?	,600	,451	,205	-,130	,131
12 Kas Eestis peaks olema võimalus soovi korral õppida ingliskeelses gümnaasiumis?	,559	,407	,152	-,183	,116
37 Kas sind häiriks, kui su õde või vend armuks mõnda kohalikku muulasesse?	-,539	,397	,275	,272	

Rotated Component Matrix[a,b]

	Component				
	1	2	3	4	5
14 Kas Eesti riik peaks finantseerima eestikeelse keskhariduse kõrval ka ingliskeelset keskharidust?	,827		-,146		
13 Kas Eestis peaks olema võimalus soovi korral omandada kõrgharidus inglise keeles?	,791				
15 Kas Eesti riik peaks finantseerima eestikeelse kõrghariduse kõrval ka ingliskeelset kõrgharidust?	,778		-,122		
12 Kas Eestis peaks olema võimalus soovi korral õppida ingliskeelses gümnaasiumis?	,727	-,119			
36 Kas sind häiriks, kui olude sunnil peaksid ülikoolis ühiselamutuba jagama muulasega?		,762			
37 Kas sind häiriks, kui su õde või vend armuks mõnda kohalikku muulasesse?		,762	,105		

TALLINNA ÜLIKOOL

NB! Tunnused, mis peale faktorite pööramist on seotud ligikaudu sama tugevalt mitme faktoriga –> halb tõlgendada -> jäta välja!

Faktoranalüüs – faktortunnuste arvutamine

5. Arvuta faktortunnuste väärtused ja pane faktoritele sisuliselt sobivad nimed!

Rotated Component Matrix[a,b]

	Component				
	1	2	3	4	5
14 Kas Eesti riik peaks finantseerima eestikeelse keskhariduse kõrval ka ingliskeelset keskharidust?	,827		-,146		
13 Kas Eestis peaks olema võimalus soovi korral omandada kõrgharidus inglise keeles?	,791				
15 Kas Eesti riik peaks finantseerima eestikeelse kõrghariduse kõrval ka ingliskeelset kõrgharidust?	,778		-,122		
12 Kas Eestis peaks olema võimalus soovi korral õppida ingliskeelses gümnaasiumis?	,727	-,119			
36 Kas sind häiriks, kui olude sunnil peaksid ülikoolis ühiselamutuba jagama muulasega?		,762			
37 Kas sind häiriks, kui su õde või vend armuks mõnda kohalikku muulasesse?		,762			

*Keelehoiakud.sav [DataSet3] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

11 : FAC4_1 -1,73147831856455

	k40	FAC1_1	FAC2_1	FAC3_1
1	2,00	1,29777	-,10567	,35252
2	1,00	.	.	.
3	1,00	,57666	1,10369	-1,2355
4	1,00	1,31840	-,94606	,09645
5	1,00	,25296	1,28055	-,14744
6	1,00	1,27485	-,13304	-,08096
7	2,00	,64318	-,58953	,36540
8	2,00	1,17733	-,39717	1,30262

TALLINNA ÜLIKOOL

Faktoranalüüs - põhisammud

Põhisammud:

1. Standardiseerida algtunnused (SPSS'is tehakse vaikimisi)
2. Moodustada algtunnustest lineaarkombinatsioonid, mis annavad meile esialgsed faktorid
3. Visata välja need lineaarkombinatsioonid, mille variatiivsus on väike ja arvutada algtunnuste kommunaliteedid
4. Pöörata järelejäänud algfaktoreid nii, et iga algtunnus oleks võimalikult tugevalt seotud ainult ühega faktoritest
5. Arvutada välja faktortunnuste väärtused ja anda faktoritele sobivad nimed

TALLINNA ÜLIKOOL

Kursus: Mitmemõõtmeline statistika

Seminar X: Tunnuste ja objektide grupeerimine
Klasteranalüüs
Faktoranalüüs

Õppejõud: Katrin Niglas
PhD, dotsent
informaatika instituut

TALLINNA ÜLIKOOL

Praktikum: tunnuste ja objektide grupeerimine

Andmestik:	Muulased.sav
Raamat:	"Vene küsimus ja Eesti valikud", toim. Mati Heidmets, TPÜ Kirjastus 1998
Artikkel:	"Usaldus ja usaldamatus rahvussuhetes", Jüri Kruusvall, lk 29-76
Faktoranalüüs:	peatükk 2.3. Eesti elanike etniline häiritus (lk 36)
Klasteranalüüs:	peatükk 2.4. Eesti elanike etnilise häirituse tüpoloogia (lk41)

NB! Tabelid on artikli lõpus!

TALLINNA ÜLIKOOL

Praktikum: tunnuste ja objektide grupeerimine

Teine näide faktor- ja klasteranalüüsi kasutanud uurimuse põhjal kirjutatud artiklitest:

Ehala, M., Niglas, K. (2004) Eesti koolinoorte keelehoiakud. *Akadeemia*, 2004, 10, lk 2115-2143.

Ehala, Martin; Niglas, Katrin (2006). Language attitudes of Estonian secondary school students. *Journal of Language, Identity and Education*, 5(3), 209 - 227.

TALLINNA ÜLIKOOL