ICS-C3000 Harjoitustyö

Tapio Friberg, 296885
Sampsa Laapotti, 293545
Hannu Huhtanen, 291288

Valitsimme harjoitustyömme aiheeksi aiheen 1. eli “Syöpäsolujen erottelu geenien ekspressioaineiston avulla”. Aloitimme aiheen tutkimisen lukemalla tehtävänantoon linketettyjä tutkimuspapereita ja testailemalla Matlabilla annetun aineiston visualisointia.

Päätimme, että järkevin tapa toimia olisi ensin valita Golub et al [1] paperin perusteella tutkittavaksi N eniten syöpätyypin kanssa korreloivaa geeniä, joilla voidaan ajaa PCA-muunnos opetusaineistolle, jonka jälkeen analysoida tämän tuloksia. Lopulta voidaan arvioida, miten opetusaineistolle ajettu muunnos poikkeaa oikealle datalle ajetusta. Jos aineistojen analyysit poikkeat toisistaan merkittävästi, ei voida sanoa, että tämä metodi olisi hyvä syöpätyypin määrittämiseen.

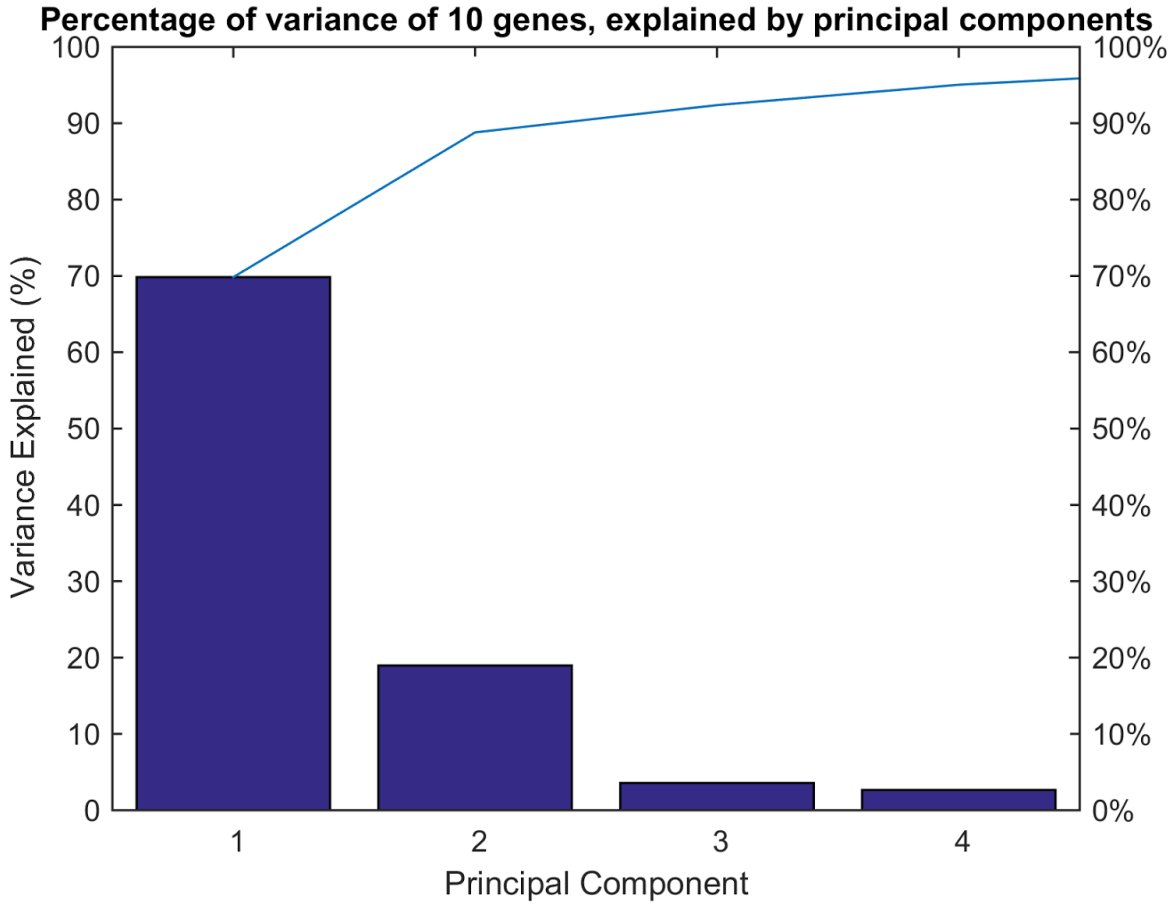
Aineistona käytimme kahta eri datasettiä. Ensimmäinen, opetusaineisto koostui 38:n ihmisen geenien ekspressiodatasta(7129 geeniä) sekä tiedon jo tunnetusta syöpätyypistä. Testiaineisto koostui samojen geenien ekspressiodatasta eri 34:ltä henkilöltä ilman tietoa syöpätyypistä.

Löyhästi artikkeliin perustuen valitsimme riittävän ‘vahvan’ korrelaation rajaksi $P(g,c)>0.3$. Opetusaineistosta luotiin luokitteluvektori c, ja datamatriisin ja luokitteluvektorin sekä luokitteluvektorin komplementin korrelaatiota käytettiin ilmaisemaan datan varianssia muuttujina toimivien geenien määrän N suhteen. Rutiini haki N/2 eniten luokitteluvektorin kanssa korreloivaa geeniä kuvaamaan ALL-tyypin leukemian varianssia, ja N/2 eniten luokitteluvektorin komplementin kanssa korreloivaa geeniä kuvaamaan AML-tyypin leukemian varianssia. Valitsimme näin vaihtelevan määrän N geenejä (N [10-400]) ja tutkimme niitä. Opetusaineistoa tutkiessamme huomasimme, että useat sadat(~900) geenit korreloivat ($P(g,c)>0.3$) syöpätyypin kanssa, eikä yli 0.3:n korrelaatiokertoimen omaavista geeneistä yksikään korreloinut rajan ylittävällä tavalla molempien syöpätyyppien kanssa, joten pidimme korrelaatiokerrointa hyvänä lähtökohtana ilmaisuvoimaisten geenien valitsemiseen.

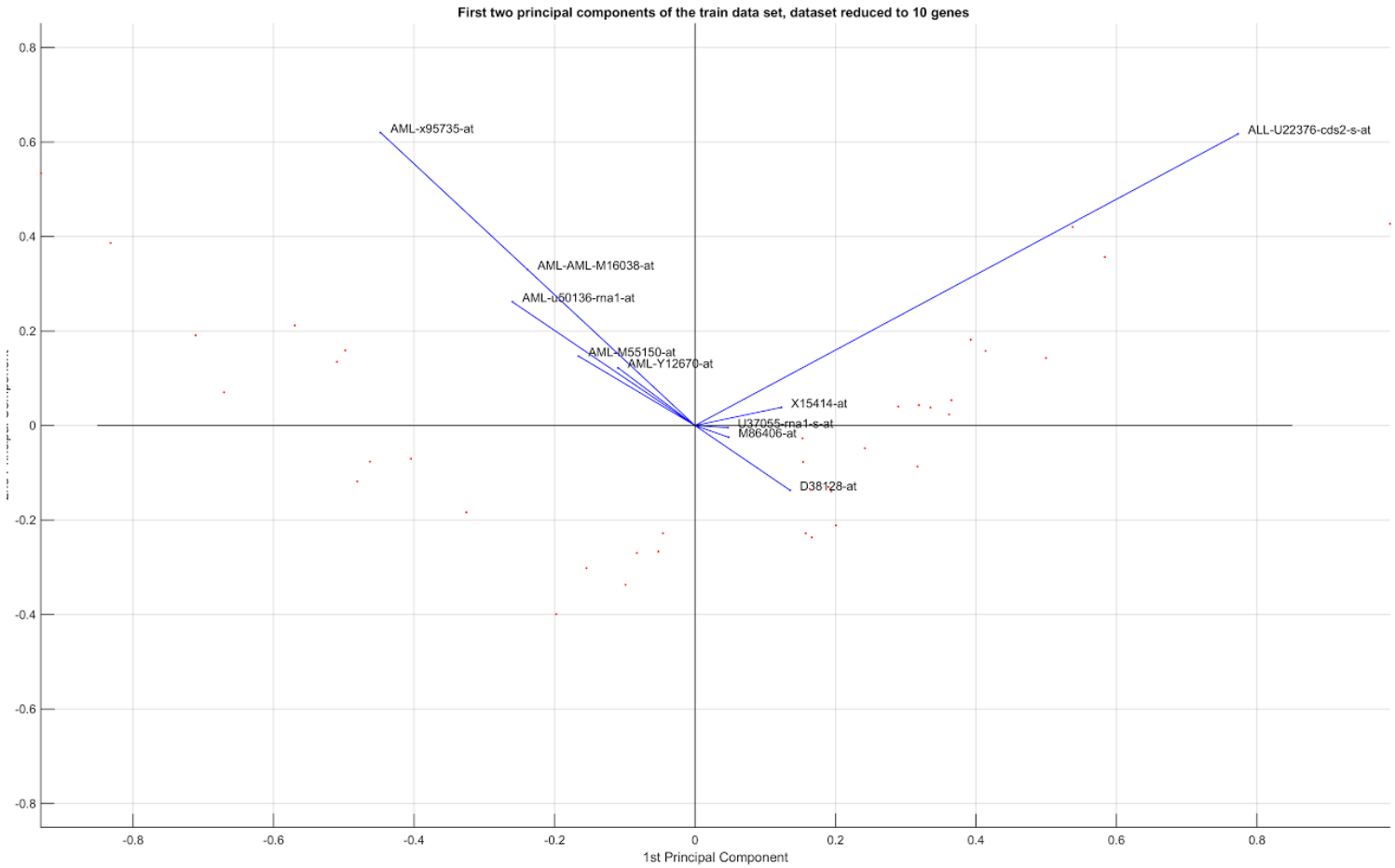
Tämän jälkeen ajoimme aineistolle pääkomponenttianalyysin ja visualisoimme kahden isoimman pääkomponentin tärkeimpien geenien sijainnin ja vertasimme niitä tutkimuksessa nostettuihin geeneihin ja löysimme paljon samoja geenejä. Lisäksi ensimmäisen pääkomponentin suuntaan saimme erottumaan selvän jakautumisen AML- ja ALL-geeneihin. Lisäksi huomasimme, että AML-geenien representaatio oli keskimäärin isompi, kun taas ALL-geenit taas olivat vähemmän representoituna, mutta selvästi havaittavissa.

Kuvissa Golubin valitsemat geenit on merkattu AML- tai ALL-etuliitteillä.

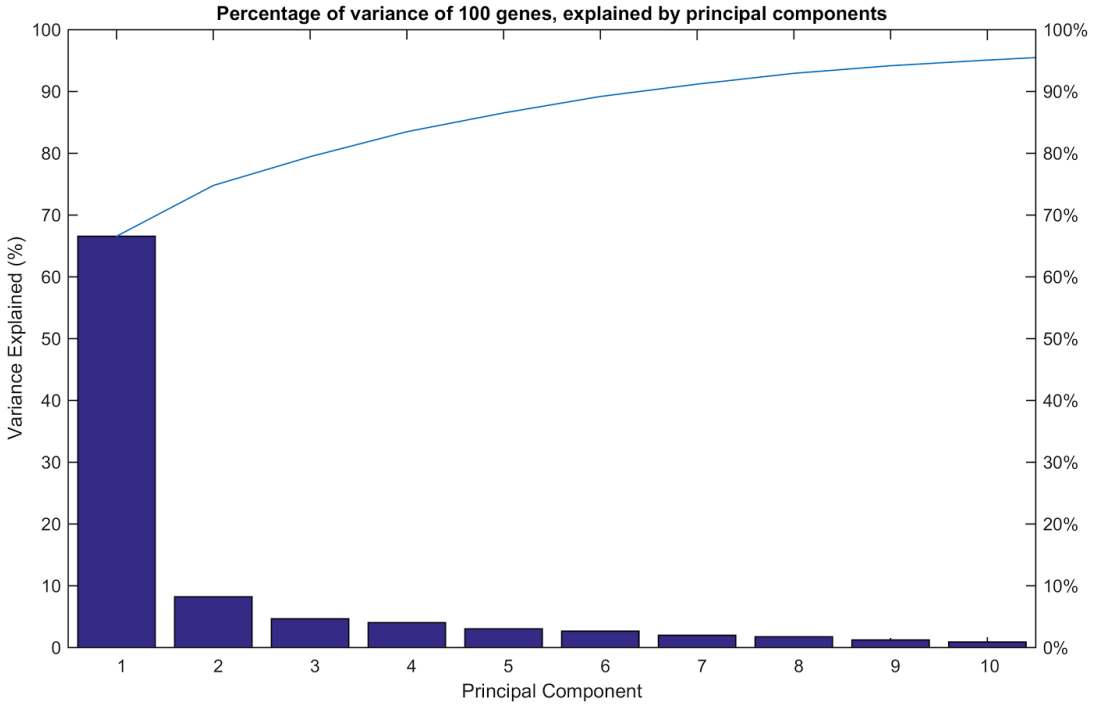
Opetusdatasetti



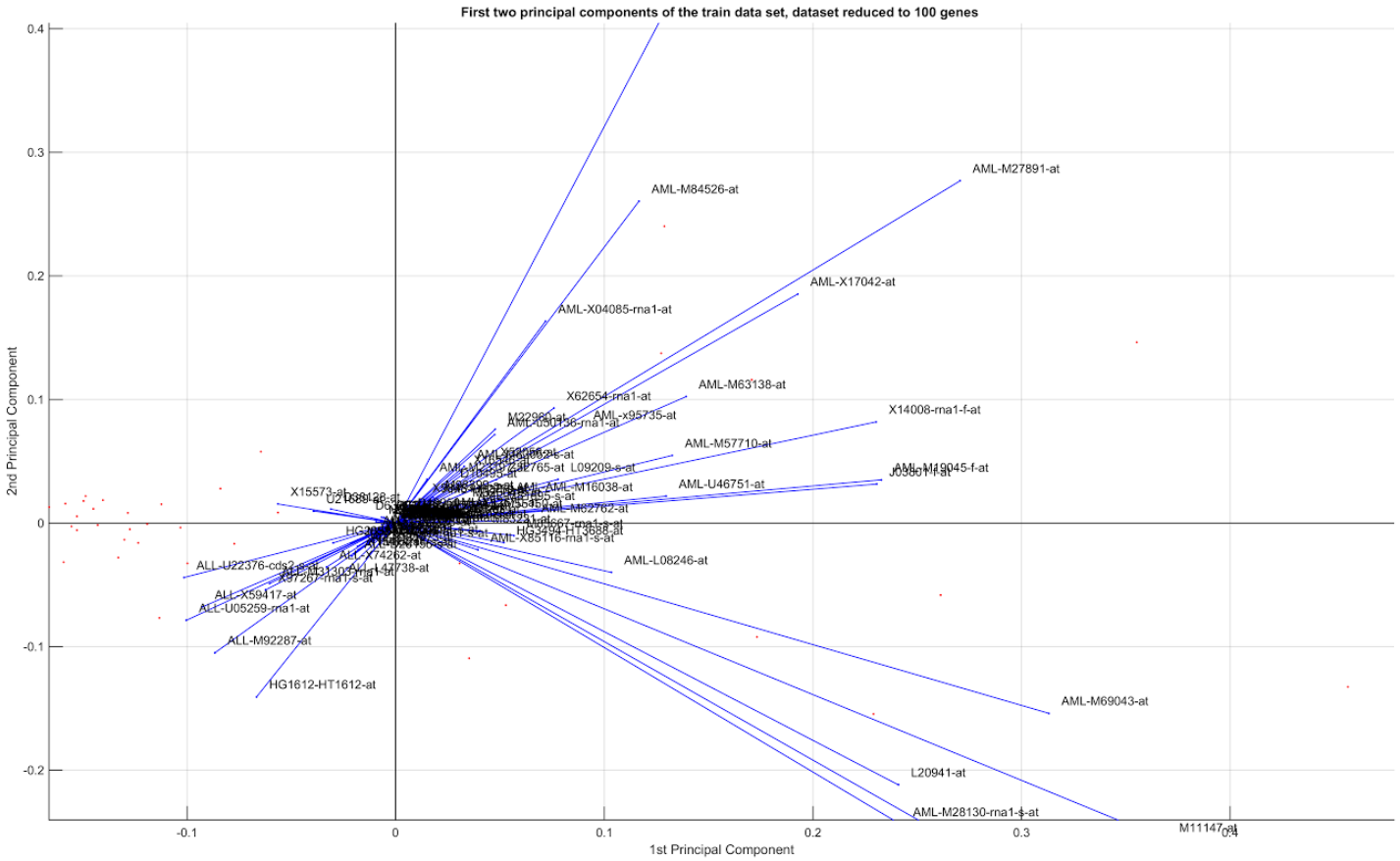
Noin 70% prosenttia 10 valitun geenin varianssista selittyy ensimmäisen pääkomponentin avulla ja kun kaveriksi otetaan toinen pääkomponentti, saadaan selitettyä jo 90% varianssista. Ei sinänsä hirveän ihmeellistä, kun kyseessä on vain kymmenen geeniä, joten periaatteessa maksimissaankin puhutaan tällöin 10:stä pääkomponentista.



Kun tarkastellaan kahta ensimmäistä pääkomponenttia, huomataan, että niiden “ulottuvuudet”, eli käytännössä geenit, joista ne koostuvat, ovat mukavan erillään. Olemme merkinneen tutkimuksessa valitut geenit etuliitteillä, jotta saamme arvioitua omien tulostemme järkevyyttä. Näyttäisi siltä, että jo pelkästään kymmenellä geenillä voidaan arvioida ainakin harjoitusdatan kohdalla aika hyvin, kumpaan syöpätyyppiin potilas kuuluu. Tässä havaitaan myös pientä dimensionaalisuuden kirousta, sillä vaikka ekat pääkomponenti selittävät ~90% varianssista, koostuvat ne silti kymmenestä geenistä. Tämän vuoksi geenien pudottaminen pois korrelaatiolla ennen visualisointia on pakollista.

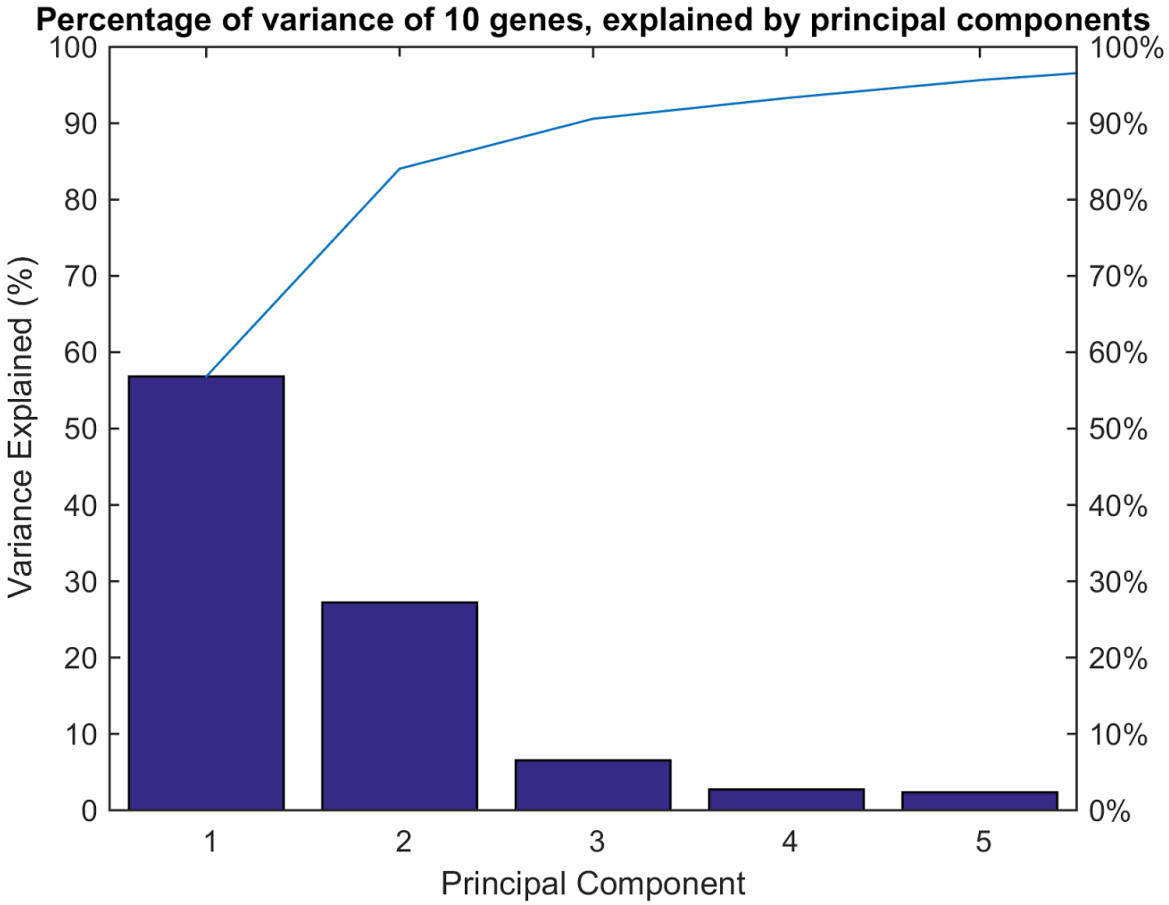


Kun N:ksi valitaan sata, alkaa data jakautua useampaa pääkomponenttiin, mutta vieläkin ensimmäinen pääkomponentti on reilusti vahvin, selittäen noin 65% varianssista. Yhdessä ensimmäinen ja toinen pääkomponentti selittävät noin 75%, joten niitä on yhä järkevää tutkia tarkemmin. Pitää kuitenkin muistaa, että nyt ulottuvuuksia on kertaluokka enemmän, mutta ensimmäisten pääkomponenttien selitysvoima ei ole pudonnut merkittävästi. On siis yhä järkevää tutkia niitä tarkemmin.

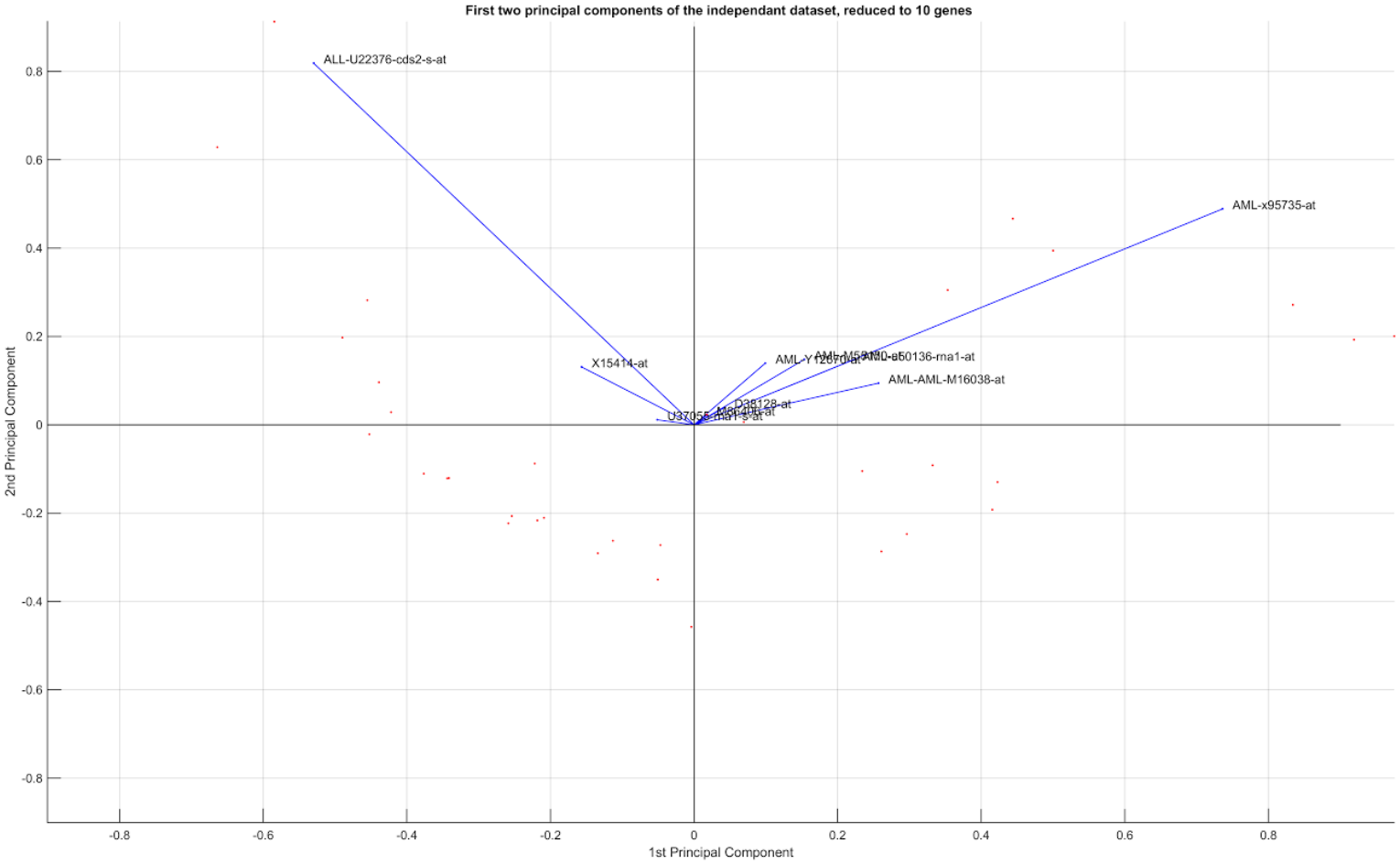


Tutkittaessa sataa eniten korreloivaa geeniä, huomataan, miten AML:ää ja ALL:ää ennustavat geenit kerääntyvät ensimmäisen pääkomponentin eri päihin. Lisäksi origon tuntumaan on ilmestynyt kasa vähemmän vahvasti expressoituvia geenejä. Huomataan myös, että AML:ää ennustavien geenien keskimääräinen expressio tuntuu olevan ALL:ää ennustavia geenejä pienempää. Jännää, mutta on linjassa Golubin havaintojen kanssa. Myös toinen pääkomponentti selittää nyt isomman osan varianssista, tämä näkyy kuvaajassa lähinnä pystyakselin suuntaisena vaihteluna.

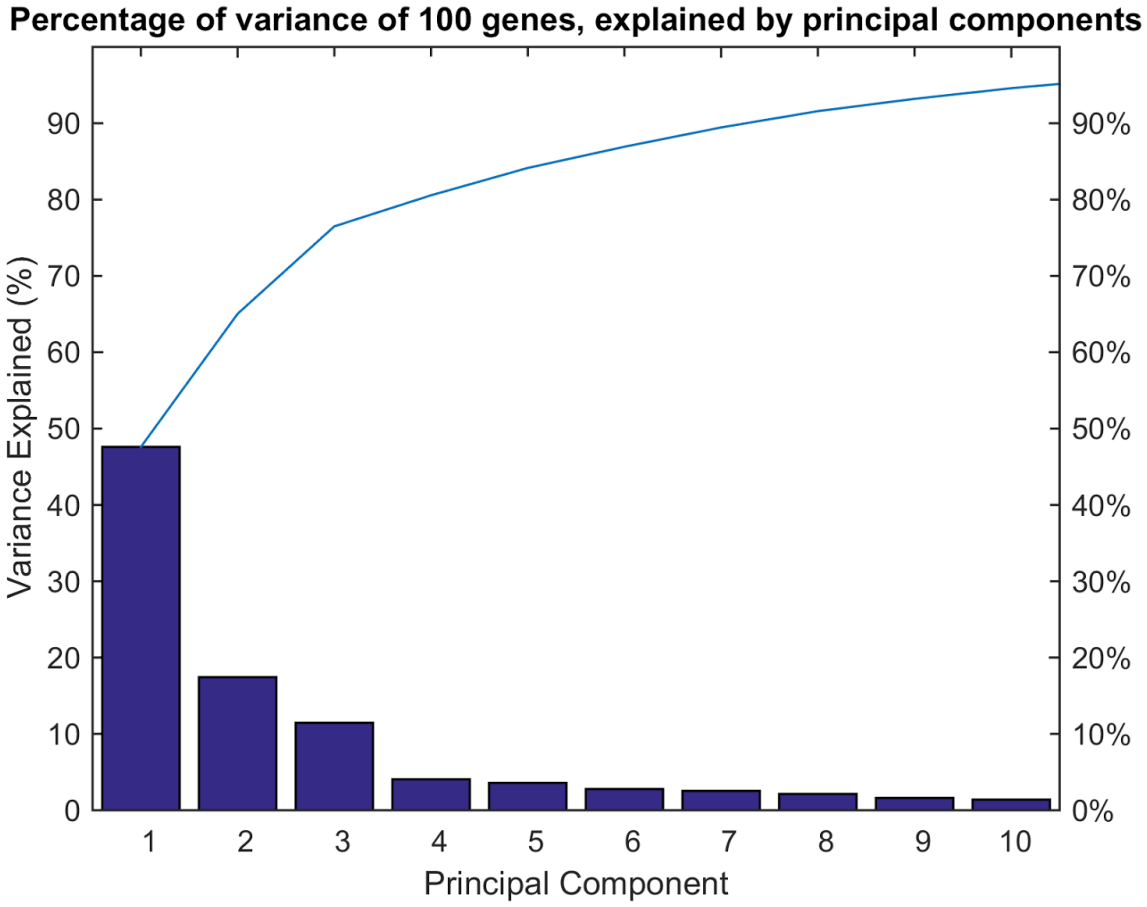
Testidatasetti



Kun siirrytään testidatan kimppuun huomataan, että kakkoskomponentin selitysvoima kasvaa, mutta toisaalta ensimmäinen ja toinen komponentti yhdessä kykenevät selittämään lähes 90% varianssista. Selvästi data on ainakin hiukan erilaista. On kuitenkin mielekästä tutkiskella kahta pääkomponenttia, joten seuraavaksi siihen.

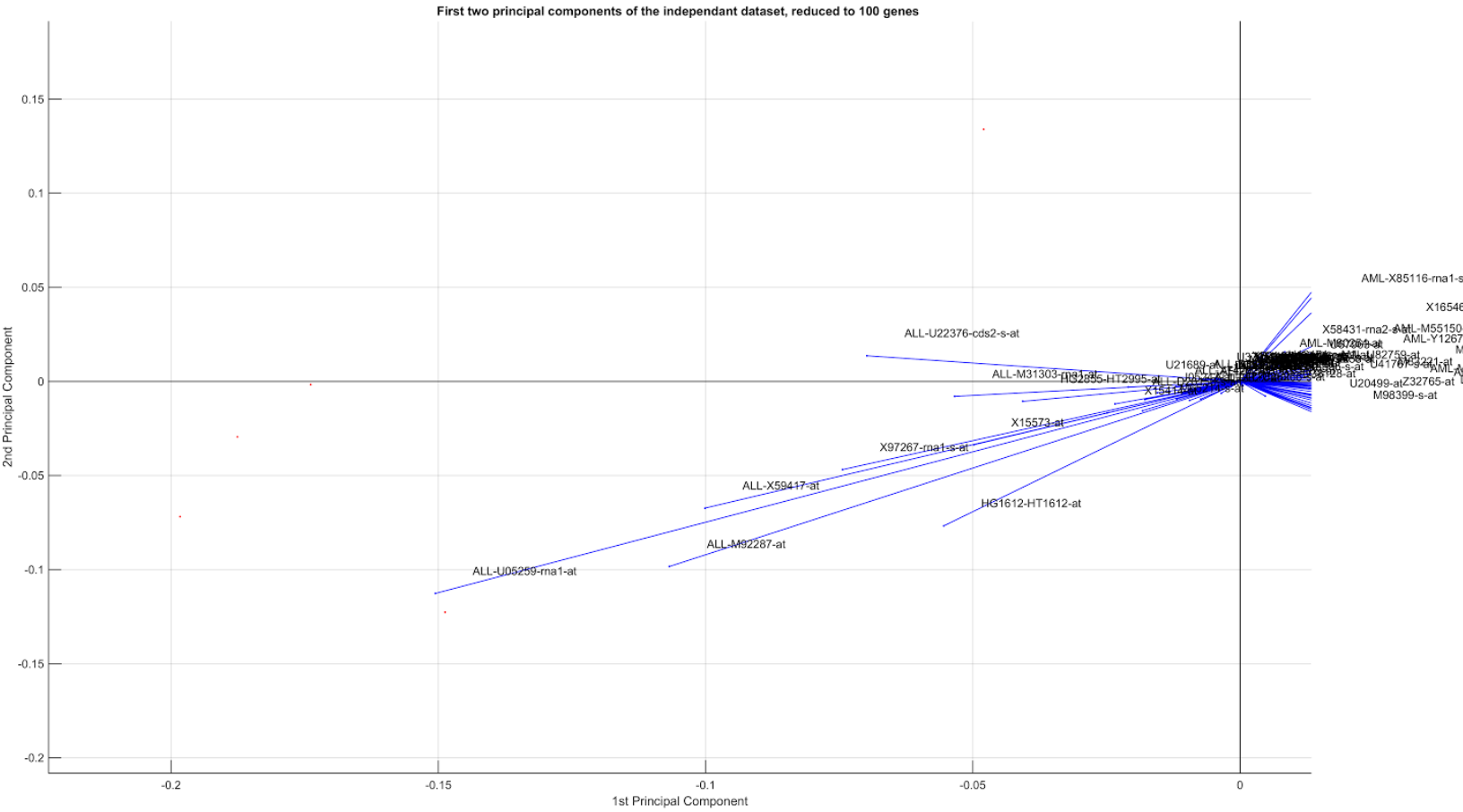


Kun tarkastellaan testidataan kymmenellä valitulla geenillä, huomataan, että havainnot ovat pitkälti samanlaisia kuin harjoitusdatalla. ALL:ää ennustaa yksi vahvasti ekspressoituva geeni ja AML:ää muutama heikommin ekspressoituva, mutta selvästi klusteroituva geeni. Jännästi matlab on plotannut x-akselin eri päin, kuin harjoitusdatalla. Kuitenkin geenit ovat pitkälti samoja. Hyvältä näyttää vielä.

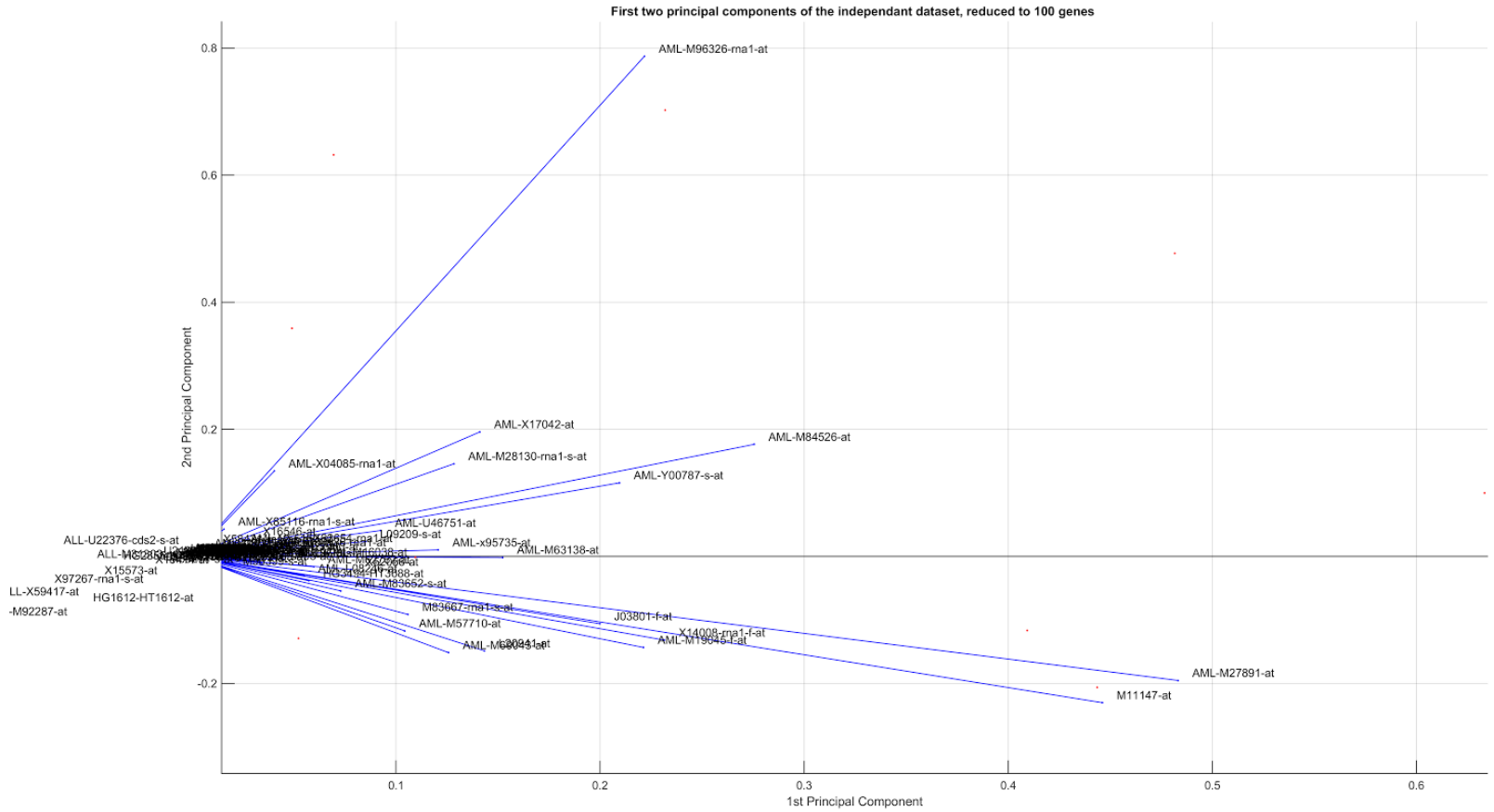


Testidata sadan geenin otoksella näyttää muuttuneen samaan suuntaan kuin kymmenen geenin otoksella, eli ekan pääkomponentin selittävyys on laskenut mutta ensimmäinen ja toinen pääkomponenti selittävät kuitenkin varianssista vielä reilusti yli puolet. Vaikuttaisi siltä, että trendi komponentteja lisätessä on se, että varianssi valuu kohti häntää. Selitysvoimaa on kuitenkin vielä riittävästi, jotta meidän kannattaa tarkastella kahta ensimmäistä pääkomponenttia lähemmin.

Testidatan ALL:ää ennustavien geenien puoli on tässä zoomattuna sisään. Kuten



huomaamme, on ALL:ää ennustavien geenien piirteet vieläkin melko samankaltaisia, muutama geeni ekspressoituu vahvasti ja pääkomponentti koostuu suuresti niistä. Huomaamme myös, että kun N:ksi on valittu 100, on mukana huomattava määrä Golubin valitsemia geenejä.



AML:n puolella pääkomponenttia havaitaan, että osa varianssista on siirtynyt toisen pääkomponentin sisään. Kuitenkin mukana näkyy suuri määrää tuttuja geenejä melko tutuissa asemissa. Kuten harjoitusaineistonkin kohdalla, huomaamme tässäkin, miten N:n koon kasvattamminen lisää geenejä lähinnä origon läheisyyteen, joten nämä geenit eivät sisällä kovin suuria määriä varianssia, ainakaan kahden ensimmäisen pääkomponentin sisällä, ja kun tiedetään, että nämä pääkomponentit selittävät suurimman osan kaikesta varianssista, huomataan että ei oikeastaan tarvita kovinkaan montaa suurta määrää geenejä.

Lopputulos:

Harjoitusaineistosta havaitut piirteet ja geenieksprессiot syöpien erottamiselle tuntuvat yleistyvän melko hyvin myös testiaineistolle. Lisäksi havaittiin, että pelkällä korrelaation määrällä geenien määrän rajoittaminen toimii melko hyvin aineiston karsimiseen. Samoin suuri osa aineiston varianssista selittyy muutaman(10-200 geenieksprессiota) varianssille, joiden seasta erottuu selvästi molemmissa aineistossa suuresti vaikutttavia ekspressioita. Melko nopeasti huomasimme, että kaikkien tuhansien ekspressioiden läpikäynti on melko turhaa ja hankalaa, kun suuri osa selitysvoimasta löytyy todella rajatusta määrästä geenieksprессioita

Liitteet:

Koko projektin koodi löytyy osoitteesta: https://github.com/sampoz/WeCuredCancerYeah

Lähteet:

[1] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science, 1999, vol 286, issue 5439, pages 531-537. http://www.broadinstitute.org/mpr/publications/projects/Leukemia/Golub_et_al_1999.pdf