

Pikajohdatus bayesilaiseen tilastanalyysiin ja monimuuttuja-analyysiin

Loppuseminaari: Terveydenhuollon uudet analyysimenetelmät (TERANA)

Aki Vehtari

`Aki.Vehtari@tkk.fi`



HELSINKI UNIVERSITY OF TECHNOLOGY
Department of Biomedical Engineering and Computational Science

1.4.2009

- Pikajohdatus bayesilaiseen tilastanalyysiin
- Gaussiset prosessit ja monimuuttuja-analyysi

Miksi bayesilaisia tilastomenetelmiä?

- Terveysthuollon ilmiöt kompleksisia
 - paljon tuntemattomia asioita
 - useita vaikeasti suoraan mitattavia asioita

Miksi bayesilaisia tilastomenetelmiä?

- Terveysthuollon ilmiöt kompleksisia
 - paljon tuntemattomia asioita
 - useita vaikeasti suoraan mitattavia asioita
- Bayesilaisen tilastotieteen menetelmät joustavia
 - johdonmukainen tapa käsitellä kaikki tuntemattomat ja epävarmuudet
 - mallin kompleksisuus voi riippua ilmiön kompleksisuudesta ja havaintojen epävarmuudesta

- Epävarmuus esitetään todennäköisyyksillä
- Todennäköisyydet päivitetään uuden tiedon avulla

Epävarmuus ja bayesilainen tilastollinen päättely

Satunnainen vs. tietämyksellinen epävarmuus

Epävarmuus voidaan jakaa

- Satunnaiseen (aleatoriseen) epävarmuuteen

- Tietämykselliseen (episteemiseen) epävarmuuteen

Epävarmuus voidaan jakaa

- Satunnaiseen (aleatoriseen) epävarmuuteen
 - emme voi saada havaintoja, jotka auttaisivat sen epävarmuuden pienentämisessä
- Tietämykselliseen (episteemiseen) epävarmuuteen

Epävarmuus voidaan jakaa

- Satunnaiseen (aleatoriseen) epävarmuuteen
 - emme voi saada havaintoja, jotka auttaisivat sen epävarmuuden pienentämisessä
- Tietämykselliseen (episteemiseen) epävarmuuteen
 - voimme saada havaintoja jotka auttavat sen epävarmuuden pienentämisessä

Epävarmuus voidaan jakaa

- Satunnaiseen (aleatoriseen) epävarmuuteen
 - emme voi saada havaintoja, jotka auttaisivat sen epävarmuuden pienentämisessä
- Tietämykselliseen (episteemiseen) epävarmuuteen
 - voimme saada havaintoja jotka auttavat sen epävarmuuden pienentämisessä
- Vertaa kolikko
 - kahdella tarkastelijalla voi olla eri tietämyksellinen epävarmuus
 - tietämyksellinen todennäköisyys muuttuu, kun informaatio muuttuu

Epävarmuus ja bayesilainen tilastollinen päättely

Esimerkki: Kahdenvärisiä nappuloita pussissa

- Jos eriväristen nappuloiden määrän suhde tunnettu
 - epävarmuutta seuraavaksi ilmestyvän nappulan väristä

Epävarmuus ja bayesilainen tilastollinen päättely

Esimerkki: Kahdenvärisiä nappuloita pussissa

- Jos eriväristen nappuloiden määrän suhde tunnettu
 - epävarmuutta seuraavaksi ilmestyvän nappulan väristä
- Jos eriväristen nappuloiden määrän suhde tuntematon
 - lisäksi tietämyksellistä epävarmuutta
 - tietämyksellinen epävarmuus muuttuu kun nappuloita nostetaan

Epävarmuus ja bayesilainen tilastollinen päättely

Esimerkki: Kahdenvärisiä nappuloita pussissa

- Jos eriväristen nappuloiden määrän suhde tunnettu
 - epävarmuutta seuraavaksi ilmestyvän nappulan väristä
- Jos eriväristen nappuloiden määrän suhde tuntematon
 - lisäksi tietämyksellistä epävarmuutta
 - tietämyksellinen epävarmuus muuttuu kun nappuloita nostetaan
- Jos yksittäin noston sijasta aikoisimme kumota koko pussin ja laskea värien määrän suhteen
 - ei satunnaista epävarmuutta
 - vain tietämyksellinen epävarmuus pussin sisällöstä

- Satunnainen epävarmuus
 - jos tiedetään riskitaso ja populaatio, kuolemien määrä vaihtelee satunnaisesti

Epävarmuus ja bayesilainen tilastollinen päättely

Esimerkki: tautiriski

- Satunnainen epävarmuus
 - jos tiedetään riskitaso ja populaatio, kuolemien määrä vaihtelee satunnaisesti
- Tietämyksellinen epävarmuus
 - riskitaso jollakin alueella ja ajanjaksona

- Satunnainen epävarmuus
 - jos tiedetään riskitaso ja populaatio, kuolemien määrä vaihtelee satunnaisesti
- Tietämyksellinen epävarmuus
 - riskitaso jollakin alueella ja ajanjaksona
- Havaintoja jotka voivat päivittää tietämyksellistä epävarmuutta
 - taustapopulaation tiedot
 - havaitut kuolemat

Epävarmuus ja bayesilainen tilastollinen päättely

Epävarmuuksien yhdistäminen?

- Merkitään
 - y havaitut nappulat (tai tautitapaukset)
 - θ nappuloiden suhde (tai tautiriski)
 - I taustatieto ongelmasta
- Satunnainen epävarmuus, jos nappuloiden suhde θ tunnettu

$$p(y|\theta, I)$$

- Tietämyksellinen epävarmuus ennen havaintoja

$$p(\theta|I)$$

Epävarmuus ja bayesilainen tilastollinen päättely

Epävarmuuksien yhdistäminen?

- Merkitään
 - y havaitut nappulat (tai tautitapaukset)
 - θ nappuloiden suhde (tai tautiriski)
 - I taustatieto ongelmasta
- Satunnainen epävarmuus, jos nappuloiden suhde θ tunnettu

$$p(y|\theta, I)$$

- Tietämyksellinen epävarmuus ennen havaintoja

$$p(\theta|I)$$

- Kuinka päivittää tietämyksellinen epävarmuus kun nappuloita havaittu?

$$p(\theta|y, I)?$$

- Kun valittu $p(y|\theta, I)$ sekä $p(\theta|I)$, voidaan laskea Bayesin kaavalla

$$p(\theta|y, I) = \frac{p(y|\theta, I)p(\theta|I)}{\int p(y|\theta, I)p(\theta|I)d\theta}$$

- Havaintomalli $p(y|\theta, I)$
 - matemaattinen kuvaus havaintomallille (satunnainen osa)
 - jos ilmiö tunnettu millä todennäköisyydellä havaittaisiin y tietyllä arvolla
 - esim. mikä on epävarmuus kuolemien määrästä, jos riskitaso tunnettu
 - esim. Poisson-mallia

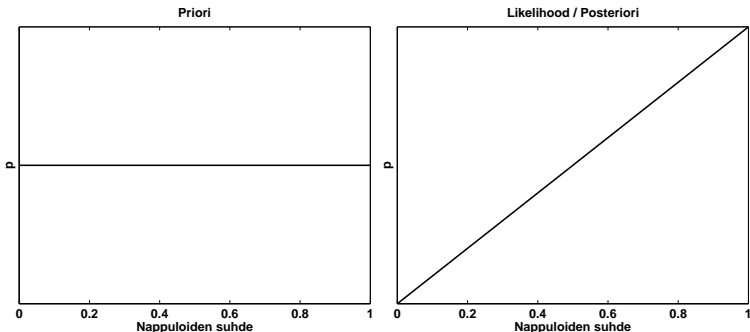
- Priori $p(\theta|I)$
 - matemaattinen kuvaus mitä tiedetään θ :sta
 - tietämyksellinen epävarmuus ennen havaintoja
 - malli ja priorit erottamattomat (kytketty mallin kautta)
 - esim. lähekkäisten alueiden riskitasot samankaltaiset
 - esim. gaussinen prosessi tai CAR-priori

- Ennusteissa otetaan huomioon posterioriepävarmuus

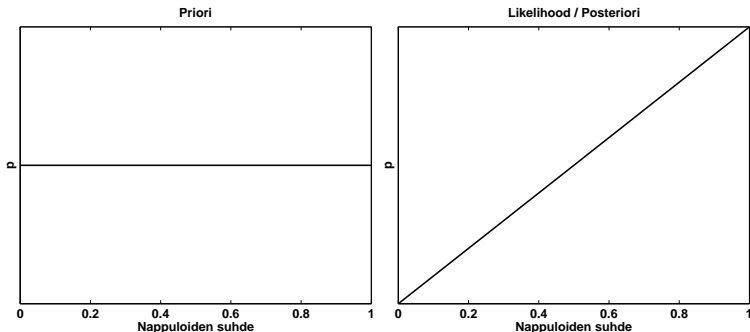
Bayesiläinen tilastollinen päättely

Ennusteet

- Ennusteissa otetaan huomioon posterioriepävarmuus
esim. nostettu pussista yksi punainen nappula

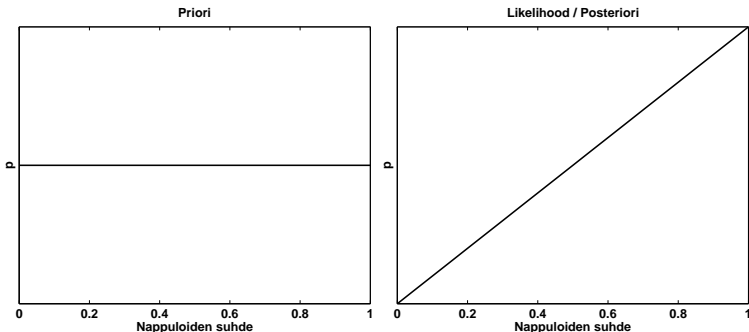


- Ennusteissa otetaan huomioon posterioriepävarmuus esim. nostettu pussista yksi punainen nappula



- Todennäköisin suhde on 1 \rightarrow seuraava on punainen todennäköisyydellä 1

- Ennusteissa otetaan huomioon posterioriepävarmuus esim. nostettu pussista yksi punainen nappula



- Todennäköisin suhde on 1 \rightarrow seuraava on punainen todennäköisyydellä 1
- Huomioidaan epävarmuus \rightarrow seuraava on punainen todennäköisyydellä $2/3$

- Eri suhdevaihtoehtoja painotetaan niiden todennäköisyydellä
 - eli integroidaan yli suhteen posterioriepävarmuuden
- Integroimalla epävarmuuksien yli otetaan epävarmuudet johdonmukaisesti huomioon
 - usein haastava osa menetelmien käyttöä
 - integrointimenetelmien kehitys osa tutkimustamme

- Malli
 - pyrkii ennustamaan ilmiön käyttäytymistä
 - voidaan käyttää ennustamaan tulevaisuutta
 - voidaan käyttää lisäämään tieteellistä ymmärrystä ilmiöstä
- Usein yksinkertaistaa todellisuutta
 - ilmiöstä saadut havainnot rajoitettuja
 - joidenkin havaittavien suureiden vaikutus voi olla paljon suurempi kuin toisten
 - yksinkertainenkin malli voi tuottaa hyödyllisiä ennusteita

- Pudotetaan palloa eri korkeuksilta ja mitataan putoamisaika sekunttikellolla käsivaralla

- Pudotetaan palloa eri korkeuksilta ja mitataan putoamisaika sekunttikellolla käsivaralla
 - Newtonin mekaniikka
 - ilmanvastus, ilmanpaine, pallon muoto, pallon pintarakenne
 - ilmavirtaukset
 - suhteellisuusteoria

- Pudotetaan palloa eri korkeuksilta ja mitataan putoamisaika sekunttikellolla käsivaralla
 - Newtonin mekaniikka
 - ilmanvastus, ilmanpaine, pallon muoto, pallon pintarakenne
 - ilmavirtaukset
 - suhteellisuusteoria
- Ottaen huomioon mittaukset, kuinka tarkka malli kannattaa tehdä?

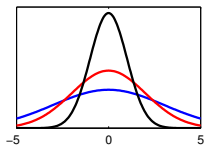
- Pudotetaan palloa eri korkeuksilta ja mitataan putoamisaika sekunttikellolla käsivaralla
 - Newtonin mekaniikka
 - ilmanvastus, ilmanpaine, pallon muoto, pallon pintarakenne
 - ilmavirtaukset
 - suhteellisuusteoria
- Ottaen huomioon mittaukset, kuinka tarkka malli kannattaa tehdä?
- On olemassa hyvin paljon tilanteita, joissa yksinkertaiset mallit hyödyllisiä ja käytännön kannalta yhtä tarkkoja kuin monimutkaisemmat

- Useita mahdollisesti ilmiötä selittäviä muuttujia (havaintotutkimuksissa tilastollinen riippuvuus ei kausaalisuus)
 - mitkä muuttujista relevantteja?
 - mikä yhteys ilmiöön?
 - onko yhteisvaikutuksia?

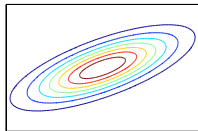
- Usein perinteisesti lineaarimalli (tai kynnystäminen)
 - jos selittävän muuttujan arvo muuttuu, muuttuu kohdemuuttujan arvo lineaarisesti (funktionaalinen muoto fiksattu)
 - epälineaarisuudet mahdollisia muuttujanmuunnoksilla (funktionaalinen muoto fiksattu)
 - mahdolliset selittävien muuttujien väliset interaktiot esitetään eksplisiittisesti lisättyinä interaktioina (esim. ikä \times sukupuoli, interaktio fiksattu)

- Usein perinteisesti lineaarimalli (tai kynnystäminen)
 - jos selittävän muuttujan arvo muuttuu, muuttuu kohdemuuttujan arvo lineaarisesti (funktionaalinen muoto fiksattu)
 - epälineaarisuudet mahdollisia muuttujanmuunnoksilla (funktionaalinen muoto fiksattu)
 - mahdolliset selittävien muuttujien väliset interaktiot esitetään eksplisiittisesti lisättyinä interaktioina (esim. ikä \times sukupuoli, interaktio fiksattu)
- Gaussiset prosessit joustava vaihtoehto
 - epälineaarisuudet ja interaktiot automaattisesti

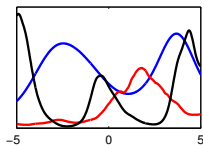
- Yleistää moniulotteisen normaalijakauman (Gaussin jakauman)
- Analyysin funktionaalinen muoto ei fiksattu, vaan asettaa priorin erilaisille funktioille



(a) Yksiulotteinen normaalijakauma

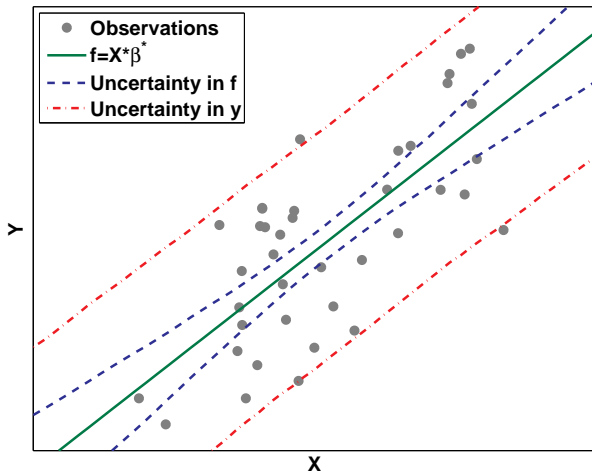


(b) Kaksiulotteinen normaalijakauma

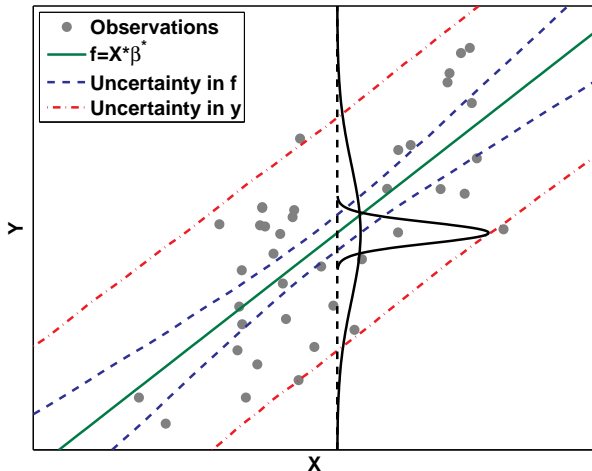


(c) Funktioita jotka poimittu gaussisesta prosessista

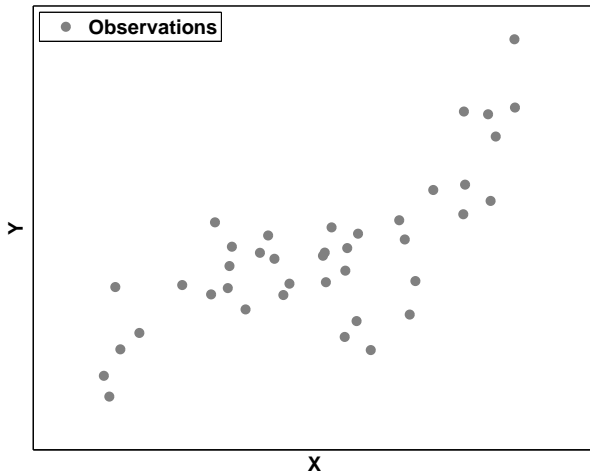
- Esimerkki lineaarisesta mallista



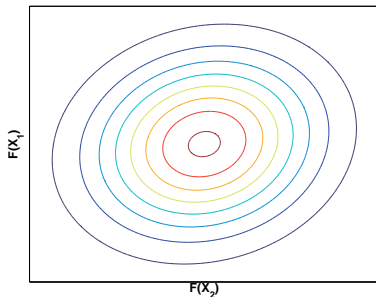
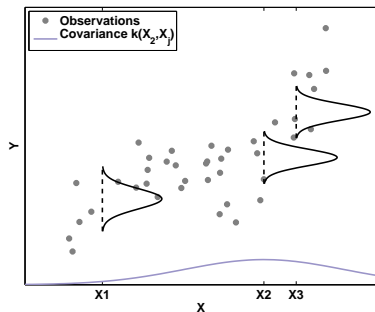
- 1) Funktioon f liittyvä epävarmuus
- 2) Havaintoihin y liittyvä epävarmuus



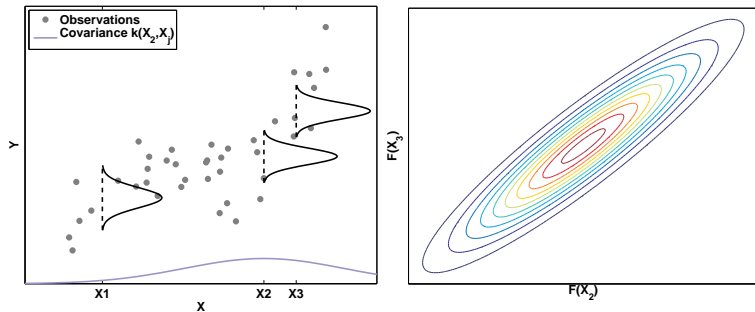
Epälineaarinen malli?



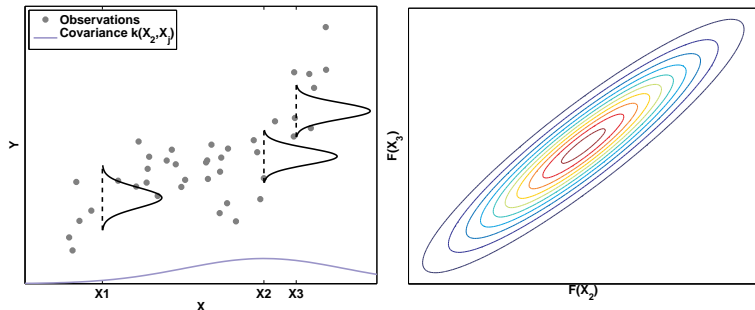
- Funktion f arvot eivät korreloi, jos x arvot kaukana toisistaan



- Funktion f arvot korreloivat, jos x arvot lähekkäiset



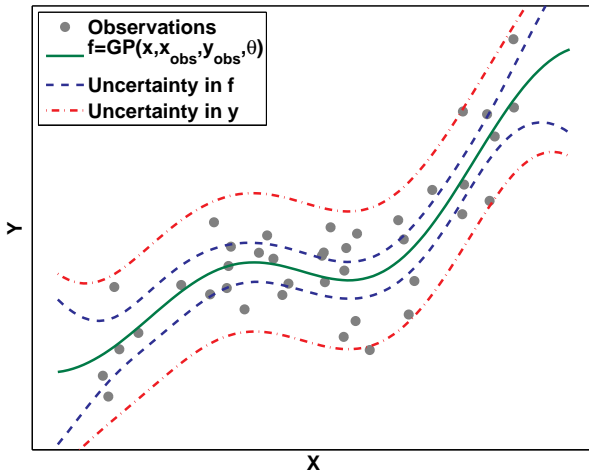
- Funktion f arvot korreloivat, jos x arvot lähekkäiset



- Korrelaation määrä kuvataan kovarianssifunktiolla

Gaussinen prosessi

- Gaussisen prosessin posterioriodotusarvo sekä funktion ja havaintojen epävarmuus



- Havaintomalli normaalijakauma

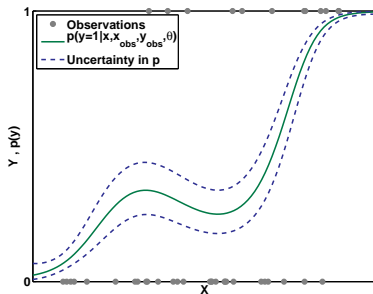
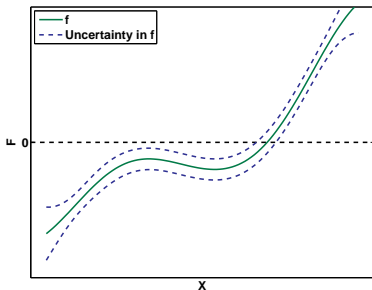
- Havaintomalli normaalijakauma
- Priori kertoo, että samankaltaisilla selittävien muuttujien arvoilla havaitaan samankaltaisia arvoja
 - havainnot korreloivat
 - voidaan ajatella myös priorina funktioille

- Havaintomalli normaalijakauma
- Priori kertoo, että samankaltaisilla selittävien muuttujien arvoilla havaitaan samankaltaisia arvoja
 - havainnot korreloivat
 - voidaan ajatella myös priorina funktioille
- Posteriori
 - havainnot lisäävät tietoa
 - edelleen voi olla epävarmuutta funktion tarkasta muodosta

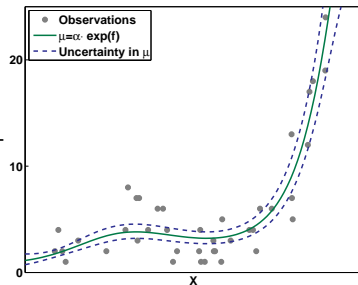
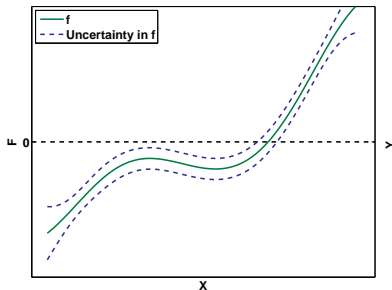
- Havaintomalli normaalijakauma
- Priori kertoo, että samankaltaisilla selittävien muuttujien arvoilla havaitaan samankaltaisia arvoja
 - havainnot korreloivat
 - voidaan ajatella myös priorina funktioille
- Posteriori
 - havainnot lisäävät tietoa
 - edelleen voi olla epävarmuutta funktion tarkasta muodosta
- Ennusteet
 - huomioidaan funktion muodon epävarmuus integroimalla

- Ei-gaussiset mallit voidaan muodostaa latenttien muuttujien avulla

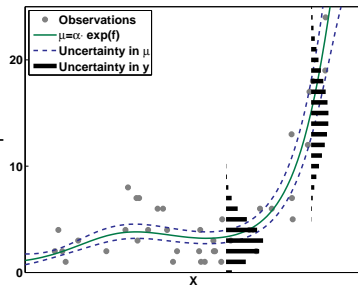
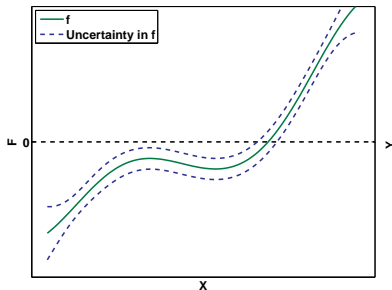
- Havaintoja kahdesta luokasta



- Lukumäärähavaintoja (esim. tautitapauksia)



- Lukumäärähavaintoja (esim. tautitapauksia)



- Gaussista prosessia, voidaan käyttää osana hierarkisia ja rakenteellisia malleja
 - esim. tiedon yhdistäminen eri tasoilta kuten potilas, ruutu, kunta, sairaanhoitopiiri, jne.

- Kovarianssifunktio kertoo monenkin muuttujan tapauksessa kuinka samankaltaisia eri muuttujan arvoja saavat alkioit ovat
 - monen muuttujan epälineaarisuudet
 - monen muuttujan väliset interaktiot
 - esimerkkejä tulevissa esityksissä:
 - spatiaaliset ja spatiotemporaaliset mallit
 - rekisterianalyysi

- Havaintopisteiden kasvaessa ongelmana laskennallinen raskaus
 - projektissa on kehitetty laskentaa nopeuttavia
 - approksimatiivisia kovarianssifunktioita
 - integrointimenetelmiä
- Kompleksisiin ilmiöihin tarvitaan joustavia malleja
 - projektissa on kehitetty
 - joustavia kovarianssimalleja
 - mallien arviointimenetelmiä

- Kompleksisiin ilmiöihin tarvitaan joustavia malleja
 - projektissa käytetty gaussisia prosesseja
- Kompleksisen ilmiön esittäminen selkeästi vaativaa
 - projektissa kehitetty menetelmiä
 - arvioimaan muuttujien relevanssia
 - arvioimaan muuttujien poisjättämisen vaikutus
 - visualisointiin
 - aliryhmien löytämiseen

- Suosittelemme bayesilaista tilastoanalyysia
 - käsittelee epävarmuudet johdonmukaisesti
 - mahdollistaa joustavien mallien turvallisemman käytön

- Suosittelemme bayesilaista tilastoanalyysia
 - käsittelee epävarmuudet johdonmukaisesti
 - mahdollistaa joustavien mallien turvallisemman käytön
- Suosittelemme gaussista prosessia monimuuttuja-analyysiin
 - mahdollistaa joustavan mallintamisen
 - menetelmäkehityksemme pääkohde, mutta muitakin erinomaisia joustavia bayesilaisia malleja on olemassa