TARTU ÜLIKOOL

MATEMAATIKA-INFORMAATIKATEADUSKOND

MATEMAATILISE STATISTIKA INSTITUUT

Joosep Raudsik

Hajuvusdiagrammid ning korrelatsioonimaatriksite illustreerimine statistikapaketis R

Bakalaureusetöö

Juhendaja:

Tanel Kaart

TARTU

2013

Sisukord

Sissejuhatus

Tihti pakub meile huvi kahe arvtunnuse omavaheline käitumine. Mõnikord on see lihtsalt mõistetav – näiteks rohkem õppides on tulemused tihti paremad ja pikemad inimesed kaaluvad enamasti rohkem. Sageli ei ole aga seos selgelt etteaimatav ning selle olemasolu, tugevuse ja suuna hindamiseks tuleb arvutada erinevaid seosekordajaid. Viimaste sisuliseks mõistmiseks on aga lisaks uuringu valdkonna tundmisele vaja teadmisi ka statistikast. Et sugugi mitte kõik uuringu tulemustest huvitatud isikud statistika-alaseid teadmisi ei oma, on statistikul vajalik osata esitada tulemusi selgelt, visuaalselt atraktiivselt ja intuitiivselt mõistetavalt.

Juhul, kui uuritavaid tunnuseid on enam kui kaks, muutub vaid seosekordajate põhjal järelduste tegemine sageli keerukaks ka piisavalt statistikateadmisi omavale inimesele. On ju N tunnuse puhul kõikvõimalikke paarikaupa seoseid N(N-1)/2 – seega vaid kümne tunnuse puhul juba 45 –, millest kompaktse ülevaate saamine vaid arve vaadates on sageli pea võimatu ning appi tuleb võtta graafiline esitus.

Hajuvusdiagrammid ja korrelatsioonimaatriksid on statistikas laialdaselt kasutatavad vahendid tunnuste vaheliste seoste kirjeldamiseks. Antud töö eesmärgiks on anda ülevaade statistikapaketi R võimalustest hajuvusdiagrammide konstrueerimiseks ja korrelatsioonimaatriksite illustreerimiseks. Bakalaureusetöö esimesed kolm peatükki tutvustavad erinevaid võimalusi visualiseerimaks tunnuste vahelisi seoseid vaid hajuvusdiagrammide abil ilma mingeid seosekordajaid arvutamata. Esimene peatükk sisaldab ülevaadet kahemõõtmeliste hajuvusdiagrammide moodustamisest funktsioonide *plot* ja *scatterplot* abil ning kolmamõõtmeliste hajuvusdiagrammide moodustamisest funktsiooni *scatterplo3d* abil, teine peatükk annab ülevaate kõrge tihedusega hajuvusdiagrammide konstrueerimisest kasutades funktsioone *hexbin* ja *sunflowermatrix* ning kolmas peatükk kirjeldab hajuvusdiagrammide maatriksite konstrueerimise võimalusi funktsioonide *scatterplotmatrix* ja *pairs* abil. Töö teises pooles annab autor ülevaate tunnuste vahelisi seoseid kirjeldavate korrelatsioonimaatriksite illustreerimisest. Neljas peatükk sisaldab põhjalikku kirjeldust funktsioonist *corrplot*. Töö viimases osas on ära toodud ülevaade paarist R-s seni realiseerimata võimalust kirjeldada korrelatsioonimaatrikseid hulknurkade abil.

Töö autor eeldab lugejalt statistikapaketi R kasutusoskust ning kasutab statistikapaketi R versiooni 2.15.2. Töös kasutatud paketid, mis ei ole vaikimisi sisse loetud, on autori poolt iga

funktsiooni puhul ära mainitud ning sisse loetud funktsiooni *library* abil. Lühemad programmikoodid on ära toodud tekstis kastiga ümbritsetult ning pikemad töö lisades.

Sõnu parameeter ja argument kasutatakse töös samatähenduslikena funktsiooni argumendi mõistes ning sõna funktsioon sünonüümina on kasutatud töös terminit käsk.

Programmikoodid alternatiivsete korrelatsioonimaatriksite illustreerimismeetodite jaoks on inspireeritud juhendaja poolt ning teostatud autori poolt.

1. Hajuvusdiagrammid

Kahe arvtunnuse omavahelist seost illustreerivatest graafilistest vahenditest kasutatakse tänapäeval enim hajuvusdiagrammi ehk korrelatsioonivälja. See joonisetüüp on vaieldamatult mitmekülgseim ja üldiselt kõige intuitiivsemalt mõistetavam. Taolist illustratiivset meetodit kasutas juba Francis Galtoni korrelatsiooni ja regressiooni kontseptsiooni väljatöötamisel ning seega võib hajuvusdiagrammi pidada esimeseks kahedimensionaalseks graafiliseks võimaluseks kirjeldamaks kvantitatiivseid andmeid [1]. Statistikapakett R pakub mitmeid funktsioone, mille abil on võimalik konstrueerida hajuvusdiagramme saamaks visuaalset ettekujutust tunnuste vahelistest seostest.

1.1. Andmestik

Jooniste konstrueerimisel kasutab autor näiteandmestikuna 25 sportlase andmeid [2] 1988. aasta Souli suveolümpiamängude seitsmevõistlusest (vt. Lisa 1). Tunnuseid on üheksa, millest seitse on SI-süsteemis mõõdetud seitsmevõistluse alade tulemused: *hurdles, highjump, shot, run200m, longjump, javelin* ning *run800m*. Sportlaste lõpptulemused punktide näol ära toodud tunnuses *score*. Lisaks on moodustatud binaarne abitunnus *kat*, kus sportlastele, kelle võistlustulemus ületas 6200 punkti, on omistatud tunnuse väärtus 1 ning mitte nii headele sportlastele 0. Järgnevais peatükkides esitatud R-i käskude puhul on eeldatud, et antud andmestik on funktsiooni *attach* abil võetud kasutusele vaikeandmestikuna, mistõttu tunnuste nimede juures ei ole andmestiku nime täiendavalt näidatud. Tabelis 1.1 on ära toodud kasutatava andmestiku esimesed read.

Tabel 1.1. Väljavõte töös kasutatavast näiteandmestikust.

	hurdles	highjump	shot	run200m	longjump	javelin	run800m	score	kat
Joyner-Kersee	12.69	1.86	15.8	22.56	7.27	45.66	128.51	7291	1
John	12.85	1.8	16.23	23.65	6.71	42.56	126.12	6897	1
Behmer	13.2	1.83	14.2	23.1	6.68	44.54	124.2	6858	1
Sablovskaite	13.61	1.8	15.23	23.92	6.25	42.78	132.24	6540	1
Choubenkova	13.51	1.74	14.76	23.93	6.32	47.46	127.9	6540	1
Schulz	13.75	1.83	13.5	24.65	6.33	42.82	125.79	6411	1

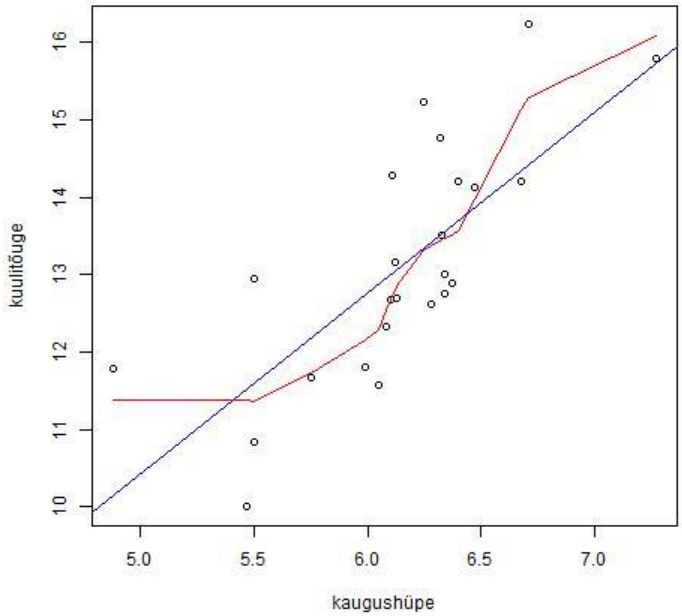
1.2. Funktsioon *plot*

Lihtsaima hajuvusdiagrammi saab R-s moodustada käsuga *plot* [3], mille argumendid on koordinaatide vektorid *x* ja *y*. Joonise konstrueerimisel kasutatavad lisaargumendid [4] nagu *col*, *xlab*, *xlim* jne töötavad tavapäraselt.

Seitsmevõistlejate kaugushüppe ja kuulitõuke tulemuste (tunnused *longjump* ja *shot*) põhjal koostatud korrelatsiooniväli on ära toodud joonisel 1.1. Tunnuste vahelist seost aitavad kirjeldada funktsiooni *plot* rakendamise tulemusena saadud graafikule lisatud regressioonisirge (sinine joon, hinnatud funktsiooniga *lm*) ja lokaalse kaalutud regressiooni abil silutud joon (punane joon, hinnatud funktsiooniga *lowess*). Et funktsioon *plot* taoliste joonte automaatset lisamist ei võimalda, tuleb need eraldi välja arvutada ja lisada käsuga *plot* konstrueeritud joonisele.

Joonis 1.1 on moodustatud järgnevate käsuridadega:

```
plot(longjump, shot, xlab="kaugushüpe", ylab="kuulitõuge")
abline(lm(shot ~ longjump), col="blue")
lines(lowess(longjump, shot), col="red")
```



Joonis 1.1. Funktsiooni *plot* abil konstrueeritud hajuvusdiagramm, millele on lisatud vastavalt funktsioonidega *lm* ja *lowess* hinnatud regressioonisirge (sinine joon) ja silutud joon (punane joon).

1.3. Funktsioon *scatterplot*

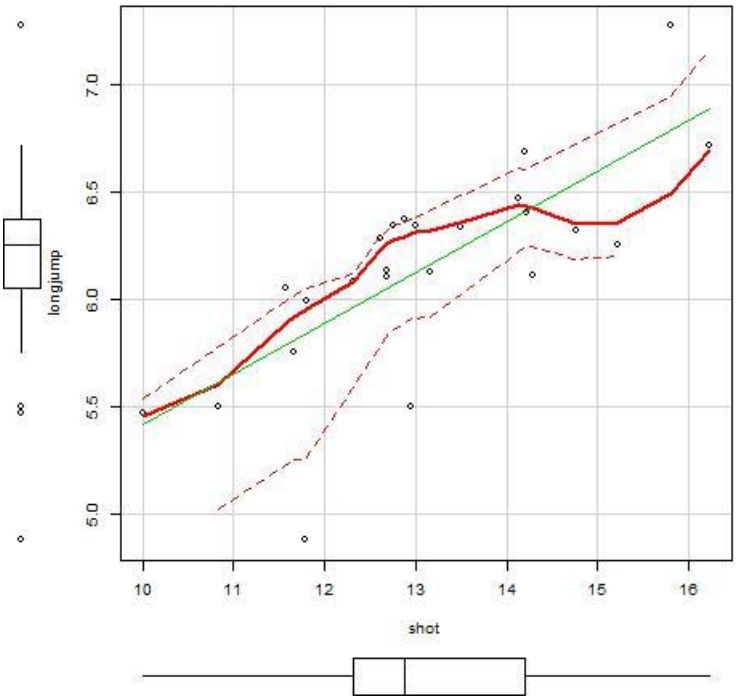
Märksa rohkem võimalusi kui funktsioonil *plot* on paketis *car* sisalduval funktsioonil *scatterplot* [5].

Vaikimisi lisatakse hajuvusdiagrammile

- lineaarne regressioonisirge (võimalik keelata käsuga *reg.line = FALSE*),
- mittelineaarset seost lähendav silutud joon (võimalik keelata käsuga *smooth = FALSE*),
- andmevektorite karpdiagrammid (võimalik keelata käsuga *boxplots = ""* või *boxplots = FALSE*; kui *boxplots = "x"* või *boxplots = "y"*, joonistub andmevektorit kirjeldav karpdiagramm vaid vastava telje juurde, *boxplots = "xy"*korral aga mõlema telje juurde – see on vaikimisi valik),
- mittelineaarse seose alusel arvutatud vahemik illustreerimaks väärtuste paiknemist (võimalik keelata käsuga *spread = FALSE*).

Joonisel 1.2 on esitatud järgneva käsu tulemusel joonistatud hajuvusdiagramm:

```
libary(car)
scatterplot(longjump ~ shot)
```



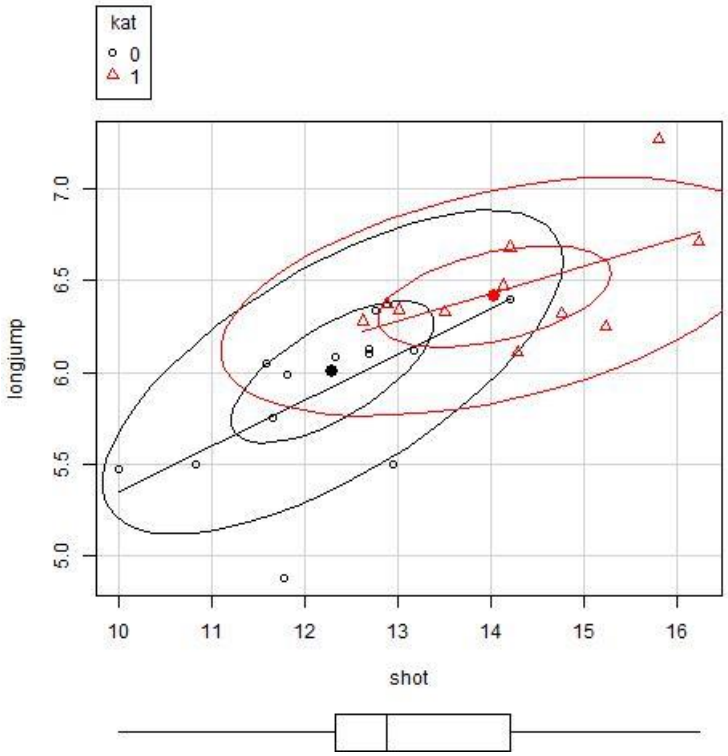
Joonis 1.2. Vaikeväärtustega funktsiooni *scatterplot* abil konstrueeritud hajuvusdiagramm.

Mõningad funktsiooni *scatterplot* lisaargumendid on järgmised:

- valiku *longjump ~ shot | kat* tulemusel tähistab R eri kategooriatesse kuuluvad (erinevate tunnuse *kat* väärtustega) punktid erinevalt ning lisab kõigile kategooriatele oma regressioonisirge ja mittelineaarse silutud joone.
- valiku *ellipse = TRUE* tulemusel joonistuvad punktide ümber andme-kontsentratsiooni ellipsid [6].

Joonisel 1.3 on esitatud järgneva, mõningaid lisavalikud kasutava käsu tulemusel joonistatud hajuvusdiagramm.

```
scatterplot(longjump ~ shot | kat, boxplots="x", ellipse=TRUE, smooth=FALSE)
```



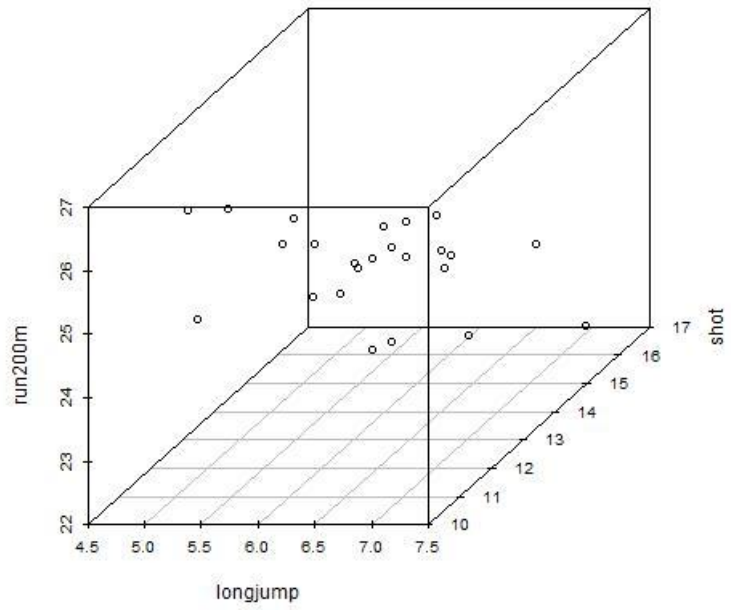
Joonis 1.3. Funktsiooni *scatterplot* abil konstrueeritud hajuvusdiagramm visualiseerimaks kahe tunnuste vahelist seost kolmanda tunnuse järgi moodustatud gruppides.

1.4. Funktsioon *scatterplot3d*

Funktsioon *scatterplot3d* samanimelises paketis annab võimaluse joonistada kolmedimensioonilisi korrelatsioonivälju [7].

Funktsioonile tuleb ette anda koordinaatide vektorid *x*, *y* ja *z*. Vaikeväärtuste korral konstrueeritav hajuvusdiagramm seitsmevõistlejate kaugushüppe, kuulitõuke ja 200 m jooksu tulemuste alusel on esitatud joonisel 1.4.

```
library(scatterplot3d)
scatterplot3d(longjump, shot, run200m)
```



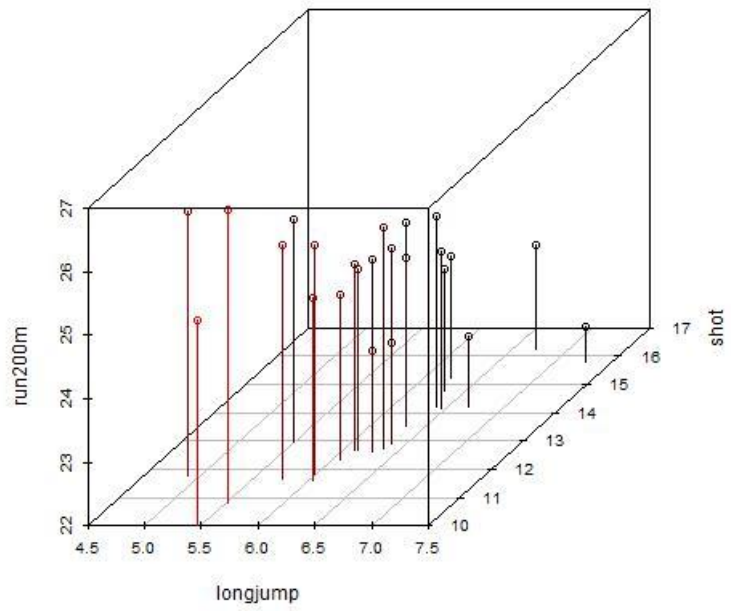
Joonis 1.4. Funktsiooni *scatterplot3d* tulemus vaikeväärtuste korral.

Mõned näited funktsiooni *scatterplot3d* lisaargumentidest on järgnevad:

- *type = "h"* tulemusel lisatakse punktidele xy-tasandini tõmmatud jooned;
- *type = "l"* tulemusel ühendatakse punktid joonega;
- *highlight.3d = TRUE* korral värvitakse punktid ja projektsioonijooned lähtuvalt tunnuse *y* väärtustest.

Seitsmevõistlejate kaugushüppe, kuulitõuke ja 200 m jooksu tulemuste vahelist seost kujutav kolmedimensionaalne, lisavalikuid *type = "h"* ja *highlight.3d = TRUE* kasutav hajuvusdiagramm on saadud järgmise käsu tulemusel ja on esitatud joonisel 1.5.

```
scatterplot3d(longjump, shot, run200m, type="h", highlight.3d=TRUE)
```

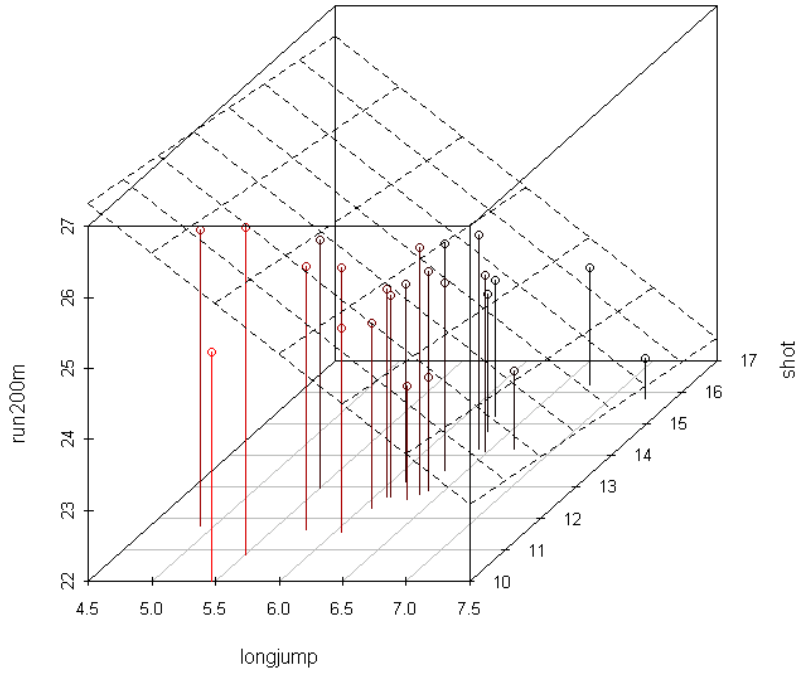


Joonis 1.5. Funktsiooni *scatterplot3d* tulemus lisaargumentide *type = "h"* ja *highlight.3d = TRUE* korral.

Funktsiooni *plane3d* abil on võimalik lisada kolmedimensionaalsele hajuvusdiagrammile tunnuste vahelist lineaarset seost kujutav tasand. Selleks tuleb funktsioonile *plane3d* ette anda funktsiooniga *lm* defineeritud mudel kujul z = x + y. Seejuures peab panema tähele, mudeli defineerimisel on tunnuste järjekord erinev funktsiooni *scatterplot3d* argumentide järjekorrast x, y, z ning funktsioon *plane3d* on ise funktsiooniga *scatterplot3d* defineeritud muutuja üks väärtustest.

Seitsmevõistlejate kaugushüppe, kuulitõuke ja 200 m jooksu tulemuste vahelist seost kujutav kolmedimensionaalne hajuvusdiagramm koos tunnuste vahelist lineaarset seost kujutava tasandiga on saadud järgmise käsu tulemusel ja on esitatud joonisel 1.6.

```
s3d <-scatterplot3d(longjump, shot, run200m, type="h", highlight.3d=TRUE)
regmodel <- lm(run200m ~ longjump + shot)
s3d$plane3d(regmodel)
```



Joonis 1.6. Funktsioonide *scatterplot3d* ja *plane3d* tulemus.

2. Kõrge tihedusega hajuvusdiagrammid

Kõrge tihedusega korrelatsooniväljadel punktide koordinaadid kattuvad ning eelnevate meetodite kasutamisel läheb osa informatsioonist kaduma (mitu punkti jääb üksteise taha ja paistab visuaalselt vaid ühe punktina). Funktsiooni *hexbin* abil on võimalik eristatavalt illustreerida punktid, mille lähedal on mitu sarnaste uuritavate tunnuste väärtustega objekti, ning funktsioon *sunflowerplot* võimaldab lisaks objektide paiknemisele tuua hajuvusdiagrammil välja ka objektide arvu, mis omavad võrdseid tunnuste *x* ja *y* väärtusi. Antud peatüki näidetes kasutatud koordinaatide vektorid on genereeritud funktsiooni *rnorm* abil.

2.1. Funktsioon *hexbin*

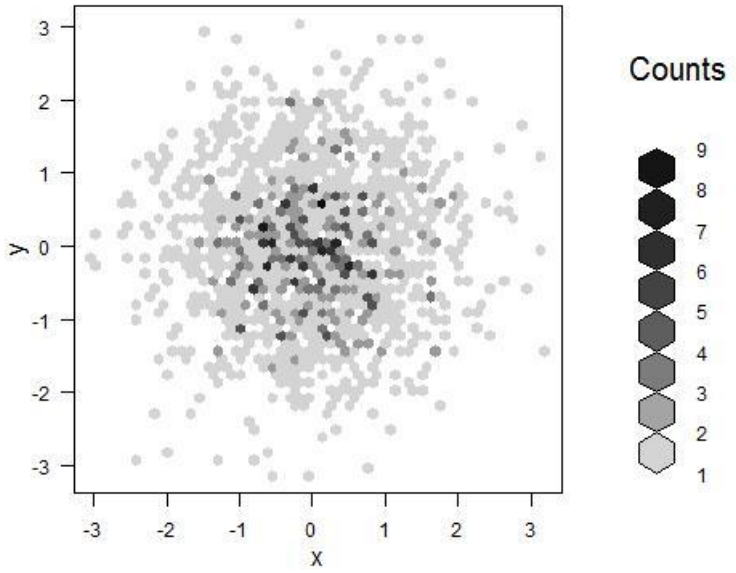
Paketi *hexbin* samanimelise funktsiooni abil väärtustatud objektide illustreerimisel jagatakse graafik kuusnurkadeks [8]. Mida rohkem on ühele kuusnurgale sattunud objekte, seda tumedamalt antud kuusnurk joonistatakse.

Funktsiooni argumentideks on koordinaatide vektorid *x* ja *y* ning muutuja *xbins*, mille väärtustamisega määratakse kuusnurkade arv x-teljel ehk see, kui mitmeks kuusnurgaks telg jagatakse. Hajuvusdiagrammi näeb, rakendades funktsiooni *hexbin* abil loodud objektile käsku *plot*.

Joonise 2.1 konstrueerimiseks kasutatud käsuread on järgmised:

```
y = rnorm(1500)
x = rnorm(1500)

library(hexbin)
bin = hexbin(x, y, xbin=50)
plot(bin)
```



Joonis 2.1. Hajuvusdiagramm, mis on konstrueeritud rakendades käsku *plot* funktsiooni *hexbin* abil loodud objektile.

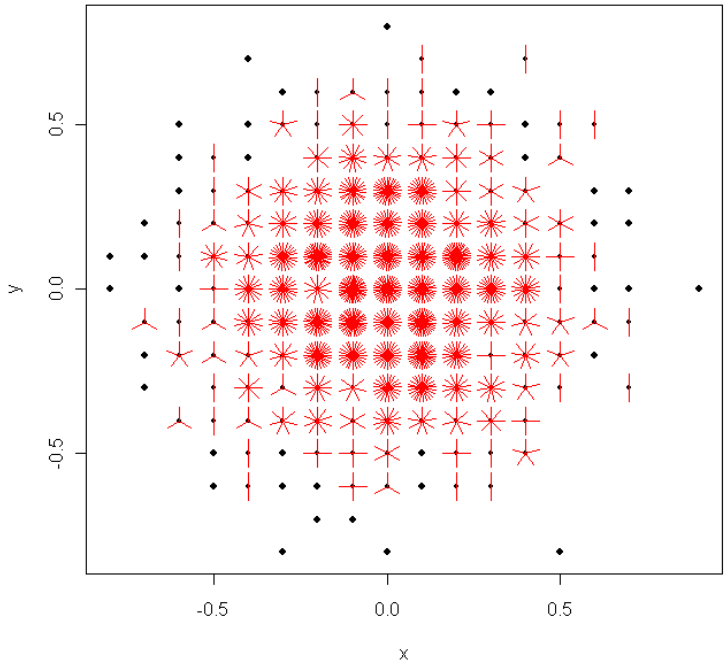
2.2. Funktsioon *sunflowerplot*

Funktsiooni *sunflowerplot* rakendamisel kirjeldatakse andmeid „päevalillede“ abil, kus „kroonlehtede“ arv näitab, kui mitmel objektil on samad koordinaadid [9].

Funktsiooni argumentideks on ka siin andmevektorid *x* ja *y*.

Järgnevate käsuridade abil on konstrueeritud joonis 2.2.

```
y = round(rnorm(1500, sd=0.25), 1)
x = round(rnorm(1500, sd=0.25), 1)
sunflowerplot(x ,y)
```



Joonis 2.2. Funktsiooni *sunflowerplot* abil konstrueeritud korrelatsiooniväli.

3. Hajuvusdiagrammide maatriks

Olgu andmestikus tunnused T_1, T_2, ..., T_k. Antud andmete põhjal koostatud hajuvusdiagrammide maatriks koosneb elementidest, mis on paarikaupa tunnuste hajuvusdiagrammid. Maatriksi i. rea j. element on tunnuste T_i ja T_j hajuvusdiagramm.

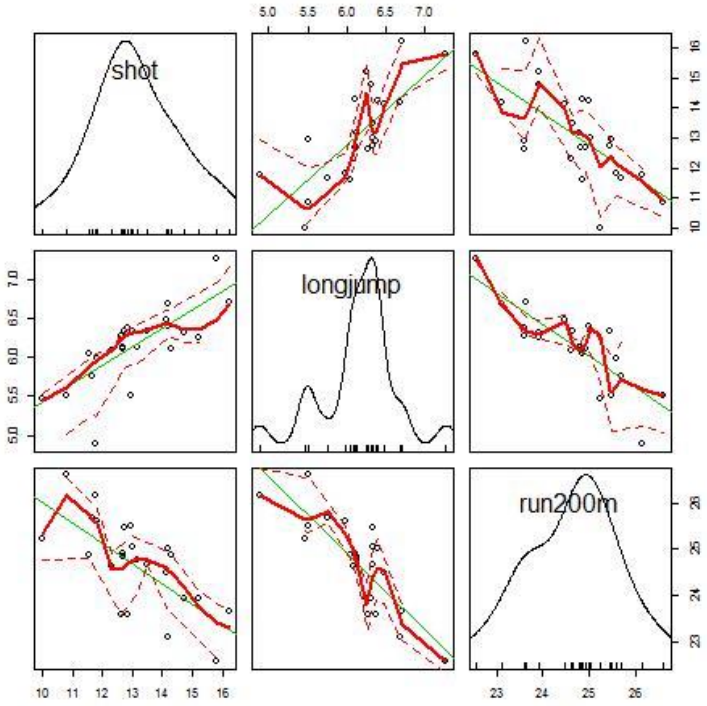
Suhteliselt väikese huvi pakkuvate tunnuste arvu korral (vähem kui 10) annab hajuvusdiagrammide maatriks suurepärase visuaalse ülevaate tunnuste vahelistest seostest [10]. Meetod näitab ära kõik paarikaupa seosed, mida on võimalik rõhutada lineaarse regressioonivõrrandi graafikute ja erinevate silutud joonte abil.

3.1. Funktsioon *scatterplotMatrix*

Funktsioon *scatterplotMatrix* paketis *car* on hajuvusdiagrammide maatriksi loomiseks [5]. Maatriksi elementideks, mis ei asu peadiagonaalil, joonistatakse funktsioonist *scatterplot* tuttavad korrelatsiooniväljad ning funktsiooni *scatterplot* argumentide rakendamine (vt peatükk 1.3) mõjub ühte moodi kõikidele joonistatud hajuvusdiagrammidele. Peadiagonaali elementideks on vaikimisi tunnuste tihedusfunktsioonide graafikud.

Joonisel 3.1 on esitatud järgneva käsu tulemusel joonistatud hajuvusdiagrammide maatriks:

```
library(car)
scatterplotMatrix(~ shot + longjump + run200m)
```



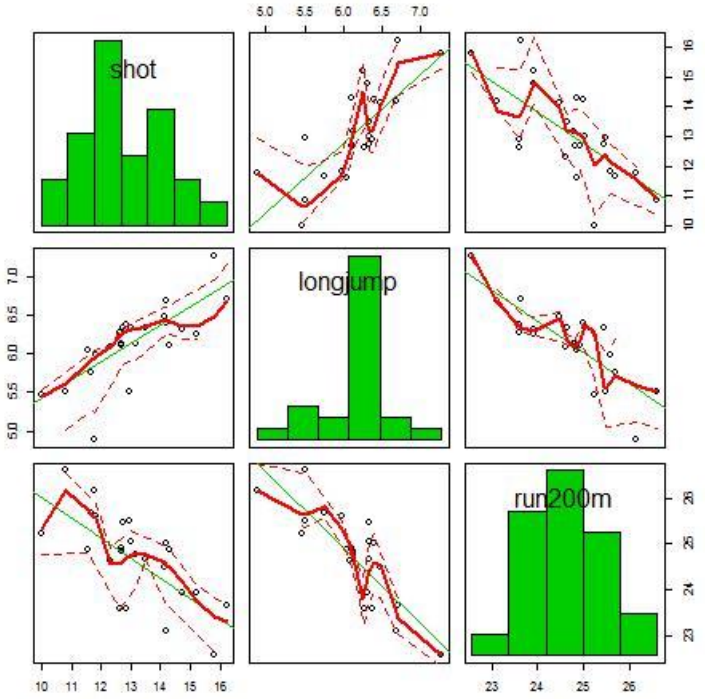
Joonis 3.1. Funktsiooni *scatterplotMatrix* vaikeväärtuste abil konstrueeritud hajuvusdiagrammide maatriks.

Mõningad funktsiooni *scatterplotMatrix* lisaargumendid on järgmised:

- valiku *diagonal="boxplot"* korral joonistatakse peadiagonaalile tunnust kirjeldavad karpdiagrammid;
- valides *diagonal="histogram"*, joonistuvad peadiagonaalile histogrammid, mille korral saab omakorda lisaargumendiga *nclas* määrata histogrammi tulpade arvu.

Joonis 3.1 on koostatud rakendades järgnevat käsurida:

```
scatterplotMatrix(~ shot + longjump + run200m, diagonal="histogram", nclass=5)
```



Joonis 3.2. Funktsiooni *scatterplotMatrix* abil koostatud korrelatsioonimaatriks. Kasutatud on lisaargumente väärtustega *diagonal="histogram"* ja *nclass=5*.

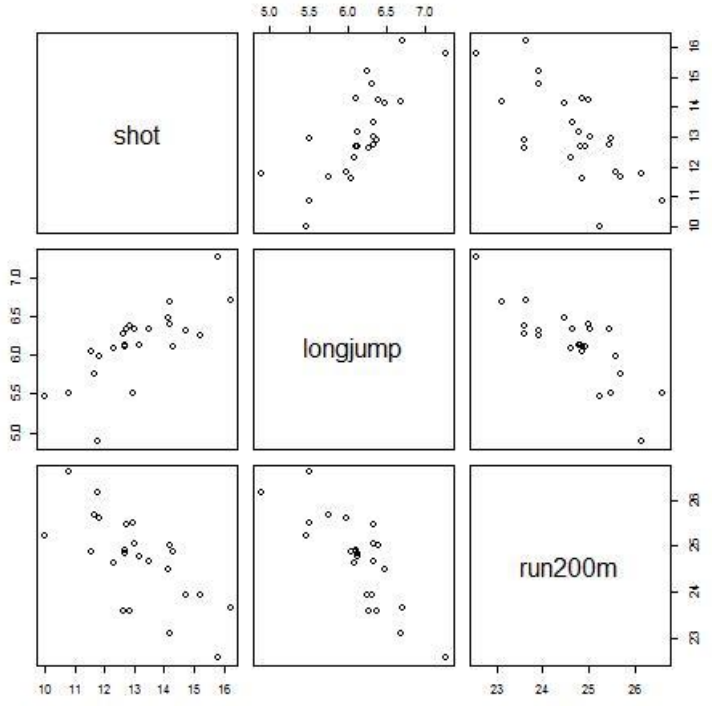
3.2. Funktsioon *pairs*

Funktsioon *pairs* võimaldab samuti luua huvi pakkuvate tunnuste vahelistest seostest hajuvusdiagrammide maatriksi, kuid on mitmekülgsem [11]. Funktsioon *pairs* jagab maatriksi kolmeks paneeliks (peadiagonaal, peadiagonaali alune ja pealne), mida on võimalik funktsioonide abil väärtustada.

Vaikimisi luuakse maatriks, mille elementidena on kujutatud tunnuste vahelisi hajuvusdiagramme. Peadiagonaalile kirjutatakse tunnuste nimed ning peadiagonaali alla ja peale joonistuvad korrelatsiooniväljad. Skaalad, millel tunnused on mõõdetud, paiknevad korrelatsiooniväljade äärtel.

Joonise 3.3 konstrueerimiseks on kasutatud järgnevat käsurida:

```
pairs(~ shot + longjump + run200m)
```

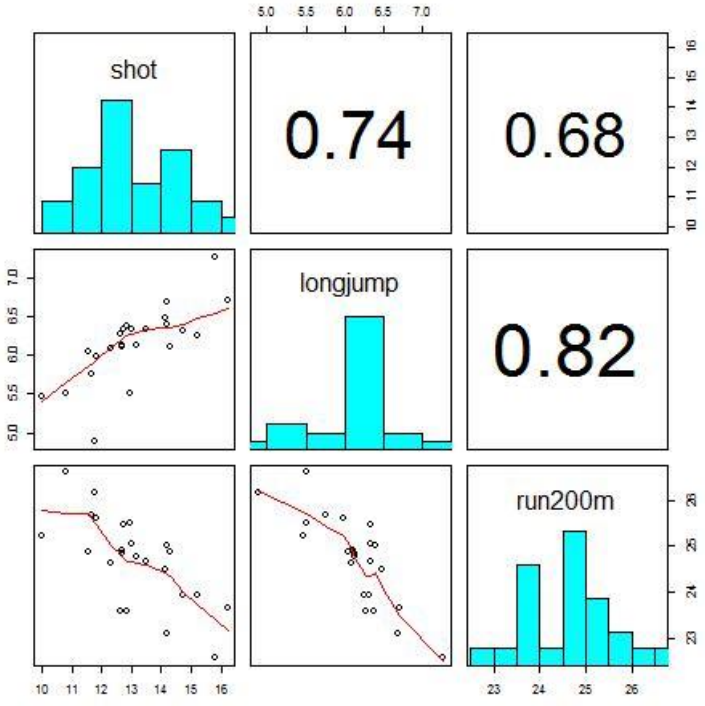


Joonis 3.3. Hajuvusdiagrammide maatriks funktsiooni *pairs* vaikeväärtuste korral.

Antud funktsioonil on suur hulk lisavõimalusi, kuna paneelide sisu muutvad argumendid *diag.panel*, *lower.panel* ja *upper.panel* on võimalik väärtustada funktsioonidega, mis on ka kasutaja poolt defineeritavad. Samas on võimalik ka rakendada funktsioone, mis on juba R-s vaikimisi olemas – näiteks *panel.smooth*.

Joonisel 3.4 on peadiagonaalil ära toodud tunnuste nimed ja histogrammid, peadiagonaalist ülevalpool korrelatsioonikordajate arvväärtused, kusjuures fondi suurus peegeldab seose tugevust, ning peadiagonaalist allpool hajuvusdiagrammid ja lokaalse kaalutud regressiooni abil silutud jooned. Joonis on konstrueeritud kasutades Lisas 2 ära toodud funktsioone ja järgnevat käsurida:

```
pairs(~ shot + longjump + run200m, lower.panel=panel.smooth,
diag.panel=panel.hist, upper.panel=panel.cor)
```



Joonis 3.4. Funktsiooni *pairs* tulemus lisaargumentide korral.

4. Korrelatsioonimaatriksi illustreerimine funktsiooniga *corrplot*

Tunnuste vahelise seose tugevust ja suunda näitav korrelatsioonikordaja on üks enamkasutatavamaid statistilisi kordajaid. Kahe tunnuse vahelist korrelatsiooni on enamasti võimalik intuitiivselt interpreteerida, vaadates hajuvusdiagrammi. Suure hulga tunnuste ja nende kõiki paarikaupa seoseid mõõtva korrelatsioonimaatriksi illustreerimine on aga keerulisem.

Selles peatükis annab autor ülevaate statistika tarkvara R funktsioonist *corrplot* [12], mis on mõeldud justnimelt korrelatsioonimaatriksite ja korrelatsioonikordajate usaldusintervallide graafiliseks kirjeldamiseks. Funktsiooni argumentideks on korrelatsioonimaatriks ja lisaparameetrid, mille abil on võimalik selgemalt välja tuua tugevamad korrelatsioonid ning rõhutada neist vaid statistiliselt olulisi. Samuti on paketis algoritmid, mida saab kasutada tunnuste järjestamiseks korrelatsioonimaatriksis, et võimalikud seoste mustrid paremini esile tuleksid.

Erinevalt hajuvusdiagrammide maatriksist ei illustreeri funktsioon *corrplot* algandmeid, vaid üksnes korrelatsioonikordajate väärtuseid, kasutades selleks erinevaid sümboleid ja värve.

Kuigi funktsiooniga *corrplot* saab illustreerida mistahes korrelatsioonikordajate (või üldisemalt – sarnasuse mõõtude) maatriksit, baseeruvad järgnevad kirjeldused Pearsoni korrelatsioonikordajate maatriksil, mis on omistatud muutujale *M* kasutades funktsiooni *cor*:

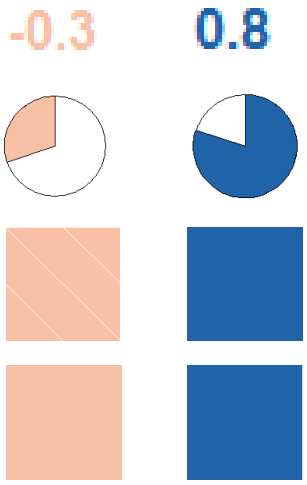
```
M <- cor(heptathlon)
```

4.1. Argument *method*

Argument *method* määrab ära, kuidas esitavad graafilise maatriksi elemendid korrelatsioonikordaja suurust. Võimalike väärtuste juures on ära toodud näited korrelatsioonidest suurustega -0,3 ja 0,8. Vaikeväärtusena kasutatakse väärtust *„circle“*.

- *„circle“*
- *„ellipse“*

- *„number“*
- *„pie“*
- *„shade“*
- *„color“*



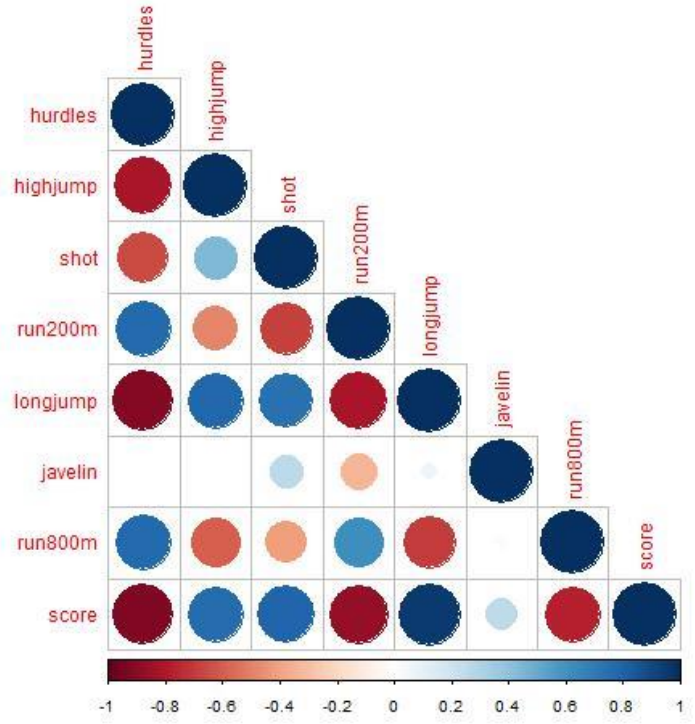
4.2. Argument *type*

Parameetri *type* määramisega on võimalik valida korrelatsioonimaatriksi kuju. Argumendi *type* võimalikud väärtused on järgmised:

- valikut *„full“* kasutatakse vaikimisi ning selle argumendi väärtuse korral kuvatakse terve korrelatsioonimaatriks;
- väärtustades parameetri suurusega *„upper“*, esitatakse ainult peadiagonaalist üleval pool asuv kolmnurk koos peadiagonaaliga;
- valiku *„lower“* korral konstrueeritakse peadiagonaalist all pool asuv kolmnurk koos peadiagonaaliga.

Joonis 4.1 illustreerib valikut *„lower“* ning on koostatud järgneva käsurea abil:

```
library(corrplot)
corrplot(M, type="lower")
```



Joonis 4.1. Funktsiooni *corrplot* abil koostatud korrelatsioonimaatriks kasutades argumendi *type* väärtust *„lower“*.

4.3. Argument *order*

Tunnuste järjestamine korrelatsioonimaatriksis muudab lihtsamini märgatavaks seoste mustrid, suunad ja anomaaliad, kui „sarnased“ tunnused paiknevad lähestikku [13].

Argument *order* sätestab tunnuste järjekorra korrelatsioonimaatriksis ning võimalikud valikud on järgmised:

- väärtust *„original“* rakendatakse vaikimisi ning see jätab tunnuste järjekorra muutmata;
- valiku *„FPC“* korral järjestatakse tunnused peakomponentide analüüsil leitud esimesele peakomponendile vastava omavektori elementide alusel;
- valides argumendi väärtuseks *„hclust“*, järjestatakse tunnused hierarhilise klasterdamise alusel;
- omistades parameetrile suurus *„alphabet“*, järjestatakse tunnused tähestiku järjekorras;

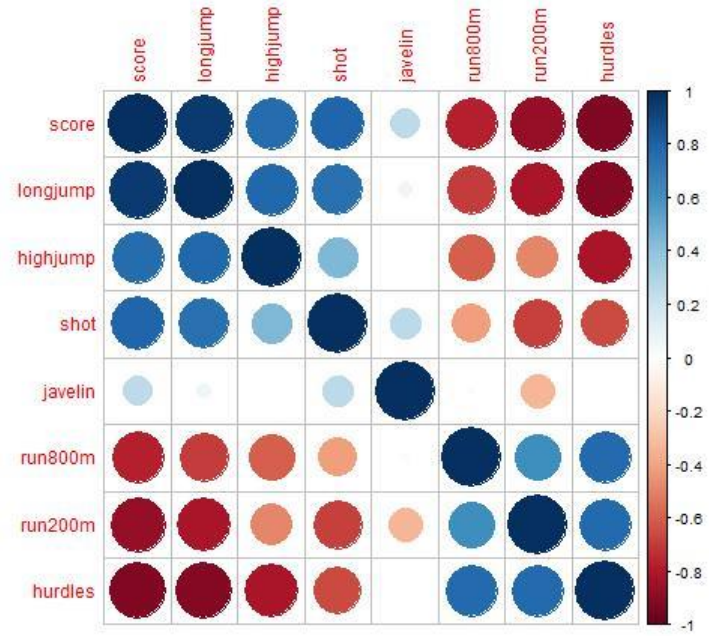
- väärtuse *„AOE“* korral kasutatakse järjestamiseks omavektorite nurga suurust a_i, mis avaldub valemist

$$a_i=\begin{cases} tan^{-1}(e_{i2}/e_{i1}), & e_i > 0 \\ tan^{-1}(e_{i2}/e_{i1}) + \pi, & muidu \end{cases}$$

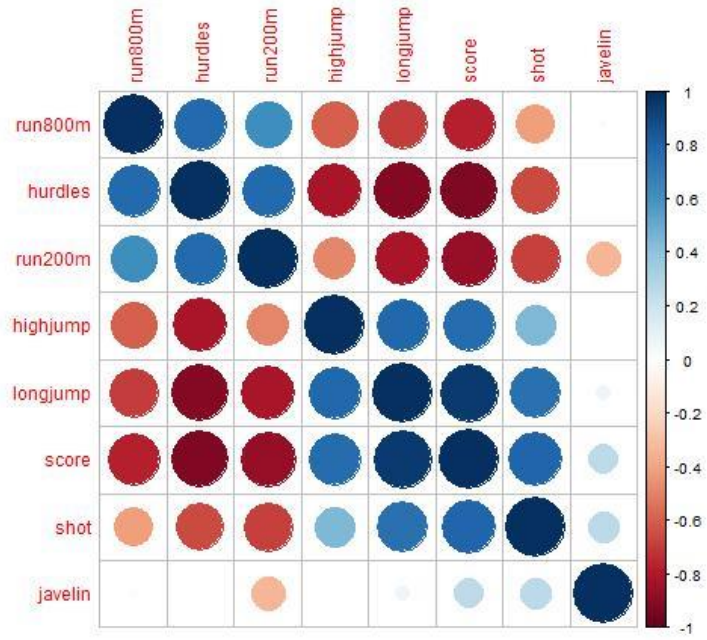
, kus e_{i1} ja e_{i2} on korrelatsioonimaatriksi kahele suurimale omaväärtusele vastavate omavektorite i. tunnusele vastavad elemendid.

Joonistel 4.2 ja 4.3 on ära toodud funktsiooniga *corrplot* illustreeritud korrelatsioonimaatriksid vastavalt parameetri *order* väärtuste *„FPC“* ja *„AOE“* korral. Joonistele eelnevad vastavad käsuread.

```
corrplot(M, order ="FPC")
corrplot(M, order ="AOE")
```



Joonis 4.2. Korrelatsioonimaatriksi illustratsioon funktsiooni *corrplot* abil. Tunnuste järjestamisel on kasutatud argumenti *order=“FPC“*



Joonis 4.3. Korrelatsioonimaatriksi illustratsioon funktsiooni *corrplot* abil. Tunnuste järjestamisel on kasutatud argumendi *order* väärtust „*AOL*“.

4.4. Olulisustõenäosuste illustreerimine

Kasutades R-i dokumentatsiooni näidetes ära toodud funktsiooni *cor.mtest* (vt. Lisa 3), näitame erinevaid võimalusi usaldusintervallide ja korrelatsioonikordajatele vastavate olulisustõenäosuste illustreerimiseks. Antud funktsioon ei ole R-s vaikimisi olemas, vaid tuleb kasutajal sisse lugeda.

Funktioon *cor.mtest* väärtustab olulisusnivoo määramisel muutuja vektoriga, mille elementideks on kolm maatriksit:

1. korrelatsioonikordajate olulisustõenäosuste maatriks,
2. korrelatsioonikordajate ülemiste usalduspiiride maatriks,
3. korrelatsioonikordajate alumiste usalduspiiride maatriks.

Järgmise käsurea abil omistame muutujale *res1* väärtuse kasutades funktsiooni *cor.mtest*:

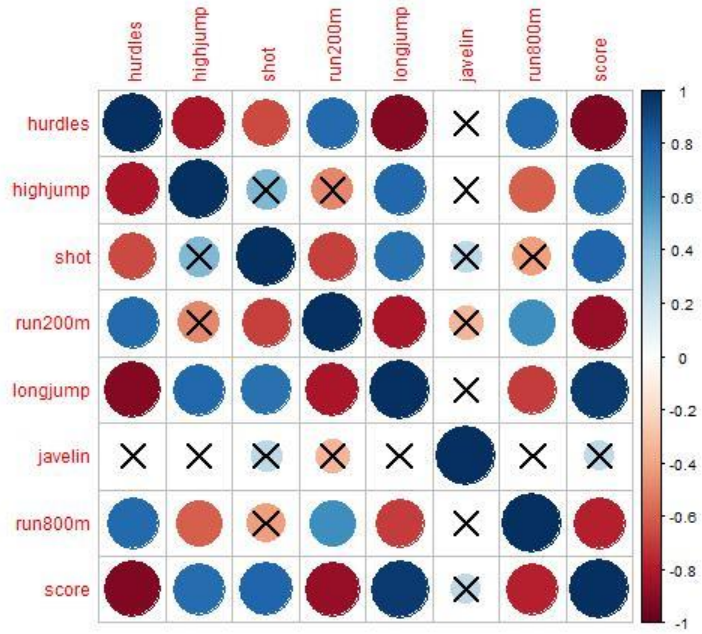
```
res1 = cor.mtest(heptathlon, 0.95)
```

Funktsiooni *corrplot* argumentide *p.mat* ja *sig.level* väärtustamisel vastavalt olulisustõenäosuste maatriksiga ja olulisusnivooga saame eristavalt illustreerida statistiliselt ebaolulised korrelatsioonikordajad. Parameeter *insig* võib omada ühte kolmest väärtusest, mille korral on võimalik erineval viisil märkida statistiliselt ebaoluliseks kujunenud korrelatsioonikordajad:

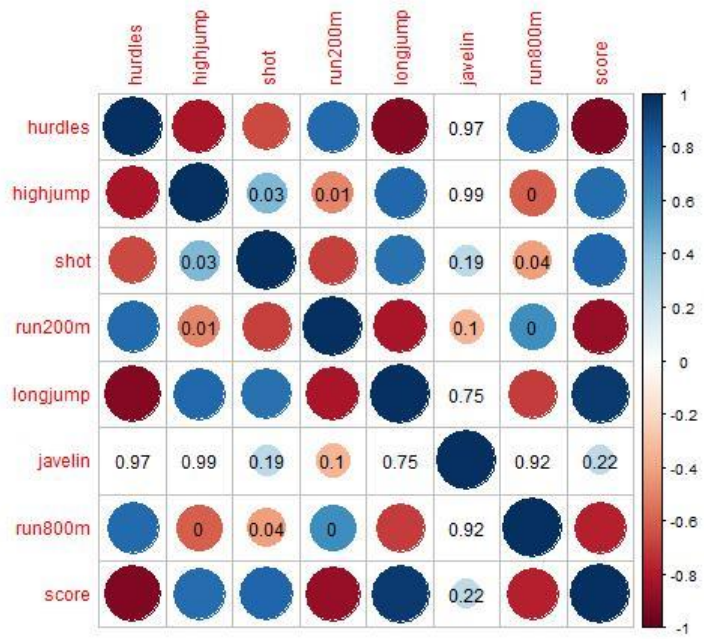
- väärtuse *„pch“* korral märgitakse statistiliselt ebaoluliseks osutunud seosed ristiga,
- väärtuse *„p-value“* korral näidatakse ära olulisustõenäosus,
- väärtuse *„blank“* korral jäetakse joonisel vastavad maatriksi elemendid tühjaks.

Järgnevate käsuridade ja joonistega 4.4 ja 4.5 on ära toodud näited argumentide *p.mat*, *sig.level* ja *insig* rakendustest.

```
corrplot(M, p.mat=res1[[1]], insig="pch", sig.level=0.001)
corrplot(M, p.mat=res1[[1]], insig="p-value", sig.level=0.001)
```



Joonis 4.4. Funktsiooni *corrplot* abil konstrueeritud korrelatsioonimaatriksi illustratsioon kasutades argumendi *insig* väärtust „*pch*“.

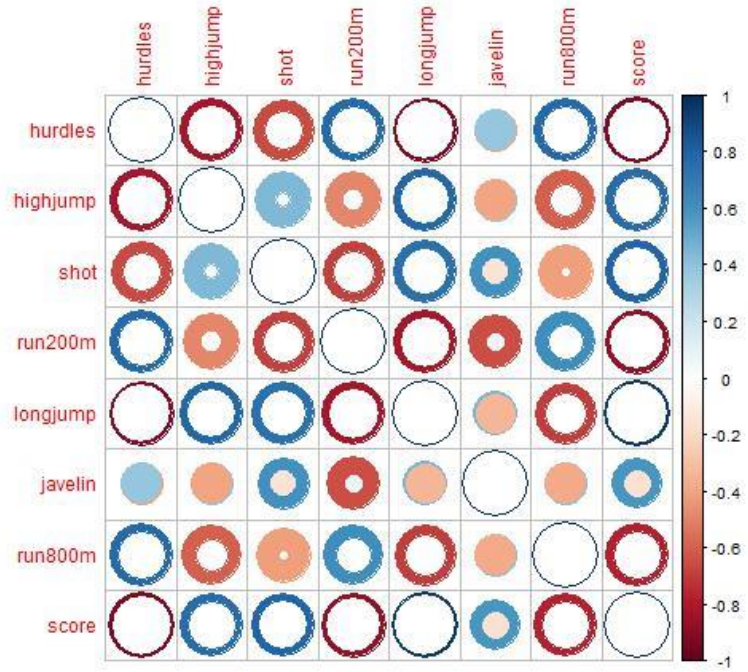


Joonis 4.5. Korrelatsioonimaatriksi illustratsioon kasutades argumendi *insig* väärtust „*p-value*“.

4.5. Usaldusintervallide illustreerimine

Usaldusintervallide visualiseerimise saab tellida agumendi *plotCI* abil. Ülemiste ja alumiste usalduspiiride maatriksite ära märkimiseks on vastavalt parameetrid *low* ja *upp*. Erinevaid võimalusi on kokku kolm, kuid kahe esimese võimaluse puhul on meetodi põhimõte väga sarnane. Omistades parameetrile *plotCI* väärtuse *„circle“* või *„square“*, joonistatakse maatriksi elementideks vastavalt seest tühjad ringid või ruudud nii, et sisemine joon illustreerib alumist ja välimine ülemist usalduspiiri. Järgneva käsurea abil on konstrueeritud joonis 4.6:

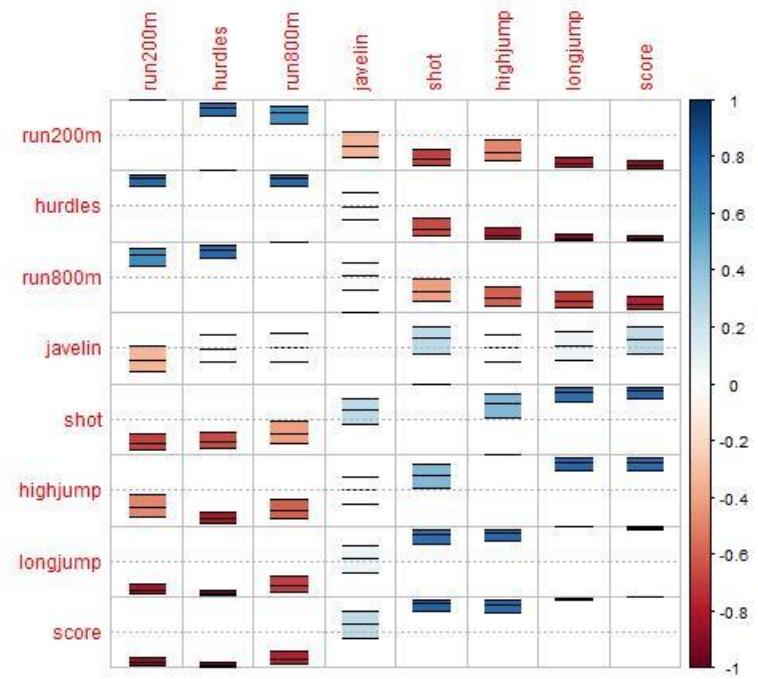
```
corrplot(M, low=res1[[2]], upp=res1[[3]], plotCI="circle")
```



Joonis 4.6. Korrelatsioonimaatriksi illustratsioon kasutades parameetri *plotCI* väärtust „circle“.

Kolmas võimalik väärtus argumendile *plotCI* on *„rect“*, mille korral iseloomustatakse usaldusintervalle ristkülikute abil. Antud kujundi ülemise ja alumise serva kõrgus illustreerivad vastavaid usalduspiire. Järgneva käsureaga on koostatud joonis 4.7:

```
corrplot(M,order="AOE", plotCI="rect",low=res1[[2]], upp=res1[[3]])
```



Joonis 4.7. Korrelatsioonimaatriksi illustratsioon kasutades parameetri *plotCI* väärtust *„rect“*.

4.6. Muud argumendid

Parameeter *col* alusel määratakse maatriksi elementide värvide valik ja muutus vastavalt korrelatsiooni tugevusele. Funktsiooni *colorRampPalette* abil on võimalik luua vastavad vektorid, mille elementideks on värvide tüüpnimetused ja/või värvikoodid. Argumendi *col* väärtustamisel vektoriga tuleb lisada sulgudesse vektori järele arv, mitut erinevat värvide ülemineku varjundit illustreerimisel kasutatakse.

Kui maatriksi tunnuste järjestamiseks on kasutatud väärtust *„hclust“*, siis argumendi *addrect* abil on võimalik valida parema eristatavuse huvides joonega ümbritsetavate klastrite arv.

Parameeter *diag* on binaarne ning omab vaikimisi väärtust TRUE. Väärtuse FALSE korral eemaldatakse maatriksilt peadiagonaali elemendid.

Argumendi *bg* väärtustamisel värvikoodi või nimetusega saab valida korrelatsioonimaatriksile taustavärvi.

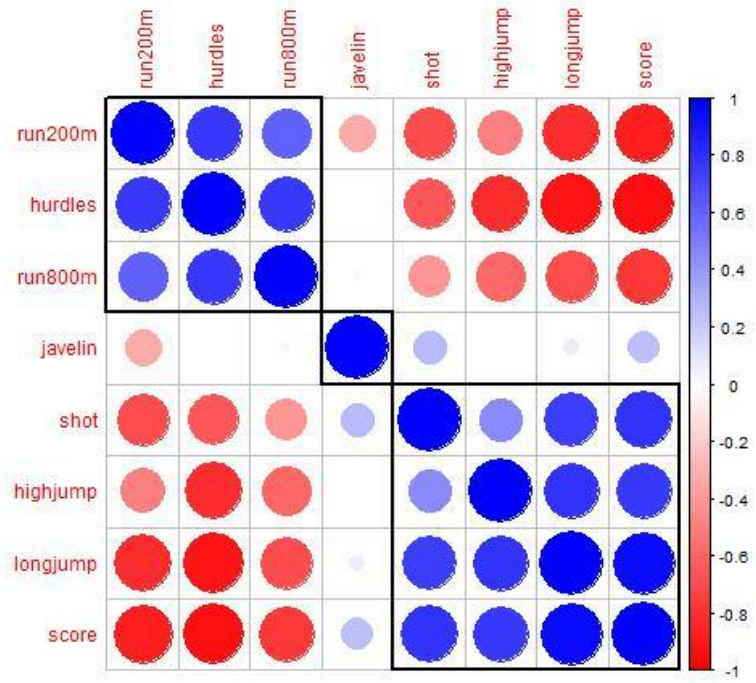
Argument *outline* on samuti binaarne tõeväärtustega parameeter, kuid vaikimisi omab see argument väärtust FALSE. Kui korrelatsioone on valitud kirjeldama ringid, ellipsid või

ruudud ja parameeteri väärtuseks on valitud TRUE, siis ümbritsetakse antud kujundid mustade piirjoontega.

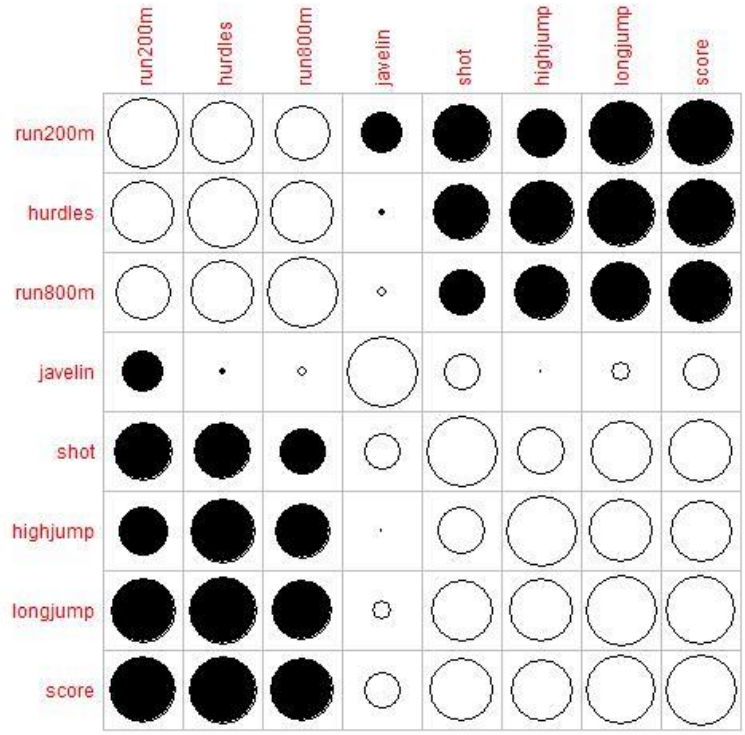
Järgnevatele käsuread ja neile vastavad joonised 4.8 ja 4.9 illustreerivad alapeatükis toodud argumentide rakendamist korrelatsioonimaatriksite illustreerimisel. Tunnuste järjekorra valimiseks on kasutatud argumendi *order* väärtust *„hclust“*.

```
col3 <- colorRampPalette(c("red", "white", "blue"))
corrplot(M, order="hclust", col=col3(100), addrect=3)

wb <- colorRampPalette(c("white", "black"))
corrplot(M, col = wb(2), order="hclust", outline=TRUE, cl.pos="n")
```



Joonis 4.8. Korrelatsioonimaatriksi illustratsioon kasutades positiivsete ja negatiivsete korrelatsioonide illustreerimiseks vastavalt sinist ja punast värvi ning ümbritsedes kolm tunnuste hierarhilisel klasterdamisel tekkinud klastrit mustade piirjoontega.



Joonis 4.9. Funktsiooni *corrplot* abil konstrueeritud korrelatsioonimaatriksi illustratsioon, kus tugevamad seosed on välja toodud suuremate ringidega. Mustad ja valged ringid märgivad vastavalt negatiivseid ja positiivseid korrelatsioone.

5. Lisavõimalused korrelatsioonimaatriksite illustreerimiseks

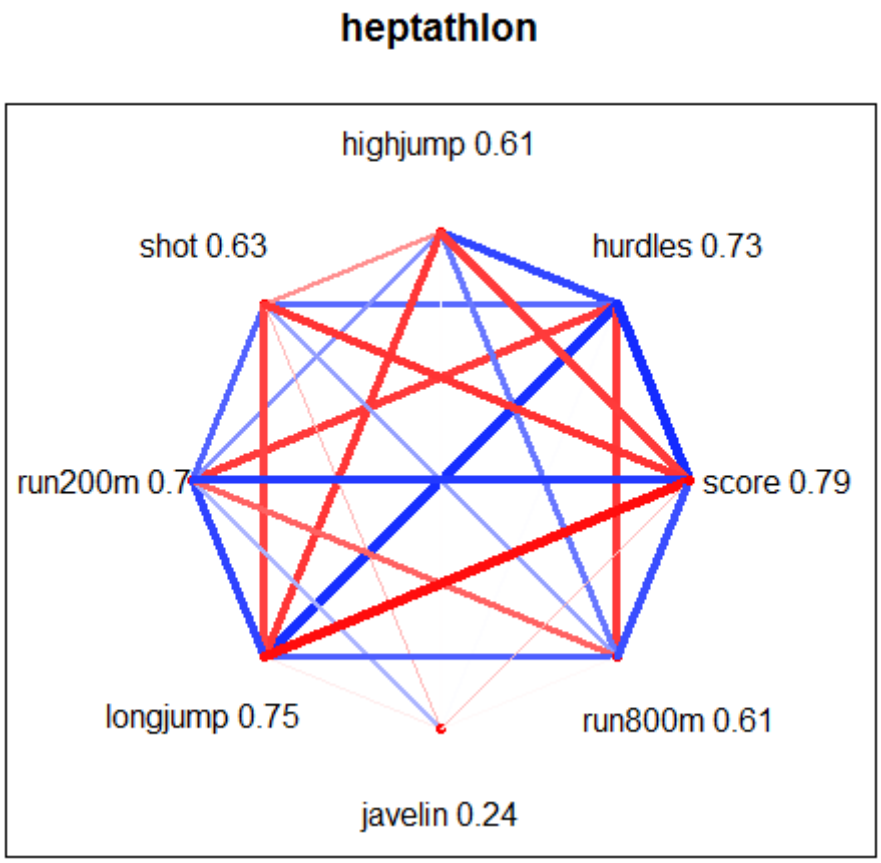
Eelnevalt kirjeldatud korrelatsioonimaatriksite illustreerimismeetoditele on olemas piiramatu arv alternatiive, mille produtseerimist piiravad vaid fantaasia ning programeerimisoskus. Selles peatükis annab autor ülevaate kahest enese programmeeritud võimalusest, mida statistikapaketi R-i pakettides ei leidu.

5.1. Korrelatsioonimaatriksi illustratsioon hulknurga kujul – funktsioon *ring.korr*

Rakendades funktsiooni *ring.korr* N tunnusega andmetabelile moodustatakse N-nurkne kujund, mille nurgad on ühendatud korrelatsiooni tugevust ja suunda kirjeldavate punaste ja siniste joontega. Mida tugevam on kahe tunnuse vaheline korrelatsioon, seda pakemalt ja tugevama tooniga on tõmmatud joon nurkade vahel. Samuti on tunnuse nime juures välja toodud antud tunnuse ja kõigi ülejäänud tunnuste vaheliste korrelatsioonikordajate absoluutväärtuste keskmine. Antud meetodi eeliseks võrreldes varem esitatudega on kompaktsus – väiksemat pinda kasutades saab edasi anda suurema hulga informatsiooni, kuna tunnuste vaheliste korrelatsioonide kirjeldamiseks kasutatakse ühist pinda. Meetodile vastav programmikood on ära toodud Lisas 4. Kasutajapoolne võimalus funktsiooni rakendamiseks on võimalik muutes ära andmestiku nime, millele soovitatakse funktsiooni rakendada, ja väärtustades soovi korral funktsiooni lisaargumendid.

Vaikimisi lisatakse joonisele ka andmestiku nimi. Järgmise käsurea abil on konstrueeritud joonis 5.1.

```
ring.korr(heptathlon)
```



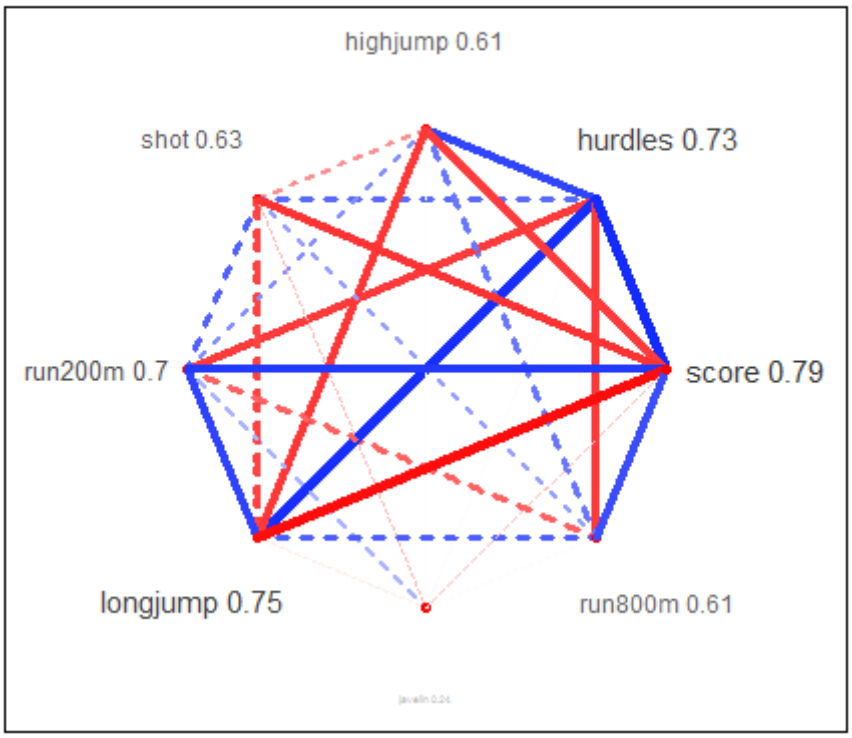
Joonis 5.1. Korrelatsioonimaatriksi esitus hulknurga kujul.

Kasutaja poolt muudetavad funktsiooni *ring.korr* lisaargumendid on järgmised:

- parameetri *main* abil saab valida joonisele pealkirja,
- argument *r* määrab ära nurga kauguse joonise keskpunktist,
- valikute *kir=TRUE* ja *col=TRUE* korral illustreeritakse vastavalt selgemalt ja suuremalt tunnuste nimed, mis on teiste tunnustega keskmiselt tugevamini korreleerunud,
- valiku *sig=TRUE* korral illustreeritakse statistiliselt ebaolulised korrelatsioonid katkendlike joontega,
- parameetri *sig.level* abil on võimalik ära määrata olulisusnivoo, millega võrreldakse korrelatsioonikordajatele vastavaid olulisustõenäosusi,
- argument *rtext* määrab ära tunnuste nimede kauguse hulknurgast.

Joonis 5.2 on konstrueeritud kasutades järgnevat käsurida:

```
ring.korr(heptathlon, main="Korrelatsioonimaatriksi illustratsioon", r=20, kir=TRUE, cex=1.1, col=TRUE, sig.level=0.00001, sig=TRUE, rtext=7.5)
```



Joonis 5.3. Korrelatsioonimaatriksi illustratsioon kasutades funktsiooni *ring.korr* lisaargumente.

5.2. Funktsiooni *ring.korr* edasiarendus

Funktsiooni *ring.korr* võimaliku edasiarendusena toob autor välja ühe alternatiivse meetodi korrelatsioonimaatriksite illustreerimiseks. Sisuline erinevus eelnevas peatükis kirjeldatust tuleneb tunnuste paigutamisest nö erinevatele tasemetele sõltuvalt tunnuste vaheliste korrelatsioonikordajate absoluutväärtuste keskmistest. Kõik tunnused, mille keskmine seotus teiste tunnustega (korrelatsioonikordajate absoluutväärtuste keskmise alusel) ületab määratud lävendi, moodustavad joonise keskele eelnevas peatükis kirjeldatud konstruktsiooni kohaselt tuumiku ning tunnused, mis lävendit ei ületa, paigutatakse tuumikust eraldi nö teisele tasemele ja kõige lähemale sellele tuumikusse kuuluvale tunnusele, millega ollakse kõige tugevamini seotud.

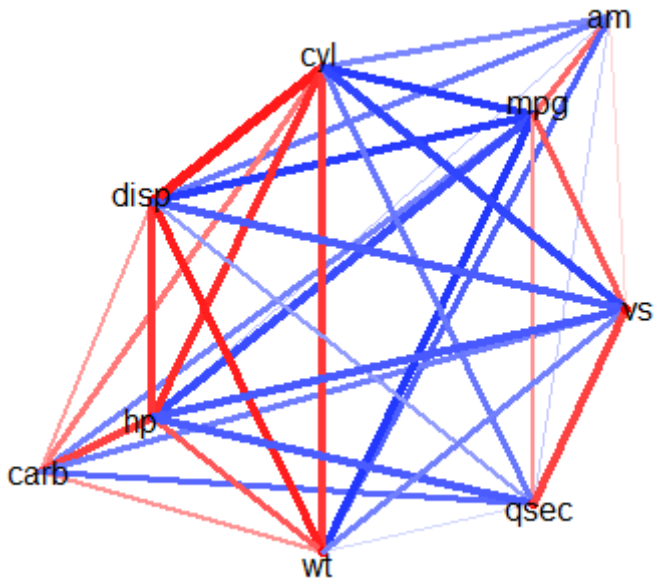
Antud meetod võimaldab paremini välja tuua omavahel tugevalt seotud tunnuste grupi ning eristada marginaalsed, vaid üksikute teiste tunnustega seostud tunnused. Meetodi rakendamine annab parema tulemuse, kui tunnuste vahelised seosed tõepoolest järgivad kirjeldatud struktuuri – leiduvad mõned omavahel tugevalt seotud tunnused pluss mõned

marginaalsed tunnused. Kui kõik tunnustevahelised seosed on sarnase tugevusega, sõltub meetodi rakendamisel saadav visuaalne pilt eelkõige määratud lävendist. Ebasobiv lävend võib tekitada andmetest ebamäärase või lausa vale ettekujutuse. Soovituslikult tuleks lävend paigutada vahemikku, mis vastab korrelatsioonikordajate absoluutväärtuste keskmiste järjestatud reas suurimale erinevusele.

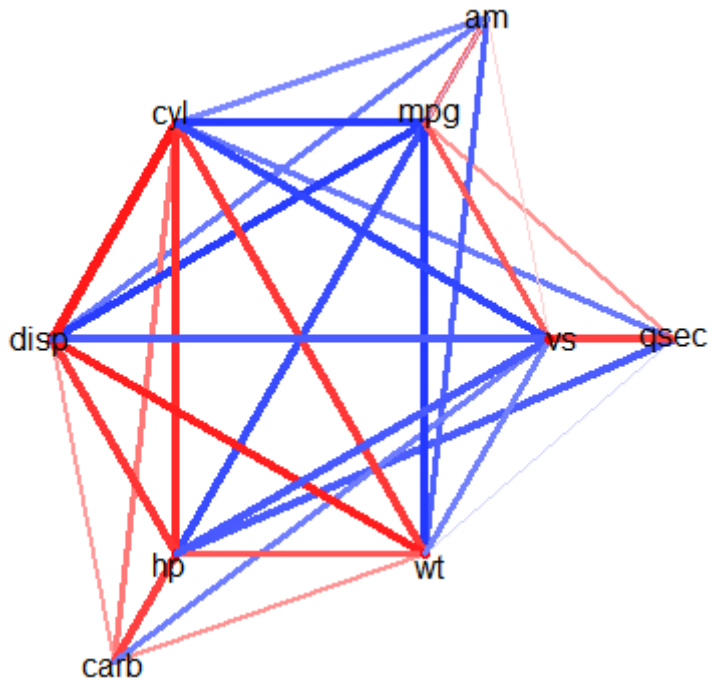
Meetodi paremaks kirjeldamiseks kasutab autor andmestiku *heptathlon* asemel andmestikku *mtcars.* Tabelist 5.1, kus on ära toodud kasutatavad tunnused ja neile vastavad korrelatsioonikordajate absoluutväärtuste keskmised, ilmneb, et pisut eristuvad tunnused *qseq, am* ja *carb,* mille keskmine seotus teiste tunnustega on nõrgem. Suurim vahe korrelatsioonikordajate absoluutväärtuste keskmiste järjestatud reas on 0,551-0,661. Seega oleks mõistlik valida lävendiks arv just sellest vahemikust. Illustreerimaks, kuivõrd mõjutab tulemust mitte kõige optimaalsema lävendi valik, on joonisel 5.3 esitatud tunnuste seotust illustreeriv diagramm lävendi 0,55 korral. Joonisel 5.4 on sama diagramm lävendi 0,56 korral. Jooniste konstrueerimiseks kasutatav programmikood on ära toodud Lisas 5. Tegu ei ole valmis funktsiooniga, mis töötaks iga ette antud andmestikuga, vaid üksnes töö autori poolt programmeeritud võimaliku funktsiooni *ring.korr* edasiarendusega.

Tabel 5.1. Illustreerimiseks kasutatavad tunnused ja neile vastavad korrelatsioonikordajate absoluutväärtuste keskmised.

mpg	cyl	disp	hp	wt	vs	qseq	am	carb
0.731	0.758	0.729	0.720	0.672	0.661	0.551	0.456	0.548



Joonis 5.3. Korrelatsioonimaatriksi illustratsioon kasutades funktsiooni *ring.korr* edasiarendust valesti määratud lävendi korral.



Joonis 5.4. Korrrelatsioonimaatriksi illustratsioon kasutades funktsiooni *ring.korr* edasiarendust õigesti määratud lävendi korral.

Scatterplots and illustration of correlation matrices in statistical package R

Bachelor

Joosep Raudsik

Summary

It is ordinary to take interest in summary of a set of bivariate data. Commonly the visual analysis gives a general idea of the relationship between two variables. This paper gives an overview of possibilities to illustrate correlation matrices and scatterplots in statistical package R. Also the paper requires basic knowledge of the mentioned program.

Plot and *scatterplot* are the most commonly used functions for illustrating scatterplots in R, in which the first one is used for more basic graphing and the second one uses nonparametric-regression smooths and regression lines to describe the relationship. Also the function lets to divide data to subsets, which are drawn on the plot with different colors and shapes to help distinguish them.

Functions *hexbin* and *sunflowerplot* are used correspondingly to illustrate high density or overlapping data. The first mentioned function divides the plot to hexagons and points high density data out by darkening them. The second function marks overlapping data points using sunflower figures and multiple points are plotted with multiple leaves.

Scatterplot matrices are constructed using functions *scatterplotmatrix* and *pairs*. The elements of the matrix, which are created using the first function, are analogical to scatterplots constructed by the function *scatterplot*. Function *pairs* is more versatile and gives the user possibility of valuing different panels of the matrix with self-written functions.

Function *corrplot* is used to illustrate correlation matrices and confidence intervals. Different arguments like *method, type, order* etc. give a wide range of possibilities to change the correlation matrix. For example the user can change the order of variables, the shapes and colors which are used to illustrate the correlation or point out the statistically insignificant correlations.

Lastly, a function named *ring.korr* is created by the author of this work to illustrate correlation matrices using polygons. The function is used to illustrate the correlations by

connecting the corners of the constructed polygon with lines of different size and color. *Ring.korr* has also arguments to point out statistically insignificant correlations or variables that are more correlated with other.

Kasutatud kirjandus

1. Friendly, M., Denis, D. 2005. The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, 41(2): 103-30. Kättesaadav: <http://www.datavis.ca/papers/friendly-scat.pdf>
2. Statistikapaketi R dokumentatsioon. *A Handbook of Statistical Analyses Using R*. Kättesaadav: <http://cran.r-project.org/web/packages/HSAUR/> [kasutatud: 23. veebruar, 2013]
3. Statistikapaketi R dokumentatsioon. *Generic X-Y Plotting*. Kättesaadav: <http://stat.ethz.ch/R-manual/R-devel/library/graphics/html/plot.html> [kasutatud: 12. veebruar, 2013]
4. Statistikapaketi R dokumentatsioon. *Draw Function Plots*. Kättesaadav: <http://stat.ethz.ch/R-manual/R-devel/library/graphics/html/curve.html> [kasutatud: 11. veebruar, 2013]
5. Abramowitz, M. Stegun, I. A. 1970. *Handbook of Mathematical Functions*. New York: Dover Publications.
6. Statistikapaketi R dokumentatsioon. *Package "car"*. Kättesaadav: <http://cran.r-project.org/web/packages/car/car.pdf> [kasutatud: 2. mai, 2013]
7. Statistikapaketi R dokumentatsioon. *Package "scatterplot3d"*. Kättesaadav: <http://cran.r-project.org/web/packages/scatterplot3d/scatterplot3d.pdf> [kasutatud 2. mai, 2013]
8. Statistikapaketi R dokumentatsioon. *Package "hexbins"*. Kättesaadav: <http://cran.r-project.org/web/packages/hexbin/hexbin.pdf> [kasutatud: 28. veebruar, 2013]
9. Statistikapaketi R dokumentatsioon. *Produce a Sunflower Scatter Plot*. Kättesaadav: <http://stat.ethz.ch/R-manual/R-patched/library/graphics/html/sunflowerplot.html> [kasutatud: 4. veebruar, 2013]
10. Murdoch, D. J., Chow, E. D. 1996. A Graphical Display of Large Correlation Matrices. *The American Statistician*, 50(2): 178-180. Kättesaadav läbi Tartu Ülikooli raamatukogu: <http://www.jstor.org.ezproxy.utlib.ee/stable/i326472> [kasutatud: 3. mai, 2013]
11. Statistikapaketi R dokumentatsioon. *Scatterplot Matrices*. Kättesaadav: <http://stat.ethz.ch/R-manual/R-patched/library/graphics/html/pairs.html> [kasutatud: 26. märts, 2013]

12. Statistikapaketi R dokumentatsioon. *Package "corrplot"*. Kättesaadav: <http://cran.r-project.org/web/packages/corrplot/corrplot.pdf> [kasutatud: 23. märts, 2013]
13. Friendly, M. 2002. Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56, 316-324. Kättesaadav: <http://www.datavis.ca/papers/corrgram.pdf>
14. Statistikapaketi R dokumentatsioon. *Motor Trend Car Road Tests*. Kättesaadav: <http://stat.ethz.ch/R-manual/R-patched/library/datasets/html/mtcars.html> [kasutatud: 23. märts, 2013]

Lisad

Lisa 1. Andmestik *heptathlon*

	hurdles	highjump	shot	run200m	longjump	javelin	run800m	score	kat
Joyner-Kersee	12.69	1.86	15.8	22.56	7.27	45.66	128.51	7291	1
John	12.85	1.8	16.23	23.65	6.71	42.56	126.12	6897	1
Behmer	13.2	1.83	14.2	23.1	6.68	44.54	124.2	6858	1
Sablovskaite	13.61	1.8	15.23	23.92	6.25	42.78	132.24	6540	1
Choubenkova	13.51	1.74	14.76	23.93	6.32	47.46	127.9	6540	1
Schulz	13.75	1.83	13.5	24.65	6.33	42.82	125.79	6411	1
Fleming	13.38	1.8	12.88	23.59	6.37	40.28	132.54	6351	1
Greiner	13.55	1.8	14.13	24.48	6.47	38	133.65	6297	1
Lajbnerova	13.63	1.83	14.28	24.86	6.11	42.2	136.05	6252	1
Bouraga	13.25	1.77	12.62	23.59	6.28	39.06	134.74	6252	1
Wijnsma	13.75	1.86	13.01	25.03	6.34	37.86	131.49	6205	1
Dimitrova	13.24	1.8	12.88	23.59	6.37	40.28	132.54	6171	0
Scheider	13.85	1.86	11.58	24.87	6.05	47.5	134.93	6137	0
Braun	13.71	1.83	13.16	24.78	6.12	44.58	142.82	6109	0
Ruotsalainen	13.79	1.8	12.32	24.61	6.08	45.44	137.06	6101	0
Yuping	13.93	1.86	14.21	25	6.4	38.6	146.67	6087	0
Hagger	13.47	1.8	12.75	25.47	6.34	35.76	138.48	5975	0
Brown	14.07	1.83	12.69	24.83	6.13	44.34	146.43	5972	0
Mulliner	14.39	1.71	12.68	24.92	6.1	37.76	138.02	5746	0
Hautenauve	14.04	1.77	11.81	25.61	5.99	35.68	133.9	5734	0
Kytola	14.31	1.77	11.66	25.69	5.75	39.48	133.35	5686	0
Geremias	14.23	1.71	12.95	25.5	5.5	39.64	144.02	5508	0
Hui-Ing	14.85	1.68	10	25.23	5.47	39.14	137.3	5290	0
Jeong-Mi	14.53	1.71	10.83	26.61	5.5	39.26	139.17	5289	0
Launa	16.42	1.5	11.78	26.16	4.88	46.38	163.43	4566	0

Lisa 2. Funktsiooni *pairs* paneelid

```
panel.hist <- function(x, ...)
{
   usr <- par("usr"); on.exit(par(usr))
   par(usr = c(usr[1:2], 0, 1.5) )
   h <- hist(x, plot = FALSE)
   breaks <- h$breaks; nB <- length(breaks)
   y <- h$counts; y <- y/max(y)
   rect(breaks[-nB], 0, breaks[-1], y, col="cyan", ...)
}
panel.cor <- function(x, y, digits=2, prefix="", cex.cor, ...)
{
   usr <- par("usr"); on.exit(par(usr))
   par(usr = c(0, 1, 0, 1))
   r <- abs(cor(x, y))
   txt <- format(c(r, 0.123456789), digits=digits)[1]
   txt <- paste(prefix, txt, sep="")
   if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
   text(0.5, 0.5, txt, cex = cex.cor * r)
}
```

Lisa 3. Funktsioon *cor.mtest*

```
cor.mtest <- function(mat, conf.level = 0.95){
	mat <- as.matrix(mat)
	n <- ncol(mat)
	p.mat <- lowCI.mat <- uppCI.mat <- matrix(NA, n, n)
	diag(p.mat) <- 0
	diag(lowCI.mat) <- diag(uppCI.mat) <- 1
	for(i in 1:(n-1)){
		for(j in (i+1):n){
			tmp <- cor.test(mat[,i], mat[,j], conf.level = conf.level)
			p.mat[i,j] <- p.mat[j,i] <- tmp$p.value
			lowCI.mat[i,j] <- lowCI.mat[j,i] <- tmp$conf.int[1]
			uppCI.mat[i,j] <- uppCI.mat[j,i] <- tmp$conf.int[2]
		}
	}
	return(list(p.mat, lowCI.mat, uppCI.mat))
}
```

Lisa 4. Funktsioon *ring.korr*

```
##Polaarkoordinaatide teisendusfunktsioonid ringy ja ringx
#vastavalt nurkade arvule n ja hulgnurga nurkade kaugusele r
#arvutatakse välja x- ja y-koordinaadid nurkadele.

ringy<-function(n,r){
        vahe=360/n
        fii=seq(1,n)
        fii=vahe*fii
        fii=2*pi*fii/360
        y=r*sin(fii)
        return (y)
}
ringx<-function(n,r){
        vahe=360/n
        fii=seq(1,n)
        fii=vahe*fii
        fii=2*pi*fii/360
        x=r*cos(fii)
        return (x)
}

##Funktsiooni definitsioon

ring.korr<-function(andmestik, sig=FALSE, col=FALSE, sig.level=0.05,  r=20, kir=FALSE,
rtext=8, cex=1, main=paste(deparse(substitute(andmestik)))){
        M=as.matrix(cor(andmestik))
        n=sqrt(length(M))

        #Värvivektori seadistamine palette abil
        n <- ncol(andmestik)

        #Korrelatsioonikordajatele vastavad p-väärtused
        p.mat <- matrix(NA, n, n)
        for(i in 1:(n-1)){
                for(j in (i+1):n){
                        tmp <- cor.test(andmestik[,i], andmestik[,j])
                        p.mat[i,j] <- p.mat[j,i] <- tmp$p.value
        }}
        diag(p.mat) <- 0

        #Statistilise olulisuse määramine
        p.mat[sig.level<p.mat ] <- 1
        if(sig==TRUE){
                p.mat[sig.level<p.mat ] <- 3
        }
```

```
        p.mat[p.mat < sig.level]<- 1

        #Polaarkoordinaatidelt üleminek
        x=ringx(n,r)
        y=ringy(n,r)

        #Sobiva suurusega joonise välja kutsumine
        plot( (-r-8):(r+8), (-r-8):(r+8), asp=1, main=main, xlab="", ylab="", type="n",
        xaxt="n", yaxt="n" )

        #Arvutatakse välja koordinaadid tunnuste nimedele
        xn=ringx(n,r+rtext)
        yn=ringy(n,r+rtext)
        maks=1/max(rowMeans(abs(M)))

        ##Kasutades arvutatud koordinaate, lisatakse tunnuste nimed joonisele.
        #Värvi tumeduse ja joone laiuse valikul kasutatakse korrelatsioonikordajate
        absoluutväärtuste maatriksit.
        abs2=abs(M)
        if(kir==TRUE){
                abs2=abs2*cex
        }
        else{
                abs2=abs2+1-abs2
        }
        absM=abs(M); Mcol=abs(M)
        if(col==FALSE){
                Mcol=Mcol-Mcol+1
        }
        text( x=xn, y=yn, labels=paste(attributes(andmestik)$names,
        round(rowMeans(absM),2)), col=rgb(1-rowMeans(Mcol), 1-rowMeans(Mcol), 1-
        rowMeans(Mcol)), cex=cex*(abs2) )

        ##Lisatakse korrelatsioonikordajaid illustreerivad jooned
        for(i in 1:n){
                for(j in 1:n){
                        if (M[i,j]<0){
                        lines( c(x[i], x[j]), c(y[i], y[j]), col=hsv(h=0.65, s=-M[i,j], v=1),
                        lwd=round(5*-M[i,j]) ,lty=p.mat[i,j] )
                        }
                        else{
                        lines( c(x[i], x[j]), c(y[i], y[j]), col=hsv(h=1, s=M[i,j], v=1),
                        lwd=round(5*M[i,j]) ,lty=p.mat[i,j])
                        }
                }
        }
}
```

Lisa 5. Funktsiooni *ring.korr* edasiarendus

```
mtcars=mtcars[,-10]
mtcars=mtcars[,-5]

#Polaarkoordinaatide teisendusfunktsioonid
ringy<-function(n,r){
        vahe=360/n
        fii=seq(1,n)
        fii=vahe*fii
        fii=2*pi*fii/360
        y=r*sin(fii)
        return (y)
}

ringx<-function(n,r){
        vahe=360/n
        fii=seq(1,n)
        fii=vahe*fii
        fii=2*pi*fii/360
        x=r*cos(fii)
        return (x)
}




ring.korr<-function(M){
        r=20
        n=sqrt(length(M))

        #Polaarkoordinaatidelt üleminek
        x=ringx(n,r)
        y=ringy(n,r)
        maks=1/max(rowMeans(abs(M)))

        l=attributes(mtcars)$names
        l=l[s!=0]

        s=rowMeans(abs(M))
        absM=abs(M)
```

```
	##JOONTE LISAMINE

	for(i in 1:n){
		for(j in 1:n){
			if (M[i,j]<0){
			lines(c(x[i], x[j]), c(y[i], y[j]), col=hsv(h = 0.65, s = -M[i,j], v =
			1),lwd=round(5*-M[i,j]))
			}
			else{
			lines(c(x[i], x[j]), c(y[i], y[j]), col=hsv(h = 1, s = M[i,j], v =
			1),lwd=round(5*M[i,j]))
			}
		}
	}
}

#Andmestik iseenesest
M=as.matrix(cor(mtcars))
s=rowMeans(abs(M))
s[s<0.56]=0
U=M[(s==0),(s!=0)]
S=M[(s!=0),(s!=0)]
maksimum=max.col(U)
r=20
vahe=360/sqrt(length(S))
fii=vahe*maksimum
fii=2*pi*fii/360
y2=(r+10)*sin(fii)
x2=(r+10)*cos(fii)
J=M[(s==0),]
n=sqrt(length(S))
x=ringx(n,r)
y=ringy(n,r)
plot((-r-8):(r+8),(-r-8):(r+8), type="n",asp=1)
#JOONTE LISAMINE
for(i in 1:(length(y2))){
	for(j in 1:(length(U[1,]))){
		if (U[i,j]<0){
		lines(c(x2[i], x[j]), c(y2[i], y[j]), col=hsv(h = 0.65, s = -U[i,j], v =
		1),lwd=round(5*-U[i,j]))
		}
		else{
		lines(c(x2[i], x[j]), c(y2[i], y[j]), col=hsv(h = 1, s = U[i,j], v =
	1),lwd=round(5*U[i,j]))
		}
```

```
        }
}

l=attributes(mtcars)$names
l=l[s!=0]        #SEES NIMED
f=attributes(mtcars)$names
f=f[s==0]        #ÜMBER NIMED

s=rowMeans(abs(M))
absM=abs(M)

ring.korr(S)
xn=ringx(n,r+1)
yn=ringy(n,r+1)
text(x=xn,y=yn,labels=paste(l) ) #SEESMISED
text(x=x2,y=y2,labels=paste(f) ) #VÄLIMISED
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina Joosep Raudsik

(*autori nimi*)

(sünnikuupäev 22.01.1991)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
Hajuvusdiagrammide ning korrelatsioonimaatriksite illustreerimine statistikapaketis R,

(*lõputöö pealkiri*)

mille juhendaja on Tanel Kaart,

(*juhendaja nimi*)

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.
3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **06.05.2013**