Matemaatiline statistika ja modelleerimine

Mitmemõõtmeline statistika

EMÜ doktorikool
DK.0007

Tanel Kaart

Peakomponent- ja faktoranalüüs
[*Principal Component Analysis*]

Faktoranalüüsi eesmärgiks on andmemahu vähendamine info koondamise teel, üldiste indeksite tekitamine, objektide/gruppide klasterdamine, …

Peakomponentanalüüs on faktoranalüüsi levinuim erijuht.

Kasutusala: kus iganes (sotsiaalteadustes, geneetikas, ökoloogias, majandusteaduses, meditsiinis, …)

Peakomponentanalüüs

• Eesmärgiks leida algsete tunnuste lineaarkombinatsioonid, mis kirjeldaksid võimalikult hästi ära algseis tunnustes sisalduva info.

• Taolisi lineaarkombinatsioone nimetatakse faktoriteks või peakomponentideks (*principal components*):

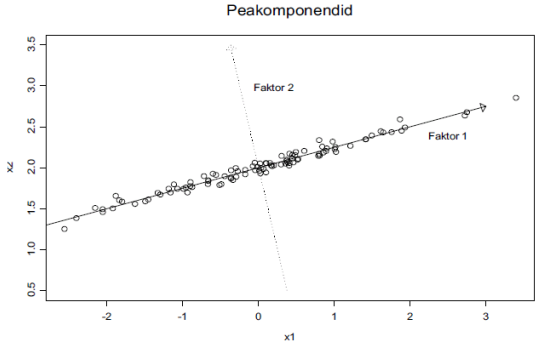
$$PC_1 = b_{11}x_1 + b_{21}x_2 + \ldots + b_{n1}x_n$$
$$PC_2 = b_{12}x_1 + b_{22}x_2 + \ldots + b_{n2}x_n$$
...
$$PC_k = b_{1k}x_1 + b_{2k}x_2 + \ldots + b_{nk}x_n$$

• Baseerub uuritavate tunnuste korrelatsiooni- või kovariatsiooni-maatriksite omaväärtusanalüüsil.

Peakomponentanalüüs

• Peakomponentanalüüs moodustab lineaarkombinatsioonid nii, et esimene peakomponent kirjeldaks ära võimalikult suure osa kõigi algtunnuste variatiivsusest, teine võimalikult suure osa alles jäänud variatiivsusest jne.

• Faktoranalüüs püüab leitud lineaarkombinatsioone täiendavalt modifitseerida, saavutamaks iga kombinatsiooni puhul maksimaalset korrelleeritust vaid ühega algtunnustest (nii on tulemusi lihtsam tõlgendada).



Peakomponentanalüüs – näide

PIIM

rasv_%	Ca_%	SOL Ca_sisaldus_%	SOL Ca_osakaal_%	P_%	kaseiin_%	pH	Valk_%	valk/rasv	RCT_min	E30_Pa	KA_%
2,40	0,1117	0,0349	31,24	0,0861	2,96	6,70	3,11	1,296	8,89	37,70	10,75
2,40	0,1062	0,0385	36,25	0,0837	3,06	6,81	3,15	1,313	7,89	47,92	10,91
2,44	0,1168	0,0329	28,17	0,0932	3,30	6,70	3,48	1,426	8,05	57,76	11,60
2,44	0,1157	0,0345	29,82	0,0923	3						
2,315	0,1115	0,0347	31,12	0,0951	3						
2,33	0,1116	0,0336	30,11	0,0857	3						

...

(kokku 12 erinevat piimaparameetrit)

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	7.15471893	4.03432229	0.5962	0.5962
2	3.12039665	2.43602205	0.2600	0.8563
3	0.68437460	0.25274238	0.0570	0.9133
4	0.43163222	0.06350002	0.0360	0.9493
5	0.36813220	0.21043953	0.0307	0.9799
6	0.15769267	0.10034312	0.0131	0.9931
7	0.05734955	0.03946902	0.0048	0.9979
8	0.01788053	0.01193299	0.0015	0.9993
9	0.00594754	0.00409184	0.0005	0.9998
10	0.00185570	0.00183630	0.0002	1.0000
11	0.00001940	0.00001940	0.0000	1.0000
12	0.00000000		0.0000	1.0000

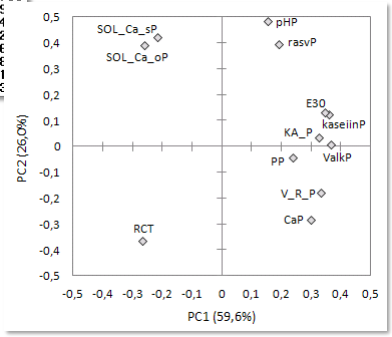
• Korrelatsioonimaatriksi omaväärtused (*eigenvalues*) näitavad faktori poolt kirjeldatavat varieeruvust ja selle osa koguvarieeruvusest.

• Et algsed tunnused analüüsi eel standardiseeritakse, on nende dispersioon võrdne ühega ja sestap omavad üksiktunnustest suuremat kirjeldusvõimet vaid faktorid, millele vastav omaväärtus > 1
(faktoranalüüsil jäetakse sisulistel kaalutlustel mõnikord mudelisse ka ühest väiksema omaväärtusega faktorid).

Peakomponentanalüüs – näide

Eigenvectors

	Prin1	Prin2	Prin3	Prin4
rasvP	0.192988	0.394139	-.363346	-.594933
CaP	0.302623	-.286834	-.100862	0.134676
SOL_Ca_sP	-.213512	0.419068	0.174607	0.379195
SOL_Ca_oP	-.258573	0.388006	0.146721	0.147925
PP	0.241582	-.046468	0.871225	-.347795
kaseiinP	0.360648	0.121031	-.103498	0.0859
pHP	0.155286	0.483246	0.112312	0.2034
ValkP	0.369573	0.004649	-.077821	0.1332
V_R_P	0.334944	-.180698	0.086889	0.4306
RCT	-.263180	-.368466	-.026361	0.0478
E30	0.346938	0.128166	-.091010	0.2481
KA_P	0.328304	0.031810	0.029155	-.1593



• Korrelatsioonimaatriksi omavektorid (*eigenvectors*) kujutavad enesest korrelatsioonikordajaid algsete tunnuste ja moodustatud faktorite vahel.

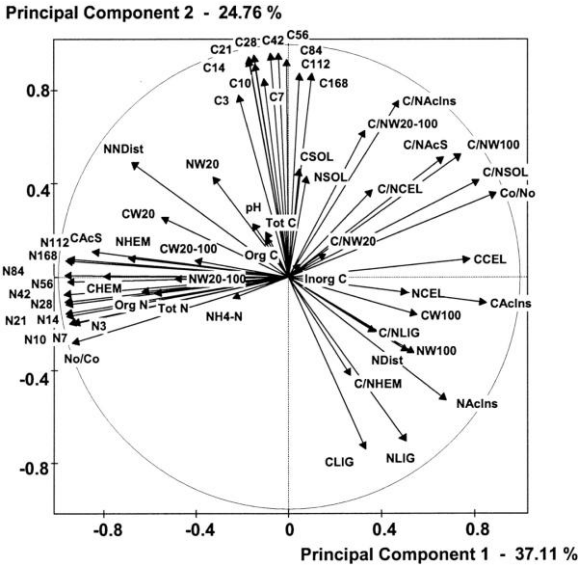
• Samuti on omavektori komponendid standardiseeritud algsete tunnuste kordajaiks peakomponendi/faktori väärtust määravas võrrandis – nn faktorlaadungid (*factor loadings*).

• Näiteks

PC_1 = 0,193*st(rasvP) + 0,303*st(CaP) – 0,214*st(SOL_CA_sp) – …,

siin st() tähendab standardiseeritud tunnust.

Peakomponentanalüüs



Virginie Parnaudeau, Bernard Nicolardota and Jérôme Pagès. Relevance of Organic Matter Fractions as Predictors of Wastewater Sludge Mineralization in Soil. Journal of Environmental Quality.

Peakomponentanalüüs

Landscape and Urban Planning 67 (2004) 175–183

Bumblebee communities as an indicator for landscape monitoring in the agri-environmental programme

Kalev Sepp [a,*], Merit Mikk [b], Marika Mänd [c], Jaak Truu [a]

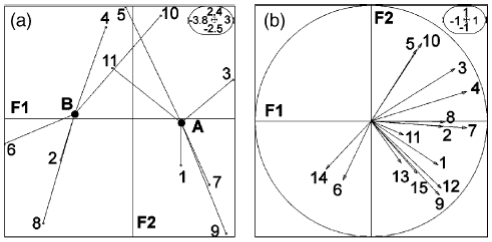
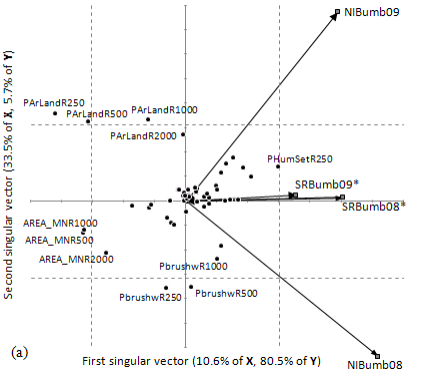
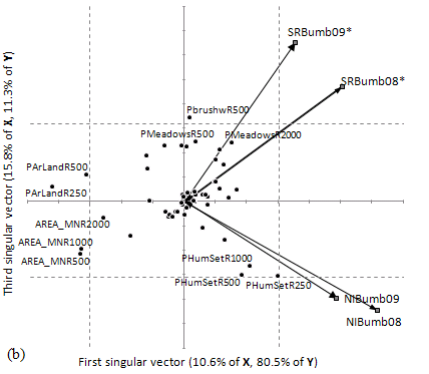


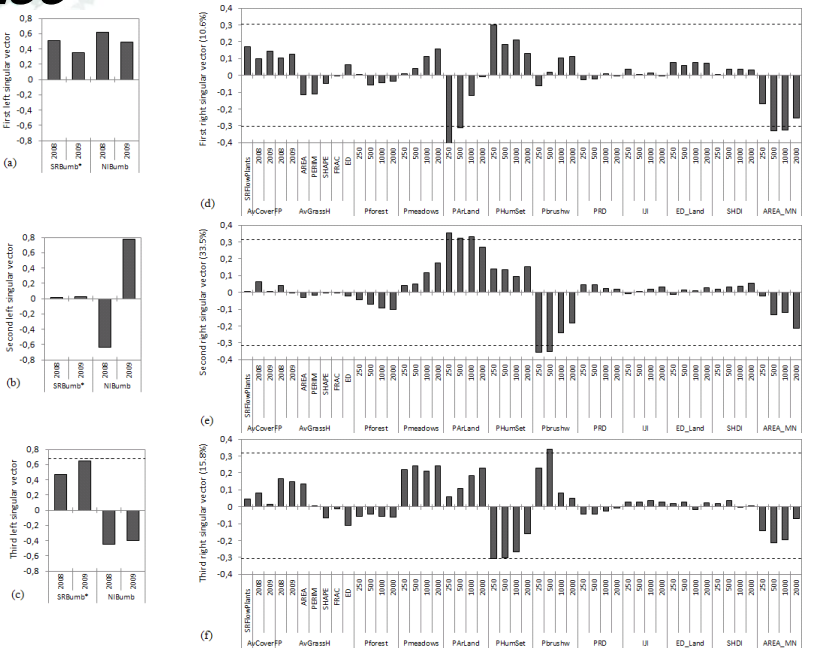
Fig. 2. (a) Ordination of studied sites according to the results of the PCA based on the regional bumblebee distribution. The numbers for the study areas are given in Table 4. A: Centroid of less intensively used agricultural area (LI) (<45% arable land); B: centroid of intensively used agricultural area (I) (>65% arable land). (b) Correlation of the bumblebee species variables with the two first axes of ($F1 \times F2$) of the PCA. (1) *B. lapidarius*, (2) *B. lucorum*, (3) *B. pascuorum*, (4) *B. veteranus*, (5) *B. schrencki*, (6) *B. ruderarius*, (7) *B. pratorum*, (8) *B. hortorum*, (9) *B. sylvarum*, (10) *B. soroeensis*, (11) *B. distinguendus*, (12) *B. hypnorum*, (13) *B. terrestris*, (14) *B. jonellus* and (15) *B. subterraneus*. The first and second principal components describe 30.9 and 18.7% of the overall data variation, respectively. The effect of the grouping factor is statistically highly significant ($P < 0.01$).

Osavähimruutude korrelatsioon (*partial least square correlation*, *PLSC*)





PLSC

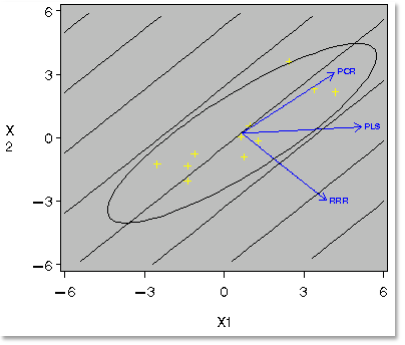


Mustrid (*patterns*)?

• Peakomponentide regressioon (*principal components regression*, *PCR*)

• Osavähimruutude korrelatsioon (*partial least squares correlation*, *PLSC*), kanooniline korrelatsioon (*canonical correlation*)

• Osavähimruutude regressioon (*partial least squares regression*, *PLSR*), *reduced rank regression* (*RRR*)



SAS Online Doc

Klasteranalüüs

Klasteranalüüsi eesmärgiks on kas tunnuste või uuritavate objektide rühmitamine. Kaks peamist algoritmi:

- hierarhiline klasterdamine;
- k-keskmiste klasterdamine.

Kasutusala: kus iganes (geneetika, ökoloogia, sotsioloogia, meditsiin, majandus, …)

Klasteranalüüs

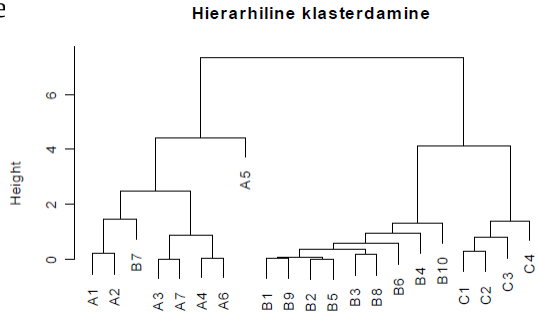
• Loodusuurija sõitis Asustamata Saarele ja avastas seal suure hulga seni teadusele tundmatuid putukaid.

• Ta mõõtis oma lühikese saarelviibimise jooksul tervel hunnikul seninägematutel putukatel igasuguseid näitajaid (igatsorti pikkuseid ja mustrielementide arvu ja paljut muudki).

• Järgmiseks soovis loodusuurija määratleda, mitmesse alamliiki leitud putukad võiksid kuuluda.

• Et saada esimest ligikaudset lähendit, kust oma uurimistööga pihta hakata, soovis ta leida sarnaste putukate rühmad – kes oleksid siis alamliikide kanditaatideks.

• Selleks söötis ta oma andmed klasteranalüüsi teostavasse programmi, mis joonistas järgmise pildi:



Märt Mölsi loengukonspektist

Klasteranalüüs

Hierarhiline klasterdamine

... on hästi kasutatav siis, kui meil on suhteliselt vähe objekte või kui on oodata, et klastrid suhteliselt selgelt üksteisest eristuvad.

Hierarhiline klasteranalüüs põhineb väga lihtsal algoritmil:

samm-sammult hakatakse omavahel kokku panema kõige sarnasemaid objekte. Näiteks, kui leidub kaks täpselt ühesuguste tulemustega objekti, siis liidetakse nad esimesel sammul üheks klastriks, peale seda võrreldakse kõiki üksikobjekte ja juba tekkinud klastreid ja liidetakse jälle kõige sarnasemad omavahel jne.

Vaatluste omavahelise kauguse määramine:

Eukleidese kaugus: $d\ (x_1, y_1), (x_2, y_2)\ = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

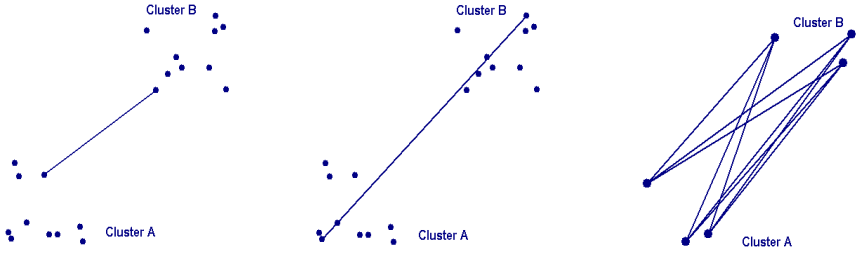
Manhattani kaugus: $d\ (x_1, y_1), (x_2, y_2)\ = |x_1 - x_2| + |y_1 - y_2|$

…

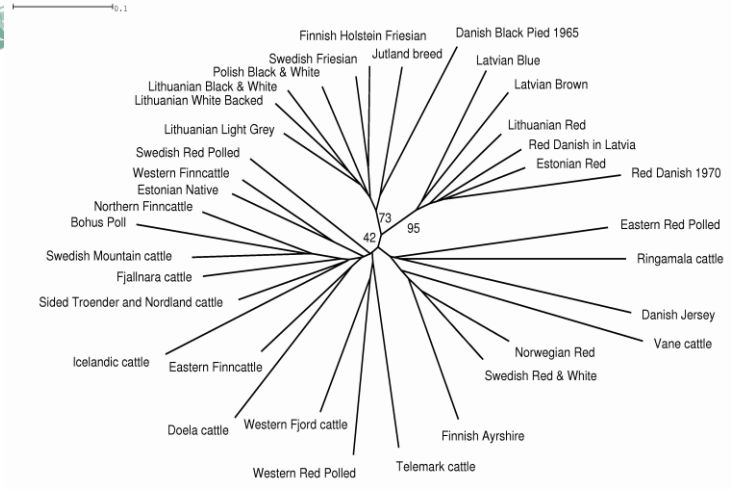
Klasteranalüüs

Klastritevahelise kauguse määramine:

"single" "complete" "average"



Hierarhiline klasterdamine – näide



Tapio et al 2006. Prioritization for Conservation of Northern European Cattle Breeds Based on Analysis of Microsatellite Data. Conservation Biology 20, 1768-1779.

Klasteranalüüs

k-keskmiste klasterdamine

... sobib grupeerimise meetodiks siis, kui objekte on nii palju, et hierarhilise klasteranalüüsi tulemus muutub ebaülevaatlikuks, kuid ka siis kui me oskame meile sobivat klastrite arvu ligilähedaselt ennustada ning ühtlasi soovime saada ka tekkivate klastrite kirjelduse nende tunnuste osas, mis on grupeerimise aluseks.

Algoritm:

• kõigepealt tuleb määrata klastrite arv, siis

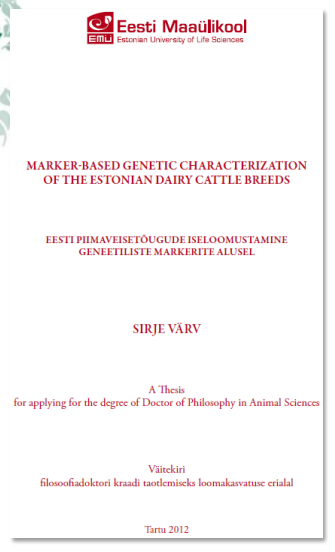
• jagada objektid esialgsetesse klastritesse,

• arvutada välja klastrite keskpunktid ning

• hakata võrdlema igat objekti kõigi klastrite keskpunktidega; kui osutub, et objekti kaugus mõne muu klastri keskpunktist on väiksem kui selle klastri keskpunktist,milles ta parasjagu asub, siis tuleb objekt teise klastrisse ümber tõsta;

• peale objekti ümbertõstmist tuleb pöörduda uuesti sammu 3 juurde ja jätkata protsessi niikaua kui kõik objektid on klastris, mille keskpunktile nad kõige lähemal asuvad.

K-keskmiste klasteranalüüs – näide


Eesti Maaülikool
Emu Estonian University of Life Sciences

MARKER-BASED GENETIC CHARACTERIZATION OF THE ESTONIAN DAIRY CATTLE BREEDS

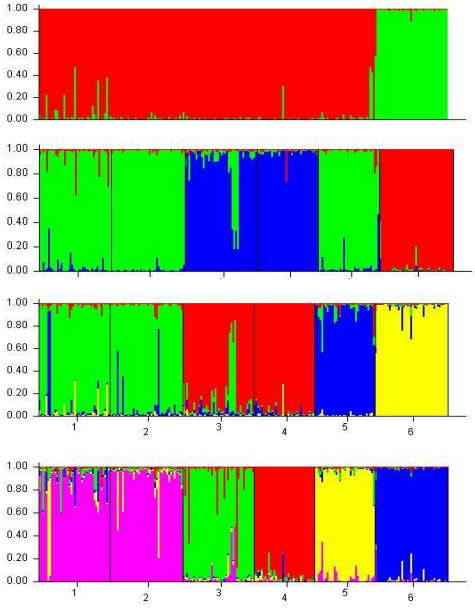
EESTI PIIMAVEISETÕUGUDE ISELOOMUSTAMINE GENEETILISTE MARKERITE ALUSEL

SIRJE VÄRV

A Thesis
for applying for the degree of Doctor of Philosophy in Animal Sciences

Väitekiri
filosoofiadoktori kraadi taotlemiseks loomakasvatuse erialal

Tartu 2012



1 – EN,
2 – WFC,
3 – ER,
4 – EHF,
5 – SwRP,
6 – DkJer

Diskriminantanalüüs

Diskriminantanalüüsi eesmärgiks on objektide rühmitamine nendel mõõdetud tunnuste alusel.

Seejuures on objektide klassidesse kuuluvus enne analüüsi teada (erinevalt peakomponent või klasteranalüüsist).

Diskriminantanalüüs

Objektidel mõõdetud tunnuste alusel kostrueeritakse nn diskremineeriv funktsioon, mis eristaks gruppe võimalikult selgelt:

$$d = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

Näiteks

2,3×"jalgade pikkus" + 6,7×"saba pikkus" – 2,8×"kere pikkus" + 5,1×"noka pikkus"

Kui saadud väärtus <10,2, siis on ilmselt tegu isase isendiga.

Täpsemalt öeldes hinnatakse tunnuste vektori $\mathbf{x}$ tihedusfunktsioon $f_t(\mathbf{x})$ igas grupis t ning arvutatakse iga objekti mingisse gruppi kuulumise tõenäosus Bayesi valemist kujul

$$\mathrm{P}(t \mid \mathbf{x}) = \frac{q_t f_t(\mathbf{x})}{\sum_u q_u f_u(\mathbf{x})},$$

misjärel määratakse iga objekt tõenäoliseimasse gruppi (suurus q_t eelnevas valemis on objekti gruppi t kuulumise algtõenäosus).

Diskriminantanalüüs – näide

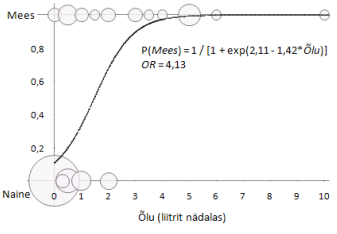
Jaanika Hõimra

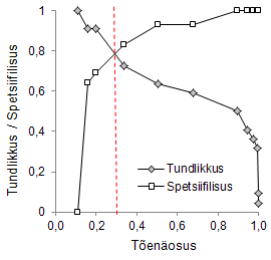
Morfomeetriliste tunnuste varieeruvuse sõltuvus keskkonnatingimustes kortslehe (*Alchemilla* L.) viiel mikroliigil eksperimendi tingimustes

Tabel 8. Klassifitseeriv diskriminantanalüüs 18 tunnuse alusel (2001). Ridades on antud empiiriliselt määratud liigid ja tulpades prognoos.

	%	KAREDA-KARVANE	VÄIKE	TERAVA-HÕLMINE	KÜÜT
KAREDAKARVANE	93,2	**262**	19	0	0
VÄIKE	99,7	1	**286**	0	0
TERAVAHÕLMINE	91,5	0	0	**259**	24
KÜÜT	87,4	0	0	36	**250**
KOKKU	92,9	**263**	**305**	**295**	**274**

Diskriminantanalüüs *vs* logistiline regressioon





Logistilise regressioonanalüüsi tulemus:
nii mehed kui ka naised identifitseeritakse õigesti 80%-lise tõenäosusega.

Et diskriminantanalüüs loeb objekti kuuluvaks suurima tõenäosusega gruppi, ei pruugi saadav klassifitseerimiseeskiri olla optimaalseim.

Diskriminantanalüüsi tulemus:
meestest identifitseeritakse õigesti 59,1%, naistest 92,9%.

Number of Observations and Percent Classified into SUGU

From SUGU	0	1	Total
0	39 92.86	3 7.14	42 100.00
1	9 40.91	13 59.09	22 100.00
Total	48 75.00	16 25.00	64 100.00

Korrespondentsanalüüs

(*correspondence analysis*)

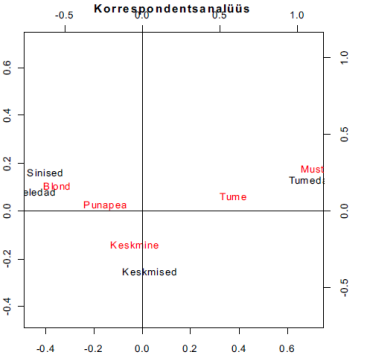
Korrespondentsanalüüs võimaldab graafiliselt kirjeldada sagedustabelite kujul esitatud tunnuste vahelisi seoseid.

Korrespondentsanalüüs

Andmestik (juuste ja silmade värv Caith'is, Šhotimaal):

Silmavärv\juuksevärv	**Blond**	**Punapea**	**Šatään**	**Brünett**	**Süsimust**
Sinine	326	38	241	110	3
Hele	688	116	584	188	4
Keskmine	343	84	909	412	26
Tumedad	98	48	403	681	85

Korrespondentsanalüüsi esitus:



Märt Mölsi loengukonspektist

Mitmemõõtmeline skaleerimine

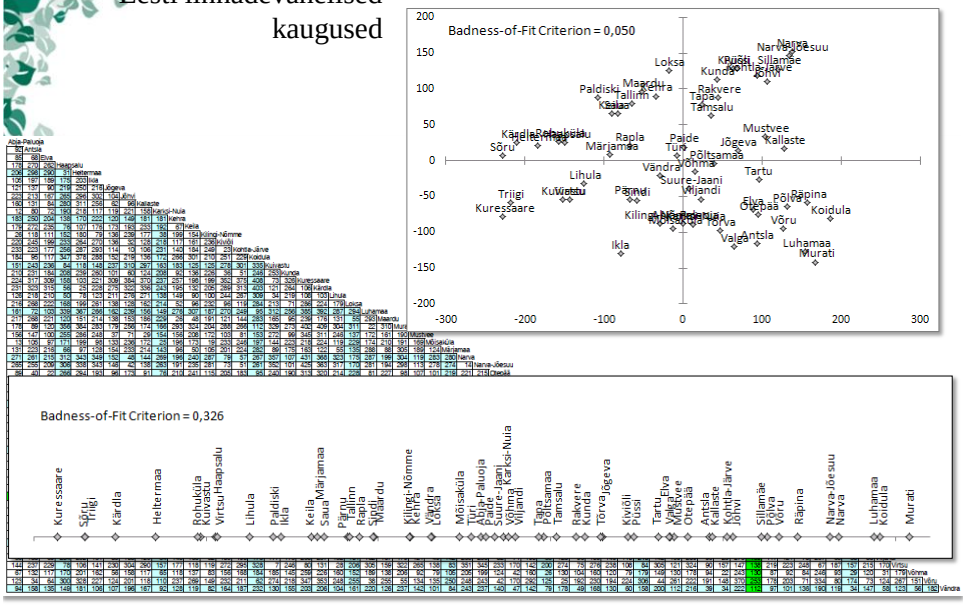
(*multidimentional scaling*)

Mitmedimensionaalse skaleerimise eesmärgiks on objektide erinevuste või sarnasuste teisendamine distantsiks mitmemõõtmelises ruumis ja saadu esitamine kaardina objektide omavahelisest paiknemisest.

Kasutusala: kus iganes (geneetikas, ökoloogias, majandusteaduses, sotsioloogias, …)

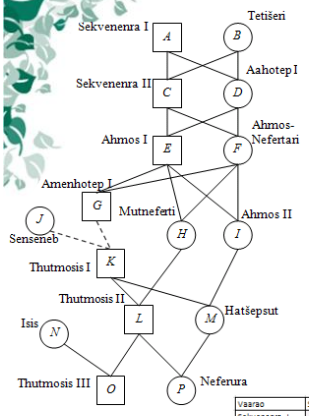
Mitmedimensionaalne skaleerimine

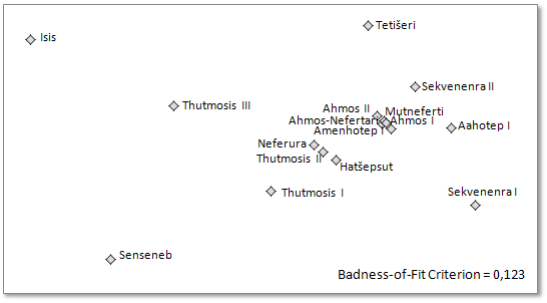
Eesti linnadevahelised kaugused



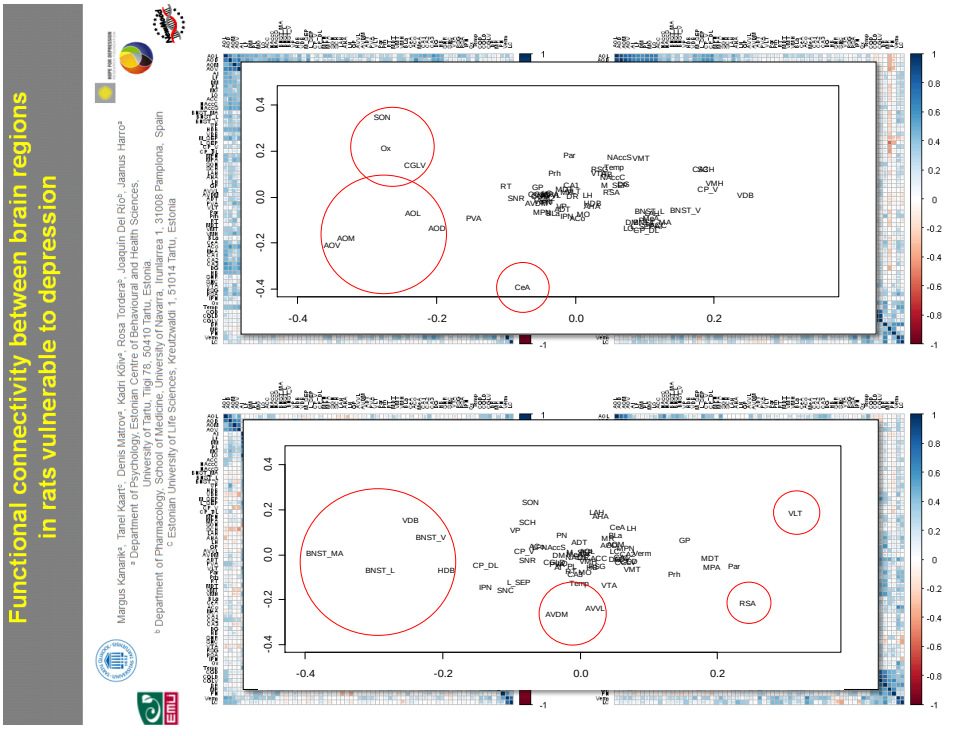
Mitmedimensionaalne skaleerimine

Egiptuse vaaraode 18. dünastia sugupuu algus





Vaarao	Sekvenenr	Tetisheri	Sekvenenr	Aahotep	Ahmos_I	Ahmos_Ne	Amenhote	Mutneferi	Ahmos_II	Senseneb	Thutmosis	Thutmosis	Hatshepsu	Isis	Thutmosis	Neferura
Sekvenenra_I	1	0	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0	0,25	0,375	0,375	0	0,1875	0,37
Tetisheri	0	1	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0	0,25	0,375	0,375	0	0,1875	0,37
Sekvenenra_II	0,5	0,5	1	0,5	0,75	0,75	0,75	0,75	0,75	0	0,375	0,5625	0,5625	0	0,28125	0,562
Aahotep	0,5	0,5	0,5	1	0,75	0,75	0,75	0,75	0,75	0	0,375	0,5625	0,5625	0	0,28125	0,562
Ahmos_I	0,5	0,5	0,75	0,75	1,25	0,75	1	1	1	0	0,5	0,75	0,75	0	0,375	0,7
Ahmos_Nefertari	0,5	0,5	0,75	0,75	0,75	1,25	1	1	1	0	0,5	0,75	0,75	0	0,375	0,7
Amenhotep_I	0,5	0,5	0,75	0,75	1	1	1,375	1	1	0	0,6875	0,84375	0,84375	0	0,421875	0,843
Mutneferti	0,5	0,5	0,75	0,75	1	1	1	1,375	1	0	0,5	0,9375	0,75	0	0,46875	0,843
Ahmos_II	0,5	0,5	0,75	0,75	1	1	1	1	1,375	0	0,5	0,75	0,9375	0	0,375	0,843
Senseneb	0	0	0	0	0	0	0	0	0	1	0,5	0,25	0,25	0	0,125	0,2
Thutmosis_I	0,25	0,25	0,375	0,375	0,5	0,5	0,6875	0,5	0,5	0,5	1	0,75	0,75	0	0,375	0,7
Thutmosis_II	0,375	0,375	0,5625	0,5625	0,75	0,75	0,84375	0,9375	0,75	0,25	0,75	1,25	0,75	0	0,625	
Hatshepsut	0,375	0,375	0,5625	0,5625	0,75	0,75	0,84375	0,75	0,9375	0,25	0,75	0,75	1,25	0	0,375	
Isis	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0,5	
Thutmosis_III	0,1875	0,1875	0,28125	0,28125	0,375	0,375	0,421875	0,46875	0,375	0,125	0,375	0,625	0,375	0,5	1	0,
Neferura	0,375	0,375	0,5625	0,5625	0,75	0,75	0,84375	0,84375	0,84375	0,25	0,75	1	1	0	0,5	1,37



Heatmap

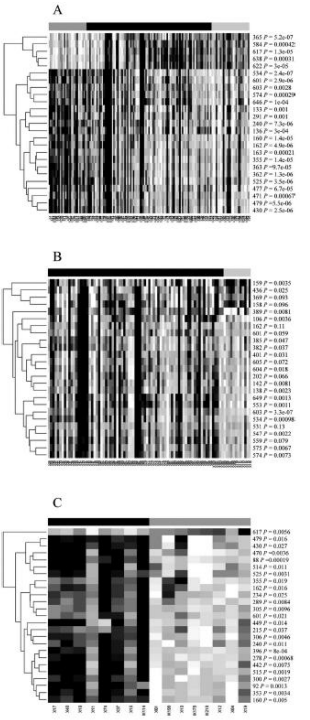


Figure 3. (A) Heat map of noncoagulating (dark gray bar on top left), intermediate (black bar on top center), and well-coagulating (light gray bar on top right) milk samples. Statistical comparison with Student *t*-test was calculated between well-coagulating and noncoagulating groups. The 25 most significantly different m/z values with respective *P*-values can be seen in the right-hand side. Rows are arranged into clusters (dendrogram on left-hand side) based on pattern similarities. The lighter the shading of a data point, the more intense the signal is compared with the mean value for the respective m/z among all analyses. Gray points are close to the mean and dark tones indicate signal intensities below the mean; (B) heat map of milks with poor (black bar) and good (gray bar) coagulation ability; (C) heat map of noncoagulating (black bar) and well-coagulating (gray bar) milks.

J. Dairy Sci. TBC:1–8
doi:10.3168/jds.2011-4468

Comparison of the metabolic profiles of noncoagulating and coagulating bovine milk

H. Harzia,*†‡[1] K. Kilk,† I. Jõudu,*‡ M. Henno,* O. Kärt,* and U. Soomets†
*Institute of Veterinary Medicine and Animal Sciences, Estonian University of Life Sciences, Kreutzwaldi 46, Tartu 51006, Estonia
†Department of Biochemistry, Medical Faculty, University of Tartu, the Centre of Excellence for Translational Medicine, Ravila 19, 50411 Tartu, Estonia
‡Bio-Competence Centre of Healthy Dairy Products, Kreutzwaldi 1, 51014 Tartu, Estonia