Luku 3
Datan visualisoinnista

Seuraavassa muutama sana datan esittämisestä kuvan muodossa. Kuvien tarkoituksena ei ole ainoastaan esittää data-analyysin tuloksia, vaan niitä tarvitaan analyysin joka vaiheessa. Kuten muuallakin tässä kurssissa, seuraavassa rajoitutaan lähinnä jatkuvista muuttujista tehtyjen mittausten käsittelyyn.

Yhden muuttujan mittaukset voidaan esittää **pistediagrammina**, jossa kutakin havaintoa vastaa janalle piirretty piste. Tämä onnistuu käytännössä vain, jos pisteitä on vähän. Muussa tapauksessa on luontevinta käyttää **histogrammia**, jossa kunkin pylvään korkeus kertoo havaintopisteiden lukumäärän pylvästä vastaavalla välillä. Histogrammin pylväiden sijainnin ja leveyden määrittämiseen käytetään usein nk. *Sturgesin sääntöjä*. Histogrammin pylväät pyritään erityisesti piirtämään niin, että kukin pylväs vastaa vähintään ~ 10 havaintopistettä. Tämä ei ole aina käytännössä mahdollista - ainakin jos kaikki histogrammin pylväät ovat samanlevyisiä. Joissakin ohjelmissa (esim. *R*) voidaan käyttää myös vaihtuvan levyisiä pylväitä, jolloin saadaan parempi resoluutio muuttujan suhteen sinne, missä dataa on paljon. Toinen vaihtoehto on tehdä muuttujanvaihdos ennen pylväiden laskemista. *Box and whiskers* – **laatikko ja viikset** (?) on eräs vaihtoehto histogrammille otoksen jakauman kuvaamiseen (kts. kuva 3.13). Kuva on periaatteessa yksiulotteinen. Piirretty laatikko kuvaa jakauman kvartaalivälin ja laatikon jakava poikkiviiva mediaanin. Laatikosta lähtevät 'viikset' ulottuvat

laatikosta ylös- ja alaspäin uloimpaan havaintopisteeseen, kuitenkin korkeintaan esim. kahta kvartaaliväliä vastaavan matkan. Näiden ulkopuolelle jäävät yksittäiset havaintopisteet piirretään myös kuvaan.

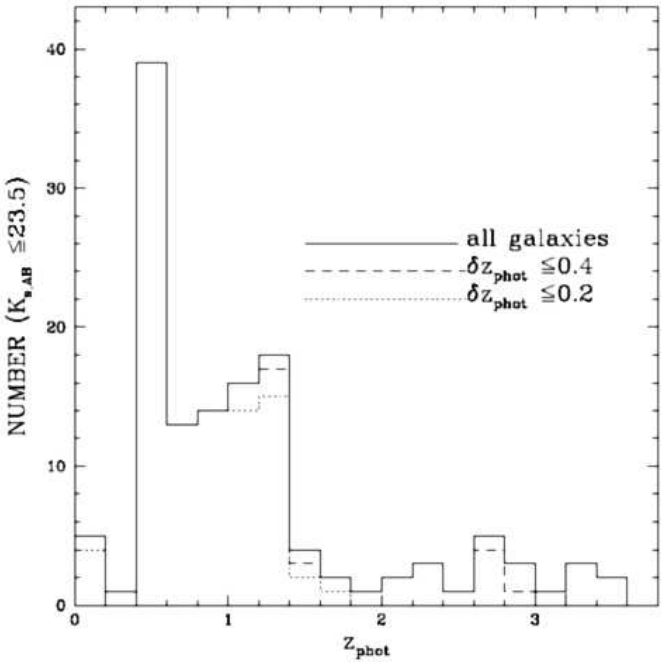


FIG. 8.—Redshift histogram of all 133 objects in our catalog with reliable redshifts (*solid line*). The two other histograms show the redshift distributions for all objects with $\delta z_{phot} \leq 0.4$ (*dashed line*) and all objects with $\delta z_{phot} \leq 0.2$ (*dotted line*), where the photometric redshift errors are the combination of those calculated using our Monte Carlo technique with the systematic errors determined from the HDF-N.

Kuva 3.1. Rudnick et al. 2001, *A K-Band Selected Photometric Redshift Catalog in the Hubble Deep Field South: Sampling the rest-frame V band to z 3*, Astronomical Journal 122, 2205.

Varsi ja lehdet (*stem and leaves*) on histogrammin versio, johon kirjataan havaintojen numeroarvot. Graafin varren muodostavat allekkain kirjoitetut lukuarvojen ensimmäiset merkitsevät numerot. Samoilla numeroilla alkavista havainnoista kirjoitetaan toinen merkitsevä numero samalle riville - näitä lehtiä on siis yhtä monta kuin samoilla varren numeroilla alkavia havaintoja on. Esimerkiksi havainnot 11, 25, 33, 24, 32, 10, ja 23 voidaan esittää diagrammina

1 10

2 543

3 32

Kaksiulotteisen datan tapauksessa useimmiten luontevin esitys on **kaksiulotteinen pistediagrammi**, johon kukin havaintopiste piirretään valitulla symbolilla. Tätä kutsutaan joskus myös hajontakuvioksi (*scatter plot*). Eri symboleita käyttämällä samaan kuvaan voidaan yhdistää eri aineistoja - olettaen, että lopputulos on vielä riittävän selkeä. Symbolien on oltava riittävän suuria, mutta ne eivät kuitenkaan peittää liiaksi toisiaan. Jos pisteitä on suuri määrä, voi eri aineistot yrittää piirtää eri väreillä. Yksittäisen pisteen symboli voi tällöin olla pienempi, ja havaintojen jakaumasta jää silti selvä yleiskuva. Edelleen on varottava symbolien piirtämistä päällekkäin - viimeksi piirretyt pisteet voivat peittää merkittävän osan aiemmin piirretyistä jolloin lopputulos on harhaanjohtava. Joskus pisteiden paikat ovat kvantittuneet, niin että suuri joukko pisteitä tulee piirretyksi samaan paikkaan. Lopputulos on jälleen harhaanjohtava, sillä lopullisessa kuvassa yksi merkki voi tarkoittaa yhtä hyvin miljoonaa kuin yhtä ainoaa mittausta. Kuvaa voi parantaa siirtämällä kutakin pistettä sattumanvaraisesti satunnaiseen suuntaan (**täristys**, *jitter*), jolloin yksittäiset pisteet hajoavat pistejoukoiksi, joista havaintojen lukumäärä voidaan arvioida.

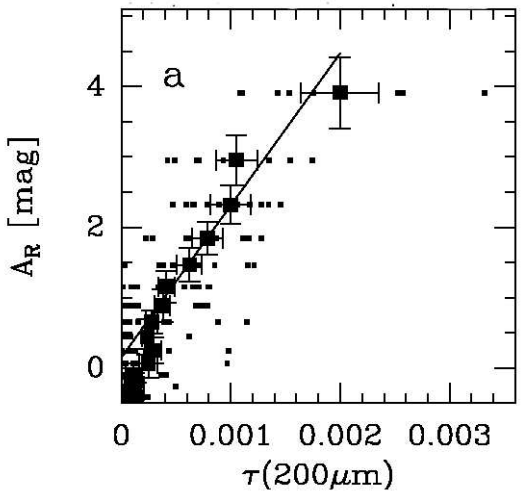
Data skaalataan molempien akselien suhteen niin, että käytettävissä oleva tila käytetään mahdollisimman hyvin hyödyksi. Jos muuttujan vaihteluväli on hyvin suuri, kannattaa harkita logaritmista asteikkoa. Myös muuttujien välille oletettu relaatio vaikuttaa akselien valintaan. Esim. potenssilakia noudattavien havaintojen tapauksessa kannattaa pystyakselille valita logaritminen asteikko, jolloin kuvassa muuttujien korrelaatio on lineaarinen, ja kulmakerroin antaa relaation eksponentin. Jos käsiteltävänä on kaksi aineistoa, joissa y-arvojen oletetaan poikkeavan vakiolla, pyritään käyttämään lineaarista asteikkoa. Jos y-arvot taasen poikkeavat vakiokertoimella, logaritmisella asteikolla ero muuntuu vakiosiirtymäksi ja on helpommin hahmotettavissa.

Kaksiulotteiseen kuvaan voidaan sisällyttää vielä tietoa useammista muuttujista. Kunkin pisteen symbolin kokoa, orientaatiota tai väriä muuttamalla saadaan kuvaan mahdollisesti useampikin lisämuuttuja. Yleisin sovellus lienee vektorikentän kuvaus, jolloin symbolina käytetyn nuolen orientaatio ja pituus kuvaavat kuhunkin pisteeseen liittyvää vektoria. Muita yleisiä tekniikoita ovat mm. **tähtipiirrokset** (*star plots, segment diagram*) tai kuuluisat, mutta (ainakin fysikaalisissa tieteissä) harvemmin käytetyt **Chernovin kasvokuvat** (*Chernov faces*). Tähtipiirroksessa kutakin havaintopistettä vastaa tähtikuvio (monikulmio), jonka sakaroiden pituudet vastaavat eri suureiden arvoja ko. pisteessä. Chernovin kasvoissa eri muuttujien arvot kuvataan tyyliteltyjen kasvojen eri piirteiden muutoksiksi (esim. silmien tai suun koko tai vinous, nenän pituus jne.). Menetelmän taustalla on sinänsä hyvä idea käyttää muuttujien koodaukseen sellaisia kuvioita, joiden analysoinnissa aivot ovat kehittyneet hyviksi. Useimmiten yksinkertainen **rinnakkaispiirros** (*parallel plot*) on hyödyllisempi. Siinä eri muuttujille piirretään vierekkäin pystysuorat koordinaattiakselit, joille havainnot merkitään – siis kukin havaintopiste jokaiselle akselille. Lopuksi samaan havaintoon liittyvät pisteet yhdistetään janoilla. Paperille piirrettynä kuvasta tulee sekava, mutta tietokoneen ruudulla kuva on melko havainnollinen, kunhan yksittäisiä pisteitä voi-

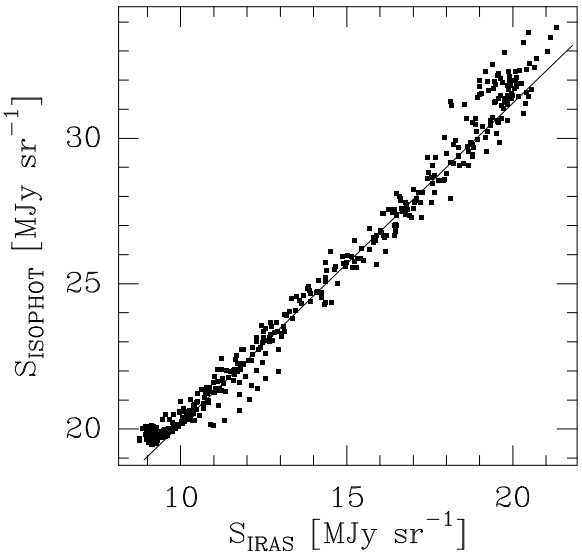
daan interaktiivisesti valita ja esim. piirtää eri väreillä (vrt. alla kuvattu *brushing*–tekniikka).

Useampiulotteisen (>3) aineiston visualisointi on aina hankalaa. Sen sijaan että kaikki muuttujat yritetään saada samaan kuvaan, voi olla parempi etsiä muuttujien kombinaatioita, joiden suhteen havainnot esitetään. Näihin **dimensionaalisuuden pienennyskeinoihin** palataan myöhemmin.

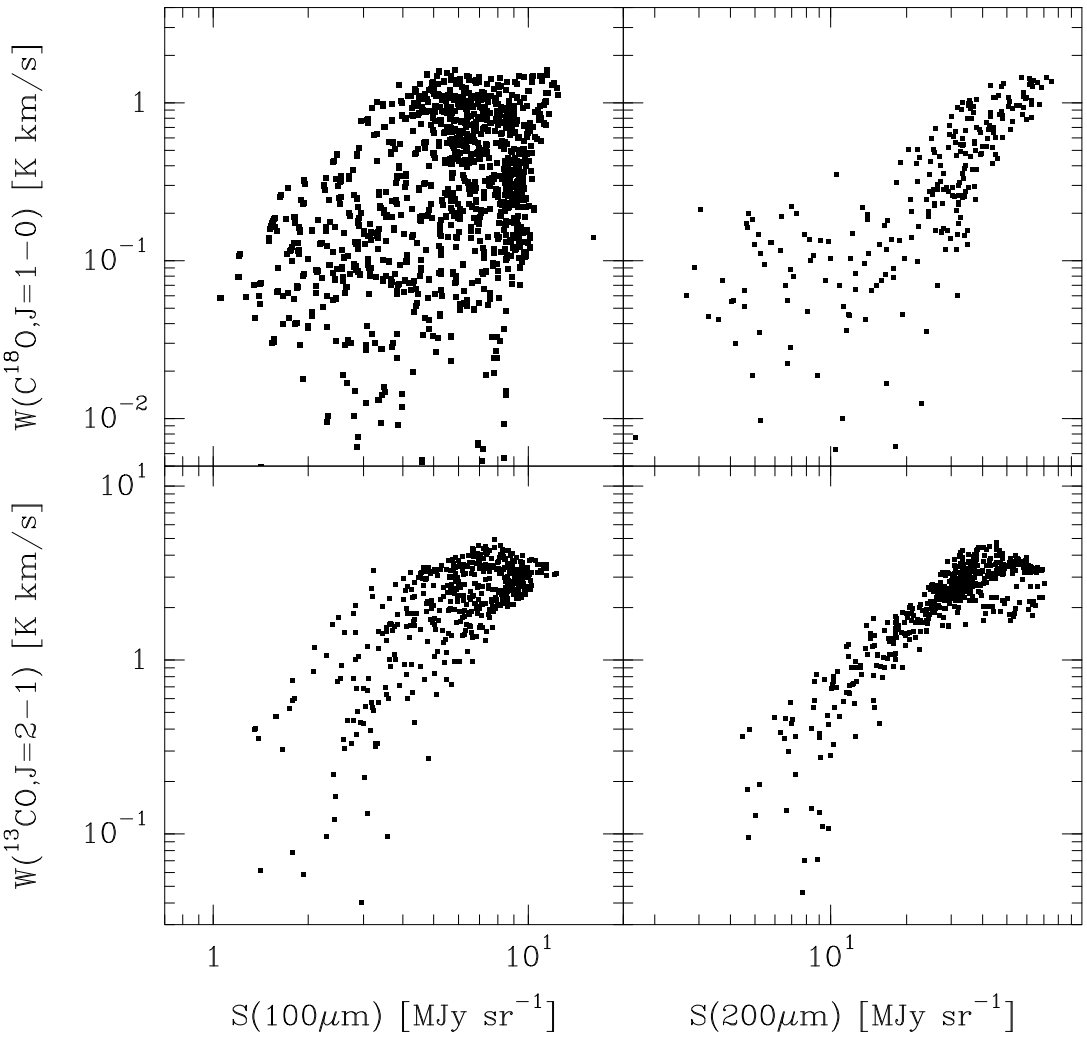
Esimerkki 3.1. Eräitä kuvia artikkelista Juvela et al. 2002, *Far-infrared and molecular line observations of Lynds L183 -studies of cold gas and dust*, A&A 382, 583.



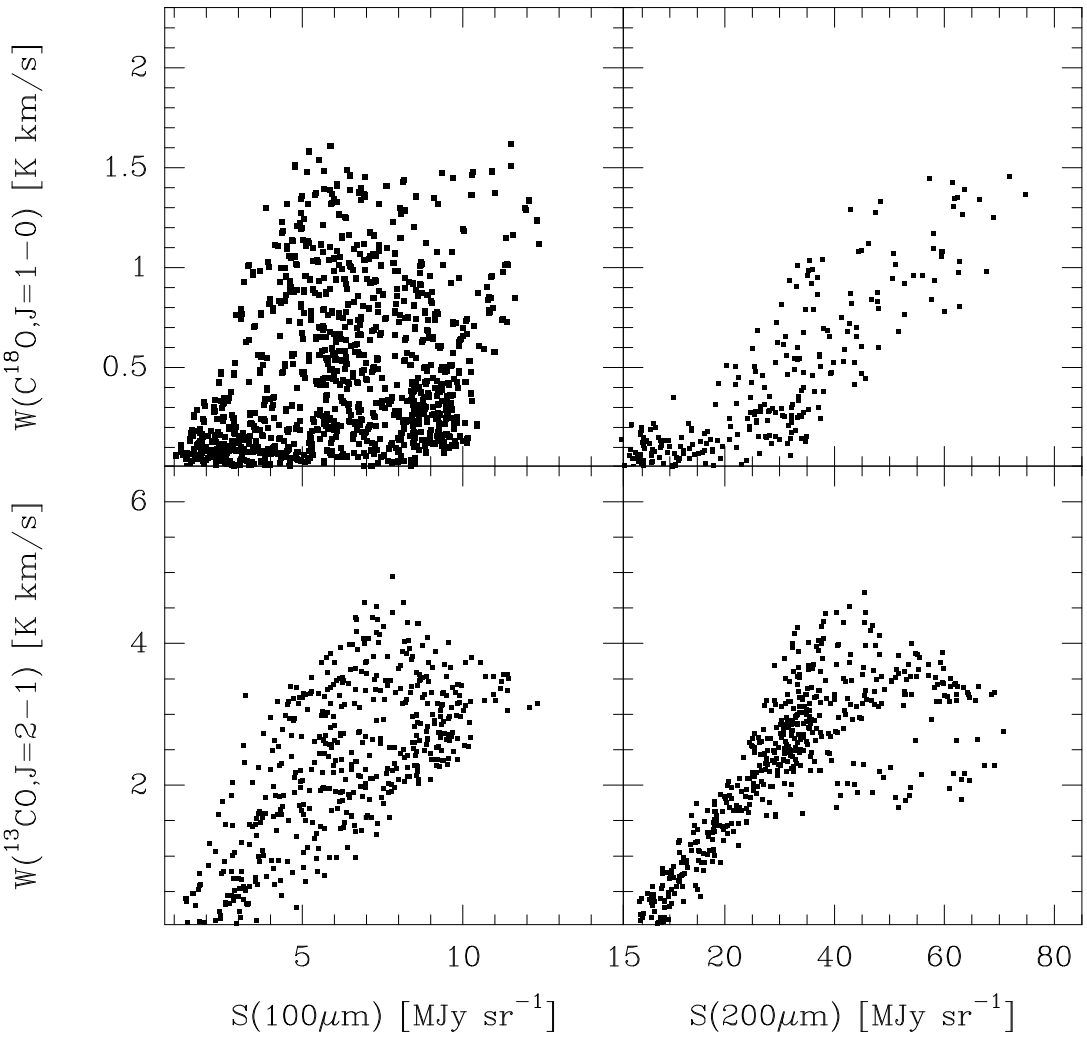
Kuva 3.2. R-kaistan ekstinktio 200μm pintakirkkauden funktiona: A_R:n mukaan luokiteltujen pistejoukkojen keskiarvot virherajoineen, sekä näihin sovitettu pns-suora. Onko koordinaattiakselien skaalaus onnistunut? Entä piirtosymbolien valinta?



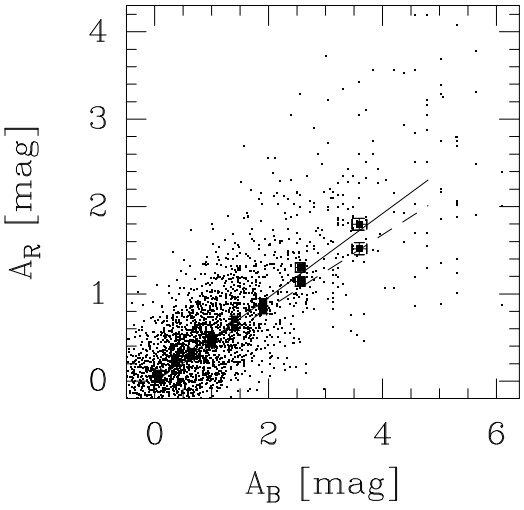
Kuva 3.3. ISOPHOT instrumentin ja IRAS satelliitin mittaamien 100μm pintakirkkauksien korrelaatio pilven L183 alueella. Suora on painotettu pns-suora.



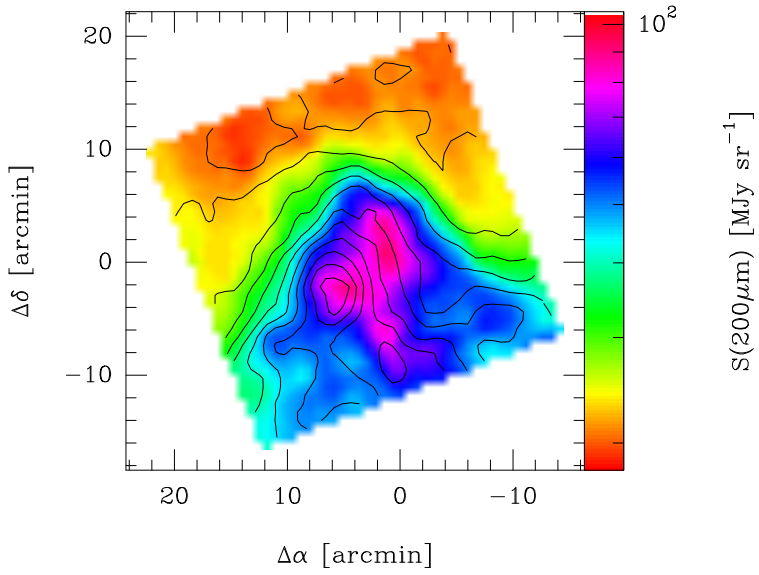
Kuva 3.4. Vertailu 100μm ja 200μm pintakirkkauksien sekä ^{13}CO ja C^{18}O molekyylien lähettämän viivasateilyn intensiteettien välillä. Artikkeliin valittiin molemmille suureille logaritmiset asteikot.



Kuva 3.5. Edellinen kuva, tällä kertaa lineaarisilla asteikoilla.



Kuva 3.6. *R*– ja *B*–kaistojen ekstinktio tähtilaskennoista johdettuna. Ekstinktio on kääntäen verrannollinen tähtien määrään yksittäisessä laskentaruudussa. Koska tähtien määrät ovat pieniä kokonaislukuja, ekstinktion arvot ovat kvantisoituneet ja piirrettävät pisteet osuisivat kuvassa harvoihin positioihin. Satunnaisen täristyksen sijasta eksinktioarvoista on laskettu keskiarvo yli kiekon, jonka koko on hieman suurempi kuin laskentaruudun. Näin pisteet on saatu poikkeamaan toisistaan, niin että todellinen jakauma tulee näkyviin. Suuremmat symbolit virherajoineen kuvaavat keskiarvoja yli laskentaruutujen, joissa on sama määrä *B*-kaistan tähtiä. Yhtenäinen viiva on pns-suora sovitettuna kaikkiin pisteisiin, ja katkonainen viiva on pns-suora sovitettuna edellä mainittuihin keskiarvoihin.



Kuva 3.7. Kuvassa korkeuskäyrät esittävät 100μm pintakirkkaujakaumaa ja taustan väritys vastaavaa 200μm emission jakaumaa. Sopivalla värikartan valinnalla ja skaalauksella saadaan selvästi harmaasävykuvaa havainnollisempi esitys.

Edelliset kuvat ovat peräisin valmiista artikkelista, mutta etsivän data-analyysin vaiheessa tehtävät kuvat ovat hyvin samankaltaisia.

Kolmiulotteinen data voidaan kuvata vielä hyvin kaksiulotteisessa kuvassa, jos kutakin pisteparia (x, y) vastaa yksi z-arvo. Piirrossymbolien vaihtelun sijasta voidaan z-muuttujan vaihtelu havainnollistaa **korkeuskäyrien** avulla. Käytännössä tämä edellyttää, että (x, y)- tasosta on riittävä määrä (mieluusti tasavälisiä) havaintoja. Käytännössä korkeuskäyrien piirtäminen edellyttää interpolointia datapisteiden väliin. Interpolointi vuoksi lopullinen kuva on aina hieman harhaanjohtava. Lopullisessa kuvassa z arvot on kuvattu kaikkialle, vaikka alunperin havainnot muodostavat vain diskreetin pistejoukon. Ongelma on sitä pahempi, mitä vähemmän dataa on alunperin. Lisäksi (ainakin joissakin ohjelmistoissa) interpolointi voi aiheuttaa oskillaatioita ym. artefakteja. Korkeuskäyrien valintaan on syytä kiinnittää erityistä huomiota. Havaintodatan tapauksessa alin korkeuskäyrä voidaan piirtää tasolle, joka vastaa vähintään 1σ tasoa havaintovirheisiin nähden. Samoin perättäisten korkeuskäyrien välin on oltava z–muuttujan suhteen vähintään saman suuruinen. Valintaan vaikuttaa myös alkuperäisten (x, y)–pisteiden lukumäärä: vierekkäisten mittauspisteiden välistä ei saisi mielellään kulkea enempää kuin yksi korkeuskäyrä. Varsinkin tietokoneella työskennellessä kannattaa käyttää värejä niin, että kukin korkeuskäyräväli väritetään eri värillä. Vielä parempaan tulokseen päästään, kun z-muuttujan arvo koodataan väreiksi jatkuvan **värikartan** avulla (*colour (lookup) table*). Sopivalla värikartan valinnalla ja skaalauksella voidaan kuvasta ottaa esiin pieniäkin z-arvon muutoksia (tai häivyttää suuriakin arvojen hyppäyksiä...). Julkaisuissa joudutaan kustannussyistä tyytymään usein harmaasävykuviin, jolloin kuvan selvyys selvästi kärsii.

Monimuuttujadatasta tehtävässä **korrelaatiodiagrammissa** ('paripiirros', *pair plot*, 'samanaikaiset hajontakuviot') kukin muuttuja piirretään 2D pistediagrammina erikseen kunkin toisen muuttujan suhteen. Diagrammit on järjestetty matriisin muotoon, jossa ainoastaan kuvat muuttujasta itsensä suhteen on poistettu. Tämä ei vielä tyhjentävästi kuvaa alkuperäistä n-ulotteista muuttuja-avaruutta, mutta paljastaa heti yksittäisten muuttujien väliset korrelaatiot.

Yleisestä kolmiulotteista dataa voidaan vielä havainnollistaa kolmiulotteisesta kohteesta tehdyillä kaksiulotteisilla (siis valokuvan tapaisilla) **projektiokuvilla**. Kolmiulotteisen muuttujan arvojen jakaumaa voidaan esittää esim. korkeuskäyrien yleistyksellä eli **tasa-arvopinnoilla** (*isocontours*). Tällöin voidaan helposti visualisoida ainoastaan yksi pinta, jolla muuttujan arvo kasvaa annettua rajaa

suuremmaksi. Usean pinnan tapauksessahan pinnat todennäköisesti peittävät toisensa näkyvistä. Tätä voi yrittää korjata tekemällä alemmat pinnat osittain läpinäkyviksi, mutta tällöinkin voidaan näitä esittää korkeintaan pari kappaletta. Vaihtoehtoisessa **tilavuusvisualisoinnissa** (*volume visualization*) jakauma esitettään kokonaisuudessaan läpikuultavana 'sumupilvenä'. Kolmiulotteisesta kuvasta on vaikea lukea myös tietyn pisteen koordinaatteja tarkasti. Jos halutaan eksaktimpi kuva myös koordinaateista, voi ainoa vaihtoehto olla esittää datan kaksiulotteisia leikkauksia. Kolmiulotteisesta jakaumasta voidaan myös ikäänkuin leikata pois pala, jolloin leikkauspinta voidaan visualisoida kuten 2D tapauksessa (esim. omat koordinaattiakselit ja korkeuskäyräkuva).

Korkeuspintojen lisäksi voidaan 2D tapauksen lailla esittää vektorisuureet **3D nuolilla**. Vektorikenttiä voidaan havainnollistaa myös piirtämällä **virtausviivat** (*streamlines*). Edellisessä tapauksessa vektorisuureen itseisarvo koodataan symbolin kokoon, jälkimmäisessä se voidaan esittää värien avulla. Väritykseen voidaan molemmissa tapauksissa käyttää myös jotakin muuta muuttujaa.

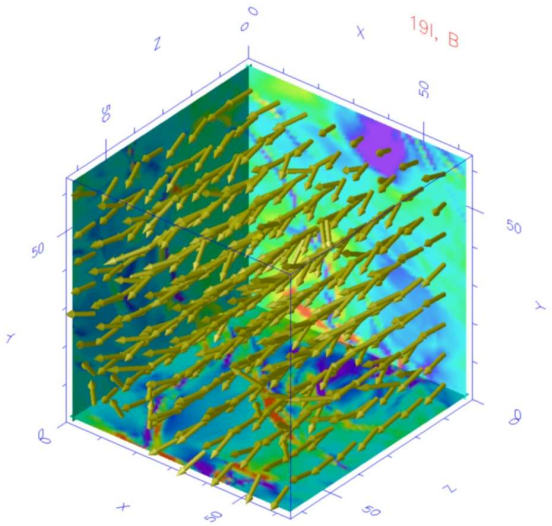
3D kuvien havainnollisuutta voi parantaa jonkin verran **stereoskooppisten** kuvien avulla. Tämän ohessa tai sijasta voidaan käyttää myös dynaamista visualisointia. Yksinkertaisimmillaan tämä voisi olla esim. 3D pistejoukon visualisointia niin, että pisteistä muodostettua kuvaa voidaan pyörittää interaktiivisesti tietokoneen kuvaruudulla. Tähän voidaan yhdistää muita tekniikoita (pisteiden symbolit, pisteiden värittäminen, pisteiden vilkkuminen). Mistä tahansa kuvasarjasta voidaan tehdä **animaatio**. Jos ohjelmisto ei itsessään tue animaatioiden luomista, voidaan yksittäiset kuvat tallettaa levylle ja koota näistä esim. gif- tai mpeg-muotoinen animaatio erillisellä ohjelmalla.

Mahdollisuus eri kuvien **linkittämiseen** ja datapisteiden **maalaamiseen** (*brushing*) ovat sangen hyödyllisiä monimuuttujadatan interaktiivisessa katselussa. Maalaamisessa yksittäisiä data-pisteitä tai pistejoukko valitaan (esim. hiirellä) yhdestä kuvasta. Valitut pisteet korostuvat automaattisesti kaikissa linkitetyissä kuvissa esim. pisteiden väriä vaihtamalla. Näin voidaan esim. korrelaatiodiagrammasta valita tiettyjen muuttujien suhteen poikkeavat arvot, ja muista kuvista nähdään, minne pisteet sijoittuvat muiden muuttujien suhteen. Kuvasta voidaan myös suoraan poimia yksittäisiä pisteitä ja selvittää, mitä tiettyä mittausta ne vastaavat.

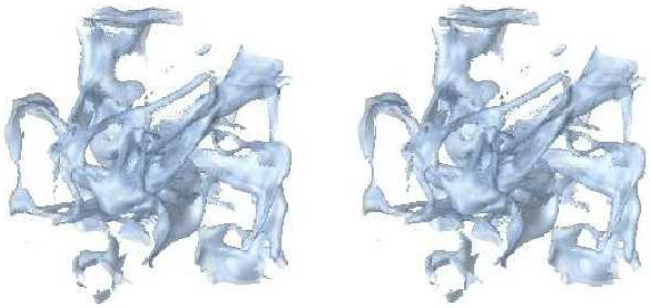
Verkosta löytyy joitakin ilmaisia ja varsin päteviä ohjelmistoja kolmi– ja useampiulotteisen datan visualisointiin. Alla on esimerkkinä joitakin **OpenDX** ohjelmalla (`www.opendx.org`) tehtyjä kuvia. Toinen ilmainen vaihtoehto visualisaatioiden tekoon on **VTK** (*Visualization Toolkit*, `public.kitware.com/VTK/`). Molemmat ohjelmistot ovat kutakuinkin samantasoisia – VTK on hieman monipuolisempi, mutta samalla vaikeammin lähestyttävä (vaatii c++– , tcl– tai java–ohjelmointia, mikä kuitenkin onnistuu esimerkkien pohjalta kohtuullisella vaivalla). OpenDX:n käyttö on vasta-alkajalle helpompaa, sillä visualisoinnin voi koota graafisen ohjelmaeditorin avulla – itse asiassa ohjelma osaa tehdä yksinker-

taisen kuvan automaattisesti, pelkän datatiedoston perusteella. Edellä annetuista verkko-osoitteista löytyy runsaasti lisäesimerkkejä. Myös IDL sisältää varsin helppokäyttöisen (mutta hieman suppeamman ?) valikoiman rutiineja 3-D visualisointiin. Vaikein askel voi olla oman datan saattaminen ohjelmistojen luettavaan muotoon.

Esimerkki 3.2. Joitakin OpenDX ohjelmalla (`www.opendx.org`) tehtyjä magnetohydrodynaamista pilvimallia esitteleviä kuvia (MHD malli: P.Padoan, JPL). Toinen vaihtoehto samantapaisten visualisaatioiden tekemiseksi on VTK (*Visualization Toolkit*, `public.kitware.com/VTK/`). Molemmat ovat kutakuinkin samantasoisia – VTK on hieman monipuolisempi, mutta vastaavasti OpenDX:n käyttö on vasta-alkajalle helpompaa. Vaikein askel voi olla oman datan saattaminen ko. ohjelmistojen

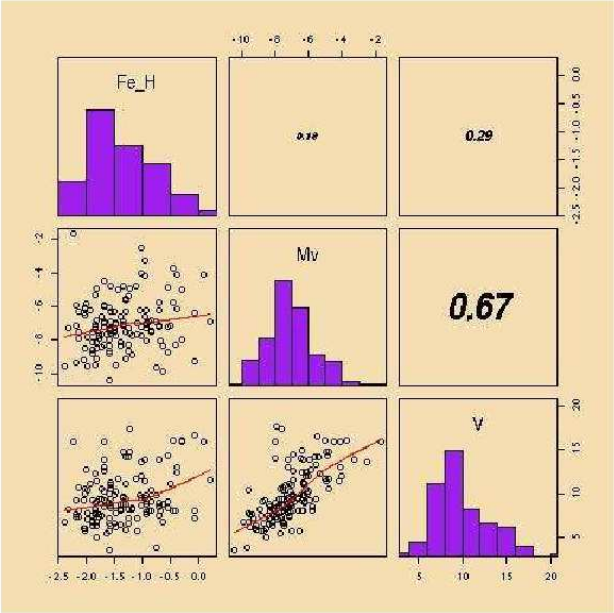


Kuva 3.8. Magneettikentän suunta ja voimakkuus 3D nuolilla piirrettynä. Magneettikentän paikallinen voimakkuus on ilmaistu nuolen pituudella. Kuution kolmella sivustalla kentän voimakkuus on esitetty väreillä.

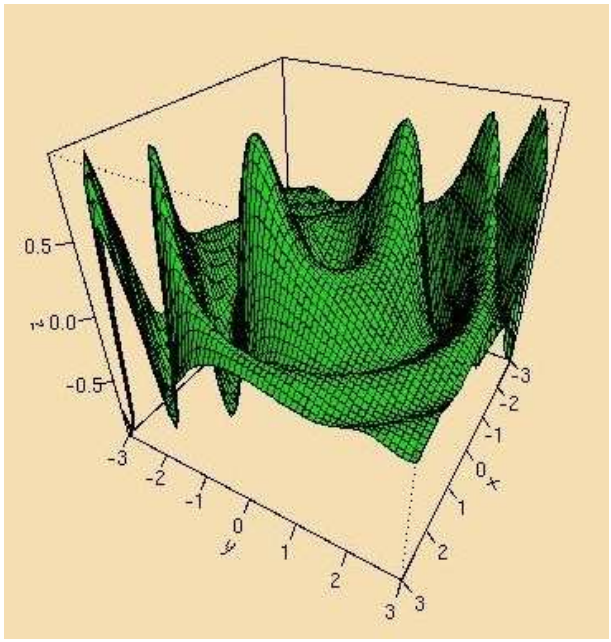


Kuva 3.9. Tiheyskenttä samassa MHD mallissa: tasa-arvopinta kätkee sisäänsä tilavuuden, jossa tiheys kohoaa annetun rajan yläpuolelle. Stereokuvassa kuvien ero vastaa mallin yhden asteen kiertoa pystysuoran akselin suhteen.

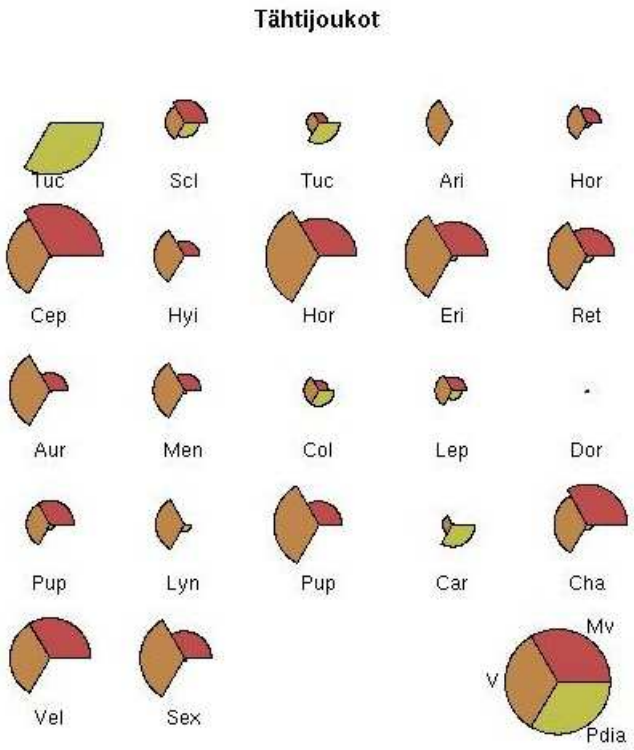
Esimerkki 3.3. Kokoelma R-ohjelmistolla tehtyjä kuvia. Kuvien tekemiseen käytetyt lähdekoodit löytyvät joko kurssin kotisivulta tai suoraan *R*:n html-manuaalista.



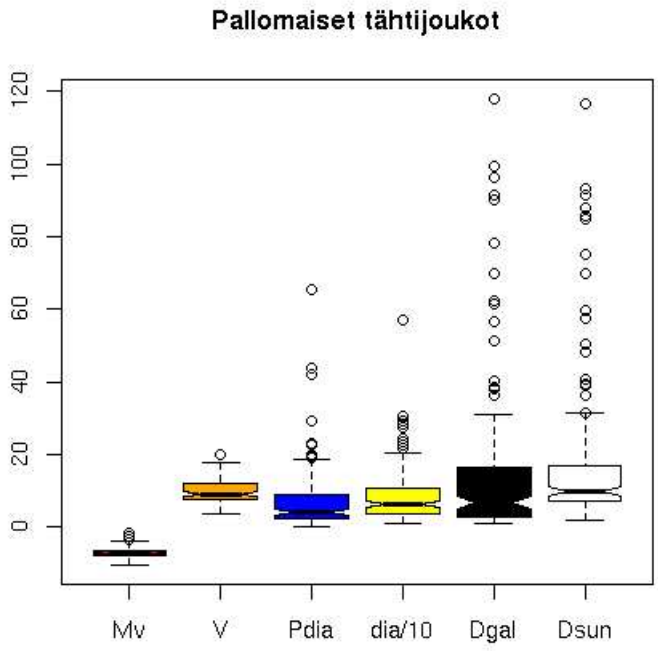
Kuva 3.10. Muuttujien väliset korrelaatiot. Kuvassa on normaali korrelaatiodiagrammi, jossa (kolme osakuvaa alhaalla vasemmalla) on kuvattu havaintopisteet vuorotellen kahden eri muuttujan suhteen. Diagonaalilla olevat osakuvat sisältävät tavallisesti vain muuttujien nimet – tässä kuvaan on lisätty yksittäisen muuttujan jakaumaa kuvaavat histogrammit. Oletusarvoisesti diagonaalin yläpuolella olevat kuvat ovat peilikuvia diagonaalin alapuolisista. Tässä näihin kuviin on tämän sijaan kirjoitettu samalla rivillä ja sarakkeella olevien muuttujien väliset korrelaatiokertoimet – myös kirjasimen koko on verrannollinen korrelaatiokertoimen arvoon.



Kuva 3.11. Rutiinilla `persp` tehty kuva kaksiulotteisesta funktiosta.



Kuva 3.12. Rutiinilla `stars` tehty esitys 22 tähtijoukon ominaisuuksista kolmen muuttujan suhteen. Kunkin sektorin säde ilmoittaa vastaavan muuttujan arvon kohteessa. Kukin tähtijoukko on kuvassa nimetty vastaavan tähdistön nimellä. Oikean alakulman ympyrä kertoo sektoreita vastaavien muuttujien nimet. Samalla rutiinilla `stars` voidaan tehdä myös varsinainen *starplot*, jossa kutakin muuttujaa vastaa yksi sakara tai monikulmion kärki (*spider plot*).



Kuva 3.13. *Box and whiskers*– piirros. Havaintopisteiden jakauma kunkin muuttujan suhteen on piirretty erikseen. Laatikko vastaa kvartaaliväliä, ja laatikon jakava viiva on jakauman mediaani. Vaakaviivoihin päättyvät 'viikset' ulottuvat äärimmäiseen havaintopisteeseen tai, tässä tapauksessa, korkeintaan 1.5 kertaa kvartaalivälin etäisyydelle laatikosta. Jos kahteen laatikkoon piirretyt lovet eivät leikkaa, ovat jakaumien mediaanit merkitsevästi toisistaan poikkeavia (α=5%).

Luku 4
Lineaariset mallit

Datan mallinnuksessa havainnoille pyritään kehittämään (laskennallinen) malli, joka selittää saadut mittaustulokset. Toisaalta mallin on syytä olla niin yksinkertainen kuin mahdollista. Seuraavassa tarkastellaan mallinnuksen ongelmaa hieman rajoittuneemmin. Oletetaan, että meillä on yksi tai useampia parametreja, joiden suhteen malli on lineaarinen (*huom: siis lineaarinen parametrien, ei riippumattomien muuttujien suhteen*). Tehtävänä on määrittää kyseisten parametrien arvot, sekä arvioida niiden luotettavuusvälit.

4.1 Sovituksen hyvyyden mitta

Ensimmäiseksi tarvitaan jokin mitta, joka kertoo kuinka hyvä mallin ja havaintojen yhteensopivuus on tietyillä parametrien arvoilla. Oletetaan ensin, että havaintojen poikkeama ($\sim$ virhe) mallin antamista arvoista noudattaa *normaalijakaumaa*. Tällöin todennäköisyys, että saadaan havainto y_i, kun oletettu mallifunktio f antaa arvon $f(x_i)$ on $\propto exp(-\frac{1}{2}(y_i - f(x_i))^2/\sigma_i^2)$, missä σ_i on kyseiselle mittaukselle arvioitu virhe ($\sim$ keskihajonta). Kokonaistodennäköisyys saadaan kertomalla yksittäisten mittausten todennäköisyydet

$$\mathcal{P} \propto \Pi_{i=1}^{N}\left[exp\left(-\frac{1}{2}(\frac{y_i - f(x_i)}{\sigma_i})^2 \right) \right]. \tag{4.1}$$

Ottamalla lausekkeen logaritmi, nähdään että todennäköisyys on suurin, kun lauseke

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2 \tag{4.2}$$

on pienin. χ^2 noudattaa luvussa 1.2 käsiteltyä χ^2-jakaumaa, jonka vapausasteiden lukumäärä on $\nu = N - M$, missä M on mallin parametrien lukumäärä. Hyvässä sovituksessa χ^2 arvo pitäisi olla lähellä vapausasteiden lukumäärää ν. Jos χ^2-arvo on suuri, malli ei selitä havaintoja hyvin tai havaintojen todelliset virheet ovat arvioituja suuremmat. Jos taasen $\chi^2 \ll \nu$, on havaintovirheiden suuruus yliarvioitu. χ^2-todennäköisyysjakauman avulla voidaan haluttaessa laskea saatua χ^2-arvoa vastaava tilastollinen todennäköisyys. Tällä ei välttämättä ole paljoakaan merkitystä, usein juuri virhearvioiden epämääräisyyden vuoksi.

4.2 Suoran sovitus

Tarkastellaan yksinkertaisena erikoistapauksena **lineaarista regressiota** eli suoran sovitusta. Malli on siis $y = a + b\,x$, ja tehtävänä on määrätä parametrien a ja b arvot sovittamalla malli havaintopisteisiin $(x_i,\, y_i)$. Tässä tehdään ero riippumattomien muuttujien (x) ja riippuvien muuttujien (y) välillä. Ainoastaan riippuvan muuttujan arvojen ajatellaan sisältävän virhettä, jolloin sovituksen hyvyyttä kuvaa χ^2-funktio

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - a - b\,x_i}{\sigma_i} \right)^2, \tag{4.3}$$

jonka minimointi antaa arvot tuntemattomille parametreille a ja b. Funktion ääriarvopisteissä sen derivaatta on nolla. Riittää siis etsiä edellisen χ^2-lausekkeen derivaattojen nollakohdat (funktiolla on tavallisesti yksi minimi eikä paikallisia maksimeja). Derivointi suoritetaaan parametrien a ja b suhteen, jolloin saadaan ehdot

$$\begin{aligned} \frac{\partial \chi^2}{\partial a} &= -2 \sum_{i=1}^{N} \frac{y_i - a - b\,x_i}{\sigma_i^2} = 0 \\ \frac{\partial \chi^2}{\partial b} &= -2 \sum_{i=1}^{N} \frac{x_i (y_i - a - b\,x_i)}{\sigma_i^2} = 0. \end{aligned}$$

Yhtälöryhmästä ratkaistaan parametrien a ja b arvot. Otetaan ensin käyttöön merkinnät

$$S = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \quad S_x = \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \quad S_y = \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}$$
$$S_{xx} = \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \quad S_{xy} = \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2}.$$

Näiden avulla ratkaisu voidaan kirjoittaa muotoon

$$a = \frac{S_{xx} S_y - S_x S_{xy}}{S\, S_{xx} - S_x^2} \qquad \text{(vakio)}$$

$$b = \frac{S\, S_{xy} - S_x S_y}{S\, S_{xx} - S_x^2} \qquad \text{(kulmakerroin)} \tag{4.4}$$

Virhearviot parametreille saadaan normaalin virheen kasautumislain avulla (kts. luku 1.3)

$$\sigma_a^2 = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial a}{\partial y_i} \right)^2 = \ldots = \frac{S_{xx}}{S\, S_{xx} - S_x^2}$$
$$\sigma_b^2 = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial b}{\partial y_i} \right)^2 = \ldots = \frac{S}{S\, S_{xx} - S_x^2}.$$

Kaavojen soveltuvuus käytännön ongelmiin on rajoitettu: (1) virheiden oletetaan olevan normaalijakautuneet (pienimmän varianssin estimaatti), (2) vain y – akselin suuntaiset virheet otetaan huomioon ja (3) virhearvioiden laskussa parametrien väliset korrelaatiot jätettiin huomiotta. Useissa ohjelmistoissa on jo mukana valmiit rutiinit niiden tapausten varalta, joissa myös pisteiden x_i virhearviot ovat käytettävissä. Luvun 5.0.2 menetelmillä sovitus voidaan tehdä yleisemmissä tapauksissa.

Oletetaan edelleen, että kaikkien riippuvan muuttujan mittausten virhearvio on σ ja virheet ovat normaalijakautuneet. Tällöin voidaan johtaa suoran parametrien todennäköisyystiheysfunktiot

$$b \sim N(\hat{b}, \sigma^2 / \sum_{i=1}^{n} (x_i - \overline{x})^2)$$

$$a \sim N(\hat{a}, \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \overline{x})^2}) \tag{4.5}$$

($\hat{a}$ ja $\hat{b}$ ovat parametreille johdetut arvot). Näiden kaavojen avulla saadaan laskettua parametrien luotettavuusrajat.

Määritettyjen parametrien avulla voidaan laskea mallin antama ennuste mielivaltaiselle x:n arvolle

$$\hat{y} = \hat{a} + \hat{b}\, x.$$

Oletetaan seuraavaksi, että sovituksen virheet ovat toisistaan riippumattomia eri x:n arvoilla ja ovat normaalijakautuneita. Tällöin voidaan t-jakauman avulla laskea **ennusteväli** (*prediction interval*) annetulle riskitasolle α. Painottamattomassa sovituksessa (kaikkien pisteiden virhearviot yhtä suuret) tulee väliksi

$$\hat{y} \pm t_{\alpha/2}\, s \sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{\sum (x_i - \overline{x})^2}}\ , \tag{4.6}$$

jossa esiintyvä keskihajonta on

$$s = \sqrt{\frac{\sum (y_i - \hat{y_i})^2}{n - 2}}\ . \tag{4.7}$$

Sovitetulla suoralla on kaksi parametria, joten vapausasteiden lukumäärä on n-2. Väli voidaan laskea kaikille x:n arvoille, jolloin sovitetun suoran ympärille voidaan piirtää vastaava **ennustevyöhyke** (*prediction band*). Tämä kuvaa yksittäisen havainnon todennäköisyyttä jäädä vyöhykkeen sisään. Kun yhtälössä 4.6 jätetään juuren alta pois tekijä '+1', saadaan vastaavat rajat riippuvan muuttujan keskiarvon jakaumalle riippumattoman muuttujan funktiona. Tämä **luotettavuusväli** (*confidence interval*) on luonnollisesti pienempi kuin ennustevyöhyke (keskiarvon tarkkuus on parempi kuin yksittäisen pisteen).

Matriisimuodossa lineaarinen malli on

$$y = Xa + \epsilon. \tag{4.8}$$

Kertomalla yhtälö puolittain matriisin X transpoosilla saadaan **normaaliyhtälöt**

$$X^T y = X^T X\, a. \tag{4.9}$$

Näiden ratkaisu on samalla ratkaisu (painottamaton) pienimmän neliösumman ongelmaan, ja se saadaan jakamalla tuntemattoman vektorin a kertoimella,

$$\hat{a} = (X^T X)^{-1} X^T y.$$

Sovituksen jäännöspoikkeamien neliösumma on

$$\mathrm{RSS} = \epsilon^T \epsilon = (y - \hat{y})^T (y - \hat{y}).$$

Tässä $\hat{y}$ tarkoittaa mallin antamaa ennustetta riippuvan muuttujan arvoille,

$$\hat{y} = X\,\hat{a}\,.$$

Tämän arvioitu varianssi on

$$\hat{\sigma} = \mathrm{RSS} \,/\, \mathrm{DF}\,,$$

jossa DF on vapausasteiden lukumäärä, eli havaintopisteiden lukumäärä vähennettynä selittävien muuttujien lukumäärällä. Sovituksen hyvyyttä kuvaava $\boldsymbol{R^2}$**–luku** määritellään tämän avulla

$$R^2 = 1 - \mathrm{RSS} \,/ \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

Jälkimmäisen termin osoittaja on siis varianssi sovitetun suoran suhteen, ja nimittäjä on kokonaisvarianssi. Luku vaihtelee välillä [0,1], ja arvo $R^2 = 1$ vastaa täydellistä sovitusta.

Jos y:n kaikkien mittausten keskihajonta on σ (ja mittaukset ovat toisistaan riippumattomia), noudattavat kertoimet $\hat{a}$ normaalijakaumaa $\mathrm{N}(a,\, \sigma^2\, (X^T X)^{-1})$. Esimerkiksi tietyn parametriarvon poikkeamaa nollasta voidaan testata testimuuttujalla

$$\hat{a_j} \,/\, \sqrt{\sigma^2\, (X^T X)^{-1}_{\mathrm{jj}}} \sim t_{\mathrm{DF}}.$$

Tässä t_{DF} tarkoittaa siis Studentin t-jakaumaa.

4.3 Yleisten kantafunktioiden sovitus

Edellä regressiossa olivat riippumattomina muuttujina havainnot. Samaa menettelyä voidaan soveltaa myös silloin, kun riippumattomat muuttujat lasketaan itse valituista funktioista. Matriisin X sarakkeet vastaavat tällöin sovituksen kantafunktioita - yhtenä (tai useampana) sarakkeena voi edelleen olla suora mittaustulos. Lineaarinen malli on

$$f(x) = y = \sum_{k=1}^{M} a_k X_k(x). \tag{4.10}$$

Mallissa tehdään edelleen selvä ero riippumattomien ja riippuvien muuttujien välillä. Kaavassa a_k ovat sovitettavat parametrit ja $X_k(x)$ valitut kantafunktiot. Funktiot X_k voivat olla mielivaltaisia, eikä niiden tarvitse olla lineaarisia *x:n suhteen*. Esimerkkinä voisi olla vaikkapa funktion

$$f(x) = a\sin 2x + be^{-x} \tag{4.11}$$

sovitus. Jos oletetaan, että y:n virheet ovat normaalijakautuneet sovituksen hyvyyttä voidaan jälleen mitata χ^2-funktiolla on

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - \sum_{k=1}^{M} a_k X_k(x_i)}{\sigma_i} \right]^2 . \tag{4.12}$$

Koska kyseessä on parametrien suhteen lineaarinen yhtälöryhmä, täydellinen sovitus voidaan kirjoittaa matriisiyhtälönä

$$A\vec{a} = \vec{y}\,, \tag{4.13}$$

missä $\vec{a}$ on parametrien muodostama vektori ja $\vec{y}$ havaintojen muodostama vektori. Tulossa $A\vec{x}$ matriisin A ensimmäisen rivin alkiot kerrotaan vuoronperää vastaavilla parametreilla. Matriisin ensimmäinen rivi sisältää siis kantafunktioiden arvot ensimmäisessä pisteessä, x_1. Matriisin ensimmäinen sarake sisältää ensimmäisen kantafunktion arvot eri pisteissä, toinen sarake toisen kantafunktion arvot jne.

Oletetaan ensin, että kaikkien havaintojen paino on sovituksessa yhtä suuri, eli y:n kovarianssimatriisi on diagonaalimatriisi, jossa kaikki diagonaalielementit ovat yhtä suuria (σ^2). Derivoimalla χ^2 parametrien suhteen ja merkitsemällä derivaatat jälleen nolliksi saadaan jälleen

$$\left(A^T A\right)\vec{a} = A^T\vec{y}\,. \tag{4.14}$$

Tämä on jo edellä mainittu normaaliyhtälö, josta parametrit a_i voidaan ratkaista esimerkiksi Gaussin eliminoinnilla tai LU-hajotelman avulla. Ratkaisu on formaalisti

$$\vec{a} = (A^T A)^{-1} A^T \vec{y}, \tag{4.15}$$

missä käänteismatriisia *ei* kuitenkaan kannata laskea eksplisiittisesti. Kyseessä on lineaarimuunnos havainnoista. Oletetaan seuraavaksi, että havainnot ovat normaalijakautuneita, jolloin myös arvot $\vec{a}$ ovat normaalijakautuneita. Näiden parametriarvojen kovarianssimatriisi on (kuten jo edellisessä luvussa todettiin)

$$C = \mathrm{cov}(\vec{a}) = \sigma^2 (A^T A)^{-1},$$

ja vastaavasti yksittäisen parametrin virhearvio on c_{ii}. Standardoitu parametriarvo

$$Z = \frac{\hat{a}_i - a_i}{\sigma\sqrt{c_{\mathrm{ii}}}} \tag{4.16}$$

noudattaa silloin jakaumaa $\sim$ N(0,1). Jos mittausten virheet σ ovat tuntemattomia, voidaan ne korvata arvioidulla keskihajonnalla

$$s = \sqrt{\mathrm{RSS}/(n-k-1)}. \tag{4.17}$$

Tässä on eksplisiittisesti oletettu, että sovitettu malli on oikea. Vapausasteita on sovituksessa havaintojen lukumäärä josta vähennetään mallin parametrien lukumäärä ja vielä yksi estimoidun keskiarvon vuoksi. Merkitään

$$T_{n-k-1} = \frac{\hat{a}_i - a_i}{s\sqrt{c_{\mathrm{ii}}}}, \tag{4.18}$$

jolloin parametrin a_i luotettavuusväli saadaan t-jakauman avulla

$$\hat{a} \pm t_{\alpha/2}\, s\, \sqrt{c_{ii}}\,. \tag{4.19}$$

Jakauman vapausasteiden lukumäärä on $n-k-1$. Testimuuttujaa T_{n-k-1} voidaan suoraan käyttää myös esim. sen testaamiseen, onko tietty parametri lainkaan tarpeen, eli onko sen arvo merkitsevästi nollasta poikkeava.

Virherajat voidaan johtaa samaan tapaan *yksittäiselle* lineaarisen mallin *ennusteelle*, $\hat{y}$,

$$\hat{y} \pm t_{\alpha/2}\, s\, \sqrt{\vec{x}_0 (X^T X)^{-1} \vec{x}_0 + 1}\,. \tag{4.20}$$

(Edelleen puhutaan painottamattomasta pienimmän neliösumman sovituksesta). Tässä $\vec{x}_0$ on piste, jossa ennuste lasketaan. Vastaavat rajat ennusteen *keskiarvolle* saadaa samasta kaavasta, jossa neliöjuuren alta on poistettu termi $+1$. Kaava ovat siis yleistys edellisen luvun kaavasta 4.6.

Usein voi herätä kysymys, selittääkö malli lainkaan havaintoja. Tätä voidaan testata asettamalla nollahypoteesi, että kaikki mallin parametrit ovat nollia. Vaihtoehtoinen hypoteesi on, että *ainakin yksi parametreistä on merkitsevä*. Testi perustuu mallin selittämän varianssin ja jäännöspoikkeamien varianssien suhteeseen. Kuten aiemmin, merkitään $\mathrm{RSS} = \sum e_i^2$. Vastaava mallin selittämä osuus on varianssista on PSS. Testimuuttuja on näiden suhde

$$\frac{\mathrm{PSS}/k}{\mathrm{RSS}/(n-k-1)} = \frac{\mathrm{PSS}/k}{s^2}, \tag{4.21}$$

ja tämä noudattaa F-jakaumaa, jonka vapausasteiden lukumäärät ovat k ja $n-k-1$. Jos testiparametrin arvo on riittävän iso, nollahypoteesi hylätään ja ainakin yksi mallin parametreista todella selittää jotakin havainnoista.

Jos halutaan testata *tietyn parametrien osajoukon riittävyyttä*, lasketaan selitetyt ja selittämättä jääneet neliösummat erikseen. Testimuuttuja on

$$F_{k-m,n-k-1} \sim \frac{(\mathrm{RSS}' - \mathrm{RSS})/(k-m)}{\mathrm{RSS}/(n-k-1)}. \tag{4.22}$$

Kaavassa m on parametrin lukumäärä valitussa osajoukossa ja RSS$'$ on jäännösneliösumma vastaavalle mallille. RSS on neliösumma, kun kaikki parametrit ovat mukana. Hypoteesit ovat $H0$: m-parametrin osajoukko on riittävä, ja $H1$: mallissa tarvitaan kaikki k parametriä. Kaavasta nähdään, että testimuuttuja saa sitä suurempia arvoja, mitä enemmän parametrien vähentäminen kasvattaa jäännösvirhettä. F-jakaumasta voidaan lukea muutoksen merkitsevyys.

Joissakin ohjelmistoissa (ainakin SAS) voidaan malliparametrien valinta tehdä automaattisesti. Ohjelma testaa kunkin parametrin tarpeellisuutta, ja jättää jäljelle vain todella tarpeelliset (*backward elimination*). Vaihtoehtoisesti valinta tehdään niin, että malliin lisätään uusia selittäviä muuttujia, kunnes sovitus ei enää oleellisesti parane (*forwards selection* tai *stepwise method*).

Edelliset esitetyt lineaariset mallin sovituksen kaavat voidaan yleistää niin, että havaintojen painotus tapahtuu ekspisiittisesti, käyttäen havaintojen kovarianssimatriisia Σ (sikäli kuin se on tiedossa tai voidaan estimoida). Kyseessä on silloin **painotettu pienimmän neliösumman sovitus**. Yhtälöt tulevat muotoon

$$G = (A^T \Sigma^{-1} A)^{-1} \tag{4.23}$$

$$\vec{a} = G A^T \Sigma^{-1} \vec{y}\,.$$

Erona edelliseen on ainoastaan se, että matriisi A on eksplisiittisesti painotettu sen kovarianssimatriisin käänteismatriisilla. Yhden muuttujan tapauksessa painotekijä olisi siis $1/\sigma^2$ (vrt. kaava 4.4). Matriisi G on parametriarvojen kovarianssimatriisi, eli sisältää parametrien virhearviot ja tiedon parametrien välisistä korrelaatioista. Kaavat antavat pienimmän neliösumman ratkaisun, mutta estimaatin varianssi on pienin mahdollinen riippumatta siitä, onko virhejakauma normaalijakauma tai ei (*best linear un-biased estimator*). Lineaarisesta mallista laskettu ennuste ja sen virhearvio ovat

$$\hat{y} = Aa, \qquad \hat{\Sigma} = AGA^T \tag{4.24}$$

Sovituksen jäännöspoikkemat sekä niiden kovarianssimatriisi ovat

$$e = y - \hat{y}\,, \quad \Sigma_e = \Sigma - \hat{\Sigma}\,. \tag{4.25}$$

Jäännöspoikkeamat kannattaa piirtää, ja tarkistaa noudattavatko standardoidut jäännökset $e/\sqrt{\Sigma_{e,\mathrm{ii}}}$ normaalijakaumaa $\sim \mathrm{N}(0,1)$. Jos virheiden oletetaan olevan gaussisia, voidaan mallin oletuksia testata. Painotettu jäännöspoikkeamien neliösumman noudattaa nimittäin χ^2-jakaumaa

$$q = e^T \Sigma^{-1} e \;\sim \chi^2_{n-k}, \tag{4.26}$$

missä vapausasteiden lukumäärä on $n - k$, sovitettavien pisteiden lukumäärä vähennettynä mallin parametrien lukumäärällä. Jos q on liian pieni, on havaintojen virhearviot yliarvioitu. Jos neliösumma on liian suuri, ei lineaarinen malli ole onnistunut selittämään havaintoja.

Edellä testeissä ja luottamusvälinvälien laskemisessa oletettiin, että havaintovirheet noudattavat normaalijakaumaa. Jos virhejakauma on todellisuudessa selvästi vino, voi myös edellisistä kaavoista määritettyjen parametrien arvoissa olla harhaa. Tilannetta voidaan yrittää parantaa tekemällä muuttujanvaihdoksen ja siirtymällä tarpeen mukaan esimerkiksi logaritmiselle asteikolle. Tällöin malli ei kuitenkin enää ole lineaarinen alkuperäisten muuttujien suhteen. Myöhemmin käsitellään yleisempää tilannetta, jossa voidaan otetaan suoraan huomioon havaintojen virhejakauma.

Esimerkki 4.1. Oliveira-Costa A., Tegmark M., Finkbeiner D.P. et al. 2002, *A New Spin on Galactic Dust*, ApJ 567, 363.

4.4 Lineaariset mallit *R*-ohjelmassa

Seuraavaksi palataan vielä lineaarisiin malleihin, joissa riippuvaa muuttujaa pyritään selittämään yhden tai useamman riippumattoman muuttujan funktiona. Kutakin riippuvasta suureesta tehtyä mittausta kohden on mittaukset yhdestä tai useammasta riippumattomasta suureesta. Esimerkit kuvaavat mallien käsittelyä *R*-ohjelmistossa, jossa lineaaristen mallien käsittelyyn liittyvät rutiinit sisältyvät `base`–pakettiin. Lisää rutiineja lineaarisen mallin testaamiseen löytyy mm. kirjastosta `lmtest`.

4.4.1 Lineaarinen malli

Perusmuodossaan lineaarisessa mallissa (*Linear Model*) on kyse yhtälön

$$y_i = \sum_{k=1}^{M} a_k \, x_i^k + e$$

sovittamisesta. Kaavassa y_i ovat riippuvasta suureesta tehdyt mittaukset, a_k ovat mallin parametrit, x_i^k on k:nnen riippumattoman muuttujan arvo i:nnessä mittauksessa ja e on sovituksen residuaali. Matriisimuodossa $y = a\,x + e$, jossa kaikki termit ovat vektoreita. Riippuvaa muuttujaa voidaan siis yrittää selittää useamman riippumattoman suureen lineaarikombinaationa, jolloin puhutaan usean muuttujan regressiosta (*multiple regression* - tilastotieteessä käyränsovituksesta käytetään tavallisesti termiä regressio). Ratkaisua haetaan pienimmän neliösumman menetelmällä, mikä rajoittaa ongelman asettelua. Riippumattomille muuttujien mahdollisia virhearvioita ei voida käyttää. Kullekin havainnolle voidaan kuitenkin antaa painokerroin, joka mahdollistaa riippuvan muuttujan virheiden huomioimisen. Kuten edellä jo painotettiin, malli on lineaarinen nimenomaan parametrien a_k suhteen. Siten käyrän

$$y(x) = a_1 \sin(x^3) + a_2 \log(x)$$

sovittaminen havaintopisteisiin y_i **on** esimerkki lineaarisesta mallista. Kaavassa x ei edusta tuntematonta, vaan on x_i ovat annettuja riippumattoman muuttujan arvoja. Kyseessä on siis itse asiassa edellisen luvun yleisten kantafunktioiden sovittamisesta.

R-ohjelmassa mallinsovitus suoritetaan yksikertaisimmillaan seuraavaan tapaan: `lm( y ~ x )`. Tilde–merkkiä käytetään ilmaisemaan riippuvuutta. Ohjelma suorittaa painottamattoman pienimmän neliösumman sovituksen, ja tuloksena saadaan sovitetun suoran vakiotermi ja kulmakerroin. Seuraavassa esimerkki painotetusta sovituksesta.

```
> x _ c(1,2,4,5,6)
> y _ c(2,4,6,7,8)
> w _ c(1,1,1,2,2)
> lm(y~x, weights=w)

Call:
lm(formula = y ~ x, weights = w)

Coefficients:
(Intercept)            x
      1.288        1.137
```

Sovituksen tuloksen mielekkyyttä voidaan tarkastella erilaisilla kuvilla. Erityisesti on syytä tarkastaa sovituksen jäännöspoikkeamat (residuaalit). Cooken etäisyys (*Cook distance*) D_i mittaa yksittäisen datapisteen (x_i, y_i) vaikutusta regressioparametreihin. Se määritetään jättämällä kyseinen piste pois sovituksesta, ja vertaamalla näin saatuja parametreja alkuperäisiin. Jos D_i:n arvo on yli ~ 0.5, on piste oleellinen (*influential*) parametrien arvojen määräytymisessä. Erityisesti on katsottava, onko Cooken etäisyys joillakin yksittäisillä pisteillä paljon muita suurempi. Jos näin on, on syytä epäillä poikkeavaa (*outlier*) pistettä tai mahdollisesti mallin puutteellisuutta. (kts. *R*-rutiini `influence.measures`). *R*-ohjelmistossa komento `plot(malli)` - missä 'malli' on muuttuja, joka sisältää sovitetun mallin - tuottaa neljä kuvaa: residuaalit sovitetun arvon funktiona, Q-Q-piirroksen (jäännösten pitäisi olla normaalijakautuneita), standardoidut residuaalit sekä Cooken etäisyydet. Komento `predict(malli, new, interval="confidence")` puolestaan laskee ennustevälin. Lisää lineaarimallin testejä löytyy *R*-kirjastosta `lmtest`. Seuraava esimerkkiohjelma tekee yksinkertaisen sovituksen, sekä tuottaa edellä mainitut kuvat.

```
# generoidaan datapisteitä
x <- rnorm(30, 10, 5)
y <- 2.0*x+1.0+rnorm(30, 0, 8)

# mallin sovitus
malli <- lm( y ~ x)

# lasketaan mallin ennusteet tasavälisesti, ja
# talletetaan uuteen muuttujaan
new <- data.frame(x = seq(0, 20, 0.5))

# piirretään sovitus ja ennusteväli
predict(malli, new, se.fit = TRUE)
pred.w.plim <- predict(malli, new, interval="prediction")
pred.w.clim <- predict(malli, new, interval="confidence")
matplot(new$x,cbind(pred.w.clim, pred.w.plim[,-1]),
      lty=c(1,2,2,3,3), type="l", ylab="predicted y",
```

```
      col=c("black", "blue", "red"))
points(x,y)

# piirretään residuaalit etc.
plot(malli)
```

4.4.2 Yleinen lineaarinen malli

Yleinen lineaarinen malli (***G****eneral* ***L****inear* ***M****odel*, GLM) yleistää edellä kuvattua usean muuttujan regressiota monellakin tavalla. Ensinnäkin malli voi sisältää *useita riippuvia muuttujia*, eli muuttujasta y tulee vektorin sijasta matriisi. Vastaavasti myös tuntemattomat regressiokertoimet muodostavat matriisin, jossa on vektori vastaten kutakin riippuvaa muuttujaa. Tämän matriisiyhtälön ratkaisu tapahtuu periaatteessa täsmälleen samoin kuin edellä lineaarisen mallin tapauksessa. Toinen GLM mallin tekemä yleistys on se, että se *sallii riippuvien muuttujien lineaariset muunnokset ja niiden lineaariset yhdistelmät*. Malli voidaan kirjoittaa muotoon

$$YM = Xa + e\,,$$

jossa matriisi M kuvaa riippuvien muuttujien lineaarisia muunnoksia. Tavalliseen tapaan voidaan kirjoittaa normaaliyhtälöt

$$X^T Xa = X^T YM$$

ja pienimmän neliösumman ratkaisu määräytyy formaalisti yhtälöstä

$$a = (X^T X)^{-1} X^T YM\,.$$

Yleistetty lineaarinen malli (*Generalized Linear Model*) laajentaa edellä kuvattua GLM-mallia vielä kahdella tavalla. Ensinnäkin *riippuvan muuttujan ei tarvitse olla normaalijakautunut* vaan sille voidaan olettaa muukin, jopa epäjatkuva jakauma. Lisäksi lineaarisen ennusteen ja riippuvan muuttujien arvon välille voidaan määritellä epälineaarinen (mutta monotoninen) muunnos *linkkifunktion* avulla. Linkkifunktio g suorittaa muunnoksen niin, että

$$y = g(Xa).$$

Olisi tietenkin mahdollista tehdä muunnos suoraan havaintoarvoihin, mutta silloin myös havaintopisteiden virhejakauma muuntuisi.

$R-$ohjelmassa yleisen ja yleistetyn lineaarisen mallin sovitukset tehdään `glm`–rutiinin avulla. Parametrin `family` avulla valitaan haluttu virhejakauma, ja parametri `link` määrittelee linkkifunktion. Normaalijakauman tapauksessa oletusarvoinen linkkifunktio on identiteetti $g(\mu) = \mu$, Poisson-jakauman tapauksessa se on logaritmifunktio, $g(\mu) = \log(\mu)$. Yleisessä tapauksessa mallin parametreja ei voida enää määrittää yksinkertaisilla matriisioperaatioilla vaan ratkaisu perustuu iteratiiviseen algoritmiin.

Jos esimerkiksi halutaan selvittää, millä nopeudella mitattu signaali vahvistuu ajan funktiona, voitaisiin sovitus suorittaa seuraavasti

```
> aika <- c(1,2,3,4,5,6,7,8,9)
> signaali <- c(1,3,2,5,4,7,6,9,8)
> glm( signaali ~ aika, family=gaussian())

Call:  glm(formula = signaali ~ aika, family = gaussian())

Coefficients:
(Intercept)         aika
     0.3333       0.9333

Degrees of Freedom: 8 Total (i.e. Null);  7 Residual
Null Deviance:      60
Residual Deviance: 7.733        AIC: 30.18

> glm( signaali ~ aika, family=poisson())

Call:  glm(formula = signaali ~ aika, family = poisson())

Coefficients:
(Intercept)         aika
     0.5028       0.1963

Degrees of Freedom: 8 Total (i.e. Null);  7 Residual
Null Deviance:      13.26
Residual Deviance: 2.549        AIC: 36.36
```

Ensimmäinen malli on itse asiassa tavallinen lineaarinen malli. Toisessa sovituksessa havaintojen oletetaan sen sijaan noudattavan Poisson-jakaumaa, ja linkkifunktio on (oletusarvoisesti) *log*(), joten sovitettava mallikin on erilainen: $log(y)=Ax$. Sovituksesta saatu suora sovittaa siis signaalin logaritmeja, ei suoraan signaalia.

Poisson-jakaumaa noudattavat havainnot voidaan sovittaa likimääräisesti myös tavallisella lineaarisella mallilla. Tällöin on kuitenkin tehtävä ensin muunnos, joka palauttaa Poisson-jakaumaa noudattavat virheet lähemmäs normaalijakaumaa. Tähän voidaan käyttää juuri esim. *log*–funktiota tai vaikka *sqrt*–funktiota.

Yleistyksistään huolimatta GLM malli on vielä melko rajoittunut. Se sallii, luonnollisesti, ainoastaan lineaariset riippuvuudet muuttujien välille. Yleistetty lineaarinen malli parantaa tilannetta jonkin verran, mutta ei suinkaan salli vielä täysin yleistä riippuvuutta muuttujien välille. Riippumattomien muuttujien todennäköisyysjakaumatkaan eivät voi olla mielivaltaisia, vaan on valittava annetuista vaihtoehdoista. Jos ongelma kuitenkin täyttää nämä vaatimukset, `glm`-rutiinin käyttö on helppoa.