Olli-Pekka Huovilainen

Aktiivisten DNA-muutosten seulonta riippuvuusmalleilla

Elektroniikan, tietoliikenteen ja automaation tiedekunta

Diplomityö, joka on jätetty opinnäytteenä tarkastettavaksi diplomi-insinöörin tutkintoa varten Espoossa 3.6.2010.

Työn valvoja:

Prof. Samuel Kaski

Työn ohjaaja:

DI Leo Lahti

Aalto-yliopisto
Teknillinen korkeakoulu

AALTO-YLIOPISTO
TEKNILLINEN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Olli-Pekka Huovilainen

Työn nimi: Aktiivisten DNA-muutosten seulonta riippuvuusmalleilla

Päivämäärä: 3.6.2010 Kieli: Suomi Sivumäärä:8+38

Elektroniikan, tietoliikenteen ja automaation tiedekunta

Professuuri: Informaatiotekniikka Koodi: T-61

Valvoja: Prof. Samuel Kaski

Ohjaaja: DI Leo Lahti

Syövän kehittymiseen liittyy geneettiset muutokset useissa solun kasvuun, jakautumiseen tai kuolemaan liittyvissä geeneissä. Näissä syöpään liittyvissä geeneissä mutaatiot aiheuttavat muutoksia geenin aktiivisuudessa syöpäsoluissa. Sekä mutaatioita että geenien aktiivisuuksia voidaan mitata geenisiruilla. Näiden kopiolukuja ilmentymämittausten avulla voidaan etsiä syöpään liittyviä geenejä.

Tässä työssä tutkittiin todennäköisyysperusteiseen kanoniseen korrelaatioanalyysiin perustuvien riippuvuusmallien käyttämistä syöpägeenien etsimisessä. Tässä menetelmässä etsintä tehdään tutkimalla kopioluku- ja ilmentymämittauksien yhteyksiä riippuvuusmalleilla kunkin geenin ympäristössä. Nämä riippuvuusmallit mahdollistavat myös etukäteistiedon hyväksikäytön rajoittamalla tutkittava riippuvuutta. Syöpägeenien etsinnässä voidaan käyttää etkäteistietona syöpägeeneihin liittyvän kopioluku- ja ilmentymämuutoksien paikkariippuvuutta. Tällä rajoitettiin menetelmän etsimä riippuvuus vain saman geenin mittausten välille. Tämä pienensi pienestä näytemäärästä johtuvaa mallin ylisovitusta. Rajoitettujen riippuvuusmallien käyttö paransi menetelmän toimivuutta selkeästi. Menetelmän todettiin toimivan parhaiten sallimalla pieni vapaus rajoitukselle. Työssä toteutettiin avoimen lähdekoodin sovellus syöpägeenien etsimiseen riippuvuusmalleilla.

Menetelmän toimivuutta verrattiin muihin ilmentymä- ja kopiolukumittausten riippuvuuksien tutkimiseen tarkotettuihin menetelmiin. Rajoitettuihin riippuvuusmalleihin perustuvan menetelmän todettiin toimivan paljon paremmin syöpägeenien etsinnässä kuin muut verratut menetelmät. Tässä työssä toteutettu menetelmä on saatujen tulosten perusteella paras menetelmä syöpägeenien etsinnässä kopioluku- ja ilmentymämittauksilla.

Avainsanat: kanoninen korrelaatioanalyysi, riippuvuusmallit, toiminnallinen genomiikka, koneoppiminen, bioinformatiikka, syöpätutkimus

AALTO UNIVERSITY
SCHOOL OF SCIENCE AND TECHNOLOGY

ABSTRACT OF THE
MASTER'S THESIS

Author: Olli-Pekka Huovilainen

Title: Screening of active DNA changes with dependency modelling

Date: 3.6.2010 Language: Finnish Number of pages:8+38

Faculty of Electronics, Communications and Automation

Professorship: Computer and Information Science Code: T-61

Supervisor: Prof. Samuel Kaski

Instructor: M.Sc.(Tech.) Leo Lahti

The development of cancer is associated with genetic abnormalities in genes which have a funtion in cell growth, division, or death. Mutations of these cancer associated genes cause changes in gene activity in cancer cells. Mutations and gene activities can be measured with microarrays. These copy number and expression measurements can be used locate cancer associated genes.

This thesis studies the use of probabilistic canonical correlationanalysis based dependency models for detecting cancer associated genes. In this method, the search is performed by examining associations between copy number and expression measurements with dependency models within each genes neighborhood. These dependency models enable the useage of priori knowledge by constraining the examined dependency. Constraining can be applied to the search of cancer associated genes with the priori knowledge of location dependencies of copy number and expression changes. This was used to restrict the modelled dependencies to within genes measurements. This reduced the overfitting caused by small sample size. The restriction improved the method considerably. An optimum for the method was found when a small freedom was allowed from the restriction. The search of cancer associated genes with dependency models was implemented as an open source application.

The effectiveness of the method was compared to other methods intended for the analyzation of dependencies between copy number and expression measurements. The method of using constrained dependency modelling was found to perform considerably better than any other compared method. The method implemented in this thesis is the best method for searching of cancer associated genes according to the results.

Keywords: canonical correlation analysis, dependency models, functional genomics, machine learning, bioinformatics, cancer research

Esipuhe

Haluan kiittää työn valvojaa professori Samuel Kaskea ja ohjaajaa Leo Lahtea hyvästä ohjauksesta. Professori Sakari Knuutilaa haluan kiittää syöpäaineistojen tarjoamisesta tutkimuksiin.

Otaniemi, 3.6.2008

Olli-Pekka Huovilainen

Sisältö

Symbolit ja lyhenteet

Symbolit

x	muuttuja
$\mathbf{x}$	vektori
$\mathbf{X}$	matriisi
Ψ	mallin kovarianssi
$\mathbf{W}$	mallin kerroinmatriisi
$\boldsymbol{\epsilon}$	mallin virhe
$\mathcal{N}(\mu, \sigma^2)$	normaalijakauma odotusarvolla μ ja varianssilla σ^2
$\tilde{\Sigma}$	näytekovarianssimatriisi
μ	odotusarvo
ℓ	negatiivinen log-todennäköisyys
$\mathbf{I}$	identiteettimatriisi
$\mathrm{P}(X)$	todennäköisyys X:lle
$\mathrm{P}(X\|y)$	todennäköisyys X:lle annettuna y
$\mathbf{z}$	piilomuuttuja
$\widehat{X}$	suurimman uskottavuuden estimaatti X:lle
$\rho(a, b)$	korrelaatio a:n ja b:n välillä
$\mathrm{var}(x)$	x:n varianssi
$\mathrm{cov}(x, y)$	x:n ja y:n välinen kovarianssi
$\exp(x)$	exponenttifunktio x:stä
$\mathrm{tr}(\mathbf{X})$	matriisin $\mathbf{X}$ jälki

Operaattorit

$\sum_i$	summa indeksin i yli
$x \sim y$	x noudattaa jakaumaa y

Lyhenteet

PCA	pääkomponenttianalyysi (principal component analysis)
pPCA	todennäköisyysperusteinen pääkomponenttianalyysi (probabilistic principal component analysis)
CCA	kanoninen korrelaatioanalyysi (canonical correlation analysis)
pCCA	todennäköisyysperusteinen kanoninen korrelaatioanalyysi (probabilistic canonical correlation analysis)
simCCA	samankaltaisuusrajoitteinen kanoninen korrelaatioanalyysi (similarity constrained canonical correlation analysis)
psimCCA	todennäköisyysperusteinen samankaltaisuusrajoitteinen kanoninen korrelaatio-analyysi (probabilistic similarity constrained canonical correlation analysis
aCGH	vertaileva genominen hybridisaatio
mRNA	lähetti-RNA
miRNA	mikro-RNA
ROC	oikeiden positiivisten määrä väärien positiivisten funktiona (receiver operating characteristics)
AUC	ROC-käyrän alle jäävä ala: testin erottelukyky (area under curvature)
EM	odotusarvon maksimointi (expectation maximization)

1 Johdanto

Syövän kehittymisen taustalla on mutaatiot geeneissä, jotka liittyvät mm. solun jakautumiseen, kasvuun, kuolemaan [Vogelstein04]. Näiden geenien mutaatioilla syöpäsoluissa on ominaista mutaation vaikutus aktiivisuuteen eli siihen kuinka paljon geenistä tehdään proteiineja. Useimmiten näissä syöpään vaikuttavissa geeneissä mutaation aiheuttama geenin monistuminen lisää geenin aktiivisuutta tai mutaation aiheuttama geenin kopion poistuminen heikentää aktiivisuutta. Tämä syöpägeeneille ominainen mutaation ja aktiivisuuden yhteys mahdollistaa syöpään vaikuttavien geenien etsinnän.

Tutkittavan genomiaineiston nopea kasvu mahdollistaa uudentyyppisten tutkimuksien tekemisen. Vertailevan genomisen hybridisaation (aCGH) kehittyminen tarjoaa korkean resoluution mittaustietoa kopiolukumuutoksista, mikä oli aiemmin mahdollista vain lähetti-RNA:sta (mRNA) geenisiruilla saaduille ilmentymämittauksille. Aiemmin on tehty runsaasti tutkimusta sekä ilmentymämittauksilla että aCGH:lla saaduilla kopiolukumittauksilla. Kuitenkin vasta viimevuosina on alkanut kehittyä molempia mittaustietoja hyödyntävät analysointityökalut. Erityinen mielenkiinnon kohde on syöpään vaikuttavien geenien etsintä yhdistelemällä kopioluku- ja ilmentymätietoa, sillä tämän yhdistelmän käyttö mahdollistaa aktiivisten kopiomuutosten eli mutaatioiden aiheuttamien ilmentymämuutosten etsinnän. Syöpägeenien luonteen takia riippuvuusmallitus on luonnollinen valinta syöpägeenien etsinnässä.

Geneettisen datan mallinnuksessa ongelmana on usein pieni näytemäärä ja iso muuttujamäärä. Yhdellä sirulla saadaan tuhansia mittaustietoja yhdestä kudosnäytteestä, kun taas näytteiden määrä on tyypillisesti joitakin kymmeniä. Toinen ongelma geneettisessä mallinnuksessa on muun tuntemattoman vaihtelun vaikutus. Mittausjoukkojen välillisessä mallintamisessa on välttämätöntä tehtävä rajoitteita, jotta pystytään välttämään liian vapaaseen malliin liittyvä ylisovittuminen. Rajoitteilla tulee myös saada muun tuntemattoman vaihtelun vaikutus estettyä. Asetettavien rajoitteiden on pohjauduttava asiasta ennalta tiedettyihin asioihin. Syöpägeenien etsinnässä luonnollinen rajoite mallille on rajoittaa mallitettava riippuvuus vain saman geenin ilmentymän ja kopiolukumuutoksen välille.

Tässä työssä tutkitaan todennäköisyysperusteiseen kanoniseen korrelaatioanalyysiin [Bach05] perustuvia riippuvuusmalleja, sekä sen pohjalta kehitettyä rajoitettua riippuvuusmallia [Lahti09]. Riippuvuusmallikehyksestä kasataan yhtenäinen avoimen lähdekoodin ohjelmistototeutus. Tähän ohjelmistopakettiin sisällytetään kaikki riippuvuusmallin erikoistapaukset sekä mallin rajoitusmahdollisuudet. Ohjelmistopakettiin lisätään työkalut koko genomin sisältävän aineiston käsittelyyn malleilla. Paketin tarkoituksena on tarjota helppokäyttöinen kokonaisuus riippuvuuksien tutkimiseen, sekä genomiaineiston tutkimisessa että yleisessä tapauksessa.

Työssä on tarkoituksena selvittää kvantitatiivisesti kuinka hyvä riippuvuusmalliperusteinen menetelmä on. Tämän selvittämiseksi menetelmää testataan syöpäaineistolla, josta tiedetään syöpään vaikuttavia geenejä. Riippuvuusmallin toimivuutta löytää syöpägeenejä testataan mallin eri parametreilla, jotta selviäisi riippuvuus-

malleihin pohjautuvan menetelmän herkkyys mallin parametreille. Tutkittavia mallin parametrejä on yhteen riippuvuusmalliin otettavan alueen koko geenin ympärillä, kohinaoletus aineistojen välisestä yhteydestä riippumattomaan vaihteluun sekä piilomuuttujan, josta mallissa oletetaan molempien mittausten muodostuvan, dimensionaalisuus. Saatuja rajoittamattomien riippuvuusmallien tuloksia verrataan rajoitetulla mallilla saataviin tuloksiin, jotta selviää miten etukäteistiedon hyödyntäminen vaikuttaa menetelmän toimivuuteen, ja miten tämä toivuuden muutos suhtautuu rajoittamattoman mallin tulosten vaihteluun eri malliparametreillä. Rajoitetusta mallista tutkitaan rajoituksen vapauden vaikutus menetelmän toimivuuteen.

Muita kirjallisuudessa olevia menetelmiä verrataan samalla aineistolla riippuvuusmalliin perustuvaan menetelmään. Tällä tavalla on tarkoitus selvittää, miten menetelmän toimivuus sijoittuu suhteessa muihin ilmentymä- ja kopiolukumittausten analysointiin tarkoitettuihin menetelmiin.

Vaikka menetelmän pääkohde on ilmentymä- ja kopiolukumittausten riippuvuuden tutkiminen, sitä testataan myös mikroRNA- ja kopiolukumittausten riippuvuuden tutkimiseen. Näillä mittauksilla on samankaltainen biologinen tausta niiden suhteesta toisiinsa kuin ilmentymä- ja kopiolukumittauksilla, joten menetelmän pitäisi soveltua myös näiden mittausten riippuvuuden tutkimiseen.

Syöpägeenien biologisesta taustasta esitellään ainoastaan tausta miten ne vaikuttavat kopioluku- ilmentymä ja mikroRNA-mittauksiin. Riippuvuusmallin rajoittaminen mahdollistaisi monen erilaisen etukäteistiedon sisällyttämisen malliin. Tässä työssä kuitenkin rajoitutaan tutkimaan ilmentymä- ja kopiolukumittausten paikkariippuvuuden hyödyntämistä etukäteisteitona mallituksessa.

Tässä työssä tutkitaan syöpägeenien mittausten etukäteistiedon hyväksikäytön vaikutus riippuvuusmallipohjaisessa syöpägeenien etsinnässä. Työssä selvitetään tämän menetelmän toimivuus suhteessa muihin samaan tarkoitukseen tehtyihin menetelmiin. Tarkoituksena on selvittää paras menetelmä syöpägeenien etsintään kopiolukuja ilmentymämittauksilla, jotta selviäisi mitä menetelmää kannattaa käyttää syöpätutkimuksessa.

2 Tilastollisen riippuvuuden mallinnus

2.1 Kanoninen korrelaatioanalyysi

Kanoninen korrelaatioanalyysi (CCA) [Hotelling33] on yleisesti käytetty työkalu tilastollisessa monimuuttuja-analyysissa kahden muuttujajoukon välisen yhteyden tunnistamisessa ja mittauksessa. CCA perustuu projektioihin, joilla molemmat muuttujajoukot projisoidaan matalampaan ulottuvuuteen ja maksimoidaan korrelaatio muuttujajoukkojen välillä tässä pienemmässä ulottuvuudessa. Näitä projisoivia vektoreita kutsutaan kanonisiksi korrelaatiovektoreiksi ja ne löydetään kahden muuttujan välisellä yhteisellä kovarianssianalyysillä. [Härdle03]

Kanoninen korrelaatioanalyysi etsii satunnaismuuttujien $\mathbf{x} \in \mathbb{R}^q$ ja $\mathbf{y} \in \mathbb{R}^p$ yhteyden lineaarikombinaatioiden $\mathbf{u}_Y^T\mathbf{x}$ ja $\mathbf{u}_Y^T\mathbf{y}$ avulla siten, että $\mathbf{u}_X$ ja $\mathbf{u}_Y$ maksimoivat korrelaation $\rho(\mathbf{u}_X, \mathbf{u}_Y)$.Oletetaan, että

$$\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} \sim \left(\begin{pmatrix}\boldsymbol{\mu}_X\\ \boldsymbol{\mu}_Y\end{pmatrix}, \begin{pmatrix}\Sigma_{XX} & \Sigma_{XY}\\ \Sigma_{YX} & \Sigma_{YY}\end{pmatrix}\right), \tag{1}$$

missä Σ_{XX} on $q \times q$ $\mathbf{x}$:n kovarianssimatriisi, Σ_{YY} on $p \times p$ $\mathbf{y}$:n kovarianssimatriisi ja $\Sigma_{XY} = \Sigma_{YX}^T$ on x:n ja y:n välinen kovarianssimatriisi. Satunnaismuuttujat $\mathbf{x}$ ja $\mathbf{y}$ ovat pystyvektoreita.Projektioiden korrelaatio on muotoa

$$\rho(\mathbf{u}_X, \mathbf{u}_Y) = \frac{\mathbf{u}_X^T\Sigma_{XY}\mathbf{u}_Y}{(\mathbf{u}_X^T\Sigma_{XX}\mathbf{u}_X)^{1/2}(\mathbf{u}_Y^T\Sigma_{YY}\mathbf{u}_Y)^{1/2}}. \tag{2}$$

Koska kaava (2) on skaala-invariantti, riittää, että maksimoidaan

$$\max_{\mathbf{u}_X, \mathbf{u}_Y} = \mathbf{u}_X^T\Sigma_{XY}\mathbf{u}_Y \tag{3}$$

rajoitteilla

$$\mathbf{u}_X^T\Sigma_{XX}\mathbf{u}_X = \mathbf{u}_Y^T\Sigma_{YY}\mathbf{u}_Y = 1. \tag{4}$$

Määritellään $q \times p$ matriisi $\mathcal{K} = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1/2}$, jolle voidaan tehdä singulaariarvohajotelma $\mathcal{K} = \mathbf{V}_X\Lambda\mathbf{V}_Y^T$. Singulaariarvohajotelmassa

$$\begin{aligned}\mathbf{V}_X &= (\mathbf{v}_{X1}, \ldots, \mathbf{v}_{Xk})\\ \boldsymbol{\Lambda} &= \mathrm{diag}(\lambda_1, \ldots, \lambda_k)\\ \mathbf{V}_Y &= (\mathbf{v}_{Y1}, \ldots, \mathbf{v}_{Yk})\end{aligned} \tag{5}$$

missä k on muuttujien välisen kovarianssimatriisin Σ_{XY} rangi eli kovarianssimatriisin Σ_{XY} lineaarisesti riippumattomien sarakkeiden määrä ja epänegatiiviset arvot $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_k$ ovat ominaisarvoja matriiseille $\mathcal{N}_1 = \mathcal{K}\mathcal{K}^T$ ja $\mathcal{N}_2 = \mathcal{K}^T\mathcal{K}$. [Härdle03]

Määritellään arvoille $i = 1, \ldots, k$ vektorit

$$\mathbf{u}_{Xi} = \Sigma_{XX}^{-1/2}\mathbf{v}_{Xi} \tag{6}$$

$$\mathbf{u}_{Yi} = \Sigma_{YY}^{-1/2}\mathbf{v}_{Yi}, \tag{7}$$

joita kutsutaan kanonisiksi korrelaatiovektoreiksi. Näiden kanonisten korrelaatiovektoreiden avulla määritellään kanoniset korrelaatiomuuttujat

$$\eta_i = \mathbf{u}_{Xi}^T \mathbf{x} \tag{8}$$
$$\psi_i = \mathbf{u}_{Yi}^T \mathbf{y}. \tag{9}$$

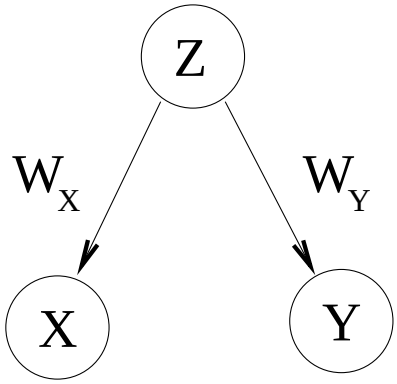
Korrelaatiomuuttujien yhdistelmä $(\eta_i, \psi_i), i = 1, \ldots, m$ on i. kanoninen muuttujapari. Suureita $\rho_i = \lambda_i^{1/2}$ arvoille $i = 1, \ldots, k$ kutsutaan kanonisiksi korrelaatiokertoimiksi. Jos merkitään $\mathbf{U}_X = (\mathbf{u}_{X1}, \ldots, \mathbf{u}_{Xk})$ ja $\mathbf{U}_Y = (\mathbf{u}_{Y1}, \ldots, \mathbf{u}_{Yk})$, saadaan

$$\begin{aligned} \mathbf{U}_X^T \Sigma_{XX} \mathbf{U}_X &= \mathbf{I} \\ \mathbf{U}_Y^T \Sigma_{YY} \mathbf{Y}_X &= \mathbf{I} \\ \mathbf{Y}_X^T \Sigma_{YX} \mathbf{X}_X &= \mathbf{P}, \end{aligned} \tag{10}$$

missä $\mathbf{P}$ on $p \times q$ diagonaalimatriisi, jonka diagonaalilla on kanoniset korrelaatiot. [Bach05], [Härdle03]

2.2 Todennäköisyysperusteinen kanoninen korrelaatioanalyysi

Kanoninen korrelaatioanalyysi ei huomioi mallin parametreihin liittyvää epävarmuutta [Lahti09]. Näiden epävarmuuksien huomioon ottamiseksi on luotu todennäköisyysperusteinen piilomuuttujamalli [Bach05], jolla on yhteys kanonisen korrelaatioanalyysiin. Siinä kahden muuttujajoukon $\mathbf{x}$ ja $\mathbf{y}$ oletetaan muodostuvan yhteisestä piilomuuttujasta $\mathbf{z}$ sekä muusta vaihtelusta kuvan 1 mukaisesti. Tätä mallia kutsutaan todennäköisyysperusteiseksi kanoniseksi korrelaatioanalyysiksi (pCCA).



Kuva 1: Todennäköisyysperusteinen kanonisen korrelaatioanalyysin (pCCA) graafinen

Havaintojen $\mathbf{x}$ ja $\mathbf{y}$ oletetaan pCCA:ssa noudattavan riippuvuusmallia

$$\begin{aligned} \mathbf{x} &= \mathbf{W}_x \mathbf{z} + \boldsymbol{\mu}_X + \boldsymbol{\epsilon}_X \\ \mathbf{y} &= \mathbf{W}_y \mathbf{z} + \boldsymbol{\mu}_Y + \boldsymbol{\epsilon}_Y, \end{aligned} \tag{11}$$

missä $\boldsymbol{\epsilon}_X$ ja $\boldsymbol{\epsilon}_Y$ ovat mallin virhe kovariansseilla $\boldsymbol{\Psi}_Y$ ja $\boldsymbol{\Psi}_Y$. Havainnot $\mathbf{x}$ ja $\mathbf{y}$ muodostuvat piilomuuttujasta $\mathbf{z}$ kerroinmatriisien $\mathbf{W}_X$ ja $\mathbf{W}_X$ kautta. Piilomuuttuja $\mathbf{z}$ vastaa pienempiulotteista avaruutta, johon projisoituna havaintojen $\mathbf{x}$ ja $\mathbf{y}$ välinen korrelaatio maksimoituu. Kovarianssimatriisit $\boldsymbol{\Psi}_Y$ ja $\boldsymbol{\Psi}_Y$ kuvaavat muuttujajoukon sisäisiä variansseja, joka on mallin mittausten välisestä yhteydestä riippumaton vaihtelu. Piilomuuttujan $\mathbf{z}$ sekä havaintojen $\mathbf{x}$ ja $\mathbf{y}$ oletetaan noudattavan jakaumia

$$\begin{aligned} \mathbf{z} &\sim \mathcal{N}(0, \mathbf{I}_d), \min\{q, p\} \geq d \geq 1 \\ \mathbf{x}|\mathbf{z} &\sim \mathcal{N}(\mathbf{W}_X\mathbf{z} + \boldsymbol{\mu}_X, \boldsymbol{\Psi}_X), \mathbf{W}_X \in \mathbb{R}^{m_X \times d} \\ \mathbf{y}|\mathbf{z} &\sim \mathcal{N}(\mathbf{W}_Y\mathbf{z} + \boldsymbol{\mu}_Y, \boldsymbol{\Psi}_Y), \mathbf{W}_X \in \mathbb{R}^{m_Y \times d}. \end{aligned} \tag{12}$$

Mallin (11) parametreille voidaan laskea suurimman uskottavuuden estimaatit analyyttisesti. Parametrien laskukaavat ovat liitteessä 6. Todelliset kanoniset korrelaatiot ja korrelaatiovektorit saadaan mallin (12) parametrien suurimman uskottavuuden estimaateista jälkikäsittelyllä paperin [Archambeau06] esitettämällä tavalla.

2.3 Todennäköisyysperusteinen pääkomponenttianalyysi

Todennäköisyysperusteinen pääkomponenttianalyysi (pPCA) [Tipping99] voidaan käsitellä mallin (11) erikoistapauksina. Se noudattavat mallia

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}. \tag{13}$$

pPCA:ssa $\mathbf{z}$ vastaa $\mathbf{x}$:n lineaariprojektiota pienempiulotteiseen avaruuteen, jossa näytteiden säilyvä informaatio maksimoituu. Tämä vastaa sitä, että virhetermi $\boldsymbol{\epsilon}$ noudattaa isotrooppista jakaumaa $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma\mathbf{I})$. Suurimman uskottavuuden estimaatti projektiomatriisille $\mathbf{W}$ virittää saman aliavaruuden kuin d vallitsevinta pääkomponenttia.

Mallista (13) voidaan johtaa riippuvuusmalli kahden muuttujajoukon välille yhdistämällä muuttujajoukot X ja Y mallin havaintoihin siten, että mallin (13) havainto $\mathbf{x}$ on molempien muuttujajoukkojen havainnot liitettynä samaan vektoriin. Tällöin malli (13) muuttuu samanlaiseksi kuin malli (11), kun merkitään $\mathbf{W} = (\mathbf{W}_X, \mathbf{W}_Y)^T$ ja $\boldsymbol{\epsilon} = \text{diag}(\boldsymbol{\epsilon}_X, \boldsymbol{\epsilon}_Y)$, lukuunottamatta oletusta virhetermien $\boldsymbol{\epsilon}$ jakaumasta. pCCA:ssa virhetermien kovarianssi ei ole rajoitettu lukuunottamatta rajoitusta, että virhetermien välinen ristikovarianssi on nolla, kun taas pPCA:ssa virhetermien kovarianssi on isotrooppinen.

Mallin (13) parametrien suurimman uskottavuuden estimaatit voidaan laskea analyyttisesti. Laskukaavat ovat liitteessä 6

2.4 Todennäköisyysperusteinen faktorianalyysi

Todennäköisyysperusteinen faktorianalyysi (pFA) muodostaa havainnoille saman mallin kuin (13). pFA eroaa kuitenkin pPCA:sta virhetermin $\boldsymbol{\epsilon}$ jakaumassa. pFA:ssa

virhetermi noudattaa diagonaalista jakaumaa $\boldsymbol{\epsilon} = \text{diag}(\epsilon_1, \ldots, \epsilon_d)$. Nämä diagonaalilla olevat alkiot yleensä estimoidaan näytteistä, ja ne ovat ehdollisesti riippumattomia annettuna piilomuuttuja $\mathbf{z}$. Piilomuuttujan tarkoitus on faktorianalyysissä selittää havaittujen muuttujien välinen korrelaatio, kun virhetermi $\boldsymbol{\epsilon}$ kuvaa muuttujalle ominaista vaihtelua. Tämän takia pFA:ssa suurimman uskottavuuden estimaatit $\mathbf{W}$:n sarakkeille eivät yleisesti vastaa pääkomponenttien muodostamaa aliavaruutta. [Tipping99]

Todennäköisyysperusteisen faktoriaalianalyysin ja todennäköisyysperusteisen pääkomponenttianalyysin mallien ero johtaa kahteen ratkaisujen ominaisuuksien eroon. Ensimmäinen ero on se, että pPCA on invariantti alkuperäisten data-akseleiden rotaatioille, kun taas pFA on invariantti komponenttikohtaiselle skaalaukselle. Toinen ero on se, että kahden faktorin mallin kumpikaan faktoreista ei ole välttämättä sama kuin yhden faktorin mallin faktori. pPCA:ssa sen sijaan pääakselit löydetään yksi kerrallaan, eivätkä aiemmat pääakselit muutu piilomuuttujan dimension kasvaessa. [Tipping99]

pFA:lle ei ole olemassa analyyttistä ratkaisua, vaan mallin parametrit pitää ratkaista odotusarvon maksimoinnille [Dempster77]. Odotusarvon maksimoinnissa pyritään löytämään suurimman uskottavuuden estimaatit mallin parametreillä iteratiivisesti. Menetelmä vaihtelee E- ja M-askelien välillä. E-askeleella lasketaan log-uskottavuuden odotusarvo senhetkisillä parametrien estimaateilla. M-askeleella lasketaan uudet estimaatit parametreille siten, että E-askeleella laskettu log-uskottavuus maksimoituu.

2.5 Rajoitettu todennäköisyysperusteinen kanoninen korrelaatioanalyysi

Kanoniseen korrelaatioanalyysiin liittyvä projektiovektoreiden vapaus johtaa helposti ylisovittumiseen [Vinod76]. Monissa sovelluskohteissa on tiedossa muuttujajoukkojen riippuvuuteen liittyvää etukäteistietoa, jota voidaan käyttää hyödyksi rajoittamalla etukäteistiedon perusteella projektioiden vapautta. Tämä vähentää ylisovitusta ja sen lisäksi näin voidaan keskittää etsintä tietyntyyppisiin riippuvuuksiin.

Projektioiden välistä riippuvuutta voidaan parametrisoida muutosmatriisilla $\mathbf{T}$ [Lahti09]. Merkitään $\mathbf{u}_X = \mathbf{u}$, jolloin voidaan kirjoittaa $\mathbf{u}_Y = \mathbf{Tu}$. Projektioiden välisen korrelaation maksimoiminen johtaa tällöin seuraavaan optimointiongelmaan:

$$\max_{\mathbf{u},\mathbf{T}} = \frac{\mathbf{u}^T \Sigma_{XY} \mathbf{Tu}}{(\mathbf{u}^T \Sigma_{XX} \mathbf{u})^{1/2} (\mathbf{Tu}^T \Sigma_{YY} \mathbf{Tu})^{1/2}}. \tag{14}$$

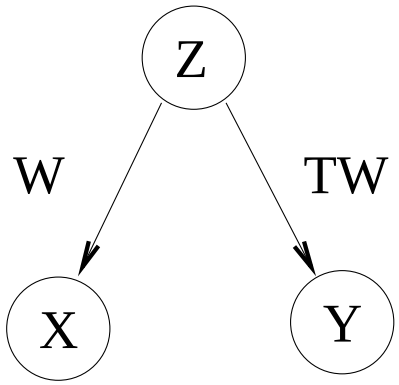
Muutosmatriisille $\mathbf{T}$ asetettavien rajoitusten avulla voidaan ohjata riippuvuusetsintää.

Vastaava todennäköisyysperusteinen rajoitettu piilomuuttujamalli voidaan muodos-

taa mallista (11) merkittäessä $\mathbf{W}_X = \mathbf{W}$ ja $\mathbf{W}_Y = \mathbf{TW}$, jolloin

$$\begin{aligned} \mathbf{x} &= \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}_X \\ \mathbf{y} &= \mathbf{TW}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}_Y, \end{aligned} \tag{15}$$

missä $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$, $\boldsymbol{\epsilon}_X \sim \mathcal{N}(0, \boldsymbol{\Psi}_X)$ ja $\boldsymbol{\epsilon}_Y \sim \mathcal{N}(0, \boldsymbol{\Psi}_Y)$. Mallia kutsutaan todennäköisyyspohjaiseksi rajoitetuksi kanoniseksi korrelaatioanalyysiksi (psimCCA) [Lahti09] ja sen graafinen esitys näkyy kuvassa 2.



Kuva 2: Todennäköisyyspohjaisen regularisoidun kanonisen korrelaatioanalyysin (psimCCA) malli

Mallin (15) parametrien ja aineiston yhteistodennäköisyydeksi saadaan

$$\mathrm{P}(\mathbf{X}, \mathbf{Y}|\mathbf{T}, \mathbf{W}, \boldsymbol{\Psi})\,\mathrm{P}(\mathbf{T}, \mathbf{W}, \boldsymbol{\Psi}), \tag{16}$$

missä $P(\mathbf{T}, \mathbf{W}, \boldsymbol{\Psi})$ on mallin priorijakauma. Oletetaan, että mallin parametrit ovat riippumattomia, jolloin saadaan yhteistodennäköisyydeksi

$$\mathrm{P}(\mathbf{x}, \mathbf{y}|\mathbf{T}, \mathbf{W}, \boldsymbol{\Psi})\,\mathrm{P}(\mathbf{T})\,\mathrm{P}(\mathbf{W})\,\mathrm{P}(\boldsymbol{\Psi}). \tag{17}$$

Tarkastellaan ainoastaan etukäteistiedon käyttämistä muutosmatriisin $\mathbf{T}$ priorijakaumassa. Muutosmatriisille $\mathbf{T}$ priorijakaumaksi määritellään matriisinormaalijakauma

$$\begin{aligned} \mathrm{P}(\mathbf{T}) &= \mathcal{N}_m(\mathbf{T}|\mathbf{M}, \mathbf{U}, \mathbf{V}) \\ &= \frac{1}{c}\exp\left(-\frac{1}{2}\,\mathrm{tr}\{\mathbf{U}^{-1}(\mathbf{T}-\mathbf{M})\mathbf{V}^{-1}(\mathbf{T}-\mathbf{M})^{\mathbf{T}}\}\right) \text{[Dutilleul99].} \end{aligned} \tag{18}$$

Asettamalla matriisinormaalijakauman rivikovarianssi $\mathbf{U}$ ja sarakekovarianssi $\mathbf{V}$ vakiokertoimiseksi identiteettimatriisiksi $\sigma_T^2\mathbf{I}$, kaavasta (18) saadaan

$$\mathrm{P}(\mathbf{T}) = \mathcal{N}_m(\mathbf{T}|\mathbf{M}, \sigma_T^2\mathbf{I}, \sigma_T^2\mathbf{I}) = \frac{1}{c}\exp\left(-\frac{1}{2\sigma_T^2}\parallel(\mathbf{T}-\mathbf{M})\parallel_F^2\right), \tag{19}$$

missä $||\quad||_F$ on Frobeniuksen normi, joka määritellään $\parallel \mathbf{X} \parallel_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|x_{ij}|^2}$. Muutosmatriisin $\mathbf{T}$ priorijakauma on

$$\mathrm{P}(\mathbf{T}) \sim \mathcal{N}_+(\parallel \mathbf{T}-\mathbf{M} \parallel_F |0, \sigma_T^2), \tag{20}$$

missä $\mathcal{N}_+$ on katkaistu normaalijakauma positiivisille arvoille. Mallin negatiivinen logaritminen todennäköisyys on tällöin,

$$\ell = \log|\Sigma| + \operatorname{tr}\Sigma^{-1}\Sigma + \frac{\| \mathbf{T} - \mathbf{M} \|_F^2}{\sigma_T^2}. \tag{21}$$

missä $\Sigma = \mathbf{W}\mathbf{W}^T + \mathbf{\Psi}$. Kaavasta (21) nähdään, että negatiivisessa log-todennäköisyydessä ainoa muuttuva termi on $\frac{||\mathbf{T}-\mathbf{M}||_F^2}{\sigma_T^2}$. Jos $\sigma_T^2 \to \infty$, muutosmatriisin muoto ei vaikuta negatiiviseen logaritmiseen todennäköisyyteen, jolloin muutosmatriisi $\mathbf{T}$ on täysin vapaa. Tällöin malli (15) pelkistyy pCCA-malliksi (11). Jos taas $\sigma_T^2 \to 0$, muutosmatriisi on $\mathbf{T} = \mathbf{M}$. $\mathbf{M}$ on siten muutosmatriisin etukäteistiedon muoto ja σ_T^2 on muutosmatriisin vapausparametri, joka ilmaisee kuinka paljon muutosmatriisi $\mathbf{T}$ voi erota etukäteistiedosta $\mathbf{M}$. [Lahti09] [1]

Mallin parametrien laskukaavat löytyvät liitteestä 6.

[1] Osa yksityiskohdista on otettu paperin [Lahti09] MLSP-esitelmästä

3 Biologinen tausta

3.1 Syöpään vaikuttavat geenimutaatiot

Artikkelissa [Vogelstein04] kuvataan syöpään vaikuttavat geenit ja niiden mutaatiot. Syövän syntyyn vaikuttavat geenit voidaan jakaa kolmeen ryhmään. Onkogeenit mutatoituvat jatkuvasti aktiivisiksi tai aktiivisiksi tilanteissa, joissa ne eivät normaalisti ole aktiivisia. Toisin sanoen geenistä transkriptoidaan RNA:ta tilanteissa, jolloin normaalisti transkriptiota ei tapahdu. Kasvurajoitegeeneissä sensijaan mutaatiot vähentävät geenin aktivaatiota. Nämä molemman tyyppiset geenien mutaatiot liittyvät prosessiin, joka lisää kasvainsoluja stimuloimalla solujen syntyä tai inhiboimalla apoptoosia. Kolmas syöpään vaikuttava geeniryhmä on stabilointigeenit, jotka liittyvät pienten virheiden korjauksiin tavallisessa DNA:n kopioimisessa, genomin rekombinaatioon mitoosissa ja kromosomin eriyttämiseen. Stabilointigeenien mutaatiot johtavat korjausmekanismin heikkenemisen takia muiden geenien mutaatioiden esiintymisen kasvuun. Mikään yksittäinen mutaatio ei riitä syövän syntymiseen, vaan siihen vaaditaan asteittaista geneettisten muutosten kasautumista.

3.2 Geneettisten mittausten tekeminen

Geenin kopioluvun mittaamiseen voidaan käyttää vertailevaa genomista hybridisaatiota (aCGH), jolla saadaan korkean resoluution mittaukset koko genomin kopiolukumuutoksista [Shinawi08]. Toisin sanoen saadaan mitattua onko tapahtunut geneettisten alueiden kopioitumista tai poistumista. Mittauksessa tutkittava ja kontrollinäyte merkitään eri väreillä ja hybridisoidaan geenisiruun, jossa on useita tuhansia koettimia. Tämä geenisiru skannataan ja koettimien kohdilla olevat väri-intensiteeti mitataan. Intensiteettien suhteet ovat verrannollisia kopiolukusuhteisiin testi- ja kontrollinäytteissä.

Geenien ilmentymämittauksissa tutkitaan geenien aktiivisuutta eli sitä kuinka paljon soluissa niistä tehdään lähetti-RNA:ta (mRNA). Tähän mittaukseen voidaan käyttää cDNA- geenisiruja. Niissä tutkittavan näytteen lähetti-RNA:lle (mRNA) tehdään käänteistranskriptio komplementaarikseksi DNA:ksi (cDNA) ja merkitään väriaineella. Yhdessä merkityn kontrollinäytteen kanssa cDNA:ksi muunnettu näyte hybridisoidaan geenisiruun, josta väri-intensiteettien erojen perusteella saadaan tutkittavan näytteen ilmentymämittaukset. Tarkempia yksityikohtia mittauksesta voi lukea [Stoughton05] ja mittauksien käsittelystä [Allison06].

Lyhyiden mikroRNA-pätkien aktiivisuuden mittaaminen voidaan tehdä geenisiruilla mRNA:lle tehtävän menetelmän kaltaisesti [Shingara05]. Muita keinoja miRNA-mittauksien tekemiseen löytyy mm. [Chen05].

3.3 Syöpägeenien vaikutus kopioluku- ja ilmentymämittauksiin

Artikkelissa [Stranger07] analysoitiin kopiolukumuutosten ja ilmentymien yhteyttä. Mutaatioiden aiheuttamat geenien kopiolukumuutosten vaikutus geenien ilmentymiseen ei ole yksiselitteistä. Useimmiten kopiolukumuutos vaikuttaa välittömässä läheisyydessä sijaitsevien geenien ilmentymään, mutta osa muutoksista vaikuttaa jopa yli 2 miljoonan emäsparin päässä. Kopiolukumuutokset voivat vaikuttaa myös useamman kuin yhden geenin ilmentymään. Kopioluvun muutoksen vaikutus kohdegeenien ilmentymään voi vaikuttaa kahteeen suuntaan. Suurimman osan merkittävistä kopiolukumuutosten ja ilmentymän riippuvuuksista on kuitenkin havaittu olevan korrelaatioltaan positiivisia erityisesti samalla alueella.

Aiemmin on käytetty paljon kopiolukumittauksia syöpään vaikuttavien geenien etsinnässä [Forozan97]. Niiden avulla on löydetty useita syöpään vaikuttavia geenejä sekä iso määrä alueellisia kromosomikasvuja ja DNA:n vahvistumisia joillakin syöpätyypeillä. Kopiolukumittauksilla on havaittu samoilla kasvaintyypeillä olevan yhteneväisiä muutoksia.

Syöpätutkimuksessa on tutkittu runsaasti geenien ilmentymisen muutoksia [Lockhart00]. Pelkkien geeni-ilmentymien tasojen on kuitenkin huomattu olevan epäluotettavia merkkejä syöpää aiheuttavista geeneistä, koska yhdelle geenille saattunut mutaatio näkyy muutoksena myös muiden geenien ilmentymässä [Miklos04]. Mutaatiot ovat paljon luotettavempia merkkejä mahdollisista syöpään vaikuttavista geeneistä kuin ilmentymien epänormaaliudet [Vogelstein04]. Syöpäsoluissa on tosin huomattu olevan huomattavasti enemmän mutaatioita kromosomitasolla verrattuna normaaleihin soluihin [Lengauer98]. Koska säätelygeenien mutaatiot lisäävät solussa kaikkien mutaatioiden määrää, on hankala erottaa mitkä kopiolukumuutokset ovat syövän syntyyn vaikuttavia ja mitkä johtuvat DNA:n korjausmekanismien heikkenemisestä.

Ilmentymä- ja kopiolukumittausten yhdistelmällä voidaan tarkastella, mitkä kopiolukumuutokset ovat aiheuttaneet syöpäsoluissa ilmentymämuutoksia. Nämä geenit ovat vahvempia ehdokkaita syövän aiheuttajiksi, sillä kopiolukumuutoksilla on havaittu olevan suora vaikutus geeni-ilmentymien voimakkuuteen syöpäsoluissa. Yhdistettyjen ilmentymä- ja kopiolukumittausten on todettu olevan hyödyllisiä syöpään vaikuttavien geenien etsinnässä [Pollack02]. Tutkimalla ilmentymä- ja kopiolukumittauksia voidaan löytää syövän kannalta tärkeitä geenejä, joita ei pelkällä kopiolukumittausten tutkimisella pystyttäisi tunnistamaan luotettavasti [Liu06].

Tätä syöpägeeneissä esiintyvää ilmentymä- ja kopiolukumittausten riippuvuutta samalla alueella pyritään tässä työssä käyttämään hyödyksi. Tämän etukäteistiedon hyödyntämistä tutkitaan syöpägeenien etsinnässä.

3.4 Mirko-RNA

MikroRNA:t ovat lyhyitä 19-25 nukleotidin mittaisia koodaamattomia RNA-pätkiä, joilla on havaittu olevan tärkeä tehtävä monissa biologisissa prosesseissa. miRNA:ta

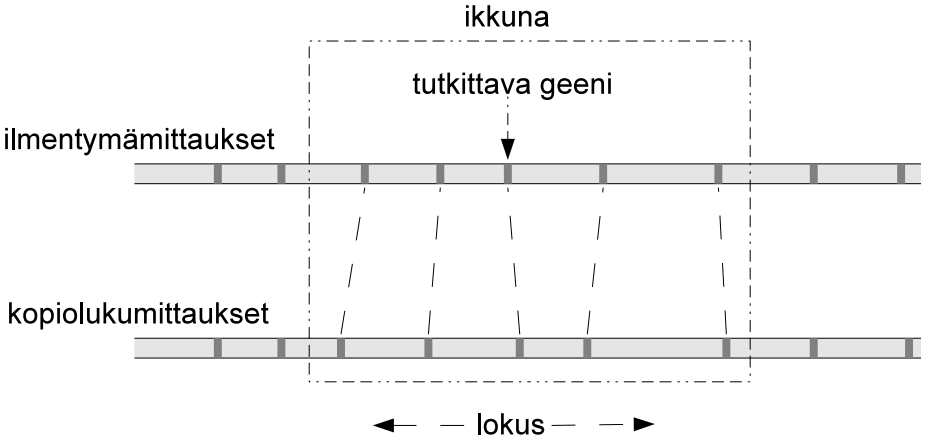
löytyy kasveista, selkärangattomista ja selkärankaisista ja monien sekvenssin on havaittu olevan säilyneitä kaukaisesti toisilleen sukua olevilla lajeilla.

Huomattava määrä miRNA:ita on havaittu sijaitsevan useisiin syöpiin liittyvien ilmentymämuutosten läheisyydessä ja niiden ilmentymämuutoksilla on havaittu olevan yhteys syövän syntyyn [Calin04]. Artikkelissa [Lu05] havaittin melkein kaikilla tutkituilla miRNA:illa olleen muutoksia ilmentymässä syöpänäytteissä. Muuttuneita miRNA-ilmentymiä havaittiin useissa kasvaintyypeissä ja miRNA:n ilmentymään perustuvan profiloinnin todettiin voivan erottaa syöpätyyppejä toisistaan. Tosin pelkillä miRNA ilmentymämittauksilla todettiin olevan hankalaa erottaa normaalia näytettä syöpänäytteestä.

Tässä työssä tutkitaan voiko syöpään vaikuttavia miRNA:ita löytää tutkimalla miRNA- ja kopiolukumittausten riippuvuutta samalla alueella.

4 Aktiivisten kopiomuutosten haku

Aktiivisilla kopiomuutoksilla tarkoitetaan geneettisen poikkeaman aiheuttamaa muutosta geenin aktiivisuudessa. Tätä voidaan tutkia tarkastelemalla geenin ilmentymä- ja koiolukumittauksia. Kappaleessa 3 todettiin kopioluku- ja ilmentymämittauksilla olevan vahva paikkariippuvuus syöpägeeneissä. Tämän takia syöpägeenejä voidaan etsiä tekemällä riippuvuusmallitus kopioluku- ja ilmentymämittauksille tutkittavan geenin ympäristössä, jota varten mittaukset paritetaan siten, että jokaista ilmentymämittausta vastaa mahdollisimman läheltä otettu opiolukumittaus. Kuvassa 3 näkyy mittausten paritus ja riippuvuusmallitukseen otettava ikkuna mittauksista tutkittavan geenin ympäriltä.



Kuva 3: Riippuvuusmallitusta varten tehtävä mittausten paritus ja tutkittavan geenin ympärille tehtävä ikkunointi. Yksittäiset mittaukset on kuvattu tummanharmaalla ja niiden paritus katkoviivalla

4.1 Mikrosiruaineiston erityispiirteiden huomiointi

Kopioluku- ja ilmentymämittauksista otettuun ikkunaan sovitetaan kappaleessa 2.2 esitettyä todennäköisyyspohjaista riippuvuusmallia

$$\begin{aligned} \mathbf{X} &\sim \mathcal{N}(\mathbf{W}_X\mathbf{z}, \mathbf{\Psi}_X), \mathbf{W}_X \in \mathbb{R}^{m_X \times d} \\ \mathbf{Y} &\sim \mathcal{N}(\mathbf{W}_Y\mathbf{z}, \mathbf{\Psi}_Y), \mathbf{W}_Y \in \mathbb{R}^{m_Y \times d}, \end{aligned} \tag{22}$$

missä $\mathbf{X}$ on ikkunoidut ilmentymämittaukset ja $\mathbf{Y}$ niihin paritetut kopiolukumittaukset, jolloin ikkunan koko on $m_X = m_Y = m$. Kopioluku- ja ilmentymämittausten välisessä riippuvuusmallituksessa piilomuuttujan $\mathbf{z}$ biologinen vastine yksiulotteisessa tapauksessa $d = 1$ on mutaatio tai muu poikkeama genomissa. Kerroinmatriisit $\mathbf{W}$ kuvaavat miten muutos heijastuu kopioluku- ja ilmentymämittauksiin. Kovarianssimatriisit $\mathbf{\Psi}$ kuvaavat jäljelle jääneen genomisesta poikkeamasta riippumattoman vaihtelun mittauksissa.

Rajoitettu todennäköisyysperustainen kanoninen korrelaatioanalyysi mahdollistaa mallin rajoittamisen siten, että riippuvuuksia haetaan vain saman lokuksen kopioluku- ja ilmentymämittausten välille. Tällöin käytetään erikoistapausta, jossa muutosmatriisina $\mathbf{T} = \mathbf{I}$ ja vapausparametrina $\sigma_T^2 = 0$. Mallin vapautta rajoitettua huomattavasti ja siten vähennettyä ylisovitusta.

MikroRNA- ja kopiolukumittausten välisessä mallituksessa mikroRNA-näytteen lokuksen ympäristöstä otetaan ikkunan koon verran lähimpiä kopiolukumittauksia. Tällöin $\mathbf{X}$ on mikroRNA-mittaukset ja $\mathbf{Y}$ kopiolukumittaukset, sekä $m_X = 1$ ja m_Y on ikkunakoko. MikroRNA:n riippuvuusmallituksessa piilomuuttujan $\mathbf{z}$ biologinen vastine on miRNA lokuksen mutaatiot ja kerroinmatriisi $\mathbf{W}_X$ kuvaa miten mutaatio vaikuttaa miRNA:n ilmentymään ja $\mathbf{W}_Y$ lähialueen kopiolukumittauksiin. Tästä muodostuu pCCA:n erikoistapaus, jossa toisessa muuttujajoukossa on vain yksi muuttuja, eikä tässä voi soveltaa psimCCA:ta.

Koko genomi voidaan seuloa tekemällä riippuvuusmallitus liukuvalla ikkunalla. Tällöin saadaan jokaiselle geenille (kromosomien päitä lukuunottamatta) riippuvuusmalli.

4.2 Jaetun signaalin tunnistus

Näytekovarianssi voidaan jakaa $\Sigma = \mathbf{W}\mathbf{W}^T + \Psi$. Tämän perusteella voidaan määrittää aineistojen väliselle riippuvuudelle mittari jaetun signaalinsuhteesta aineiston sisäisiin vaikutuksii.

$$\frac{\mathrm{tr}(\mathbf{W}\mathbf{W}^T)}{\mathrm{tr}(\mathbf{\Psi})}\text{[Lahti09].} \tag{23}$$

Tällä geenien pisteytyksellä saadaan selville todennäköiset aktiiviset kopiomuutokset. Tosin, koska kyse on ikkunoiduille näytteille tehdystä riippuvuusmallituksesta, saatu mahdollinen korkea pisteytys ei välttämättä liity suoraan ikkunan keskellä olevaan geeniin, vaan kertoo, että muodostetun ikkunan ympäristössä on aktiivisia kopiomuutoksia. Riippuvuuden mittari ei ole yksiselitteinen vaan siinä voitaisiin käyttää mm. permutoiduilla näytteillä laskettuja p-arvoja.

4.3 Avoimen lähdekoodin toteutus: pint

Edellä esitetty algoritmi toteutettiin pint-nimisenä avoimen lähdekoodin lisäpakettina [2] R-ohjelmointikielelle [R Development Core Team09], ja se on julkaistu Bioconductorin versiossa 2.6. Tämä lisäpkaetti pitää sisällään työkalut pCCA kehyksen riippuvuusmallitukseen sisältäen erikoistapaukset ja regularisoinnin. Siinä on myös työkalut koko genomin läpikäymiseen ikkunoimalla sekä kopioluku- ja ilmentymämittausten riippuvuusmallitukseen, että miRNA- ja kopiolukumittausten riippuvuusmallitukseen.

[2] http://www.bioconductor.org/packages/2.6/bioc/html/pint.html

Tämän lisäpaketin tarkoituksena on tarjota helppokäyttöinen työkalu syöpoäaineistojen tutkimiseen. Helppokäyttöisyyden takia pint mahdollistaa riippuvuusmallien soveltamisen geneettisiin mittauksiin ilman tarkempaa teknistä tietämistä tai osaamista.

R on tilatollisen laskennan ja grafiikan järjestelmä, joka koostuu kielestä sekä suoritusvaiheen ympäristöstä grafiikoilla. R perustuu S- ja Scheme-kieliin ja sen jakelusovellus sisältää ison määrän toimintoja tilastollisiin käsittelyihin. [Hornik10]

Bioconductor on vuonna 2001 aloitettu avoimen lähdekoodin ja kehityksen ohjelmistoprojekti, jonka tarkoituksena on tarjota työkaluja genomiaineiston analyysiin ja ymmärtämiseen. Bioconductor pohjautuu pääasiassa R-ohjelmointikieleen ja sillä on kaksi julkaisujaksoa vuosittain. Suurin osa Bioconductorin ohjelmistopaketeista jaetaan R-kielellä tehtyinä paketteina ja suurin osa ohjelmistopaketeista keskittyy DNA mikrosiruanalyysiin. Myöhemmin Bioconductor on laajentunut koskemaan myös muita genomiaineistoja kuten SAGE-, sekvenssi- ja SNP-aineistoja. [Gentleman08]

4.4 Vaihtoehtoiset menetelmät

Kirjallisuudessa on esitelty muita ilmentymä- ja kopiolukumittausten väliseen mallitukseen soveltuvia menetelmiä. Näistä menetelmistä iCluster [Shen09] on näytteiden luokitteluun mittausten perusteella tehty menetelmä. intCNGEan [Wieringen09] perustuu tilastolliseen testiin geenin kopiolukumuutosten aiheuttamista ilmentymämuutoksista. edirassa [Schäfer09] lasketaan korrelaatio ilmentymä- ja kopiolukumittausten välille erikseen syöpä- ja kontrollinäytteille ja järjestyslukutestin avulla saadaan p-arvo muutokselle. Korrelaatiopohjaisessa menetelmässä käytetään korrelaatiota mallittamaan mittausten yhteyttä, ja siinä lasketaan p-arvo muutokselle geenin ilmentymän ja sen alueen kopiolukumuutosten välille permutoimalla kopiolukuikkunaa. GSVD-pohjaisesssa menetelmässä [Berger06] iteratiivisesti projisoidaan näytteet suurinta singulaariarvoa vastaavaan ominaisvektoriin ja poistetaan kerralla pieni määrä pienintä rinnakkaista assosiaatiota sisältävät geenit. SODEGIR-menetelmässä [Bicciato09] pehmennetään suodatuksella näytteiden mittaukset ja luokitellaan jokaisessa näytteessä erikseen geenit, joilla samansuuntaisesti muuntuneita mittauksia. Näiden luokittelujen perusteella etsitään tilastollisesti koko aineistossa esiintyvät muuntuneet geeenit.

Tässä osiossa esitellään tarkemmin vaihtoehtoiset menetelmät sekä miten niitä testattiin syöpägeenien löytämisessä.

4.4.1 iCluster

iClusterin [Shen09] perusideana on yhteisesti etsiä piilossa olevia kasvaintyyppejä esimerkiksi kopioluku- ($\mathbf{X}_1$), DNA:n metylointi- ($\mathbf{X}_2$), mRNA-ilmentymämittauksista

($\mathbf{X}_3$). iClusterissa mallitetaan havaintoja seuraavalla mallilla

$$\begin{aligned} \mathbf{X}_1 &= \mathbf{W}_1\mathbf{z} + \boldsymbol{\epsilon}_1 \\ \mathbf{X}_1 &= \mathbf{W}_2\mathbf{z} + \boldsymbol{\epsilon}_2 \\ &\vdots \\ \mathbf{X}_m &= \mathbf{W}_m\mathbf{z} + \boldsymbol{\epsilon}_n, \end{aligned} \tag{24}$$

missä m on erilaisten genomimittausten määrä samoille näytteille. $\mathbf{z}$ on piilokomponentti joka yhdistää m osaa mallista ja sisältää eri mittausten väliset riippuvuudet. Virhetermit $(\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_m)$ sisältävät jäljelle jäävän jokaisen mittaustyypin sisäisen varianssin. Virhetermien kovarianssimatriisi on diagonaalinen. Kerroinmatriisit $(\mathbf{W}_1, \ldots, \mathbf{W}_m)$ projisoivat piilokomponentin mittauksiksi maksimoimalla mittaustyyppien välisen korrelaation.

Useiden mittaustyyppien sisällyttäminen korostaa mallin (24) ongelmaa muuttujien lukumäärän suhteesta näytteiden määrään. Tämä ratkaistaan iClusterissa regularisoimalla kerroinmatriiseja lisäämällä kustannustermi, joka pakottaa ison osan $\mathbf{W}$:n termeistä nollaksi. iCluster luokittelee näytteet k:n lähimmän naapurin menetelmällä mallille (24) lasketun piilokomponentin $\mathbf{z}$ avulla.

Malli (24) on vastaava mallin (22) kanssa muutamin poikkeuksin. Kappaleessa 4.1 esitetty psimCCA:han pohjautuva menetelmä jakaa genomin tasakokoisiin ikkunoihin joihin sovitetaan piilomuuttujamalli, kun taas iCluster sovittaa mallin koko genomiin. Menetelmien erona on siten niiden lähestymistapa ratkaista ison muuttujamäärän ja pienen näytemäärän ongelma. Toinen ero on mallin virheen kovarianssin muodossa, joka on iClusterissa diagonaalineen ja psimCCA:ssa vapaa. iCluster eroaa myös eri aineistojen lukumäärän suhteen, kun taas pintissä ei ole implementoituna kuin vain kahden aineiston riippuvuusmallitus. Siinä olevat riippuvuusmallit soveltuvat kyllä monen aineiston mallittamiseen.

Myös iClusterin päämäärä eroaa kappaleessa 4.1 esitetystä menetelmästä. Kappaleessa 4.1 esitetyn menetelmän tarkoituksena on etsiä geenejä tai kohtia genomista, joissa kopiolukumuutos aiheuttaa ilmentymämuutoksia. iClusterin tarkoituksena on sen sijaan luokitella näytteet ryhmiin. Tätä menetelmää voi myös soveltaa muutosalueiden hakuun tarkastelemalla minkä geenien perusteella näytteet luokittellaan.

Menetelmien syöpägeenien löytämistä testattiin tekemällä aineiston jokaiselle kromosomille näytteiden luokittelu, ja katsomalla millä geeneillä on $\mathbf{W}$:ssä nollasta poikkeavia arvoja. Menetelmän implementointiin liittyvien muistiongelmien takia luokittelua ei voitu tehdä koko genomille kerralla. Menetelmän estimoista $\mathbf{W}$:stä etsittiin ne geenit, joiden avulla oli suoritettu luokittelu.

4.4.2 intCNGEan

intCNGEan -menetelmässä [Wieringen09] kopiolukumittaukset muutetaan todennäköisyyksiksi siten onko kyseisen geenin kohdalla kopioluku pienentynyt, pysynyt

normaalina vai kasvanut. Toisin sanoen jokaiselle kopiolukunäytteelle annetaan todennäköisyys jokaiselle näistä kolmesta vaihtoehdosta. Menetelmässä mallinnetaan havaittuja geeni-ilmentymien muutoksia näiden muutostodennäköisyyksien perusteella.

Kopiolukumuutosten aiheuttamien ilmentymämuutosten testaamiseksi menetelmässä päätellään ensin kopiolukunäytteiden muutostodennäköisyyksien perusteella verrataanko kopioluvun aiheuttamaa ilmentymän kasvua normaalina pysymiseen vai kopioluvun aiheuttamaa ilmentymän pienenemistä normaalina pysymiseen. Menetelmässä voi käyttää joko CramÄér-Von Mise -tyyppistä testiä tai painotettua Mann-Whitney -testiä.

Tilastollisia testejä varten estimoidaan empiirinen nollajakauma permutoimalla jokaiselle geenille satunnaisesti valittuja ilmentymiä pitämällä geenikopioluku samana. Koska nollahypoteesina on ettei ole mitään yhteyttä kopioluvun ja ilmentymän välillä, tämä permutointi antaa satunnaisen käyttäytymisen. Nollahypoteesin p-arvon lasku suoritetaan yksisuuntaisella p-testillä.

Menetelmällä voidaan ottaa myös lähialueiden kopiolukumuutokset huomioon. Tätä perustellaan sillä, että naapurigeenien kopiolukumuutokset vaikuttavat kohdegeeniin samalla tavalla. inCNGEan käyttää hyödyksi menetelmää [Allison06], jossa huomioitavan kopiolukualueen koko määräytyy sarjana vierekkäisten kopiolukumittauksina, joilla on samankaltainen ilmentymä. Kopioluvun aiheuttamat ilmentymämuutokset testataan kuitenkin jokaiselle geenille erikseen.

Väärien löytöjen määrän hallitsemiseksi osa geeneistä hylätään ennen testiä. Hylättävät geenit määritellään kopiolukumuutostodennäköisyyksistä. Geeni hylätään, jos sen kopiolukumuutostodennäköisyyden odotusarvo on yksipuolinen. Toisin sanoen suurella osalla näytteistä on iso todennäköisyys kopioluvun pysymisenä samana. Tällöin menetelmässä käytetty testaus ei todennäköisesti ei ole tarpeeksi tehokas löytämään muutoksia geeneissä. Geeni hylätään myös, jos monella yksittäisellä näytteellä on tasaisesti jakautunut muutostodennäköisyys.

Menetelmän testauksessa tilastollisen testauksen menetelmänä käytettiin CramÄér-Von Mise tyyppistä testiä. Tuloksena saadut geenit järjestettiin ensin p-arvon perusteella siten, että p-arvon 0.05 alittavat geenit järjestettiin menetelmän estimoiman muutoksen suuruuden perusteella. Tämän p-arvon ylittävät geenit järjestettiin p-arvon perusteella ja tasatilanteessa nämä järjestettiin muutoksen suuruuden perusteella.

4.4.3 edira

edirassa [Schäfer09] etsitään kopiolukumuutosten aiheuttamia geeni-ilmentymämuutoksia vertailemalla kontrollinäytteitä ja potilasnäytteitä. Kopiolukumittaukset jaetaan kromosomeissa erikseen molemmissa ryhmissä yhtäjaksoisiin alueisiin, joissa kopiolukumuutokset ovat samanlaisia. Alueen sisällä kaikille näytteille annetaan näytteiden kopiolukumittauksien mediaani. Alueet määritellään pienimmän neliösumman avulla käyttämällä alueiden lukumäärän määrittämisessä Bayesilaista informaatio-

kriteeriä.

Kopioluvun ja ilmentymän välisessä analyysissä käytetään edira:ssa muokattua Pearsonin korrelaatiota. Korrelaatiossa keskitetty korrelaatiokerroin määritellään

$$r_{EC} = \frac{\sum_{j=1}^{m} (X_j - A)(Y_j - B)}{\sqrt{\sum_{j=1}^{m} (X_j - A)^2} \sqrt{\sum_{j=1}^{m} (Y_j - B)^2}} \tag{25}$$

missä $Y_j, j = 1, \ldots, m$ on potilaiden ilmentymämittaukset ja $X_j, j = 1, \ldots, m$ niitä vastaavien aluiden kopiolukuarvot. A ja B ovat kontrolliryhmän näytteistä laskettut mediaanit. Poikkeamat referenssimediaaneista kopioluku- ja ilmentymäarvoissa, $X_j - A$ ja $Y_j - B$, ovat samansuuntaiset jos niillä on sama etumerkki. Tämä johtaa positiiviseen arvoon r_{EC}:ssä. r_{EC} saa arvoja -1 ja 1 välillä kuten tavallinen Pearsonin korrelaatiokerroin.

Menetelmä laskee näytteille p-arvot samansuuntaisista kopioluku- ja ilmentymämuutoksista Wilcoxonin merkkisen järjestyslukutestin avulla. Näiden p-arvojen avulla menetelmä etsii alueita, joista löytyy poikkeavuuksia. Tämä alueiden etsimisalgoritmi jakaantuu kahteen osaan: iteratiiviseen alueiden hakuun ja testaamiseen löydettyjen alueiden poikkeavuudesta.

Kappalessa 4.1 esitettyyn malliin verrattuna edirassa on muutama oleellinen ero. edirassa näytteet jaetaan kontrolli- ja potilasnäytteisiin ja poikkeamien etsintä tehdään vertaamalla potilasnäytteiden mittauksia kontrollinäytteisiin. Kappaleessa 4.1 esitetty menetelmä sen sijaan ei tee eroa eri näytteille, ja poikkeamien etsintä tehdään kaikille näytteille samanaikaisesti. Riippuvuuksien mallitus tehdään edirassa muokatulla Pearsonin korrelaatiolla, kun taas kappaleen 4.1 menetelmä mallintaa kanoniseen korrelaatioon pohjautuvilla malleilla. Viereisten geenien näytteiden huomioimisessa on myös eroja. Edira käyttää Bayesilaista informaatiokriteeriä ja iteratiivista p-arvoihin perustuvaa alueiden etsintää. Kappaleen 4.1 menetelmä sen sijaan käyttää vakiokokoista ikkunaa.

Menetelmän testataamiseksi kopiolukumittaukset jaettiin alueisiin ja suoritettiin poikeavien geenien etsintä menetelmän vakioparametreilla. Geenit järjestettiin ensin muutoksen suunnan mukaan siten, että molemmissa mittauksissa samansuuntaisesti muuttuneet geenit järjestettiin ensin. Tämän jälkeen geenit järjestettiin p-arvojen mukaan.

4.4.4 Korrelaatioperusteinen riippuvuushaku

Korrelaatioperusteisessa riippuvuushaussa [Lipson04] naapurigeenien kopiolukumuutosten vaikutus geenin ilmentymään otetaan huomioon ottamalla geenin ympäriltä vakiokokoinen ikkuna kopiolukumittauksista. Geenin ilmentymämittausvektorista $X(i, \cdot)$ lasketaan korrelaatioiden keskiarvo kopiolukumittasten vektoreille geenin i k-kokoisessa naapurustossa $\Gamma_k(i) = (i - k, \ldots, i + k)$

$$r(i, \Gamma_k(i)) = \frac{1}{2k+1} \sum_{j=i-k}^{i+k} r(i, j), \tag{26}$$

missä $r(i,j)$ on mikä tahansa korrelaatiomitta ilmentymävektorin $X(i,\cdot)$ ja kopiolukuvektorin $Y(i,\cdot)$ välillä.

Korrelaatiomitoista saadaan laskettua p-arvot permutoimalla satunnaisia ilmentymävektoreita kopiolukumittasten naapurustolle kullekin geenille. Alkuperäinen ja permutoinnilla lasketut korrelaatioarvot järjestetään ja p-arvoksi saadaan geenin ilmentymällä lasketun korrelaation järjestysnumero jaettuna permutaatioiden määrällä.

Menetelmä käyttää genomiyhtenäisiä alimatriiseja löytääkseen aktiivisten kopiomuutosten haussa. Menetelmässä käytetään biologista mallia, jonka mukaan tietyllä kromosomialueella tietyssä näytteessä genomimuutos vaikuttaa suurinpaan osaan kopiolukumittauksia, mutta vain muutamiin ilmentymämittauksiin.

Korrelaatioperusteinen käyttää vakiokokoista ikkunointia kuten pint. Korrelaatioperusteisessa menetelmässä ikkunoidaan kuitenkin vain kopiolukumittaukset, eikä myös ilmentymämittauksia, kuten pintissä. Menetelmät eroavat myöskin ilmentymä- ja kopiolukumittausten välisen riippuvuuden mittarissa. Korrelaatioperusteisessa menetelmässä lasketaan permutoinnilla empiiriset p-arvot, kun taas pintissä ei ole tällä hetkellä implementoituna p-arvojen laskemista, vaan siinä lasketaan vain sovitetun mallin parametreistä riippuvuusarvo. Suurin ero menetelmien välillä on kuitenkin tilastollisen malllin puute korrelaatioperusteisessa menetelmässä. Korrelaatioperusteinen menetelmä muistuttaa pint:ssä olevaa miRNA sovellusta, jossa vain toinen mittausaineisto ikkunoidaan.

4.4.5 GSVD-perusteinen menetelmä

GSVD-perusteinen menetelmä [Berger06] perustuu geenien "leikkaamiseen", jossa iteratiivisesti prjoisoidaan mittaukset suurinta singulaariarvoa vastaavaan ominaisvektoriin. Tämä projektio vastaa suuntaa, jolla mittauksilla on suurin variaatio. Suurimmat rinnakkaiset assosiaatiot säilytetään tulevia projektioita varten, kun pieni osa geeneistä, joilla pienin variaatio, leikataan pois. Leikattu aineisto järjestetään uudelleen ja iteratiivista prosessia jatketaan, kunnes geeneistä on jäljellä pieni ennalta määrätty osa. Tämä geenien leikkaamiseen singulaariarvohajotelmalla perustuva menetelmä tuottaa lopputuloksena sisäkkäiset klusterit sen perustella, miten paljon geeneillä on variaatiota eri näytteissä.

GSVD-perusteisessa menetelmässä kaksi mittausaineistoa yhdistetään yhteen matriisiin $\mathbf{R} = (\mathbf{R}_A\mathbf{R}_B)^T \in \mathbb{R}^{(n+p)\times m}$. Tälle matriisille lasketaan lasketaan yleistetty singulaariarvohajotelma

$$\mathbf{R}_A = \mathbf{U}(\mathbf{\Sigma}_A, \mathbf{0})\mathbf{X}^{-1} \tag{27}$$

$$\mathbf{R}_B = \mathbf{U}(\mathbf{\Sigma}_B, \mathbf{0})\mathbf{X}^{-1}, \tag{28}$$

missä

$$\mathbf{\Sigma}_A = \begin{pmatrix} \boldsymbol{I}_A & & \\ & \mathbf{D}_A & \\ & & \mathbf{0}_A \end{pmatrix} \tag{29}$$

$$\mathbf{\Sigma}_B = \begin{pmatrix} \boldsymbol{I}_B & & \\ & \mathbf{D}_B & \\ & & \mathbf{0}_B \end{pmatrix}. \tag{30}$$

Tämän hajotelman avulla poistetaan ennalta määrätty osa geeneistä tarkastelemalla matriisien $\mathbf{D}_A$ ja $\mathbf{D}_B$ diagonaaleilla olevia arvoja $\mathbf{D}_A = \text{diag}(\alpha_1, \vdots, \alpha_m$ ja $\mathbf{D}_B = \text{diag}(\beta_1, \vdots, \beta_m$. Näillä arvoilla voidaan määritellä kulma

$$\theta_i = \arctan\left(\frac{\alpha_i}{\beta_i}\right) - \frac{\pi}{4}. \tag{31}$$

Tämä kulma määrittää suhteellisen merkittävyyden geeniavaruuden projektiossa ensimmäisen mittausaineistolla suhteessa toiseen mittausaineistoon. Kulma 0 tarkoittaa, että geenit voivat olla yhtä merkittäviä molemmissa mittausaineistoissa. Kulmalla $+\pi/4$ toisella mittausaineistolla ei ole mitään merkittävyyttä suhteessa ensimmäiseen, ja kulmalla $+\pi/4$ ensimmäisellä mittausaineistolla ei ole mitään merkittävyyttä suhteessa toiseeen.

Menetelmässä valitaan tavoitesuunta θ:lle vaihtoehdoista θ_{min}, θ_{max} ja $\theta = 0$. Aineisto projisoidaan $\mathbf{X}$:n tavoitesuunnan perusteella valittua suurinta varianssia vastaavaan sarakkeeseen. Tästä projektiosta valitaan molemmista aineistoista ennalta määrätty osa $\mathbf{R}_A$ ja $\mathbf{R}_B$ geeneistä, jotka edustavat suurinta rinnakkaista riippuvuutta. Tämän jälkeen matriisit $\mathbf{R}_A$ ja $\mathbf{R}_B$ muodostetaan uudelleen säilytetyillä geeneillä. Tätä jatketaan kunnes jäljella olevia geenejä on vähemmän kuin näytteitä.

GSVD-menetelmää muokattiin testausta varten järjestämällä leikkaamisella poistetut geenit. Jokaisella iteraatiolla poistetut geenit järjestettiin projektion avulla samalla parusteella kuin menetelmä säilyttää geenit. Näin saadaan lista kaikista pois leikatuista geeneistä järjestettynä menetelmän geenin poistamisen perusteella. Viimeisessä iteraatiossa geenejä ei poistettu, vaan kaikki jäljellä olleet geenit järjestettiin projektion perusteella. GSVD-menetelmää testattiin kaikilla kolmella taivoitekulmalla. Jokasella iteraatiolla jätettävien geenien suhteeksi valittiin vakioarvo 95%. Koska menetelmä käsittelee molempien mittausaineistojen geenejä erikseen, lopputuloksena tulee jokaiselle kulmalle kaksi järjestettyä listaa geeneistä, joista toinen on järjestetty kopiolukumittausten mukaan ja toinen ilmentymämittausten mukaan.

4.4.6 SODEGIR

SODEGIR [Bicciato09] on kolmivaiheinen menetelmä genomilaajuiseen kopiolukuja ilmentymämuutosten tunnistamiseen yhdessä näytteessä tai koko aineistossa. Ensimmäisessä vaiheessa tunnistetaan tilastollisesti alueelliset epätasapainot ilmentymä- ja kopiolukumittauksissa. Toisessa vaiheessa tilastollisesti määritellään geenit, joilla

on yhdenmukaisia muutoksia ilmentymä- ja kopiolukumittauksissa yhdessä näytteessä. Näitä geenejä kutsutaan SODEGIR:ksi. Viimeisessä vaiheessa yksittäisten näytteiden SODEGIR:ien perusteella määritellää koko aineiston SODEGIR tunnusmerkki.

Ensimmäisessä vaiheessa kopioluku- ja ilmentymämittaukset muutetaan normaalinäytteistä poikkeaviksi koppioluku- ja ilmentymäarvoiksi vertailemalla tutkittavia näytteitä ja kontrollinäytteitä. Nämä arvot tasoitetaan suodattamalla kernelifunktioilla, jolloin saadaan ilmentymäarvoja tasoitettua ja kopiolukuarvojen saaminen geenien kohdalta mahdollistuu.

Toisessa vaiheessa määritellään tilastollinen merkitsevyys samansuuntaisille muutoksille kopioluku- ja ilmentymäarvoille. Muutoksille lasketaan tilastolliset luottamusvälit permutoinnin avulla. Näiden luottamusvälien avulla määritellään geenin tila, onko geenillä kopioluku- tai ilmentymämuutos. Tästä tilasta käytetään mentelmässä nimitystä SODEGIR. Arvolla 1 se vastaa samansuuntaista heikkenemistä arvoissa, arvolla 3 se vastaa samansuuntaista vahvistumista arvoissa ja arvon 2 se saa kaikissa muissa tilanteissa.

Menetelmä suorittaa aiemmat vaiheet jokaiselle tutkittavan aineiston näytteelle erikseen. Viimeisessä vaiheessa jokaiselle geenille lasketaan p-arvo kaikkien näytteiden sen geenin SODEGIR-arvojen perustella. Näiden p-arvojen perustella saadaan q-arvot, joita käytetään mittana muutoksen merkittävyydelle.

4.5 Analysoitava aineisto

4.5.1 Mahasyöpäaineisto

Menetelmien testauksessa käytetään kahta eri aineistoa: mahasyöpä- ja keuhkosyöpäaineistoa. Mahasyöpäaineisto koostuu 38 mahasyöpänäytteestä ja 8 normaalin kudoksen näytteestä otetuista kopioluku- ja ilmentymämittauksista [Myllykangas08]. Mahasyöpäaineisto esikäsiteltiin kuten [Myllykangas08] paperissa, ja esikäsittelyn jälkeen aineisto sisälsi 5596 paritettua mittausta kustakin näytteestä. Mahasyöpäaineisto sisälsi biolääketieteen asiantuntijan kokoaman listan 59 tunnetusta mahasyöpään vaikuttavasta geenistä [Lahti09].

4.5.2 Keuhkosyöpäaineisto

Keuhkosyöpäaineisto koostuu 16 potilaan keuhkosyöpäkasvaimesta otettuista näytteistä, jotka sisältävät ilmentymä-, kopioluku- ja mikroRNA -mittaukset [Wikman07], [Nymark06]. Ilmentymämittaukset on käsitelty MAS5-algoritmilla [Hubbell02]. Kopiolukumittausten koetinarvot normalisoitiin nollakeskiarvolle ja yksikkövarianssiin. miRNA-data on kvantiilinormalisoitu ja log2-muunnettu. Esikäsittelyn jälkeen keuhkosyöpäaineisto sisälsi 354 miRNA-mittausta ja 8521 kopiolukumittausta. Geeni-ilmentymä- ja kopiolukumittausten välisessä riippuvuusmallitusta varten molempien mittausten koettimet paritettiin. Keuhkosyöpäaineistolle saatiin 5123 paritet-

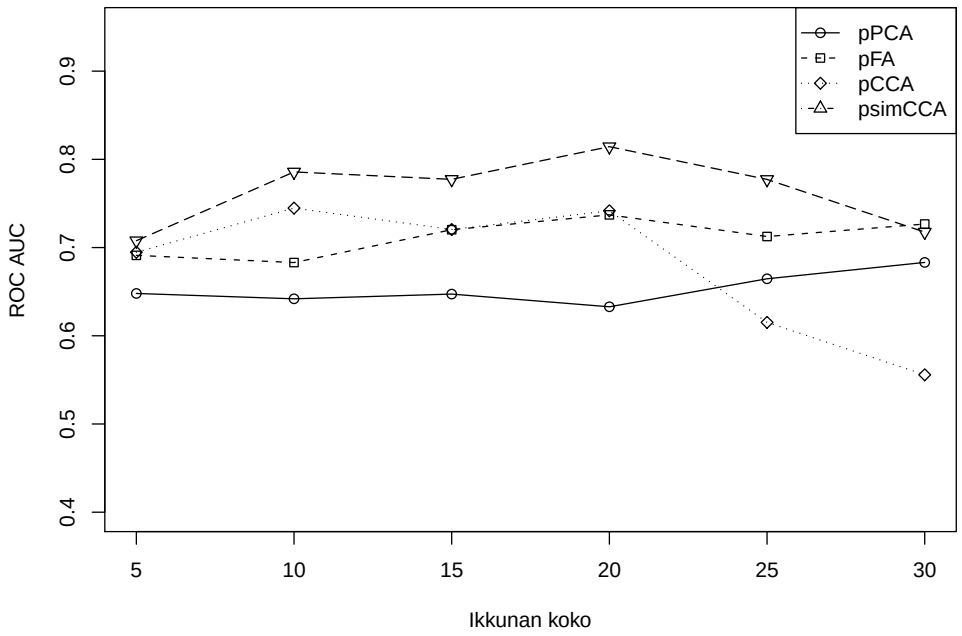
tua kopioluku- ja ilmentymämittausta. Keuhkosyöpäaineisto sisälsi biolääketieteen asiantuntijan kokoaman listan 44 tunnetusta keuhkosyöpään vaikuttavasta geenistä.

4.5.3 Aineistojen käyttö

Suuremman näytemäärän takia menetelmien testauksessa käytetään pääasiallisesti mahasyöpäaineistoa. Mahasyöpäaineisto ei sisällä miRNA-mittauksia, jonka vuoksi miRNA:n ja kopioluvun välisen riippuvuusmenetelmien tutkimiseen käytetään keuhkosyöpäaineistoa. Ilmentymämittausten ja kopiolukumittausten välisen riippuvuusmenetelmien tuloksia verrataan myös keuhkosyöpäaineistolla.

5 Tulokset

5.1 Parametrien vaikutus korrelaatioanalyysissä

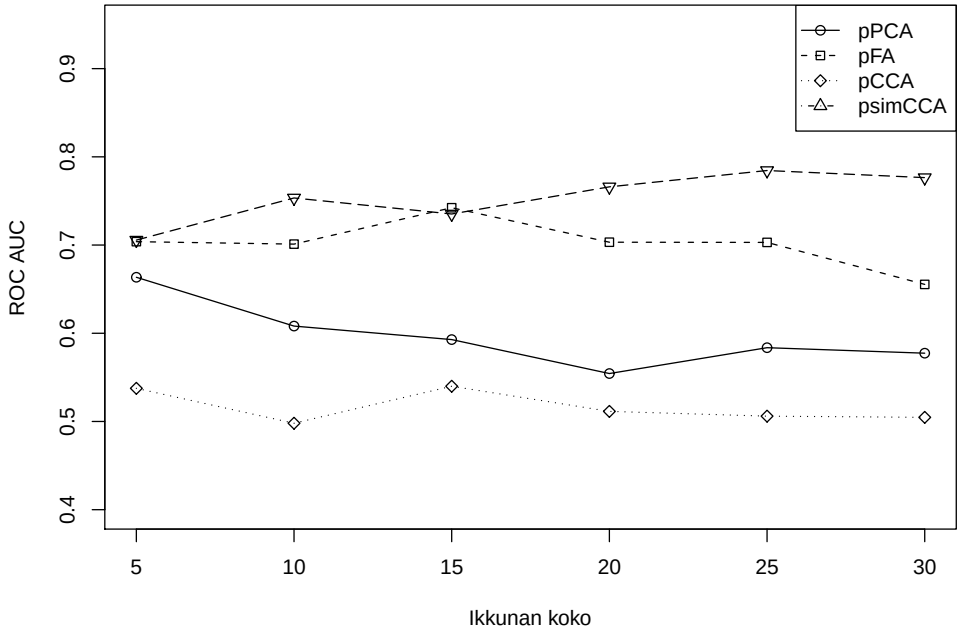


Kuva 4: Mallien AUC eri ikkunako'oilla piilomuuttujan ulottuvuuden ollessa 1

Mallin (22) toimivuutta löytää tunnettuja syöpägeenejä mahasyöpäaineistosta testattiin eri kovarianssirajoituksilla ja kerroinmatriisin $\mathbf{W}$ regularisoinnilla sekä ilman regularisointia. Myös piilomuuttujan dimensionaalisuuden vaikutusta testattiin. psimCCA:sta testattiin tässä vain erikoistapaus, jossa $\sigma_T^2 = \infty$ ja $\mathbf{M} = \mathbf{I}$, jolloin $\mathbf{T} = \mathbf{I}$ ja siten $\mathbf{W}_X = \mathbf{W}_Y$. Yksiulotteisella piilomuuttujalla $\mathbf{z} \in \mathbb{R}^{1\times m}$ mallien ROC AUC -arvot näkyvät kuvassa 4. Täyden dimension piilomuuttujalla, jolloin piilomuuttuja $\mathbf{z}$ on kokoa $m \times m$, missä m on ikkunan koko, ROC AUC -arvot näkyvät kuvassa 5.

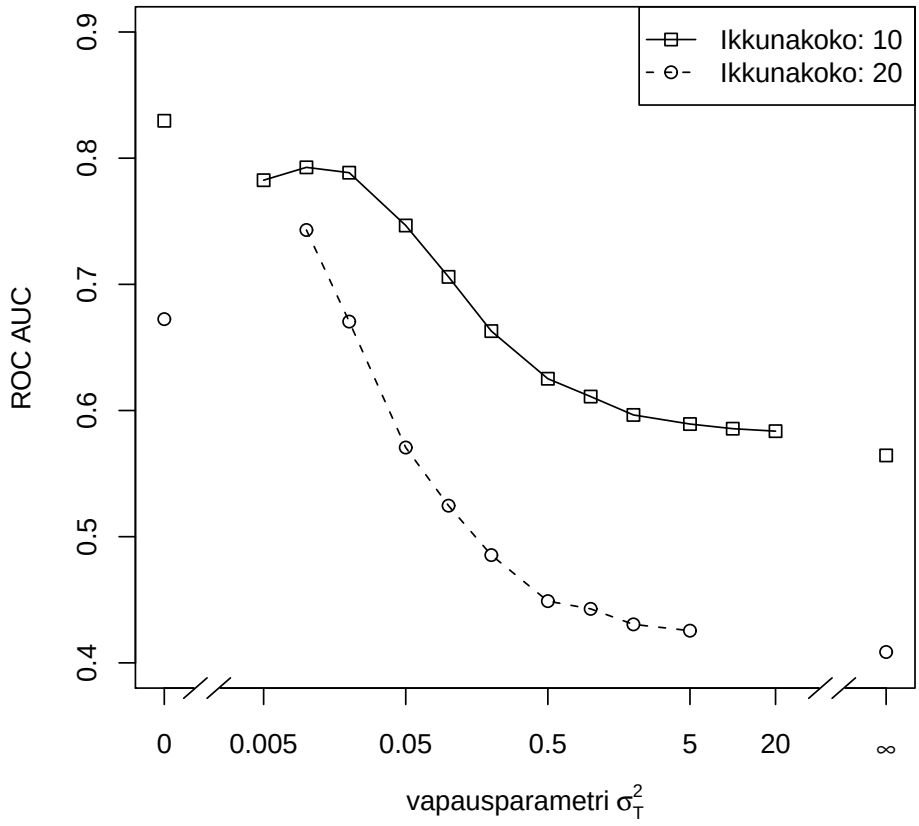
Kuvista 4 ja 5 näkyy, että psimCCA toimii paremmin sekä yksiulotteisella että täyden dimension piilomuuttujalla paremmin tai vähintään yhtä hyvin kuin mikään muu menetelmä kaikilla ikkunako'oilla. pFA tosin toimii muutamilla parametreilla yhtä hyvin kuin psimCCA. Molemmissa piilomuuttujatapauksissa pFA toimii kaikilla parametreilla paremmin kuin pPCA. Yksiulotteisen piilomuuttujan tapauksessa pCCA toimii pienillä ikkunako'oilla paremmin tai yhtä hyvin kuin pFA, mutta toimivuus heikkenee selvästi ikkunakoon kasvaessa yli 20:n. Myös psimCCA:n toimivuus alkaa heiketä ikkunakoon kasvaessa yli 20:n.

Täyden $\mathbf{z}$ dimension malleissa näkyy selvästi reguloimattomien pFA ja pPCA mal-



Kuva 5: Mallien AUC eri ikkunako'oilla piilomuuttujan ulottuvuuden ollessa k

lien ylioppimisen vaikutus ikkunan kasvaessa. pCCA:ssa ylioppiminen näkyy kaikilla ikkunako'oilla ja se ei toimi juurikaan satunnaisotantaa paremmin. psimCCA:ssa ikkunakoon kasvu ei heikennä mallin toimivuutta.



Kuva 6: Vapausparametrin σ_T^2 vaikutus toimivuuteen kahdella ikkunako'illa 10 ja 20 sekä psimCCA:n erikoistapaukset 700:lla satunnaisikkunalla (tähän kuvaan tulee vielä lisää sigma-arvoja kunhan laskennat tulee valmiiksi, tulokset voi vähän muuttua)

psimCCA-mallista testattiin yleisempi tapaus mahasyöpäaineistolla, jossa vapausparametri on välillä $0 < \sigma_T^2 < \infty$. Koska yleisessä tapauksessa psimCCA:n parametrien estimointi sisältää kaksi sisäkkäistä EM-iteraatiota, se on laskennallisesti huomattavasti raskaampi kuin psimCCA:n erikoistapaukset, joissa vapausparametri saa arvot $\sigma_T^2 = 0$ ja $\sigma_T^2 = \infty$. Tämän takia vapausparametrin σ_T^2 vaikutuksen testaamisessa ei käytetty koko mahasyöpäaineistoa, vaan aineistosta arvottiin 700 satunnaista määrätyn kokoista ikkunaa riippuvuusmallitusta varten. Eri ikkunako'oilla satunnaisikkunat ovat eri, jolloin niiden tulokset eivät ole verrattavissa keskenään, kuten ei myöskään satunnaisikkunoilla saadut tulokset ole verrattavissa koko genomilla saatuihin tuloksiin. Tällä tavalla saadaan kuitenkin verrattua miten vapausparam-

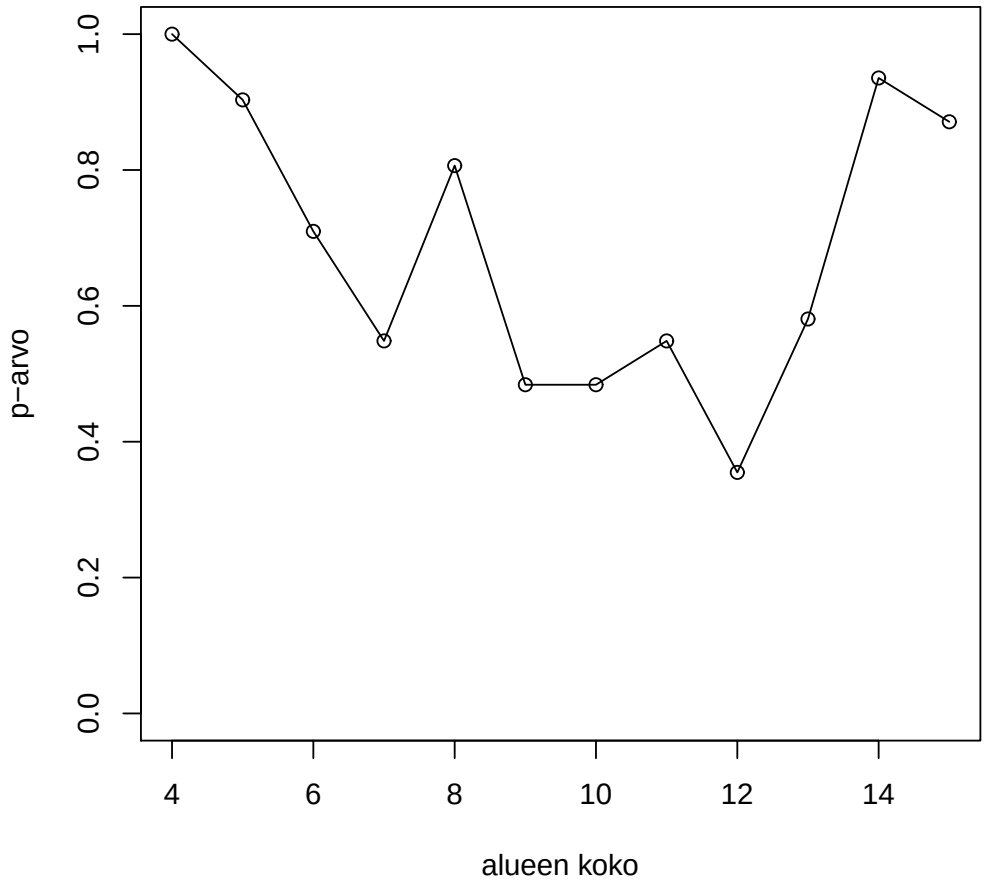
terin σ_T^2 muutos vaikuttaa kohdegeenien löytymiseen. Piilomuuttujan $\mathbf{z}$ dimension kasvattaminen lisäsi myös laskentaa tarvittavaa aikaa huomattavasti, joten testaaminen tehtiin vain yksiulotteiselle tapaukselle. Vapausparametrin vaikutus syöpägeenien löytymiseen näkyy kuvassa 6. Kuvaan on lisätty myös erikoistapaukset samoilla satunnaisikkunoilla. Mallin erikoistapauksesta, jossa vapausparametrilla on arvo $\sigma_T^2 = \infty$, ei ole tehty sovellustoteutusta, jossa mallinkovarianssi $\mathbf{\Psi}$ olisi isotrooppinen, joten kuvassa 6 oleva erikoistapaus $\sigma_T^2 = \infty$ ei ole täysin vertailukelpoinen. Käytetyssä erikoistapauksessa ei ollut mallin kovarianssille rajoitetta. Oletettavasti vapaalla kovarianssilla ikkunakoolla 10 tulee isompi AUC, koska kuvassa 4 pCCA saa isomman AUC:n kuin pFA. Ikkunakoolla 20 pFA ja pCCA saa saman AUC:n, joten sillä ikkunakoolla ei oletettavasti rajoitettukaan riippuvuusmalli eroa eri kovariansseilla.

Kuvasta 6 näkee, että mallin toimivuus paranee vapausparametrin pienentyessä tiettyyn rajaan saakka. Ikkunakoolla 10 toimivuus alkaa heiketä vapausparamterin ollessa $\sigma_T^2 < 0.02$. Tämän perusteella voidaan todeta, että menetelmä toimii parhaiten syöpägeenien etsinnässä, kun sille annetaan rajoituksena $\mathbf{T} = \mathbf{I}$, sekä vapausparametrilla sallitaan pieni poikkeama tästä rajoituksesta. Myös ikkunakoolla 20 menetelmä paranee vapausparametrin pienentyessä. Tälle ikkunakoolle ei laskettu tarpeeksi pienille vapausparametrin arvoille menetelmän toimivuutta, jotta nähtäisiin samanlaista notkahdusta kuvaajassa kuin ikkunakoolla 10. Jos ikkunakoolla 10 menetelmän toimivuudessa on yhtä iso ero sekä rajoitetussa tapauksessa että rajoittamattomassa tapauksessa eri kovariansseilla, menetelmälle löytyy vapausparametrin arvo $\sigma_T^2 = 0.01$, jolla se toimii parhaiten. Tämä testaus osoittaa kuinka syöpägeenien etsinnässä etukäteistiedon hyväksikäyttö kopioluku- ja ilmentymämuutosten paikkariippuvuudessa parantaa syöpägeenien löytymistä.

pCCA-menetelmän toimivuutta syöpään vaikuttavien mikroRNA:n löytämiseksi testattiin p-testillä. Tässä testissä selvitettiin mikroRNA:n riippuvuusmallien satunnaiskäyttäytyminen tekemällä riippuvuusmalli jokaiselle mikroRNA:lle satunnaisilla samankokoisilla kopiolukuikkunoilla. Näitä satunnaismalleja tehtiin jokaisen mikroRNA:n ympäristössä yhteensä 100 ja niille laskettiin AUC:t. Näiden avulla saadaan empiiriset p-arvo menetelmän toimivuudelle $p = (k+1)/(n+1)$, missä k on niiden satunnaismallien määrä, joilla tuli isompi AUC kuin paritetulla aineistolla, ja k on kaikkien satunnaismallien lukumäärä. Nämä empiiriset p-arvot näkyvät kuvassa 7. Tästä kuvasta näkee, että pienillä ja isoilla ikkunoilla menetelmä toimii yhtä huonosti kuin satunnaisikkunoilla, ja ikkunako'oilla 6-13 vain marginaalisesti paremmin, tosin niilläkin p-arvo on $p > 0,2$. Tästä voidaan päätellä, että tämä menetelmä ei toimi syöpään vaikuttavien mikroRNA:n löytämisessä. Mahdollisista syistä menetelmän huonolle toimivuudelle on keskusteltu osiossa 6.

5.2 Vaihtoehtoisten menetelmien vertailu

Vaihtoehtoisten menetelmien toimivuutta vertailtiin psimCCA-menetelmään, jonka ikkunakoko on 25. Ikkunakoko valittiin karkeasti puolesta välistä molempia mahdollisia ääripäitä, jotta vertailua varten ei käytettäisi mallin valinnassa hyödyksi

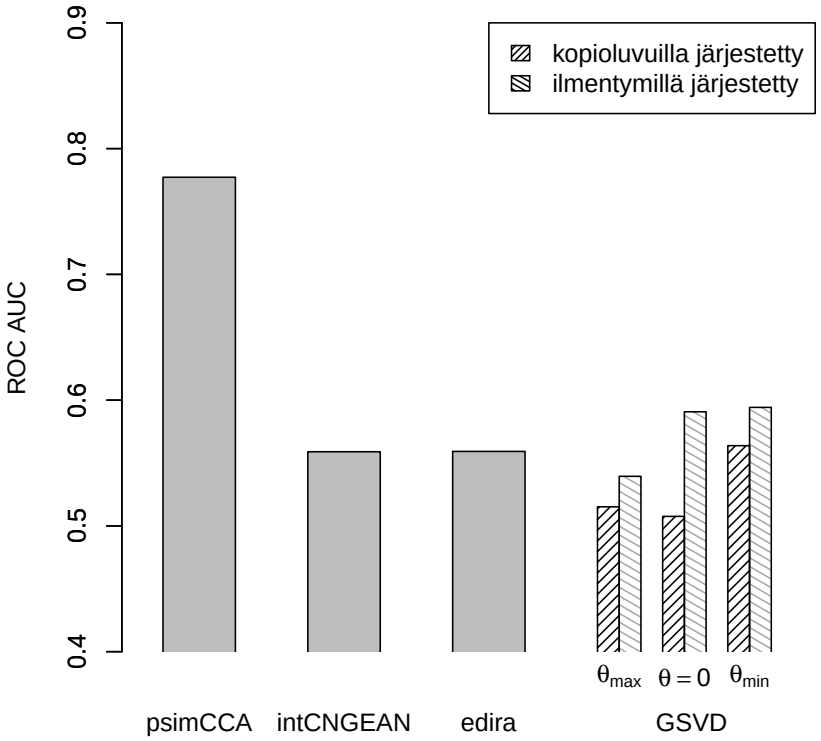


Kuva 7: Ikkunakoon vaikutus syöpägeenien löytymiseen aCGH- ja miRNA-datan välisessä riippuvuusmallituksessa

testauksessa optimoituja mallin parametrejä, koska vaihtoehtoisissa menetelmissä ei myöskään optimoitu parametrejä vaan käytettiin niille annettuja vakioarvoja. edira, inCNGEan ja GSVD-perusteisessa menetelmissä saadaan ROC-käyrä, joten niiden vertaaminen on suoraviivaista psimCCA:han. Näiden menetelmien sekä psimCCA-menetelmän AUC:t näkyvät kuvasta 8.

iClusterissa sen sijaan saadaan vain tietty määrä löydettyjä geenejä, joille ei anneta järjestystä. Niiden vertailemiseksi muista menetelmistä otettiin sama määrä korkeimman prioriteetin saaneita geenejä ja verrattiin löydettyjen syöpägeenien määrää. Kaikkien menetelmien 858 merkittävimpien löydettyjen geenien sisältämä syöpägeenien määrä näkyy taulukossa 1.

SODEGIR-menetelmän toimivuutta ei voitu verrata testauksessa ilmenneiden ongelmien takia. Korrelaatioperusteista menetelmää ei voitu testata toteutuksen puutteen takia.



Kuva 8: psimCCA:n edira:n, intCNGEan:n ja GSVD-perusteisen menetelmän vertailu AUC:lla

Kuvasta 8 näkyy selvästi kuinka psimCCA-menetelmä toimii huomattavasti paremmin kuin edira, intCNGEan ja GSVD -perusteinen menetelmä keuhkosyöpäaineistolla. Taulukon 1 mukaan ensimmäisten 858 geenin perusteella iCluster ja psimCCA

on selkeästi parempia kuin edira, intCNGEan ja GSVD -perusteinen menetelmä. Näistä kahdesta psimCCA on jonkin verran parempi menetelmä. Näiden tulosten perusteella voidaan sanoa, että psimCCA on tehokkain menetelmä syöpägeenien etsimisessä.

Taulukko 1: Löydettyjen syöpägeenien määrä menetelmien löytämän 858 merkittävimmän geenin joukossa

menetelmä	löydetyt syöpägeenit	löydettyjen syöpägeenien osuus
psimCCA	32	$3,73\%$
edira	10	$1,16\%$
iCluster	26	$3,03\%$
intCNGEan	11	$1,28\%$
GSVD (isoin ROC AUC)	12	$1,40\%$

6 Pohdinta ja yhteenveto

Menetelmien tulokset eroavat jonkin verran [Lahti09] saamista tuloksista, jotka on laskettu samalla aineistolla. psimCCA ja pCCA saavat kuitenki melko samat tulokset. Kuvassa 4 näkyvät tulokset ovat parempia kuin [Lahti09] saamat. Molemmissa on todettu psimCCA:n toimivan paremmin kuin pCCA:n, mikä osoittaa rajoittamisen parantavan tässä sovelluskohteessa menetelmän toimivuutta ja siten puolestapuhuu etukäteistiedon hyväksikäytöstä.

Mahdollisia syitä eri tuloksiin on useita. Todennäköisyysperusteiset menetelmät ovat luonteeltaan satunnaisia, jolloin lasketut mallit voivat vaihdella jonkin verran kerrasta toiseen. Tämän vaihtelun ei kuitenkaan pitäisi olla suurta. Paperissa [Lahti09] käytetty implementaatio pCCA:ssa käyttää kaavoissa (A1) sijoituksia $\mathbf{M}_X = \mathbf{I}$ ja $\mathbf{M}_X = \mathbf{P}_d$, kun taas pint-paketin implementoinnissa sijoitukset on toisinpäin. Molemmat noudattavat paperissa [Bach05] annettuja rajoituksia, mutta silti tämä näkyy riippuvuusmallien parametreissa ja siten myös menetelmän tuloksissa. Tämä on todennäköisin syy eroihin pCCA-mentelmän tuloksissa [Lahti09] paperiin. Nämä implementaation erot osoittavat kuinka herkkiä menetelmät saattavat olla toteutukselle, vaikka ne noudattaisivat taustalla olevaa hyvin perusteltua teoriaa.

Tässä työssä tehdyt testit psimCCA-menetelmälle sai kaikilla tutkituilla ikkunako'oilla paremmat tulokset kuin [Lahti09] paperissa. Tälle ei löydy mitään muuta syytä kuin menetelmän satunnaisuus, sillä mallin implementoinnin pitäisi olla identtinen. Yksi mahdollisuus on laskennassa tapahtunut virhe joko tässä työssä tai [Lahti09] paperissa saaduissa tuloksissa.

miRNA:n ja kopioluvun riippuvuuteen perustuvan menetelmän huono toimivuus voi selittyä tunnettujen miRNA:den löytötavalla. Tunnettujen syöpään vaikuttavien miRNA:n löytäminen on perustunut aiemmin pääasiassa ilmentymämittauksiin, jolloin niiden ilmentymäarvot ovat selkeästi isommat kuin muilla. Tämän takia psimCCA on löytänyt ison riippuvuuden tunnetuille syöpään liittyville miRNA:lle permutoinnista huolimatta. Pelkkien ilmentymien perusteella järjestetyillä miRNA:lla saa ROC AUC-arvon $0,9$. Menetelmän löytämät miRNA:t, jotka eivät ole tunnettujan miRNA:den listassa ei välttämättä tarkoita ettei ne olisi syöpään vaikuttavia vaan, että niitä ei ole mahdollisesti vielä löydetty pelkästään ilmentymään perustuneen etsimisen takia. Tästä varmuuden saamiseksi menetelmän löytämiä miRNA:ita pitäisi tutkia muilla keinoin, jotta selviää onko kyseessä aiemmin löytymättömiä syöpään vaikuttavia miRNA:ta vai onko kyseessä se, ettei riippuvuusmallitus sovellu syöpään vaikuttavien miRNA:den etsimiseen tässä käytetyllä tavalla.

Useissa muissa vertailtavien menetelmien papereissa oli tutkittu kyseisiä menetelmiä eri tavoin normalisoiduille aineistoille. Tämä on mahdollisesti vaikuttaa jonkin verran näiden menetelmien tuloksiin. Menetelmien eroihin voi vaikuttaa myös mahasyöpäaineiston mahdolliset erityiset ominaisuudet, joiden takia psimCCA toimii siinä erityisen hyvin tai muut menetelmät erityisen huonosti. Kuitenkin psimCCA:n AUC on huomattavasti suurempi kuin muilla menetelmillä, joten esikäsittelyn erot

eivät riitä selittämään mentelmien toimivuuksien eroja. Voidaan siis todeta, että psimCCA on paras vertailluista menetelmistä syöpägeenien etsinnässä.

iClusterin löytämät syöpägeenien määrät olivat samassa suuruusluokassa psimCCA:n löytämien kanssa. Koska iClusterissa mallinnetaan aineistojen välistä riippuvuutta samankaltaisella todennäköisyysperusteisella mallilla, voidaan todeta, että tämä malli soveltuu hyvin aktiivisten kopiomuutosten hakuun.

Muiden etukäteistietojen lisääminen psimCCA-menetelmään voisi parantaa sen toimivuutta vielä entisestään. Biologisia metaboliareittejä voisi hyödyntää etukäteistietona vertailemalla vain niitä geenejä, jotka ovat samalla reitillä. Toinen mahdollinen hyödynnettävä olisi geenisäätelyverkot. Näiden avulla voitaisiin tutkia riippuvuuksia geenin ilmentymän ja sen säätelyjaksojen välillä. MikroRNA:n ja kopiolukumittauksien riippuvuusmallituksessa voitaisiin hyödyntää muutoksen suuntaa antamalla muutosmatriisille sopiva etukäteistieto, joka ohjaisi riippuvuusmallitusta samansuuntaisten muutosten löytämiseen.

Korrelaatioanalyysiperusteista menetelmää voisi parantaa pint:ssä lisäämällä p-arvon määrityksen geenien muutoksille esimerkiksi permutoimalla toista ikkunaa. Tällä tavalla voitaisiin varmistua siitä, että menetelmän löytämä riippuvuus liittyy nimenomaan kytseisen alueen mittauksiin. Samalla voitaisiin myös havaita ylisovitus aineistolla, jolle ei voida laskea AUC:ta. Pakettiin voitaisiin myös lisätä menetelmä adaptiiviseen ikkunanmääritykseen. Ikkuna voitaisiin määrittää jokaiselle geenille erikseen käyttämällä esimerkiksi Bayesilaista informaatiokriteeriä ikkunakoon määrittämisessä. Menetelmää voitaisiin laajentaa koskemaan myös useampaa kuin kahta eri aineistoa. Muissa kuin psimCCA:ssa se onnistuu suoraviivaisesti ja psimCCA:ssa pitäisi määritellä kerroinmatriisien suhde uudella tavalla.

Tässä työssä on osoitettu, että korrelaatioanalyysiperusteinen riippuvuushaku toimii paremmin kuin muut vertaillut mentelmät. Riippuvuushaun todettiin parantuvan käyttämällä etukäteistietoa kopioluku- ja ilmentymämittausten paikkariippuvuudesta. Tässä tutkittiin vain yksinkertaisen etukäteistiedon hyväksikäyttämistä riippuvuushaussa. Menetelmää on kuitenki mahdollista laajentamaan huomattavasti tutkittavien aineistojen, kätettävien etukäteistietojen ja genomin seulontaan liittyvien ominaisuuksien osalta. Korrelaatiopohjaisen riippuvuushakua odotetaan käytettävän jatkossa genomiaineistojen tutkimuksessa. Tässä työssä tehtyä ohjelmistopakettia onkin ladattu toukokuun puoliväliin mennessä 115:sta eri IP-osoitteesta.

Viitteet

[Allison06] David B. Allison, Xiangqin Cui, Grier P. Page ja Mahyar Sabripour. *Microarray data analysis: From disarray to consolidation and consensus. Nature Reviews Genetics*, 7: 55–65, May 2006

[Archambeau06] Cédric Archambeau, Nicolas Delannay ja Michel Verleysen. *Robust probabilistic projections.* Teoksessa W.W. Cohen ja A. Moore, toimittajat, *Proceedings of the 23rd International conference on machine learning*, sivut 33–40. ACM, 2006

[Bach05] Francis R. Bach ja Michael I. Jordan. *A probabilistic interpretation of canonical correlation analysis.* Tekninen Raportti 688, Department of Statistics, University of California, Berkeley, 2005

[Berger06] John A. Berger, Sampsa Hautaniemi, Sanjit K. Mitra ja Jaakko Astola. *Jointly analyzing gene expression and copy number data in breast cancer using data reduction models. IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3: 2–16, 2006

[Bicciato09] Silvio Bicciato, Roberta Spinelli, Mattia Zampieri, Eleonora Mangano, Francesco Ferrari, Luca Beltrame, Ingrid Cifola, Clelia Peano, Aldo Solari ja Cristina Battaglia. *A computational procedure to identify significant overlap of differentially expressed and genomic imbalanced regions in cancer datasets. Nucl Acids Res*, 37(15): 5057–5070, 2009

[Calin04] George Adrian Calin, Cinzia Sevignani, Calin Dan Dumitru, Terry Hyslop, Evan Noch, Sai Yendamuri, Masayoshi Shimizu, Sashi Rattan, Florencia Bullrich, Massimo Negrini ja Carlo M. Croce. *Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. Proceedings of the National Academy of Sciences of the United States of America*, 101(9): 2999–3004, 2004

[Chen05] Caifu Chen, Dana A. Ridzon, Adam J. Broomer, Zhaohui Zhou, Danny H. Lee, Julie T. Nguyen, Maura Barbisin, Nan Lan Xu, Vikram R. Mahuvakar, Mark R. Andersen, Kai Qin Lao, Kenneth J. Livak ja Karl J. Guegler. *Real-time quantification*

of microRNAs by stem-loop RT-PCR. Nucl Acids Res, 33(20): e179–, 2005

[Dempster77] Arthur P. Dempster, Nan M. Laird ja Donald B. Rdin. *Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the royal statistical society, series B*, 39(1): 1–38, 1977

[Dutilleul99] Pierre Dutilleul. *The mle algorithms for the matrix normal distribution. Journal of Statistical Computation and Simulation*, 64: 105–123, September 1999

[Forozan97] Farahnaz Forozan, Ritva Karhu, Juha Kononen, Anne Kallioniemi ja Olli-Pekka Kallioniemi. *Genome screening by comparative genomic hybridization. Trends in Genetics*, 13(10): 405 – 409, 1997

[Gentleman08] Robert Gentleman, A.J. Rossini, Sandrine Dudoit ja Aboyoun Patrick. *The bioconductor FAQ*. `http://www.bioconductor.org/docs/faq`, June 2008

[Hornik10] Kurt Hornik. *The R FAQ*, 2010

[Hotelling33] Harold Hotelling. *Analysis of a complex of statistical variables into principal components. The Journal of Educational Psychology*, 24(6): 417–441, September 1933

[Hubbell02] Earl Hubbell, Wei-Min Liu ja Rui Mei. *Robust estimators for expression analysis bioinformatics. Bioinformatics*, 18: 1585–1592, 2002

[Härdle03] Wolfgang Härdle ja Léopold Simar. *Applied multivariate statistical analysis*. Springer, toinen painos, 2003

[Lahti09] Leo Lahti, Samuel Myllykangs, Sakari Knuutila ja Samuel Kaski. *Dependency detection with similarity constraints*. Teoksessa *In Proceedings IEEE International Workshop on machine learning for signal processing*. IEEE Signal Processing Society, 2009

[Lengauer98] Christopher Lengauer, Kenneth W. Kinzler ja Bert Vogelstein. *Genetic instabilities in human cancers. Nature*, 396(6712): 643–649, 1998

[Lipson04] Doron Lipson, Amir Ben-Dor, Elinor Dehan ja Zohar Yakhini. *Joint analysis of dna copy numbers and gene expression levels*. Teoksessa Inge Jonassen ja Junhyong Kim, toimittajat, *WABI*, *Lecture Notes in Computer Science*, osa 3240, sivut 135–146. Springer, 2004

[Liu06] Fenghua Liu, Peter J. Park, Weil Lai, Elizabeth Maher, Arnab Chakravarti, Laura Durso, Xiuli Jiang, Yi Yu, Amanda Brosius, Meredith Thomas, Lynda Chin, Cameron Brennan, Ronald A. DePinho, Isaac Kohane, Rona S. Carroll, Peter M. Black ja Mark D. Johnson. *A Genome-Wide Screen Reveals Functional Gene Clusters in the Cancer Genome and Identifies EphA2 as a Mitogen in Glioblastoma. Cancer Research*, 66(22): 10815–10823, 2006

[Lockhart00] David J. Lockhart ja Elizabeth A. Winzeler. *Genomics, gene expression and dna arrays. Nature*, 405: 827–836, 2000

[Lu05] Jun Lu, Gad Getz, Eric A. Miska, Ezequiel Alvarez-Saavedra, Justin Lamb, David Peck, Alejandro Sweet-Cordero, Benjamin L. Ebert, Raymond H. Mak, Adolfo A. Ferrando, James R. Downing, Tyler Jacks, H. Robert Horvitz ja Todd R. Golub. *MicroRNA expression profiles classify human cancers. Nature*, 435: 834–838, 2005

[Miklos04] George L Gabor Miklos ja Ryszard Maleszka. *Microarray reality checks in the context of a complex disease. Nature Biotechnology*, 22: 615–521, 2004

[Myllykangas08] Samuel Myllykangas, Siina Junnila, Arto Kokkola, Reija Autio, Ilari Scheinin, Tuula Kiviluoto, Marja-Liisa Karjalainen-Lindsberg, Jaakko Hollmen, Sakari Knuutila, Pauli Puolakkainen ja Outi Monni. *Integrated gene copy number and expression microarray analysis of gastric cancer highlights potential target genes. International Journal of Cancer*, 123: 817–825, August 2008

[Nymark06] Penny E. H. Nymark, Harriet Wikman, Salla T. Ruosaari, Jaakko Hollmén, Esa Vanhala, Antti Karjalainen, Sisko L. Anttila ja Sakari Knuutila. *Identification of specific gene copy number changes in*

asbestos-related lung cancer. *Cancer Research*, 66: 5737–5743, June 2006

[Pollack02] Jonathan R. Pollack, Therese Sørlie, Charles M. Perou, Christian A. Rees, Stefanie S. Jeffrey, Per E. Lonning, Robert Tibshirani, David Botstein, Anne-Lise Børresen-Dale ja Patrick O. Brown. *Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors*. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20): 12963–12968, 2002

[R Development Core Team09] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009

[Rubin82] Donald B. Rubin ja Dorothy T. Thayer. *Em algorithms for ml factor analysis*. *Psychometrika*, 47: 69–76, March 1982

[Schäfer09] Martin Schäfer, Holger Schwender, Sylvia Merk, Claudia Haferlach, Katja Ickstadt ja Martin Dugas. *Integrated analysis of copy number alterations and gene expression: a bivariate assessment of equally directed abnormalities*. *Bioinformatics*, 25(24): 3228–3235, October 2009

[Shen09] Ronglai Shen, Adam B. Olshen ja Marc Ladanyi. *Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis*. *Bioinformatics*, 25(22): 2906–2912, September 2009

[Shinawi08] Marwan Shinawi ja Sau Wai Cheung. *The array cgh and its clinical applications*. *Drug Discovery Today*, 13(17-18): 760 – 770, 2008

[Shingara05] Jaclyn Shingara, Kerri Keiger, Jeffrey Shelton, Walairat Laosinchai-Wolf, Patricia Powers, Richard Conrad, David Brown ja Emmanuel Labourier. *An optimized isolation and labeling platform for accurate microRNA expression profiling*. *RNA*, 11(9): 1461–1470, 2005

[Stoughton05] Roland B. Stoughton. *Applications of dna microarrays in biology*. *Annual Review of Biochemistry*, 74(1): 53–82, 2005

[Stranger07] Barbara E. Stranger, Matthew S. Forrest, Mark Dunning, Catherine E. Ingle, Claude Beazley, Natalie Thorne, Richard Redon, Christine P. Bird, Anna de Grassi, Charles Lee, Chris Tyler-Smith, Nigel Carter, Stephen W. Scherer, Simon Tavare, Panagiotis Deloukas, Matthew E. Hurles ja Emmanouil T. Dermitzakis. *Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. Science*, 315(5813): 848–853, 2007

[Tipping99] Michael E. Tipping ja Christopher M. Bishop. *Probabilistic principal component analysis. Journal of the Royal Statistical Society, Series B*, 61(3): 611–622, 1999

[Vinod76] H. D. Vinod. *Canonical ridge and econometrics of joint production. Journal of Econometrics*, 4(2): 147 – 166, 1976

[Vogelstein04] Bert Vogelstein ja Kenneth W Kinzler. *Cancer genes and the pathways they control. Nature Medicine*, 10: 789–799, 2004

[Wieringen09] Wessel N. van Wieringen ja Mark A. van de Wiel. *Nonparametric testing for dna copy number induced differential mrna gene expression. Biometrics*, 65: 19–29, 2009

[Wikman07] Harriet Wikman, Salla T. Ruosaari, Penny E. H. Nymark, VK Sarhadi, J Saharinen, Esa Vanhala, Antti Karjalainen, Jaakko Hollmén, Sakari Knuutila ja Sisko L. Anttila. *Gene expression and copy number profiling suggests the importance of allelic imbalance in 19p in asbestos-associated lung cancer. Oncogene*, 26: 4730–4707, February 2007

Suurimman uskottavuuden estimaatit riippuvuusmalleille

Suurimman uskottavuuden estimaatit todennäköisyysperustaiselle kanoniselle korrelaatioanalyysille

Mallin (11) parametreille voidaan määrittää seuraavat suurimman uskottavuuden estimaatit

$$\begin{aligned}
\widehat{\mathbf{W}}_X &= \tilde{\Sigma}_{XX}\mathbf{U}_{Xd}\mathbf{M}_X \\
\widehat{\mathbf{W}}_Y &= \tilde{\Sigma}_{YY}\mathbf{U}_{Yd}\mathbf{M}_Y \\
\widehat{\mathbf{\Psi}}_X &= \tilde{\Sigma}_{XX} - \widehat{\mathbf{W}}_X\widehat{\mathbf{W}}_X^T \\
\widehat{\mathbf{\Psi}}_Y &= \tilde{\Sigma}_{YY} - \widehat{\mathbf{W}}_Y\widehat{\mathbf{W}}_Y^T \\
\hat{\boldsymbol{\mu}}_X &= \tilde{\boldsymbol{\mu}}_X \\
\hat{\boldsymbol{\mu}}_Y &= \tilde{\boldsymbol{\mu}}_Y,
\end{aligned} \tag{A1}$$

missä $\tilde{\Sigma}$ on näytekovarianssimatriisi ja $\tilde{\boldsymbol{\mu}}$ näytekeskiarvo. Matriisit $\mathbf{M}_X, \mathbf{M}_Y \in \mathbb{R}^{d\times d}$ ovat mielivaltaisia matriiseja site, että $\mathbf{M}_X\mathbf{M}_Y^T = \mathbf{P}_d$, ja matriisien $\mathbf{M}_X$ ja $\mathbf{M}_Y$ spektraalinormit ovat pienempiä kuin yksi. Matriisit $\mathbf{U}_{Xd}$ ja $\mathbf{U}_{Yd}$ ovat d ensimmäistä kanonista suuntaa ja matriisi $\mathbf{P}_d$ on diagonaalimatriisi, jonka diagonaalilla on d ensimmäistä kanonista korrelaatiota. [Bach05]

Ehdollisiksi odotusarvoiksi ja variansseiksi saadaan

$$\begin{aligned}
\mathrm{E}(\mathbf{z}|\mathbf{x}) &= \mathbf{M}_X^T\mathbf{U}_{Xd}^T(\mathbf{x}-\boldsymbol{\mu}_X) \\
\mathrm{E}(\mathbf{z}|\mathbf{y}) &= \mathbf{M}_Y^T\mathbf{U}_{Yd}^T(\mathbf{y}-\boldsymbol{\mu}_Y) \\
\mathrm{var}(\mathbf{z}|\mathbf{x} &= \mathbf{I}-\mathbf{M}_X\mathbf{M}_X^T) \\
\mathrm{var}(\mathbf{z}|\mathbf{y} &= \mathbf{I}-\mathbf{M}_Y\mathbf{M}_Y^T) \\
\mathrm{E}(\mathbf{z}|\mathbf{x},\mathbf{y}) &= \begin{pmatrix}\mathbf{M}_X\\ \mathbf{M}_Y\end{pmatrix}^T \begin{pmatrix}(\mathbf{I}-\mathbf{P}_d^2)^{-1} & (\mathbf{I}-\mathbf{P}_d^2)^{-1}\mathbf{P}_d \\ (\mathbf{I}-\mathbf{P}_d^2)^{-1}\mathbf{P}_d & (\mathbf{I}-\mathbf{P}_d^2)^{-1}\end{pmatrix} \begin{pmatrix}\mathbf{U}_{Xd}^T(\mathbf{x}-\boldsymbol{\mu}_X)\\ \mathbf{U}_{Yd}^T(\mathbf{y}-\boldsymbol{\mu}_Y)\end{pmatrix} \\
\mathrm{var}(\mathbf{z}|\mathbf{x},\mathbf{y}) &= \mathbf{I} - \begin{pmatrix}\mathbf{M}_X\\ \mathbf{M}_Y\end{pmatrix}^T \begin{pmatrix}(\mathbf{I}-\mathbf{P}_d^2)^{-1} & (\mathbf{I}-\mathbf{P}_d^2)^{-1}\mathbf{P}_d \\ (\mathbf{I}-\mathbf{P}_d^2)^{-1}\mathbf{P}_d & (\mathbf{I}-\mathbf{P}_d^2)^{-1}\end{pmatrix} \begin{pmatrix}\mathbf{M}_X\\ \mathbf{M}_Y\end{pmatrix}.
\end{aligned} \tag{A2}$$

Riippumatta siitä, mitä $\mathbf{M}_X$ ja $\mathbf{M}_Y$ ovat, d-ulotteiset aliavaruudet $\mathbb{R}^p$ ja $\mathbb{R}^Q$, mihin $\mathbf{x}$ ja $\mathbf{y}$ ovat projisoitu laskettaessa posteriooriodotusarvoja $\mathrm{E}(\mathbf{z}|\mathbf{x})$ ja $\mathrm{E}(\mathbf{z}|\mathbf{y})$, ovat samat kuin kanonisessa korrelaatioanalyysissä. Mallin (11) ratkaisuille, jotka minimoivat $-\log|\mathbf{\Psi}| = -\log|\mathbf{\Psi}_X| - \log|\mathbf{\Psi}_Y|$, saadaan sisältävät ominaisuuden Mallin (11) ratkaisuilla, jotka minimoivat $-\log|\mathbf{\Psi}| = -\log|\mathbf{\Psi}_X| - \log|\mathbf{\Psi}_Y|$, ovat niitä, joissa

$$\mathbf{M}_X = \mathbf{M}_Y = \mathbf{M} = \mathbf{P}_d^{1/2}\mathbf{R}, \tag{A3}$$

missä $\mathbf{R}$ on d kokoinen rotaatiomatriisi. Ratkaisut ovat tällöin

$$\widehat{\mathbf{W}}_X = \tilde{\Sigma}_{XX}\mathbf{U}_{Xd}\mathbf{P}_d^{1/2}\mathbf{R} \tag{A4}$$

$$\widehat{\mathbf{W}}_Y = \tilde{\Sigma}_{YY}\mathbf{U}_{Yd}\mathbf{P}_d^{1/2}\mathbf{R}.\text{[Bach05]} \tag{A5}$$

Suurimman uskottavuuden estimaatit todennäköisyysperustaiselle pääkomponenttianalyysille

pPCA:n mallin parametrien suurimman uskottavuuden estimaatit saadaan laskettua analyyttisesti. Kerroinmatriisin estimaatti on

$$\widehat{\mathbf{W}} = \mathbf{U}_d(\mathbf{\Lambda}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}, \tag{A6}$$

missä d sarakevektoria $p \times d$ matriisista $\mathbf{U}_d$ ovat pääkomponentit näytekovarianssimatriisista $\tilde{\Sigma}$ vastaavilla ominaisarvoilla $\lambda_1, \ldots, \lambda_q$ $q \times q$ diagonaalimatriisissa $\mathbf{\Lambda}_q$. $\mathbf{R}$ on mielivaltainen $q \times q$ ortogonaalinen rotaatiomatriisi. Kun $\mathbf{W} = \widehat{\mathbf{W}}$, suurimman uskottavuuden estimaattoriksi σ^2:lle saadaan

$$\widehat{\sigma}^2 = \frac{1}{p-d}\sum_{j=d+1}^{p}\lambda_j. \tag{A7}$$

Suurimman uskottavuuden estimaatit todennäköisyysperustaiselle faktorianalyysille

pFA:n mallin parametrien lasku tehdään EM-algoritmillä. E-askeleella etsitään odotusarvo täyden datan uskottavuuden logaritmille annetuilla havaituilla arvoilla $\mathbf{X}$ ja sen hetkisillä estimoiduilla parametreilla

$$\ell = -\frac{p}{2}\sum_{j=1}^{p}\log\epsilon_j^2 - \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{p}\frac{(X_{ij} - W_j z_i)^2}{\epsilon_j^2} - \frac{1}{2}\sum_{i=1}^{k} z_i z_i^t, \tag{A8}$$

missä X_{ij} on muuttujan i näytteen j arvo, z_i on piilomuuttujan i. sarakevektori ja W_j on kerroinmatriisin j. rivivektori. Täyden datan uskottavuuden laskemiseksi annetuilla mallin (13) parametreilla $\boldsymbol{\epsilon}^2$ ja $\mathbf{W}$ täytyy laskea odotusarvo ehdollisille näytekovariansseille. Muuttujien

$$\begin{aligned}\delta_t &= \mathbf{W}_t(\boldsymbol{\epsilon}^2 + \mathbf{W}_t\mathbf{W}_t^T)^{-1} \\ \Delta_t &= \mathbf{I} - \mathbf{W}_t^T(\boldsymbol{\epsilon} + \mathbf{W}_t\mathbf{W}_t^T)^{-1}\mathbf{W}_t\end{aligned} \tag{A9}$$

avulla saadaan laskettua mallin parametreilla ehdolliset odotusarvot näytekovarianssimatriiseille

$$\mathrm{E}(\tilde{\Sigma}_{XX}|\mathbf{X}, \epsilon_t^2, \mathbf{W}) = \tilde{\Sigma}_{XX}$$

$$\mathrm{E}(\tilde{\Sigma}_{XZ}|\mathbf{X}, \epsilon_t^2, \mathbf{W}) = \tilde{\Sigma}_{XX}\delta_t \tag{A10}$$

$$\mathrm{E}(\tilde{\Sigma}_{ZZ}|\mathbf{X}, \epsilon_t^2, \mathbf{W}) = \delta_t^T\tilde{\Sigma}_{XX}\delta_t + \Delta_t. \tag{A11}$$

Kaavat (A9) ja (A10) määrittävät odotusarvon maksimoinnin E-askeleen. [Rubin82]

M-askeleella maksimoidaan uskottavuuden logaritmin odotusarvoa olettaen, että se perustuu täydelle datalle. Tällä askeleella estimoidaan mallille uudet parametrit. Uudet mallin parametrit saadaan laskettua

$$\mathbf{W}_{t+1} = \tilde{\Sigma}\delta_t(\delta_t\tilde{\Sigma}^T\delta_t^T + \Delta_t^T)^{-1} \tag{A12}$$

$$\boldsymbol{\epsilon}_{t+1} = \operatorname{diag}\left\{\tilde{\Sigma} - \tilde{\Sigma}\delta_t(\delta_t^T\tilde{\Sigma}\delta_t + \Delta_t)^{-1}(\tilde{\Sigma}\delta_t)^T\right\}. \tag{A13}$$

Kaavat (A12) ja (A13) ovat odotusarvon maksimoinnin M-askel. Estimaatit mallin parametreille saadaan iteroimalla E- ja M-askelien välillä, kunnes haluttu tarkkuus parametreille on saavutettu. [Rubin82]

Suurimman uskottavuuden estimaatit todennäköisyysperustaiselle rajoitetulle kanoniselle korrelaatioanalyysille

psimCCA -mallin paramterit saadaan laskettua pCCA:n odotusarvon maksimointi -algoritmilla. pCCA:n parametrien odotusarvon maksimoinnin M-askel on

$$\mathbf{W}_{t+1} = \tilde{\Sigma}\boldsymbol{\Psi}_t^{-1}\mathbf{W}_t\mathbf{M}_t(\mathbf{M}_t + \mathbf{M}_t\mathbf{W}_t^T\boldsymbol{\Psi}_t^{-1}\tilde{\Sigma}\boldsymbol{\Psi}_t^{-1}\mathbf{W}_t\mathbf{M}_t)^{-1} \tag{A14}$$

$$\boldsymbol{\Psi}_{t+1} = \begin{pmatrix}(\tilde{\Sigma} - \tilde{\Sigma}\boldsymbol{\Psi}_t^{-1}\mathbf{W}_t\mathbf{M}_t\mathbf{W}_{t+1}^T) & 0 \\ 0 & (\tilde{\Sigma} - \tilde{\Sigma}\boldsymbol{\Psi}_t^{-1}\mathbf{W}_t\mathbf{M}_t\mathbf{W}_{t+1}^T)\end{pmatrix} \tag{A15}$$

$$\mathbf{M}_{t+1} = \mathbf{I} + \mathbf{W}_t^T\boldsymbol{\Psi}_t^{-1}\mathbf{W}_t)^{-1}. \tag{A16}$$

psimCCA:n erikoistapaus, jossa vapausparametri $\sigma_T^2 = 0$, saadaan suoraan laskettua kaavoilla (A14, A15, A16) asettamalla $\mathbf{W} = (\mathbf{W}_X, \mathbf{T}\mathbf{W}_X)^T$. Yleisessä tapauksessa jokaisella parametrien $\mathbf{W}$ ja $\boldsymbol{\Psi}$ iteaariolla maksimoidaan mallin logaritminen todennäköisyys (21) optimoimalla $\mathbf{T}$. [Bach05], [Archambeau06]