TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
MATEMAATILISE STATISTIKA INSTITUUT

Kaspar Märtens

Pärilike tunnuste täpne prognoosimine DNA põhjal

Magistritöö matemaatilise statistika erialal (30 EAP)

Juhendaja: PhD Leopold Parts

Tartu 2015

Pärilike tunnuste täpne prognoosimine DNA põhjal

Lühikokkuvõte:

Indiviidi DNA järjestuse põhjal on võimalik ennustada pärilikke tunnuseid, nagu juuksevärv, kehakaal ja erinevad haigusriskid. Käesolevas töös uurime fenotüüpide prognoosimist suure pärmipopulatsiooni näitel. Meie eesmärgiks on välja selgitada, kui täpselt on võimalik fenotüüpe prognoosida ainult genotüübi põhjal ning kuivõrd saab ennustustäpsust suurendada, kui on teada indiviidi teiste fenotüüpide mõõtmistulemused. Modelleerimisel võtame arvesse eksperimendidisainist tulenevaid indiviididevahelisi sõltuvusi. Töös võrdleme lineaarsete segamudelite, mitmemõõtmeliste lineaarsete segamudelite ning juhuslike segametsade prognoosivõimet. Meil õnnestub saavutada ennustustäpsus, mis ületab pärilikkusel põhinevad klassikalised piirid. Lisaks pakume prognoosimiseks välja uue meetodi, mis kombineerib lineaarsete segamudelite ning juhuslike segametsade tugevaid külgi. Näitame simuleeritud andmetel, et selle prognoositäpsus ületab alternatiivsete mudelite oma.

Märksõnad: Lineaarne segamudel, mitmemõõtmeline lineaarne segamudel, juhuslik segamets

Accurate genomic prediction of heritable traits

Abstract:

Knowing the genome sequence of an individual makes it possible to predict heritable phenotypes, such as hair color, body weight, and disease risk. In this thesis, we focus on genomic prediction in a large yeast population. Our goal is to explore how well can phenotypes be predicted from genomes alone, and how much can we benefit from including data from other phenotypes. Due to the experimental setup, the individuals are not independent, so we must correct for confounding from the population structure. We compare the performance of a broad range of models: Linear Mixed Models with increasing model complexity, Multi-Trait Linear Mixed Models and Mixed Random Forest. As a result, we achieve prediction accuracy beyond the heritability limit. In addition, we propose a new method for genomic prediction which combines the strong aspects of Linear Mixed Models and Mixed Random Forests. We carry out a simulation study and show that our model outperforms alternative approaches.

Keywords: Linear Mixed Model, Multi-Trait Linear Mixed Model, Mixed Random Forest

Sisukord

Sissejuhatus

Kui palju saame inimese kohta öelda ainult tema DNA põhjal? Kas lisainformatsiooni (näiteks haigusloo) olemasolul suudame teha paremaid ennustusi tema kohta? Käesolevas töös uurime statistiliste meetodite abil, kuidas pärilikke tunnuseid võimalikult täpselt prognoosida.

Personaalse meditsiini idee kohaselt on iga inimese geneetilise informatsiooni ehk genotüübi arvessevõtmisel võimalik haiguseid täpsemalt diagnoosida ning patsientidele efektiivsemat ravi määrata. Inimese DNA järjestuse sekveneerimine on muutunud kättesaadavaks: USAs on plaanis lähiaastatel sekveneerida 1 000 000 ameeriklast, Ühendkuningriigis on käigus analoogiline 100 000 genoomi projekt, Eestis on koostöös *Broad* instituudiga alustatud mitme tuhande indiviidi sekveneerimist. TÜ Eesti Geenivaramu on kogunud rohkem kui 50 000 geenidoonori andmed, mille kõrvutamine e-tervise haiguslugude registriga võimaldab tuvastada erinevate haigustega seotud riskifaktoreid. Siin mängivad rolli nii keskkond, meie harjumused kui ka pärilik soodumus.

Lisaks haigusriskidele leidub palju indiviidi iseloomustavaid tunnuseid ehk fenotüüpe, nagu näiteks juuksevärv, pikkus ja kehamassiindeks, mis on pärilikud. Fenotüübi väärtused varieeruvad indiviidide vahel ning osa sellest variatsioonist on seletatav genotüübi põhjal, st see on pärilik. Ent enamasti ei määra geneetilised faktorid mingit pärilikku tunnust täielikult, vaid ka keskkonnategurid omavad vähemal või suuremal määral mõju, vastasel juhul oleksid näiteks ühemunakaksikud igas mõttes identsed. *Pärilikkus* näitabki, kui suure osa fenotüübi dispersioonist erinevate indiviidide vahel seletab genotüüp. On selge, et selle väärtus sõltub nii uuritavast populatsioonist kui ka fenotüübist.

Indiviidide vahel esineb teatud geneetilisi erinevusi, kuigi kahe suvalise inimese DNA on suures osas identne. Bioloogiliste mehhanismide tundmaõppimise seisukohast on oluline uurimisküsimus, millised genoomi piirkonnad on põhjuslikult seotud meid huvitava tunnusega. Kahte erinevat versiooni mingist DNA piirkonnast (näiteks geenist või ühestainsast nukleotiidist) nimetatakse alleeliks. Üksikuid nukleotiide, kus esineb variatsiooni piisavalt paljude indiviidide vahel, nimetatakse SNPdeks (*Single Nucleotide Polymorphism*). Et tuvastada seoseid alleelide arvu ja meid huvitava fenotüübi vahel, viiakse läbi ülegenoomseid assotsiatsiooniuuringuid (*Genome-Wide Association Study*, GWAS). Nende tulemusena on seostatud konkreetseid SNPe selliste haigustega nagu diabeet või autoimmuunhaigused [1].

Kuigi GWASides on tuvastatud seoseid genoomi piirkondade ja huvipakkuva fenotüübi

vahel, on esile kerkinud puuduva pärilikkuse (*missing heritability*) probleem: uuringutes tuvastatud SNPd seletavad ainult väikese osa fenotüübi dispersiooni pärilikust komponendist [2], [3]. Näiteks inimese pikkuse pärilikkuseks on hinnatud 80%, aga rohkem kui 100 000 inimest hõlmanud suures GWAS uuringus leiti pikkuse seos 180 SNP-ga, mis seletavad kokku ainult 10% pikkuse dispersioonist [4]. Ülejäänud 70 protsendipunkti on justkui seletamata. Selle põhjuseks võib olla asjaolu, et fenotüüpi mõjutab suur arv SNPe, mille efektisuurused on liiga väikesed, et neid saaks tuvastada siiani kasutatud valimimahtude juures [5], [6]. Samuti võivad seda seletada geneetilised koosmõjud, st et põhjuslike SNPde mõjud ei ole sõltumatud [3].

Teisalt, põhjuslike seoste otsimise asemel võime püstitada küsimuse, kui hästi on võimalik prognoosida huvipakkuva fenotüübi väärtust, kasutades indiviidi genotüübi informatsiooni. Kuigi esmapilgul võib tunduda, et kellegi fenotüüpi (näiteks pikkust) on lihtsam teada saada kui genotüüpi, siis iga fenotüüpi pole meil võimalik mõõta ega vaadelda. Näiteks kui tahame prognoosida, kas inimesel esineb mingi pärilik haigus, siis haiguse avaldumine ei pruugi olla mõõdetav, aga seda võib saada prognoosida geneetilise informatsiooni abil. Veelgi enam, võime saada haigusriski prognoosi täpsustada, kui meil on teada mõne teise fenotüübi väärtus, näiteks võib kõrge "halva" kolesterooli tase suurendada riskiskoori. Inimese varasemast haigusloost leiab analoogilist fenotüübi informatsiooni, mis võib ennustustäpsust suurendada. Magistritöös keskendumegi eelkõige fenotüübi prognoosimisele.

Selle asemel, et tegeleda fenotüübi prognoosimisega inimeste näitel, uurime käesolevas töös pagaripärmi (*Saccharomyces cerevisiae*). Kuna laias plaanis pole vahet, millise organismi geno- ja fenotüübist räägime, nimetame ka pärmi isendeid *indiviididekс*. Uuritavateks fenotüüpideks on pärmiraku pooldumisajad erinevates keskkondades (näiteks hakkavad mõned indiviidid suhkrulahuses kiiremini kasvama, mõned taluvad kuumust paremini kui teised). Magistritöö põhineb eksperimendil, mida viiakse läbi rahvusvahelise koostööna Euroopa molekulaarbioloogia laboratooriumi (EMBL), Nice'i ülikooli ja Göteborg'i ülikooli vahel. Kuigi pärmi fenotüüpide kogumise protsess veel pooleli ning meil on kasutada alamhulk kõigist tulevastest andmetest, siis projekti lõppemisel tekib seni teadaolevalt kõige suurem sekveneeritud ning fenotüpiseeritud indiviidide kogum. Seega on meil ainulaadne võimalus uurida pärmi näitel pärilikkust ning saada selgust, kui täpselt üleüldse on võimalik fenotüüpi prognoosida.

Magistritöö eesmärgiks on uurida pärilike tunnuste prognoosimiseks erinevaid mudeliklasse ning võrrelda nende ennustustäpsust. Selle põhjal mõõdame, kui suur osa pärilikkusest on mudelite abil seletatav. Modelleerimisel peame arvesse võtma eksperimendi-

disainist tulenevat populatsioonistruktuuri, st et indiviidid pole sõltumatud. Kõigepealt tutvume lineaarse segamudeliga, mis on lineaarse mudeli analoog olukorras, kus klassikaline sõltumatuse eeldus pole täidetud. Mitme fenotüübi ühisjaotuse modelleerimiseks vaatleme mitmemõõtmelisi lineaarseid segamudeleid. Nende abil uurime, kas teiste fenotüüpide teadmine võimaldab genotüübil põhinevaid prognoose täpsustada. Samuti anname ülevaate juhuslikust segametsast ning võrdleme selle ennustustäpsust segamudelite omaga. Lõpuks pakume prognoosimiseks välja uue meetodi, mis seisneb juhusliku segametsa edasiarenduses.

Töö koosneb kolmest peatükist. Esimeses anname ülevaate mudelitest. Teises kirjeldame bioloogilise eksperimendi disaini ning selgitame, kuidas on saadud genotüübi ja fenotüübi andmed. Kolmandas osas rakendame mudeleid andmetele ning võrdleme nende prognooside headust. Tulemustena näeme, et kasutades ennustamisel nii genotüüpi kui ka teisi fenotüüpe, saavutame täpsuse, mis ületab pärilikkusel põhinevad piirid. Lisaks näitame simuleeritud andmetel, et meie välja pakutud mudeli prognoositäpsus ületab alternatiivsete meetodite oma.

Käesolev töö põhineb eksperimentaalsetel andmetel, mille kogumiseks vajalikud katsed disainisid ja viisid laboris läbi Johan Hallin, Gianni Liti ja Jonas Warringer. Töös esitatud tulemusteni olen jõudnud meeldivas koostöös nii nende kui juhendaja Leopold Partsiga. Eriti tänan Leopoldi tema põhjalike seletuste, väärtuslike nõuannete ning igakülgse abi eest.

1 Metoodika

Kuna meie põhieesmärk on tegeleda fenotüübi prognoosimisega genotüübi põhjal, tutvumegi selles peatükis erinevate mudelitega, mida kasutada fenotüüpide modelleerimiseks. Ühelt poolt, see ülesanne on lähedalt seotud põhjuslike seoste tuvastamisega, sest ennustamisel oleks loomulik kasutada just neid SNPe, millel on leitud oluline seos fenotüübiga. Teisalt, alljärgnevas peatükis näeme, et on ka teisi võimalusi: näiteks saame prognoosida ka ainult indiviididevahelist geneetilist sarnasust arvesse võttes.

Olgu y_i fenotüübi mõõtmistulemus i-nda indiviidi jaoks ning x_{ij} tema genoomi j-nda positsiooni genotüüp. Teame, et meie uuritav pärmipopulatsioon on põlvnenud kahest isendist: üks NA (*North American*) ja teine WA (*West African*) tüvest. Seega indiviidi genoomi iga positsiooni jaoks on kaks võimalikku alleeli: NA ja WA. Arvestades, et indiviidil on igast kromosoomist kaks koopiat (üks kummaltki vanemalt), on tal igas positsioonis NA ja WA esinemiste arv kokku kaks: ühelt vanemalt võib olla saanud kas NA või WA alleeli ning teiselt vanemalt sama moodi kas NA või WA. Kodeerides x_{ij} nii, et see tähistab WA alleelide esinemise arvu, saame võimalike genotüüpide NA/NA, NA/WA, WA/NA ja WA/WA jaoks vastavalt x_{ij} kodeeringud 0, 1, 1, 2. Tähistame N indiviidi jaoks fenotüüpide vektori $\mathbf{y} := (y_1, \ldots, y_N)^T$ ning genotüübi maatriksi $\mathbf{X} := (\mathbf{x}_1, \ldots, \mathbf{x}_M)$, mille veergudes on genoomi positsioonid, mida nimetame SNPdeks.

Kogu töös kasutame notatsiooni, kus vektorid ja maatriksid on paksus kirjas (vektorid väiketähtedega, maatriksid suurtähtedega), arve tähistame tavaliste väiketähtedega.

Matemaatiliselt tähendab mudel fenotüübi jaoks lihtsalt uuritava tunnuse $\mathbf{y} \in \mathbb{R}^N$ modelleerimist argumentide $\mathbf{X} \in \{0, 1, 2\}^{N \times M}$ abil. Kuid sobiva mudeli valik ei ole triviaalne ülesanne: enamasti on tunnuste arv M palju suurem kui indiviidide arv N, samuti tuleb arvesse võtta populatsioonistruktuurist tulenevat indiviididevahelist sõltuvust (klassikaline *i.i.d.* eeldus ei kehti). Lisaks on tihti teada ühe fenotüübi $\mathbf{y}$ asemel mitu erinevat tunnust $\mathbf{y}_1, \ldots, \mathbf{y}_P$, seega võime tahta arvesse võtta nendevahelist korrelatsiooni. Probleemi aktuaalsust näitab asjaolu, et viimase paari aasta jooksul on töötatud välja mitmeid uusi lähenemisi fenotüübi modelleerimiseks populatsioonistruktuuri olemasolul [7]–[11].

Mudelite sobitamist alustame mudelitest, kus vaatleme SNPide efekte aditiivsetena. Nimetame SNPi $\mathbf{x}_j$ efekti aditiivseks, kui $\mathbf{y}$ sõltub argumendist $\mathbf{x}_j$ lineaarselt. Peame silmas aditiivsust alleelide arvu suhtes: kui teatud positsioonis ühe WA alleeli omamine suurendab fenotüübi väärtust teatud kordaja võrra, siis kahe WA alleeli omamisel on efekt

kaks korda suurem. Aga näiteks dominantsuse esinemisel näeme efekti ainult kahe WA alleeli omamisel: see ei ole aditiivne. Mudelisse saame lisada ka geneetilisi koosmõjusid ehk interaktsioone. Tunnuste $\mathbf{x}_i$ ja $\mathbf{x}_j$ interaktsiooni all peame silmas nende korrutise mudelisse lisamist.

1.1 Lineaarne segamudel

Alustame lihtsaimast võimalikust mudelist, milleks on pideva y-tunnuse korral lineaarse regressiooni mudel, ning arutame selle puuduseid enne keerulisemate mudelite juurde minekut. Näiteks võime kasutada järgnevat lineaarset mudelit

$$\mathbf{y} = \mu\mathbf{1} + \sum_{j=1}^{M} \beta_j \mathbf{x}_j + \boldsymbol{\varepsilon}\,,$$

kus $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ ning kus j-nda SNP-i efekt on β_j. Siin $\mathbf{1}$ tähistab N-mõõtmelist ühtede vektorit ning $\mathbf{I}$ ühikmaatriksit. Kuna selles mudeli esituses on esindatud kõik SNPid, võiksid paljud kordajad β_j võrduda nulliga, sest arvatavasti on neist ainult vähesed põhjuslikult seotud uuritava fenotüübiga. Sellisel lähenemisel on mitmeid probleeme. Näiteks kuna $M > N$, siis pole kordajate β_j hinnangud üheselt määratud. Selle probleemi vältimiseks oleks alternatiiv lisada mudelisse ainult üks argument $\mathbf{x}_j$ korraga (st võtta $M = 1$). Sellist lähenemist kasutatakse GWAS uuringutes, kus põhiliseks eesmärgiks on leida uuritava fenotüübiga põhjuslikult seotud SNPid. Tüüpiliselt sobitatakse iga $\mathbf{x}_j$ jaoks eraldi ühe argumendiga mudel ning järjestatakse SNPid nende efekti suuruse põhjal. Nii leiame üles küll kõige tugevama signaali, ent ainult ühe tunnuse põhjal prognoosimine ei taga head tulemust. Idee poolest võiksime rakendada ka edaspidist mudelivalikut (*forward selection*), st lisada mudelisse samm-sammult mingi teatud arvu m argumenti, kus $m < M$. Ometi ei ole tavaline lineaarne regressioon siinkohal sobiv mudel, sest populatsioonistruktuur genereerib $\mathbf{y}$ elementide vahel sõltuvuse, mida ei seleta ära väike arv argumente $\mathbf{x}_1, \ldots, \mathbf{x}_m$. Seetõttu ei ole täidetud mudeli jääkide $\boldsymbol{\varepsilon}$ mittekorreleerituse eeldus.

Lineaarne segamudel (*Linear Mixed Model*, LMM) võimaldab modelleerida indiviididevahelist sõltuvust juhuslike efektide abil. Selleks defineerime geneetilisel sarnasusel põhineva juhuslike efektide vektori $\mathbf{g} = (g_1, \ldots, g_N)^T$, kus i-nda indiviidi geneetiline efekt g_i võtaks arvesse kõigi SNPde mõju (täpsemalt defineerime g_i hiljem). Lubades $\mathbf{g}$ ele-

mentidel olla korreleeritud, saame mudeli esituse

$$\mathbf{y} = \mu\mathbf{1} + \sum_{j=1}^{m} \beta_j \mathbf{x}_j + \mathbf{g} + \boldsymbol{\varepsilon}\,,$$

kus juhuslik efekt $\mathbf{g}$ modelleerib seda osa indiviidide vahelisest korrelatsioonist, mis on SNP-ide $\mathbf{x}_1, \ldots, \mathbf{x}_m$ poolt seletamata, ning $\boldsymbol{\varepsilon}$ võtab arvesse sõltumatut müra, mis on tekkinud $\mathbf{y}$ mõõtmisel. Et edaspidi kasutada lühemat kirjapilti, koondame kõik fikseeritud efektid ühte mudelimaatriksisse $\mathbf{X}$ ning esitame lineaarse segamudeli kujul

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\varepsilon}\,. \tag{1}$$

Valemi (1) erijuhuks on mudel, mis sisaldab geneetilisi faktoreid ainult juhuslike efektidena:

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{g} + \boldsymbol{\varepsilon}\,. \tag{2}$$

Mudeli esituses kasutasime juba juhuslikku efekti $\mathbf{g}$, mille defineerime nüüd järgnevalt

$$g_i := \sum_{j=1}^{M} W_{ij} u_j\,,$$

kus $\mathbf{W}$ on saadud genotüübi maatriksi $\mathbf{X}$ veergude tsentreerimisel ja standardiseerimisel ning u_j on j-nda SNP-i efekt $u_j \sim \mathcal{N}(0, \sigma_g^2)$ [12], [13]. Teisisõnu, g_i avaldub kõigi SNPde lineaarkombinatsioonina, ainult et siin vaatleme nende efekte juhuslikuna, mitte fikseerituna. Selle tulemusena on mudel hinnatav, sest juhuslike efektide u_j jaoks on vaja hinnata ainult üks parameeter σ_g^2. Seda parameetrit on lihtne interpreteerida mudeli (2) jaoks, sest siis jaguneb $\mathbf{y}$ dispersioon komponentideks σ_g^2 ja σ_e^2, kus esimene näitab geneetiliste efektide osatähtsust ning teine mõõtmismüra oma. Maatrikskujul esitub $\mathbf{g} = \mathbf{W}\mathbf{u}$, kus $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_g^2\mathbf{I})$. Järelikult

$$\mathbf{g} \sim \mathcal{N}\left(\mathbf{0},\ \sigma_g^2 \mathbf{W}\mathbf{W}^T\right)\,.$$

Tähistades realiseerunud suguluse maatriksi $\mathbf{K} := \mathbf{W}\mathbf{W}^T$ (*realized relationship matrix, kinship matrix*), ning eeldades, et geneetilised efektid $\mathbf{g}$ on sõltumatud mõõtmisvigadest $\boldsymbol{\varepsilon}$, saame dispersiooni-kovariatsioonimaatriksi esituse

$$\mathrm{Var}\,(\mathbf{y}) = \mathrm{Var}\,(\mathbf{g}) + \mathrm{Var}\,(\boldsymbol{\varepsilon}) = \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I}\,.$$

Seega saame mudelitele (1) ja (2) vastavad marginaaljaotused

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta},\ \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I}\right) \tag{3}$$

ja

$$\mathbf{y} \sim \mathcal{N}\left(\mu\mathbf{1},\ \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I}\right). \tag{4}$$

Suguluse maatriksi $\mathbf{K}$ elemendid k_{ij} näitavad indiviidide i ja j poolt jagatud genoomi osakaalu. Selle ligikaudse lähendi saaksime leida indiviidide põlvnemise põhjal: näiteks kui kaks indiviidi jagavad täpselt ühte vanemat, siis nad jagavad ligikaudu 50% genoomist. Aga realiseerunud $\mathbf{K}$ ehk täpse genotüübi informatsiooni kasutamisel saame täpsemad prognoosid [12].

Kokkuvõttes erinevadki segamudelid (3) ja (4) tavaliselt lineaarsest mudelist selle poolest, et diagonaalse dispersiooni-kovariatsioonimaatriksi $\sigma_e^2\mathbf{I}$ asemel kasutame mittediagonaalset $\sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I}$. Kuna dispersioon on jagatud kaheks liidetavaks, nimetatakse neid dispersioonikomponentideks: geneetiline komponent ning müra. Mudelis (4) saame suhet $\sigma_g^2/(\sigma_g^2 + \sigma_e^2)$ interpreteerida, kui suure osa fenotüübi dispersioonist seletavad geneetilised aditiivsed komponendid. Seda nimetatakse *kitsaks pärilikkuseks* (*narrow sense heritability*) ja tähistatakse h^2.

Saadud kahe dispersioonikomponendiga mudelit võime ka üldistada, jagades genoomi regioonideks (näiteks kromosoomide kaupa) ning omistades igale regioonile oma dispersioonikomponendi. Hinnates iga regiooni r jaoks eraldi suguluse maatriksi $\mathbf{K}_r$, saame mudeli

$$\mathrm{Var}\left(\mathbf{y}\right) = \sum_{r=1}^{R} \sigma_r^2\mathbf{K}_r + \sigma_e^2\mathbf{I}\,,$$

mis sisaldab igale regioonile (kokku R tükki) vastavat dispersioonikomponenti. Sel juhul näitaks suhe $\sigma_r^2/(\sum_i \sigma_i^2 + \sigma_e^2)$, kui suure osa fenotüübi dispersioonist kirjeldab regioon r.

1.2 Mitmemõõtmeline lineaarne segamudel

Siiani oleme käsitlenud mudeleid ühele fenotüübile korraga, kuid tihti on indiviidide kohta teada mitme fenotüübi väärtused. Isegi kui huvi pakub konkreetse fenotüübi prognoosimine, võime teiste teadaolevate fenotüüpide abil saada täpsemaid prognoose. Seetõttu tutvume siin alapeatükis mitmemõõtmelise lineaarse segamudeliga (*Multi-Trait Linear Mixed Model*, MT LMM), mis võimaldaks mitme fenotüübi samaaegset modelleeri-

mist.

Idee seisneb kõigi tunnuste $\mathbf{y}_1, \ldots, \mathbf{y}_P$ jaoks korraga ühe segamudeli kasutamises, mis võtaks arvesse fenotüüpidevahelisi kovariatsioone nende ühisjaotuse modelleerimisel. Selleks defineerime järgneva mudeli fenotüüpide $\mathbf{y}_j$ jaoks iga $j \in \{1, \ldots, P\}$ korral

$$\mathbf{y}_j = \mathbf{f}_j + \mathbf{g}_j + \boldsymbol{\varepsilon}_j \,, \tag{5}$$

kus $\mathbf{f}_j = \mathbf{X}_j \boldsymbol{\beta}_j$ on fikseeritud efektide vektor, mis avaldub mudelimaatriksi $\mathbf{X}_j$ ja parameetrite vektori $\boldsymbol{\beta}_j$ kaudu, $\mathbf{g}_j$ on geneetiliste juhuslike efektide vektor ning $\boldsymbol{\varepsilon}_j$ tähistab sõltumatut mõõtmismüra. Juhuslike efektide vektori $\mathbf{g}_j$ elementidevahelisi korrelatsioone näitab geneetilise suguluse maatriks $\mathbf{K}$. Saadud mudel (5) näib sarnane ühemõõtmelise segamudeliga (1). Kui eeldaksime, et erinevatele fenotüüpidele vastavad juhuslike efektide vektorid $\mathbf{g}_i$ ja $\mathbf{g}_j$ või $\boldsymbol{\varepsilon}_i$ ja $\boldsymbol{\varepsilon}_j$ on sõltumatud, võiksime (5) asemel sobitada iga fenotüübi jaoks eraldi mudeli. Kui aga fenotüübid on korreleeritud, siis peame lubama sõltuvust erinevatele fenotüüpidele $\mathbf{y}_i$ ja $\mathbf{y}_j$ vastavate juhuslike efektide vektorite $\mathbf{g}_i, \mathbf{g}_j$ või $\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j$ vahel.

Olgu juhuslikud efektid $\mathbf{g}_1, \ldots, \mathbf{g}_P$ ja $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_P$ sellised, et

$$\begin{aligned} \mathrm{Cov}(\mathbf{g}_i, \mathbf{g}_j) &= c_{ij}\mathbf{K} \\ \mathrm{Cov}(\boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_j) &= s_{ij}\mathbf{I} \end{aligned}$$

ning $\mathbf{g}_i$ ja $\boldsymbol{\varepsilon}_j$ sõltumatud. Nüüd on lubatud fenotüüpide i ja j vaheline sõltuvus. Täpsemalt, nendevaheline kovariatsioon on jagatud kaheks osaks, kus komponendid c_{ij} ja s_{ij} näitavad, kui suur osa sellest kovariatsioonist on seletatud vastavalt geneetiliste aditiivsete ning ülejäänud efektide poolt. Iga fenotüübi $\mathbf{y}_j$ dispersiooni-kovariatsioonimaatriks jääb analoogiliseks ühemõõtmelise segamudeli omaga

$$\mathrm{Var}\left(\mathbf{y}_j\right) = c_{jj}\mathbf{K} + s_{jj}\mathbf{I}.$$

Et spetsifitseerida fenotüüpide ühisjaotuse jaoks mitmemõõtmeline mudel eelneva arutelu põhjal, esitame selle maatrikskujul. Selleks defineerime fenotüüpide maatriksi $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_P) \in \mathbb{R}^{N \times P}$, fikseeritud efektide maatriksi $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_P) \in \mathbb{R}^{N \times P}$, geneetiliste juhuslike efektide maatriksi $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_P)$ ja juhusliku müra $\mathbf{E} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_P)$. Mitmemõõtmelise segamudeli saame siis esitada järgnevalt

$$\mathbf{Y} = \mathbf{F} + \mathbf{G} + \mathbf{E} \,. \tag{6}$$

Juhuslikud maatriksid $\mathbf{G}$ ja $\mathbf{E}$ saame fenotüüpidevaheliste kovariatsioonimaatriksite $\mathbf{C} := (c_{ij})$, $\mathbf{S} := (s_{ij})$ abil esitada nii, et

$$\begin{aligned} \operatorname{vec} \mathbf{G} &\sim \mathcal{N}(\mathbf{0}, \mathbf{C} \otimes \mathbf{K}) \\ \operatorname{vec} \mathbf{E} &\sim \mathcal{N}(\mathbf{0}, \mathbf{S} \otimes \mathbf{I}), \end{aligned}$$

kus $\operatorname{vec}(\cdot)$ tähistab vektoriseerimise operaatorit, mis transformeerib maatriksi veeruvektoriks. Näiteks kui maatriks $\mathbf{G}$ on mõõtmetega $N \times P$, siis vektor $\operatorname{vec}(\mathbf{G})$ mõõtmetega $NP \times 1$ on saadud $\mathbf{G}$ veergude $\mathbf{g}_1, \ldots, \mathbf{g}_P$ üksteise alla ladumise teel. Operaator $\otimes$ tähistab kahe maatriksi otsekorrutist (*Kronecker product*). Näiteks maatriksite $\mathbf{C}$ ja $\mathbf{K}$ otsekorrutis on järgnev $NP \times NP$ plokkmaatriks

$$\mathbf{C} \otimes \mathbf{K} = \begin{pmatrix} c_{11}\mathbf{K} & c_{12}\mathbf{K} & \cdots & c_{1P}\mathbf{K} \\ c_{21}\mathbf{K} & c_{22}\mathbf{K} & \cdots & c_{2P}\mathbf{K} \\ \vdots & \vdots & \ddots & \vdots \\ c_{P1}\mathbf{K} & c_{P2}\mathbf{K} & \cdots & c_{PP}\mathbf{K} \end{pmatrix}.$$

Eelnevast järeldub, et $\operatorname{vec}(\mathbf{Y})$ marginaaljaotuseks on järgnev mitmemõõtmeline normaaljaotus

$$\operatorname{vec} \mathbf{Y} \sim \mathcal{N}(\operatorname{vec} \mathbf{F},\ \mathbf{C} \otimes \mathbf{K} + \mathbf{S} \otimes \mathbf{I}). \tag{7}$$

Niisiis, maatriksid $\mathbf{K}$ ja $\mathbf{I}$ (mõõtmetega $N \times N$) näitavad vastavate juhuslike efektide *indiviididevahelisi* sõltuvusi (või nende sõltuvuste puudumist), aga maatriksid $\mathbf{C}$ ja $\mathbf{S}$ (mõõtmetega $P \times P$) näitavad sõltuvusi *fenotüüpide* vahel.

Kui $\mathbf{K}$ saame arvutada eelnevalt genotüübi andmetelt, siis $\mathbf{C}$ ja $\mathbf{S}$ tuleb hinnata mudeli sobitamise käigus. Efektiivsed algoritmid suudavad mudeli (6) parameetreid lineaaralgebra trikkide abil mõistliku arvutusliku keerukuse ja ajaga hinnata [14]. Selliste mudelite sobitamine on implementeeritud paketis LIMIX[1] [9].

1.3 Segamudeli argumenttunnuste valik

Ühe- või mitmemõõtmelise segamudeli sobitamisel kasutame kahest komponendist koosnevat kovariatsioonistruktuuri, vastavalt kas $\sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I}$ või $\mathbf{C} \otimes \mathbf{K} + \mathbf{S} \otimes \mathbf{I}$. Aga milliseid tunnuseid lisada fikseeritud efektidena? Kuna soovime peale vabaliikme veel fikseeritud efekte mudelisse lisada, kirjeldame järgnevalt ühte võimalust mudeli argumentide valikuks.

[1] http://pmbio.github.io/limix/

Alustame mudelivalikuga ühemõõtmelise segamudeli jaoks. Kuna optimaalne tunnuste arv mudelis pole teada, hakkame edaspidise mudelivalikuga (*forward selection*) mudelisse argumente lisama. Kui oleme mudelisse lisanud juba m tunnust, siis järgmise argumendi valimiseks lisame olemasolevasse mudelisse ühekaupa kõik kandidaattunnused. Saadud $(m+1)$ argumendiga mudeleid võrdleme tõepärasuhte testi abil mudeliga, kus oli m tunnust. Tunnuse, mis suurendas mudeli tõepära enim, lisame mudelisse järgmiseks argumendiks. Tähistades m-ndaks sammuks mudelisse lisatud argumentide efektid $\mathbf{X}\boldsymbol{\beta}$ ning kandidaattunnuste indeksite hulga $\mathcal{I}_m$, saame kriteeriumiks valida järgmine tunnus indeksiga

$$j^{m+1} := \arg\max_{j \in \mathcal{I}_m} \left\{ \frac{L(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta} + \beta_j \mathbf{x}_j,\ \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I})}{L(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta},\ \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I})} \right\},$$

kus $L(\cdot)$ tähistab normaaljaotuse tõepära ning parameetrid $\boldsymbol{\beta}, \beta_j, \sigma_g^2, \sigma_e^2$ on hinnatud suurima tõepära meetodil.

Ehitades nii mudelit ühe tunnuse kaupa, vajame lõpetamiskriteeriumit, millisest hetkest alates enam tunnuseid mitte lisada. Otsustasime mitte seada tõepärasuhte testi p-väärtustele eeldefineeritud piiri, sest mitmese testimise tõttu (iga järgmise argumendi valikul teeme suurusjärk M testi) on p-väärtus kaotanud oma esialgse interpretatsiooni. Samuti pole päris selge, milline piir tagaks parima prognoosivõime. Seetõttu kasvatame mudelit järk-järgult, lisades järjest argumente, ja uurime, milline mudel tagab täpseimad prognoosid väljaspool treeningvalimit.

Alternatiiv edaspidisele mudelivalikule on kasutada mõnda regulariseerimisel põhinevat lähenemist, nagu LASSO analoog lineaarsete segamudelite jaoks [7]. Nii saaksime ühe sammuga hinnata kõigi kordajate väärtused, kusjuures enamik neist hinnataks nullideks. Sellel lähenemisel on aga puudus, et erinevalt edaspidisest mudelivalikust pole koosmõjude mudelisse lisamine praktikas teostatav (kõigi võimalike interaktsioonide arv on liiga suur). Kuna üldiselt on geenide omavahelised koosmõjud oodatavad, on mõistlik lubada mudelisse lisada interaktsioone. Seetõttu jäime edaspidise tunnuste valiku juurde.

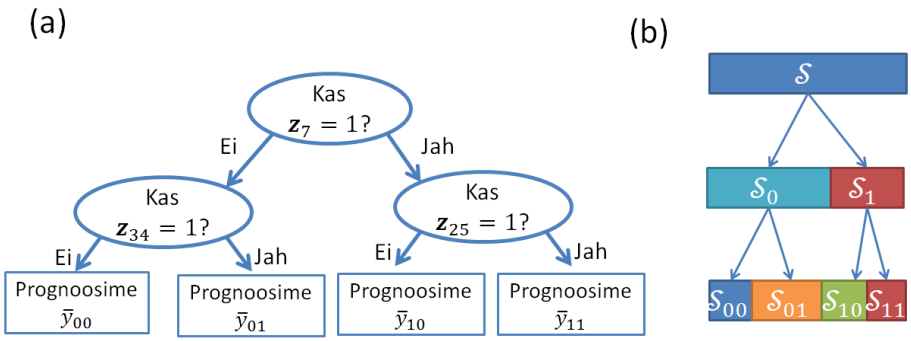
Mitmemõõtmelise segamudeli jaoks ei hakka me eraldi argumentide valikut läbi viima. Kui oleme iga fenotüübi jaoks eraldi leidnud optimaalse konfiguratsiooniga mudeli, lisame nendesamade mudelite argumendid vastavate fenotüüpide fikseeritud efektideks mitmemõõtmelises mudelis.

1.4 Juhuslik segamets

Kui modelleeritav tunnus sõltub argumentidest mittelineaarselt, võib lineaarse mudeli asemel olla prognoosimiseks efektiivsem kasutada otsustuspuudel (*decision tree*) põhinevat lähenemist. Pideva y-tunnuse korral nimetatakse vastavat otsustuspuud regressioonipuuks (*regression tree*). Tavaliselt kombineeritakse mitmed puud üheks komplektiks (*ensemble*), sest nii saadakse täpsemad prognoosid kui üksiku puu abil. Kui üks puu pakub suhteliselt lihtsa rusikareegli prognoosimiseks, siis paljude puude komplekt on juba paindlikum mudel. Lisaks aitab see ülesobitumise(*overfitting*) vastu. Ühte sellistest algoritmidest, mis kombineerib üksikute puude ennustused lõpliku tulemuse saamiseks, nimetatakse juhuslikuks metsaks (*Random Forest*, RF).

Järgnevalt tutvume kõigepealt regressioonipuuga ning selle üldistusega, mis võtab arvesse populatsioonistruktuuri. Nii nagu lineaarse mudeli üldistust nimetatakse lineaarseks *sega*mudeliks, nimetame juhusliku metsa üldistust juhuslikuks *sega*metsaks (*Mixed Random Forest*, MRF). Ülevaade regressioonipuust ja juhusliku metsa algoritmist põhineb materjalidel [15], [16] ning selle üldistus juhuslikuks segametsaks artiklil [17].

Siin peatükis lähtume puu konstrueerimisel binaarsetest argumenttunnustest $\mathbf{z}_1, \ldots, \mathbf{z}_M$. Kuna meie genotüübi tunnused $\mathbf{x}_1, \ldots, \mathbf{x}_M$ on väärtustega $\{0, 1, 2\}$, saame need binariseerida näiteks tingimuste $x_{ij} > 0$ või $x_{ij} < 2$ järgi. Juhusliku metsa ja segametsa rakendamisel kasutasimegi binaarseid argumenttunnuseid $\mathbf{x}_j > \mathbf{0}$ ja $\mathbf{x}_j < \mathbf{2}$, nii on kogu tunnuste arv on kaks korda suurem kui esialgsete SNPde arv M.

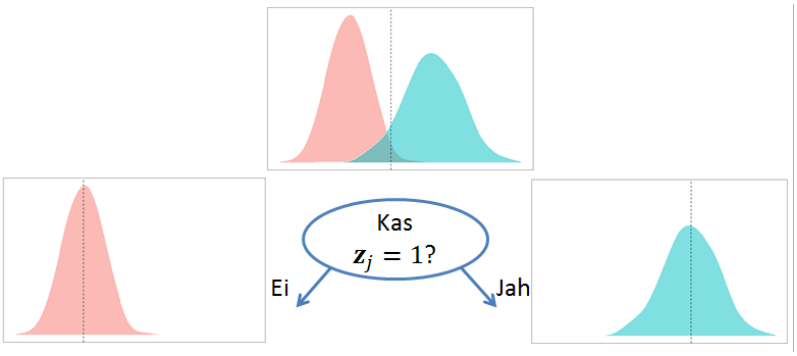


Joonis 1. (a) Näide regressioonipuust (puu kõrgus 2) binaarsete argumentide $\mathbf{z}_j$ korral. Igas tipus jagatakse valim kaheks: esimene jagamine toimub $\mathbf{z}_7$ järgi, teine jagamine $\mathbf{z}_{34}$ või $\mathbf{z}_{25}$ järgi (sõltuvalt, kas tunnuse $\mathbf{z}_7$ väärtus oli 0 või 1). (b) Otsustuspuu jagab valimi elementide indeksid $\mathcal{S}$ alamhulkadeks $\mathcal{S}_{00}, \mathcal{S}_{01}, \mathcal{S}_{10}, \mathcal{S}_{11}$. Prognoosiks tagastame iga hulga elementide keskmise $\bar{y}_{00}, \ldots, \bar{y}_{11}$.

Regressioonipuu (näide joonisel 1) igas tipus jaotame valimi kahte ossa teatud kriteeriu-

mi põhjal. Näiteks pärast esimest jagamist oleme valimi elementide indeksid $\mathcal{S}$ jaganud kaheks hulgaks $\mathcal{S}_0$ ja $\mathcal{S}_1$, sõltuvalt sellest, kas indiviidi i jaoks kehtib $z_{ij} = 0$ või $z_{ij} = 1$. Nii jagame valimi alamhulkadeks (nende arv sõltub puu kõrgusest) ning prognoosiks raporteerime $\mathbf{y}$ keskmise väärtuse igas alamhulgas.

Kuidas aga puud konstrueerida? Näiteks kuidas valida esimene tunnus $\mathbf{z}_j$, mille põhjal valimi kõik elemendid $\mathcal{S}$ jagada kaheks osaks $\mathcal{S}_0$ ja $\mathcal{S}_1$? See tunnus peaks olema selline, et jagamise tulemusena paraneks prognooside täpsus võimalikult palju. Arvestades, et prognoosiks on vaatluste keskmine nendes hulkades $\bar{y}_0$ ja $\bar{y}_1$, on loomulik seada kriteeriumiks, et prognoosivigade ruutude summa oleks minimaalne (joonis 2).



Joonis 2. Illustratsioon n-ö kasulikust kaheksjagamisest. Värviga on näidatud tunnuse $\mathbf{z}_i$ väärtused, katkendjoonega antud tipu elementide keskmine ehk prognoos. Andmete kaheksjagamisel eraldame kaks alamklassi, mille tulemusena väheneb märgatavalt prognoosivigade ruutude summa.

Üldiselt, kui n-ndal sammul on meil vaja indeksite hulk $\mathcal{S}$ jagada kaheks, siis prognoosivigade ruutude summa minimeerimine annab järgneva kriteeriumi

$$j^{n+1} := \arg\min_j \left\{ \sum_{i \in \mathcal{S}_0^j} (y_i - \bar{y}_0)^2 + \sum_{i \in \mathcal{S}_1^j} (y_i - \bar{y}_1)^2 \right\},$$

kus $\mathcal{S}_0^j = \{i : z_{ij} = 0\}$ ja $\mathcal{S}_1^j = \{i : z_{ij} = 1\}$. Et saaksime seda lähenemist üldistada ning võtta arvesse populatsioonistruktuuri, esitame prognoosivigade ruutude summa alternatiivsel kujul, et tuua välja analoogia lineaarse mudeli sobitamisega. Tähistades $\beta_0 := \bar{y}_0$ ning $\beta_1 := \bar{y}_1 - \beta_0$, saame kriteeriumi

$$j^{n+1} := \arg\min_j \left\{ \sum_{i \in \mathcal{S}_0^j} (y_i - \beta_0)^2 + \sum_{i \in \mathcal{S}_1^j} (y_i - \beta_0 - \beta_1)^2 \right\}.$$

Pidades silmas, et hulgas $\mathcal{S}_0^j$ kehtib $z_{ij} = 0$ ning hulgas $\mathcal{S}_1^j$ vastavalt $z_{ij} = 1$, saame

kriteeriumi esitada

$$j^{n+1} := \arg\min_{j} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 z_{ij})^2 \,. \tag{8}$$

Eeldusel, et pärast kaheksjagamist on vaatlused kummaski hulgas normaaljaotusega, st

$$\mathbf{y} \sim \mathcal{N}(\beta_0 \mathbf{1} + \beta_1 \mathbf{z}_j,\, \sigma_e^2 \mathbf{I}) \,, \tag{9}$$

on kriteerium (8) samaväärne normaaljaotuse (9) log-tõepära maksimeerimisega. Niisiis saame tunnuse $\mathbf{z}_j$ valiku ülesande lahendada lineaarsete mudelite sobitamise abil.

Populatsioonistruktuuri olemasolul saame valemi (9) n-ö loomuliku üldistuse, nii nagu üldistasime tavalise lineaarse mudeli lineaarseks segamudeliks,

$$\mathbf{y} \sim \mathcal{N}(\beta_0 \mathbf{1} + \beta_1 \mathbf{z}_j,\, \sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I}) \,, \tag{10}$$

kus $\mathbf{K}$ on peatükist 1.1 tuttav geneetilise sarnasuse maatriks.

Eelnevas arutelus andsime ülevaate, kuidas konstrueerida kaheksjagamist puu esimeses tipus. Rakendades nüüd analoogilist protseduuri rekursiivselt soovitud arv kordi nii puu vasaku kui ka parema tipu jaoks, saame tulemuseks soovitud kõrgusega puu.

Juhusliku metsa algoritmi idee on keskmistada mitmete puude ennustused ning selle tulemusena saada parem prognoosivõime. Et kõik puud ei tuleks ühesugused, sobitame iga puu treeningandmetest võetud *bootstrap* valimil. Et puude sarnasust veelgi vähendada, ei valita puu tippudes potentsiaalseteks kandidaattunnusteks mitte kõiki m tunnust, vaid iga tipu jaoks juhuslik alamhulk nendest. Juhusliku metsa prognoosiks on kõigi puude ennustuste keskmine.

Juhuslik segamets on analoogiline meetod, kus puude konstrueerimisel kasutame traditsioonilise ruutude summa minimeerimise asemel mudelit (10). Sedasama mudelit kasutame ka prognoosimisel. See tähendab, et iga tipu elementide keskmisele lisandub veel populatsioonistruktuuril põhineva juhusliku efekti prognoos.

1.5 Parameetrite hindamine

Peatükkides 1.1 ja 1.2 nägime, et segamudelite sobitamine taandub normaaljaotuse parameetrite hindamisele. Hinnata on vaja nii fikseeritud efektide kordajad $\boldsymbol{\beta}$ kui ka dispersiooniparameetrid. Näiteks mitmemõõtmelise segamudeli korral tuleb hinnata feno-

tüüpidevahelised kovariatsioonimaatriksid $\mathbf{C}$ ja $\mathbf{S}$, mis sisaldavad kumbki $P(P+1)/2$ elementi.

Paketis LIMIX on segamudelite sobitamine realiseeritud suurima tõepära (*Maximum Likelihood*, ML) printsiibil: gradiendil põhineval numbrilise optimeerimise meetodil maksimeeritakse mitmemõõtmelise normaaljaotuse log-tõepära. Lineaaralgebra trikkide abil (vt [14]) on võimalik parameetrite hindamine $O(N^3 + P^3)$ ajalise keerukusega ning $O(N^2 + P^2)$ mälukeerukusega [9].

Tutvume lähemalt ühemõõtmelise lineaarse segamudeliga, mida kasutame edaspidise mudelivaliku protsessis, kui testime korduvalt tõepärasuhte testiga kõiki kandidaattunnuseid. Tähistagu $\mathbf{X}\boldsymbol{\beta}$ mudelisse eelnevalt lisatud fikseeritud efekte. Efektiivseks hindamiseks esitatakse mudel (3) reparametriseeritud kujul

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta},\ \sigma_g^2(\mathbf{K} + \delta\mathbf{I})\right), \tag{11}$$

kus δ näitab suhet σ_e^2/σ_g^2. Selle mudeli parameetrite $\boldsymbol{\beta}$ ning σ_g^2 suurima tõepära hinnangute jaoks eksisteerib suletud kujul lahend, mis võimaldab mudeli sobitamise asendada ühe muutuja δ funktsiooni maksimeerimisega (see optimeerimisülesanne ei ole kumer) [11]. Tarkvaras on kandidaattunnuste $\mathbf{x}_1, \ldots, \mathbf{x}_m$ testimine realiseeritud nii, et mudeli

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta} + \beta_j\mathbf{x}_j,\ \sigma_g^2(\mathbf{K} + \delta\mathbf{I})\right)$$

parameeter δ sobitatakse nullmudelil (11) ning kasutatakse sedasama väärtust kõigi argumentidega $\mathbf{x}_j$ mudelites. See lähend teeb võimalikuks suure arvu kandidaattunnuste testimise mõistliku aja jooksul. Veel peame arvestama, et LIMIXi leitavad dispersioonikomponentide hinnangud võivad olla teatava nihkega, kuna seal hinnatakse parameetreid suurima tõepära, mitte REML (*Restricted Maximum Likelihood*) meetodil. See ei ole suur probleem, kuna valimimahu kasvades dispersioonikomponentide ML hinnangute nihe väheneb ning me hindame mudeli parameetreid rohkem kui 1000 indiviidi pealt.

1.6 Prognoosimine

Siiani oleme tutvunud lineaarsete segamudelitega, nende mitmemõõtmelise üldistusega ning näinud nende kasutusvõimalust juhusliku segametsa algoritmis. Samuti saime põgusa ülevaate, millisel põhimõttel mudelite parameetreid hinnatakse. Nüüd uurime, kuidas kasutada hinnatud segamudelit fenotüüpide prognoosimiseks.

1.6.1 Prognoosimine ühemõõtmelise segamudeli abil

Olgu $\mathbf{y} \in \mathbb{R}^N$ vaadeldud fenotüübi väärtused (st need, mida kasutasime mudeli sobitamisel) ning $\mathbf{y}^* \in \mathbb{R}^{N^*}$ uute indiviidide omad (st need, mida soovime prognoosida). Kuna kõigi $(N + N^*)$ indiviidi kohta on olemas genotüübi informatsioon, saame arvutada nende suguluse maatriksi $\tilde{\mathbf{K}}$. Olgu selle read ja veerud järjestatud nii, et esimesed N tükki vastavad vaadeldud $\mathbf{y}$-ga indiviididele ning ülejäänud N^* rida ja veergu uutele indiviididele.

Eeldame, et nii vaadeldud $\mathbf{y}$ kui ka uued $\mathbf{y}^*$ järgivad ühiselt mudelit (3). See tähendab, et nende ühisjaotus on mitmemõõtmelise normaaljaotusega

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{y}^* \end{pmatrix} \sim \mathcal{N}(\tilde{\mathbf{X}}\boldsymbol{\beta},\ \sigma_g^2\tilde{\mathbf{K}} + \sigma_e^2\tilde{\mathbf{I}}) \,,$$

kus $\tilde{\mathbf{X}}$ tähistab kõigi indiviidide ühist mudelimaatriksit. Eraldame nüüd selle normaaljaotuse keskväärtusvektorist ning kovariatsioonimaatriksist esimesele N vaatlusele ning ülejäänud N^* indiviididele vastavad osad.

Kõigepealt jagame suguluse maatriksi $\tilde{\mathbf{K}}$ plokkideks

$$\tilde{\mathbf{K}} = \begin{pmatrix} \mathbf{K} & \mathbf{K}^{*T} \\ \mathbf{K}^* & \mathbf{K}^{**} \end{pmatrix} \,,$$

kus $\mathbf{K}$ tähistab N vaadeldud fenotüübiga indiviidide suguluse maatriksit, $\mathbf{K}^*$ tähistab sugulusi N vaadeldud ja N^* uue indiviidi vahel ning $\mathbf{K}^{**}$ uute indiviidide vahelisi sugulusi. Nüüd saame esitada $\mathbf{y}$ ja $\mathbf{y}^*$ ühisjaotuse

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{y}^* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{X}^*\boldsymbol{\beta} \end{pmatrix},\ \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right) \,, \tag{12}$$

kus $\mathbf{X}\boldsymbol{\beta}$ ja $\mathbf{X}^*\boldsymbol{\beta}$ on fikseeritud efektid vastavalt $\mathbf{y}$ ja $\mathbf{y}^*$ jaoks ning kovariatsioonimaatriks esitub järgnevalt

$$\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} = \begin{pmatrix} \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I} & \sigma_g^2\mathbf{K}^{*T} \\ \sigma_g^2\mathbf{K}^* & \sigma_g^2\mathbf{K}^{**} + \sigma_e^2\mathbf{I} \end{pmatrix} .$$

Märgime, et praktikas on lihtsam kõigepealt välja arvutada terve kovariatsioonimaatriks $\sigma_g^2\tilde{\mathbf{K}} + \sigma_e^2\tilde{\mathbf{I}}$ ning seejärel eraldada sealt vastavad plokid $\boldsymbol{\Sigma}_{11}$ jne.

Seosest (12) järeldub tinglikustamise abil $(\mathbf{y}^*|\mathbf{y})$ jaotus:

$$\mathbf{y}^*|\mathbf{y} \sim \mathcal{N}\left(\mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),\ \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right) \tag{13}$$

Fenotüüpide $\mathbf{y}^*$ väärtusteks prognoosime selle tingliku jaotuse moodi (normaaljaotuse korral langeb see kokku keskväärtusega). Nii saame $\mathbf{y}^*$ prognoosideks

$$\hat{\mathbf{y}^*} := \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Selle avaldise esimene liidetav koosneb fikseeritud efektidest uute indiviidide jaoks. Teine liige tuleneb suguluse maatriksi poolt tekitatud kovariatsioonistruktuurist. Sellist segamudeli prognoosi nimetatakse ka BLUPiks (*Best Linear Unbiased Prediction*).

1.6.2 Prognoosimine mitmemõõtmelise segamudeli abil

Nüüd vaatleme olukorda, kus $\mathbf{y}_1, \ldots, \mathbf{y}_P$ on vaadeldud fenotüübi väärtused ning $\mathbf{y}_1^*, \ldots, \mathbf{y}_P^*$ need, mida soovime prognoosida. Eeldame, et kõik need fenotüübid järgivad ühiselt mitmemõõtmelist segamudelit (7). Siis saame nende ühisjaotuseks

$$\begin{pmatrix} \mathbf{y}_1 \\ \cdots \\ \mathbf{y}_P \\ \mathbf{y}_1^* \\ \cdots \\ \mathbf{y}_P^* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{X}_1\boldsymbol{\beta}_1 \\ \cdots \\ \mathbf{X}_P\boldsymbol{\beta}_P \\ \mathbf{X}_1^*\boldsymbol{\beta}_1 \\ \cdots \\ \mathbf{X}_P^*\boldsymbol{\beta}_P \end{pmatrix},\ \boldsymbol{\Sigma} \right),$$

kus $\mathbf{X}_j\boldsymbol{\beta}_j$ ja $\mathbf{X}_j^*\boldsymbol{\beta}_j$ on j-nda fenotüübi fikseeritud efektid vastavalt vaadeldud ja uute indiviidide jaoks ning kovariatsioonimaatriks $\boldsymbol{\Sigma}$ on selline, et $\boldsymbol{\Sigma} = \mathbf{C} \otimes \tilde{\mathbf{K}} + \mathbf{S} \otimes \tilde{\mathbf{I}}$, kus $\tilde{\mathbf{K}}$ ja $\tilde{I}$ on mõlemad $(N + N^*)$ rea ning veeruga, nagu eelnevas alapeatükis 1.6.1.

Tingliku jaotuse

$$(\mathbf{y}_1^*, \ldots, \mathbf{y}_P^*)^T|(\mathbf{y}_1, \ldots, \mathbf{y}_P)^T \tag{14}$$

ning selle keskväärtuse saame leida analoogiliselt, nagu leidsime tingliku jaotuse (13) (ptk 1.6.1).

Vahel on meil teada uute N^* indiviidide kohta teada ka mõnede fenotüüpide väärtused. Vaatleme näiteks olukorda, kus meil on teada kõigi teiste fenotüüpide väärtused peale

$\mathbf{y}_j^*$. Siis saame leida tingliku jaotuse

$$\mathbf{y}_j^* | (\mathbf{y}_1, \ldots, \mathbf{y}_P, \mathbf{y}_1^*, \ldots, \mathbf{y}_{j-1}^*, \mathbf{y}_{j+1}^*, \ldots, \mathbf{y}_P^*)^T \tag{15}$$

See käib analoogiliselt eelnevate tinglikustamistega. Kuna sel juhul võtame prognoosimisel arvesse teisi fenotüüpe, nimetame edaspidi (15) abil saadud prognoosi tinglikuks ning (14) marginaalseks.

1.7 Mudelite prognoosivõime hindamine

Erinevate mudelite prognoosivõime võrdlemiseks on kõigepealt vaja need mudelid sobitada, seejärel kasutada neid ennustamiseks ning iseloomustada prognooside täpsust. Kui prognoosida neidsamu väärtuseid, millele sobitasime mudeli, võime saada liiga optimistliku hinnangu ülesobitumise (*overfitting*) tõttu. Seepärast kasutame mudeli sobitamiseks üht osa andmestikust (nn treeningandmed) ning hindame mudeli headust teisel osal (nn testandmed).

Antud töös jagame andmestiku kolmeks (joonis 3): nimetame neid treening-, valideerimis- ja testandmestikuks, proportsioonidega vastavalt 60%, 20%, 20%. Igast mudeliklassist (näiteks ühemõõtmeline segamudel aditiivsete efektidega) parima mudeli leidmiseks kasutame esimesi 60% ja 20% tükke: mudelid sobitame treeningandmetel ning kasutame neid mudeleid valideerimisandmestikul prognoosimiseks. Nii saame leida näiteks viia läbi mudelivalikut ühemõõtmeliste segamudelite jaoks, valides välja täpseimate prognoosidega mudeli.

Joonis 3. Kogu andmestiku jagame osadeks: treeningandmed (60%), valideerimisandmed (20%) ja testandmed (20%).

Eelneva protseduuri tulemusena oleme igast mudeliklassist välja valinud parima konfiguratsiooniga mudeli. Aga valideerimisandmestikul saadud täpsuse hinnangud võivad ikkagi olla ülehinnangud, kuna lõpliku mudeli väljavalimise kriteeriumiks oli maksimaalne täpsus valideerimisandmestikus. Et saada mudelile lõplik täpsuse hinnang, sobitame vastava mudeli uuesti 80% andmetest (treening- ja valideerimisandmetel) ning hindame selle prognooside täpsust testandmestikul.

Mudeli prognoosivõime hindamisel kasutasime väljaspool treeningvalimit arvutatud determinatsioonikordajat

$$R^2 := 1 - \frac{\sum_{i \in \mathcal{I}} (y_i - \hat{y}_i)^2}{\sum_{i \in \mathcal{I}} (y_i - \bar{y})^2},$$

kus indeksite hulk $\mathcal{I}$ indikeerib kas valideerimis- või testandmetesse kuulumist, $\hat{y}_i$ näitab mudeli prognoosi indiviidile i ning $\bar{y}$ keskmist y_i väärtust hulgas $\mathcal{I}$. Niisiis, R^2 näitab, kui suure osa fenotüübi dispersioonist saab seletada antud mudeli abil. Kui edasises räägime prognooside täpsusest, peamegi silmas R^2 väärtust.

Selle näitaja alusel saame mudeleid omavahel võrrelda, aga lisaks tekib ka võrdlusmoment kitsa ja laia pärilikkusega. Kitsas pärilikkus (*narrow-sense heritability*, tähis h^2) näitab, kui suur osa fenotüübi dispersioonist on seletatav geneetiliste aditiivsete efektidega. Lai pärilikkus (*broad-sense heritability, repeatability*, tähis H^2) näitab, kui suur osa fenotüübi dispersioonist on võimalik seletada geneetiliste faktorite abil (sealhulgas nii aditiivsed, dominantsuse kui ka interaktsiooniefektid).

Aditiivsete efektidega mudeli ennustustäpsuse ülemiseks tõkkeks on kitsas pärilikkus ning selle hindamist tutvustasime peatükis 1.1. Eeldades, et laia pärilikkuse poolt seletamata dispersiooni allikas on katsepõhine mõõtmismüra, seab lai pärilikkus teoreetilise piiri, kui täpselt on üleüldse võimalik progoosida fenotüüpi. Laia pärilikkuse hinnangu leidmiseks hindame suurust $(1 - H^2)$, mis näitab, kui suure osa fenotüübi dispersioonist moodustab mõõtmistest tulenev müra.

2 Andmed

Käesolevas töös uurime pärilike tunnuste prognoosimist pagaripärmi näitel. Pagaripärm (*Saccharomyces cerevisiae*) on ainurakne olend, kes pooldub kiiresti erinevates lahustes (näiteks õllevirdes, veinihakatises, tainas). Tema geneetiline informatsioon sisaldub DNA järjestuses, mis jaguneb 16 kromosoomiks. Pärm on selles mõttes eriline organism, et ta saab elada ja kasvada kahes erinevas vormis: *haploidne* (kellel on igast kromosoomist üks koopia) ning *diploidne* (igast kromosoomist kaks koopiat).

Pärmi kasutatakse laialdaselt geneetilise mudelorganismina selle mitmete eeliste tõttu [18], [19]. Esiteks, on võimalik läbi viia suuremahulisi eksperimente suhteliselt väikese maksumusega. Nii saame piisavalt suure valimimahu jaoks nii genotüübi andmed kui ka erinevate fenotüüpide mõõtmistulemused. Teiseks, pärmiga on vajadusel lihtne teha lisaeksperimente. Näiteks võime tahta kontrollida uute indiviidide jaoks prognoositud fenotüübi väärtust või modifitseerida indiviidi genoomi, veendumaks leitud seoste põhjuslikkuses. Meil võib tekkida ka hüpotees, millist fenotüüpi peaksime veel mõõtma, et prognoosida täpsemini huvipakkuvat fenotüüpi.

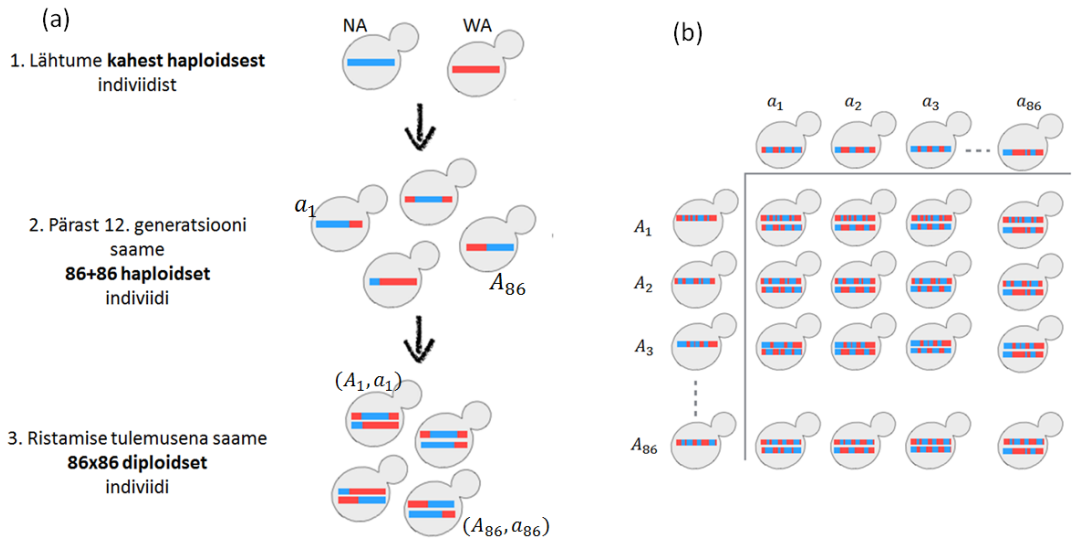
Magistritöö põhineb eksperimendil, kus uuritakse 7396 indiviidist koosnevat pärmipopulatsiooni ning mille lõplikul valmimisel tekib seni suurim sekveneeritud ning fenotüpiseeritud indiviidide kogum. Varasemalt on pärmi fenotüüpide pärilikkust ning prognoosimist uuritud väiksema valimimahu juures näiteks artiklites [20], [21].

Järgnevalt anname ülevaate eksperimendidisainist, st millise protsessi tulemusena saadi genotüübi ja fenotüübi andmed. See aitab aru saada tugeva populatsioonistruktuuri tekkemehhanismist.

2.1 Eksperimendidisain ja genotüübid

Haploidsel pärmil on kaks tüüpi A ja a (alfa), millel on analoogia inimese sooga: A-tüüpi indiviidid saavad paarituda ainult alfa-tüüpi indiviididega.

Uuritav pärmipopulatsioon saadi 86 A-tüüpi haploidse indiviidi $(A_1, \ldots, A_{86})$ ning 86 alfa-tüüpi haploidse indiviidi $(a_1, \ldots, a_{86})$ ristamisel, vt joonis 4 (b). Need 86+86 vanemat on valitud kahe pärmitüve (*North American, NA* ja *West African, WA*) ristandi 12. generatsioonist. Niisiis pärinevad kõik 86×86 diploidset indiviidi kahest haploidsest eellasest NA ja WA, vt joonis 4 (a).



Joonis 4. (a) Kahe haploidse NA ja WA indiviidi 12. generatsioonist valiti välja haploidsed $(A_1, \ldots, A_{86})$ ja $(a_1, \ldots, a_{86})$. Nende ristamisel skeemi (b) järgi saadi diploidsete indiviidide populatsioon. Punased ning sinised fragmendid tähistavad vastavalt WA ja NA eellastelt pärit olevaid alleele. (b) Ristamise skeem: haploidsete indiviidide A_i ja a_j paarikaupa ristamisel saame diploidse järglase (A_i, a_j). Joonis on kohandatud Gianni Liti loodud graafikust.

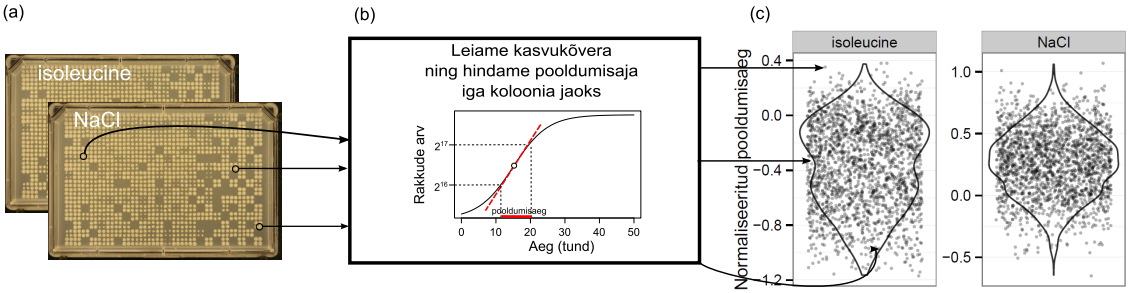
Sellise katsedisaini korral ei ole 7396 indiviidi genotüübi saamiseks vaja neid kõiki sekveneerida. Piisab, kui sekveneerida nende 86+86 haploidset vanemat, sest vanemate genotüübi põhjal saame konstrueerida diploidsete järglaste genotüübid. Näiteks kui vanemal A_i on üks genoomi positsioon pärit WA eellaselt ning teisel vanemal a_j on seesama positsioon pärit NA eellaselt, teame, et indiviidil (A_i, a_j) on selle koha peal üks WA ja üks NA alleel. Kasutades peatükis 1 tutvustatud kodeeringut, konstrueerisime genotüübi maatriksi $\mathbf{X} \in \{0, 1, 2\}^{N \times M}$, kus $M \approx 10000$ on unikaalsete SNPde arv.

2.2 Fenotüübid

Uuritavad fenotüübi andmed kirjeldavad pärmi kasvukiirust erinevates keskkondades: *glycine, isoleucine, NaCl, rapamycin, galactose, hydroxyurea, phleomycine*. Need keskkonnad on valitud nii, et nad kataksid erinevad energiaallikad (näiteks *galactose*), osmootilise stressi (*NaCl*) ja vähiravimid (*rapamycin*).

Et mõõta pärmirakkude pooldumisaega mingis keskkonnas, paigutati pärmi isendid laboris plaadile ning plaati töödeldi vastava keemilise ühendi lahusega. Plaadil on 32×48 ruudustik, mille igas positsioonis on üks pärmikoloonia (joonis 5 (a)). Jälgides iga koloonia suuruse muutumist ajas, saame eraldada iga ajahetke jaoks koloonia rakkude arvu

ning neile andmetele sobitada kasvukõvera (joonis 5 (b)). Fenotüübi väärtuseks võtame kõige kiiremale kasvu hetkele vastava pooldumisaja. See tähendab, et leiame ajahetke, millal kasvukõvera puutuja tõus on maksimaalne, ning arvutame sellele kasvukiirusele vastava aja, mis kulub koloonia rakkude arvu kahekordistumiseks. Normaliseeritud fenotüübi (joonis 5 (c)) saame, kui rakendame logaritmitud pooldumisaegadele normaliseerimise töövoogu, mis eemaldab plaadi mõju jms.



Joonis 5. Fenotüüpide genereerimise protsess. Pärmirakud kasvavad plaatidel (a), kus näeme iga koloonia pindala. Iga ajahetke jaoks arvutame koloonia pindala põhjal koloonia rakkude arvu. Sobitame iga indiviidi jaoks kasvukõvera (b), et hinnata pooldumisaeg, mis vastab ajahetkele, mil kasvukõvera tõus on maksimaalne. Arvutades nii pooldumisaja kõigi indiviidide jaoks, saame fenotüüpide jaotused (c).

Mõõtmised, millest magistritöös lähtume, on läbinud eelnevalt kirjeldatud normaliseerimise töövoo. Meil on kasutada iga indiviidi kohta normaliseeritud pooldumisaegade 4 kordusmõõtmist. Tähistame i-nda indiviidi fenotüübi mõõtmistulemused $y_{i1}, \ldots, y_{i4}$. Eeldades, et tegelik fenotüübi väärtus sel indiviidil on μ_i ning et mõõtmisvead on normaaljaotusest dispersiooniga σ_i^2, saame

$$y_{i1}, \ldots, y_{i4} \sim \mathcal{N}(\mu_i, \sigma_i^2).$$

Kuna soovime hinnata indiviidide fenotüübi tegelikku väärtust μ_i, defineerime fenotüübi y_i, mida edasises modelleerima hakkame, kui mõõtmiste keskmise

$$y_i := \frac{1}{4}(y_{i1} + y_{i2} + y_{i3} + y_{i4}).$$

Siis y_i jaoks kehtib

$$y_i \sim \mathcal{N}(\mu_i, \frac{1}{4}\sigma_i^2).$$

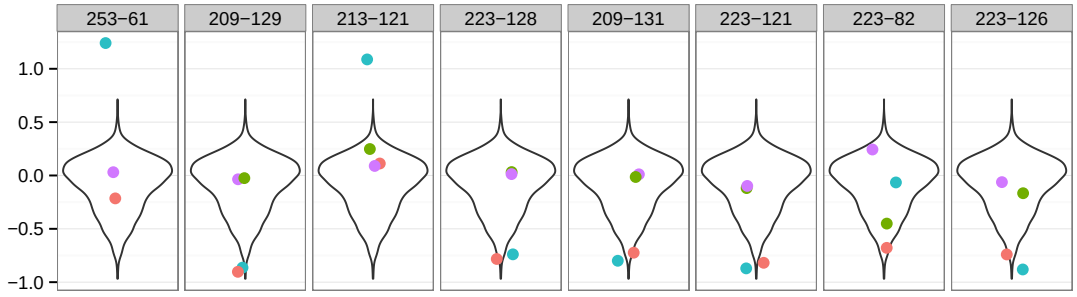
Et avaldada μ_i üldkeskmise μ kaudu, eeldame lisaks, et kehtib $\mu_i \sim \mathcal{N}(\mu, \sigma^2)$, millest järeldub

$$y_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2 + \sigma_i^2}{4}\right), \tag{16}$$

kus kogu dispersioon on jagatud kaheks komponendiks: üks osa näitab kui varieeruvad on indiviidide tegelikud fenotüübid μ_i ning teine osa näitab mõõtmisest tulenevat ebatäpsust.

2.3 Andmete puhastamine

Enne kordusmõõtmiste keskmistamist indiviidi kaupa, soovisime veenduda, et mõõtmised on piisavalt usaldusväärsed. Iga indiviidi 4 vaatluse põhjal saame hinnata y_i mõõtmisvea dispersiooni σ_i^2. On selge, et mida suurem see on, seda ebatäpsem on indiviidi fenotüübi hinnang. Et probleemi olemasolust ülevaadet saada, visualiseerisime iga keskkonna kõige suurema dispersiooniga indiviidide mõõtmisi (näide joonisel 6).



Joonis 6. Kõige suurema mõõtmistulemuste variatsiooniga indiviidid keskkonnas *phleomycine*, y-teljel on normaliseeritud pooldumisaeg. Viiuldiagramm näitab fenotüübi jaotust üle kõigi indiviidide, värvilised täpid näitavad antud indiviidi (näiteks indiviidi 253-61) kordusmõõtmiste väärtuseid.

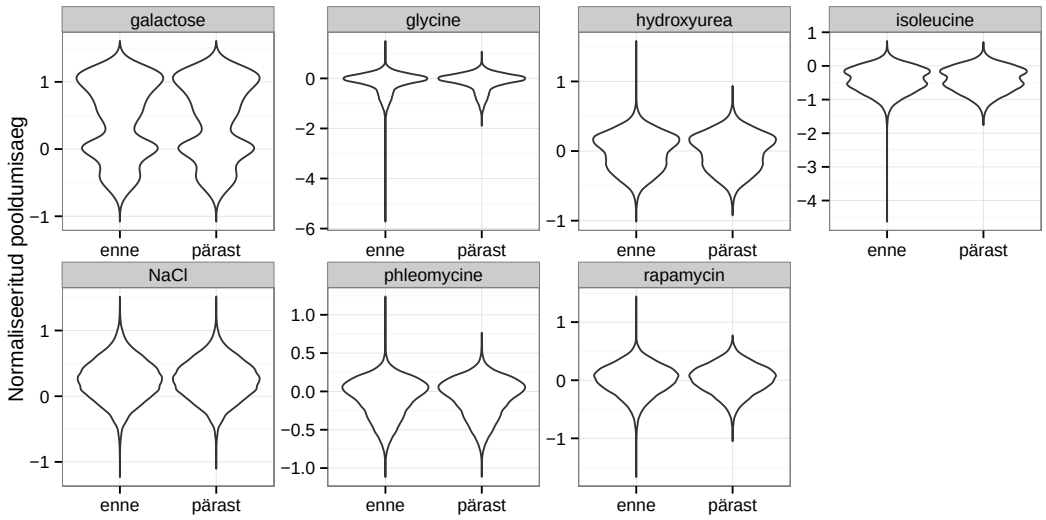
Kui ühe indiviidi mõõtmiste ulatus katab ära suurema osa fenotüübi jaotusest, tekib kahtlus, kuivõrd usaldusväärsed need on. Et fikseerida kriteerium, mille järgi suure variatsiooniga indiviide välja filtreerida, otsustasime kasutada järgnevat rusikareeglit: järjestades indiviidid kordusmõõtmiste dispersiooni kasvamise järgi, leiame optimaalse n nii, et indiviidide $\{1, \ldots, n\}$ pealt hinnatud keskmise fenotüübi hinnang $\frac{1}{n}\sum_{i=1}^{n} y_i$ oleks võimalikult täpne (st võimalikult väikese dispersiooniga). Arvestades seost (16), seisneb see järgmise avaldise minimeerimises

$$\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Var}\left(y_i\right) = \frac{1}{4n^2}\left(n\sigma^2 + \sum_{i=1}^{n}\sigma_i^2\right).$$

Üldiselt tagab suurem n täpsema keskmise hinnangu, aga väga ebatäpsete mõõtmistulemuste lisamine enam täpsust ei suurenda, vaid toob kaasa müra. Eelneva rusikareegli

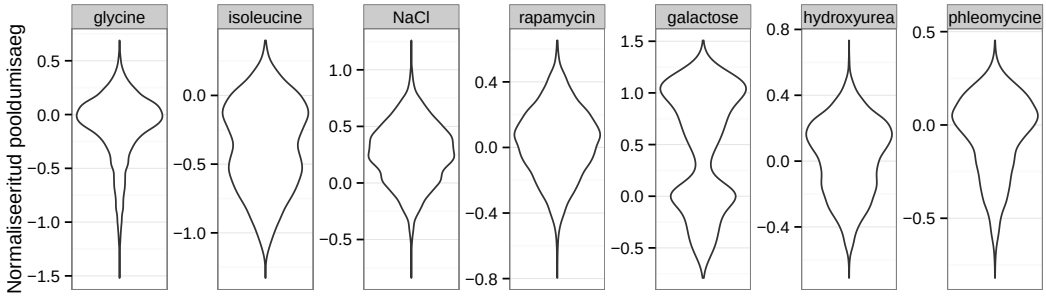
idee on leida teatas mõttes optimaalne n, millest alates toob vaatluste arvu suurendamine kaasa kaotuse täpsuses.

Rusikareegli rakendamisel jäi erinevatest keskkondadest välja 8 kuni 475 indiviidi. Nii vabanesime erinditest, mis olid märgatavalt aeglasema või kiirema kasvuga kui ülejäänud (joonis 7).



Joonis 7. Fenotüüpide kordusmõõtmiste jaotused enne ja pärast erindite eemaldamist.

Seejärel keskmistasime väärtused kõigis keskkondades indiviidide kaupa ning jätsime alles ainult need indiviidid, kelle kohta olid olemas kõigi fenotüüpide väärtused. Nii saime fenotüüpide lõplikud jaotused (joonis 8), mida hakkame edasises modelleerima. Kuna indiviidide arv on suhteliselt suur, ning me sobitame mitmeid erinevaid mudeleid kõigile fenotüüpidele, otsustasime mõistliku arvutusaja huvides kogu järgnevas töös kasutada juhuslikku 2000 indiviidi suurust alamhulka kõigist andmetest.



Joonis 8. Lõplike fenotüüpide jaotused, mis on saadud pärast andmete puhastamist ja keskmistamist.

3 Tulemused

Peatükis 1 andsime ülevaate mudelitest, mida saame kasutada fenotüübi prognoosimiseks, ning peatükis 2 tutvusime pärmipopulatsiooniga, mille fenotüüpe hakkame modelleerima. Järgnevalt rakendame neid mudeleid andmetele, et võrrelda nende ennustustäpsust ning uurida, kuivõrd hästi suudame prognoosida fenotüüpi, võrreldes pärilikkusel põhineva teoreetilise piiriga. Vaatleme järgnevaid mudeliklasse: lineaarsed segamudelid, mitmemõõtmelised lineaarsed segamudelid ning juhuslikud segametsad. Viimases alapeatükis (ptk 3.5) tutvustame metoodilist edasiarendust: pakume prognoosimiseks välja uue meetodi, mis üldistab juhusliku segametsa algoritmi.

3.1 Lineaarsed segamudelid

Prognoosimist alustame lineaarsete segamudelite abil: iga fenotüübi jaoks sobitame mudeleid eraldi. Lisame järk-järgult mudelisse $0, 1, \ldots, 25$ SNPi, kasutades edaspidist mudelivalikut (ptk 1.3), ja uurime, milline arv argumente tagab täpseimad prognoosid valideerimisandmestikul (ptk 1.7). Mudelite sobitamisel vaatleme kahte eraldi klassi: ainult aditiivsete efektidega (ptk 1) mudelid ning sellised, mis võivad sisaldada ka tunnuste koosmõjusid. Viimasel juhul lisame kandidaattunnuste sekka interaktsioonid kõigi varasemalt mudelisse lisatud argumentidega. Muuhulgas lubame ka tunnuse interaktsiooni iseendaga, mis võimaldab modelleerida mitteaditiivset (dominantsuse/retsessiivsuse) efekti.

Sobitasime erinevate argumentide arvuga segamudeleid ning hindasime iga mudeli prognooside täpsust valideerimisandmestikul R^2 abil (joonis 9). Kõigis keskkondades saime interaktsiooniefekte sisaldavate mudelitega täpsemad prognoosid kui aditiivsete mudelite abil (arvulised väärtused tabelis 1).

Tabel 1. Parimate lineaarsete segamudelite prognooside täpsus valideerimisandmetel (jooniselt 9 valitud suurima täpsusega mudelid klassidest LMM (+) ja LMM(*)).

	glycine	isoleucine	NaCl	rapamycin	galactose	hydroxyurea	phleomycine
(+)	0.66	0.66	0.78	0.75	0.83	0.71	0.76
(*)	0.69	0.73	0.79	0.77	0.93	0.77	0.82

Kuigi peatükis 1.1 arutlesime, miks populatsioonistruktuuri olemasolul ei ole õigustatud tavalise lineaarse mudeli kasutamine, võrdlesime siiski selle prognoosivõimet segamudelite omaga. Näeme, et väikese arvu argumentide korral on ennustustäpsus madal, aga näiteks 25 tunnusega mudel on juba võrreldav segamudelitega (näiteks keskkonnas

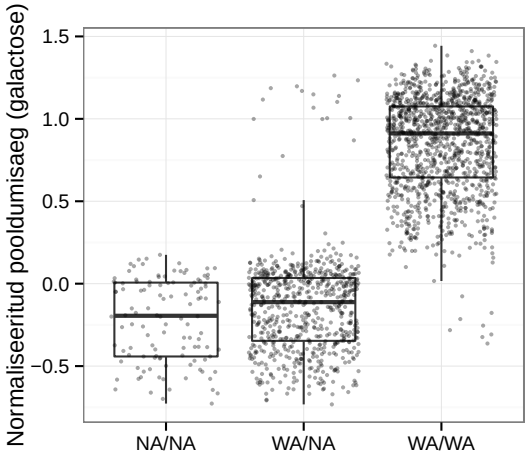
Joonis 9. Lineaarsete mudelite (LM) ja lineaarsete segamudelite (LMM) prognooside täpsus valideerimisandmetel. (+) tähistab aditiivsete efektidega ja (*) interaktsiooniefektidega mudelit. Mudeli nimele järgnev arv näitab fikseeritud efektide arvu. Näiteks "LMM (*) 5" tähendab, et LMM mudelis oli 5 aditiivset või interaktsiooniefekti. Katkendjoontega on näidatud kitsas (vasakpoolne) ja lai pärilikkus (parempoolne).

glycine on parimad LM ja LMM täpsused vastavalt 0.57 ja 0.69). See ei ole üllatav, kuna suur arv argumente fikseeritud efektidena peaks lähendama segamudelis kasutatavat suguluse maatriksit $\mathbf{K}$.

Segamudeli prognoositäpsus on pärilikkuse lähedal juba väikese arvu fikseeritud efektidega. Mõnes keskkonnas (näiteks *glycine, rapamycin*) on aditiivsete efektidega segamu-

delite seas parimaks mudeliks selline, mis sisaldab ainult vabaliiget. See on mõnevõrra üllatav, et ainuüksi geneetilisel sugulusel põhinevad juhuslike efektide prognoosid võimaldavad nii hästi prognoosida ning et ühegi konkreetse SNPi lisamine ei pruugi võimaldada täpsemat tulemust.

Keskkond *galactose* on teistest erinev selle poolest, et juba ühe argumendiga lineaarne mudel saavutab suhteliselt hea prognoositäpsuse (0.67), sest ühe SNPi põhjal (joonis 10) saame eristada, kas tegu on pigem aeglase- või kiirekasvulise indiviidiga (see fenotüüp on bimodaalse jaotusega). Segamudelite seas toimub täpsuse suurenemine teise interaktsiooniefekti lisamisel (enne 0.83, pärast 0.93, mis ületab kitsa pärilikkuse piiri). Selgub, et tegu on mudelisse esimesena lisatud SNPi interaktsiooniga iseendaga: joonisel 10 näeme tõepoolest, et efekt on mitteaditiivne, sest aeglaselt kasvavad ainult need indiviidid, kellel on antud positsioonis WA/WA genotüüp.

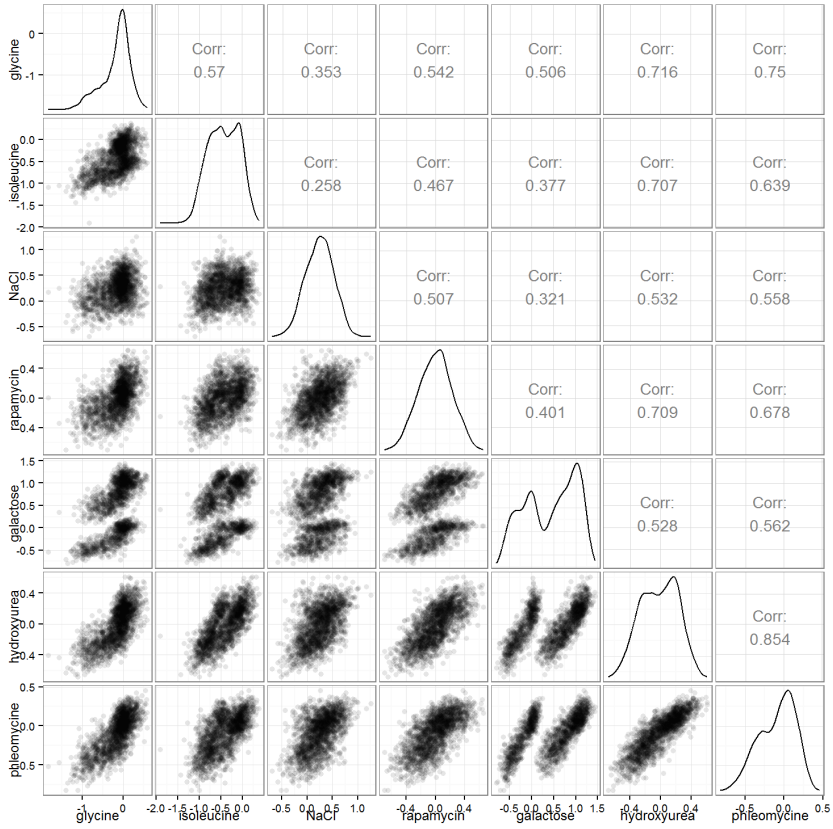


Joonis 10. Keskkonnas *galactose* kõige tugevama seosega SNP (kromosoom IV, positsioon 463889. Iga genotüübi kohta (x-teljel on WA alleelide arv kas 0, 1 või 2) on näidatud *galactose* keskkonnas normaliseeritud pooldumisaja jaotus karpdiagrammi abil. Üksikud vaatlused on tähistatud punktikestega.

Alates teatud arvu fikseeritud efektide lisamisest hakkab mudeli ennustustäpsus langema. See indikeerib ülesobitumist: kuigi treeningandmetel saime viimase argumendi lisamisel paremini sobiva mudeli, siis see seos enam ei üldistu treeningvalimist väljapoole. Iga keskkonna jaoks saame leida optimaalse mudeli, st sellise, mis saavutab maksimaalse R^2 väärtuse valideerimisandmestikul.

3.2 Mitmemõõtmelised lineaarsed segamudelid

Eelmises alapeatükis vaatlesime fenotüüpe teineteisest sõltumatult ning leidsime iga fenotüübi jaoks parima mudeli. Aga arvestades, et fenotüübid on omavahel korreleeritud (joonis 11), võime nendevahelise korrelatsiooni modelleerimisel saada täpsemaid prognoose. Selleks kasutame mitmemõõtmelist lineaarset segamudelit (*Multi-Trait Linear Mixed Model*, MT LMM), mis hindab kõigi fenotüüpide ühisjaotuse.



Joonis 11. Fenotüüpidevahelised seosed. Punktipilved indikeerivad, et kasvukiirus erinevates keskkondades on positiivselt korreleeritud. Tugevaim seos on *hydroxyurea* ja *phleomycine* vahel.

Iga fenotüübi jaoks lisame fikseeritud efektideks parajasti need argumendid, mida kasutasime selle tunnuse jaoks täpseimate prognoosidega mudelis. Eristame kolme tüüpi mitmemõõtmelisi segamudeleid: ainult vabaliikmega mudel, aditiivsete efektidega mudel ning interaktsioone sisaldav mudel. Lisaks fikseeritud efektidele sisaldavad kõik mudelid ka geneetilisel sarnasusel põhinevat juhuslikku efekti.

3.2.1 Mitmemõõtmelised prognoosid

Sobitasime andmetele kolm eelnevalt mainitud mudelit ning kasutasime neid prognoosimiseks. Kuigi prognoositavaks on fenotüüpide ühisjaotus, vaatleme ennustustäpsust iga fenotüübi kohta eraldi.

Näeme, et mitmemõõtmelise mudeli kasutamine ei anna suurt võitu ennustustäpsuses võrreldes parimate LMM mudelitega (tabel 2). Aga näiteks vabaliikmega MT LMM prognoosid on täpsemad võrreldes vabaliiget sisaldavate segamudeliga (*glycine* korral 0.69 ja 0.66). Märkame, et mitmemõõtmelises mudelis võime fikseeritud efektide lisamisel täpsuses kaotada, mis võib indikeerida ülesobitumist (aditiivsete efektidega MT LMM prognoositäpsus jääb kõigis keskkondades alla vabaliiget sisaldavale MT LMM mudelile).

Tabel 2. Mitmemõõtmeliste segamudelite prognooside täpsus valideerimisandmetel: ainult vabaliikmega (0), aditiivsete efektidega (+) ning interaktsiooniefektidega (*) mudel.

	glycine	isoleucine	NaCl	rapamycin	galactose	hydroxyurea	phleomycine
(0)	0.69	0.72	0.79	0.76	0.93	0.77	0.81
(+)	0.66	0.66	0.78	0.75	0.83	0.71	0.76
(*)	0.69	0.72	0.79	0.76	0.93	0.77	0.81

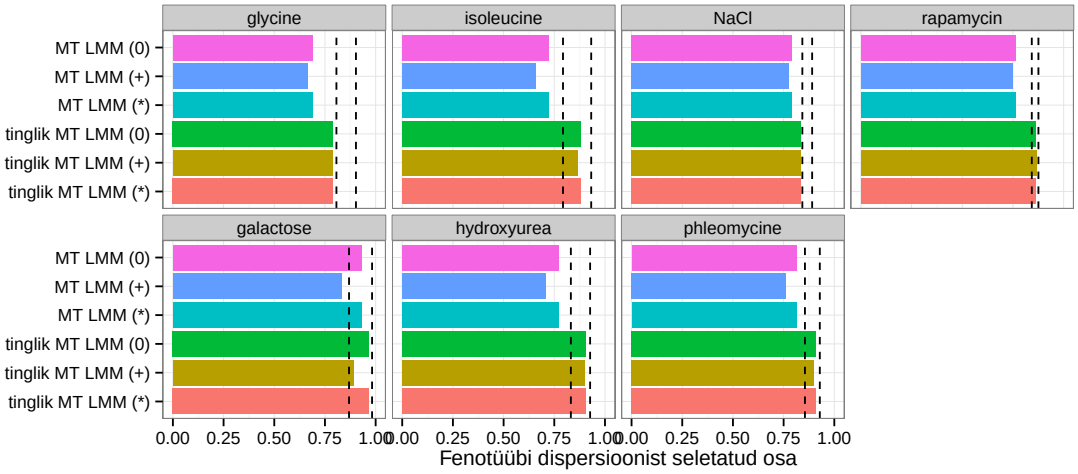
3.2.2 Tinglikud prognoosid

Siiani oleme vaadelnud mudeleid, mille fenotüübi prognoosid põhinevad genotüübil. Juhul, kui meil on teada mõne teise fenotüübi väärtused, võime saada täpsemaid prognoose. Uurimaks, kas ja kui palju täpsust see lisab, leidsime MT LMM mudeli tinglikud prognoosid iga fenotüübi jaoks tingimusel, et kõigi teiste fenotüüpide väärtused on teada (tabel 3). Näeme, et tinglikustamine on kõigis keskkondades märgatavalt suurendanud prognooside täpsust.

Tabel 3. Tinglikustatud mitmemõõtmeliste segamudelite prognooside täpsus valideerimisandmetel: ainult vabaliikmega (0), aditiivsete efektidega (+) ning interaktsiooniefektidega (*) mudel.

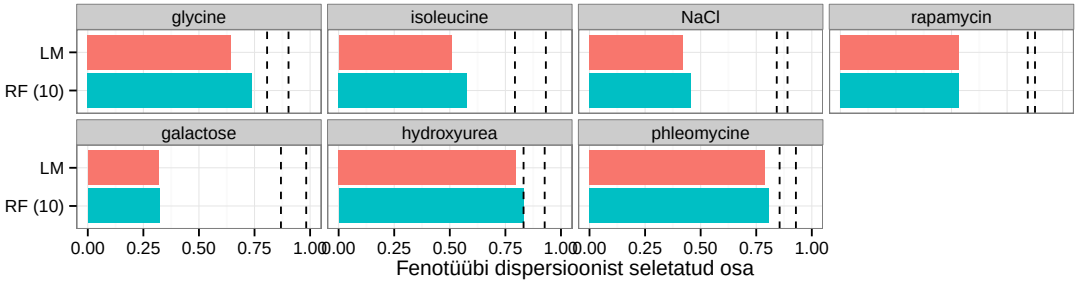
	glycine	isoleucine	NaCl	rapamycin	galactose	hydroxyurea	phleomycine
(0)	0.79	0.88	0.84	0.86	0.97	0.90	0.91
(+)	0.79	0.87	0.83	0.87	0.89	0.90	0.90
(*)	0.79	0.88	0.84	0.86	0.97	0.90	0.91

On märkimisväärne, et enamikes keskkondades ületab tinglike prognooside täpsus kitsast pärilikkust ning mõnes keskkonnas (*galactose, phleomycine*) on peaaegu võrdne laia pärilikkuse hinnanguga (joonis 12).



Joonis 12. Mitmemõõtmeliste segamudelite (MT LMM) prognooside täpsus valideerimisandmetel: ainult vabaliikmega (0), aditiivsete efektidega (+) ning interaktsiooniefektidega (*) mudel. Märksõna "tinglik" näitab, et prognoosimisel on võetud arvesse kõiki teisi fenotüüpe, st tinglikustatud on ülejäänud fenotüüpide järgi. Katkendjoontega on näidatud kitsas (vasakpoolne) ja lai pärilikkus (parempoolne).

Arvestades, et teiste fenotüüpide teadmine võimaldab täpsemaid prognoose, tekib loomulik küsimus, kas üldse on vaja genotüüpi teada või saaks sama täpselt ennustada teiste fenotüüpide abil. Selleks kasutasime tavalist lineaarset mudelit ning juhuslikku metsa, ilma populatsioonistruktuuri arvesse võtmata (imiteerimaks olukorda, kus meil genotüübi andmeid pole). Nägime, et enamikes keskkondades oli prognoositäpsus märgatavalt madalam võrreldes genotüübil põhinevate mudelitega (joonis 13), aga näiteks *hydroxyurea* korral saavutasime kitsa pärilikkusega võrreldava täpsuse.

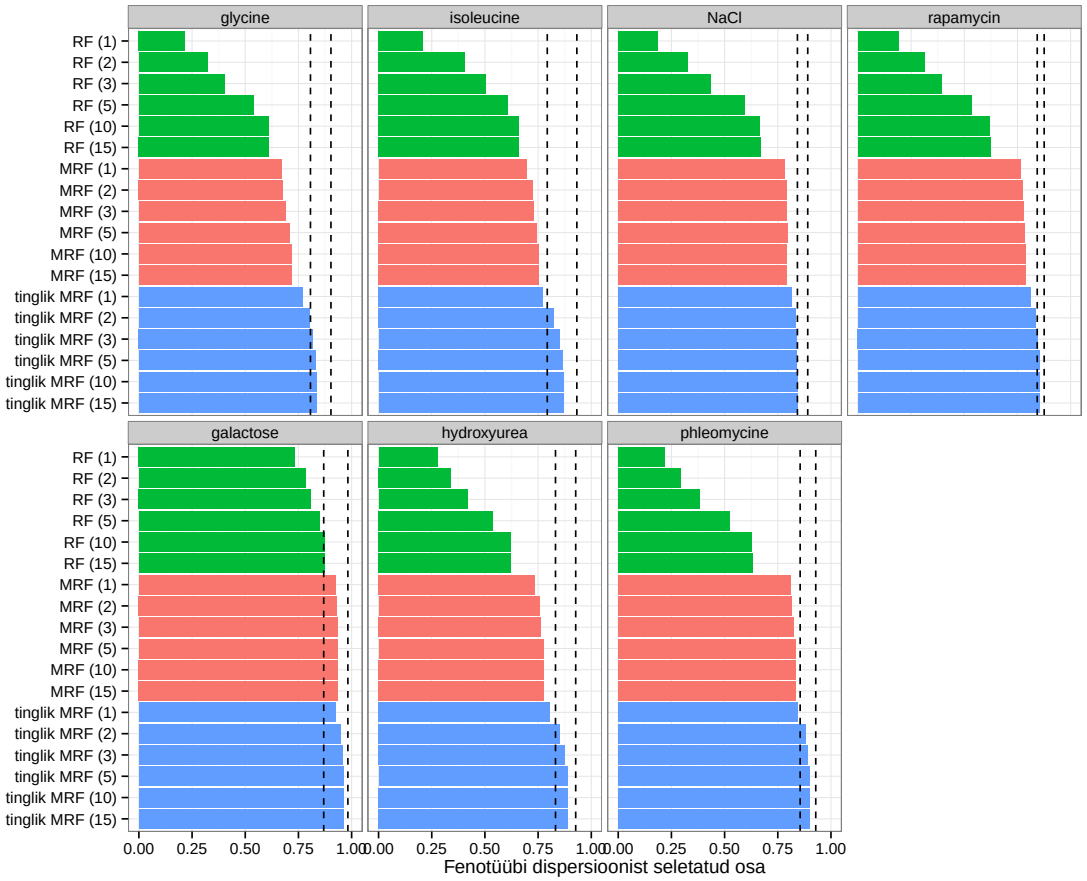


Joonis 13. Ainult teistel fenotüüpidel põhinevate mudelite prognooside täpsus valideerimisandmetel: lineaarne mudel (LM) ning juhuslik mets (RF) kõrgusega 10.

Parima täpsuse saavutame siiski geno- ja fenotüübi mõõtmiste kombineerimisel (tingliku MT LMM abil). Teisisõnu, teiste tunnuste kasutamine võimaldab täpsemalt prognoosida, aga ainult nendest ei piisa.

3.3 Juhuslikud segametsad

Järgnevalt uurime, milline on juhusliku segametsa prognoositäpsus võrreldes ühe- ja mitmemõõtmeliste segamudelitega. Oma olemuselt sobib regressioonipuu hästi mittelineaarsete efektide arvessevõtmiseks, näiteks dominantsuse efekti saame modelleerida ühe hargnemise abil, aga lineaarsete mudelite korral oleks vaja mudelisse lisada kaks liiget ning hinnata kaks kordajat. Samas võib arvata, et juhuslik segamets ei ole efektiivne tugevate lineaarsete efektide korral.



Joonis 14. Juhuslike metsade (RF), juhuslike segametsade (MRF) ning sellist segametsade, kus argumentidena kasutatakse lisaks genotüübile ka teisi fenotüüpe (tinglik MRF), prognooside täpsus valideerimisandmetel. Sulgudes olev arv näitab puude kõrgust metsas.

Juhusliku segametsa algoritmi rakendamiseks on kasutajal vaja teha mitmeid valikuid: puude arv metsas, puude kõrgus ning kui suurt osa SNPidest kasutame tippudes argumenttunnustena puu kasvatamise käigus. Puu kõrgus määrab teatud mõttes mudeli keerukuse: suurem arv hargnemisi võimaldab modelleerida keerukamat sõltuvusstruktuuri. Otsustasime sobitada juhuslikke segametsi erinevate kõrgustega $1, \ldots, 15$. Ülejäänud

kahe näitaja jaoks fikseerisime parameetrite väärtused: metsas 100 puud ning igas tipus kasutame kandidaattunnustena juhuslikku 66% suurust alamhulka kõigist tunnustest.

Nagu segamudelitegi korral, sobitasime võrdlustaseme saamiseks kõigepealt populatsioonistruktuuri mitte arvesse võtva tavalise juhusliku metsa. Madalate puude korral on selle prognoosivõime märgatavalt kehvem kui juhuslikul segametsal (joonis 14) ning 15 hargnemisest jääb väheks, et tulemus oleks võrreldav kõige lihtsama segametsaga. Segametsade prognoositäpsus kasvab enamikes keskkondades algul, aga hiljemalt kõrguse 10 korral oleme jõudnud tasemele, kust edasi täpsus ei suurene.

Võrreldes ainult genotüübil põhinevaid mudeleid, on parimate juhuslike segametsade (tabel 4) prognoositäpsus on võrreldav nii ühe- kui ka mitmemõõtmeliste segamudelite omaga. Aga mõnes keskkonas, näiteks *phleomycine* korral on juhusliku segametsa prognoosid täpsemad kui MT LMM omad (R^2 vastavalt 0.83 ja 0.81).

Kuna mitmemõõtmeliste mudelite prognoositäpsus suurenes tinglikustamisel ülejäänud fenotüüpide väärtustele, uurisime, millist võitu annaks teistes keskkondades tehtud mõõtmiste arvessevõtmine juhusliku segametsa korral. Selleks lisasime need fenotüübid argumenttunnuste sekka. Parimad tulemused on enamike keskkondade korral võrreldavad tinglike MT LMM mudelitega (tabel 3) ning ei saa öelda, et üks meetod annaks ühtlaselt paremaid tulemusi kui teine. Näiteks *glycine* keskkonnas olid täpsused segametsa ja MT LMM jaoks vastavalt 0.84 ja 0.79, aga *galactose* korral kummalgi 0.96 ja 0.96.

Tabel 4. Parima juhusliku metsa (RF), juhusliku segametsa (MRF) ning teisi fenotüüpe modelleeriva juhusliku segametsa (t MRF) prognooside täpsus valideerimisandmetel.

	glycine	isoleucine	NaCl	rapamycin	galactose	hydroxyurea	phleomycine
RF	0.61	0.66	0.67	0.62	0.87	0.62	0.63
MRF	0.72	0.75	0.79	0.79	0.94	0.78	0.84
t MRF	0.84	0.87	0.84	0.85	0.96	0.89	0.90

3.4 Mudeliklasside võrdlus testandmetel

Igast mudeliklassist oleme kõigi keskkondade jaoks välja valinud ennustustäpsuse poolest parima mudeli. Sobitasime vastava konfiguratsiooniga mudelid nüüd uuesti 80% andmestikust (treeningandmete ja valideerimisandmete ühendil), et hinnata lõplik täpsus testandmetel.

Kui vaatleme ainult genotüübil põhinevaid mudeleid (tabel 5), on interaktsioone sisaldavad mudelid täpsemad kui ainult aditiivsete efektidega, kusjuures LMM ja MT LMM

annavad enam-vähem sama täpsed prognoosid. Kuigi ei eristu ühte mudelit, mis oleks kõigis keskkondades teistest parem, näeme, et juhuslik segamets annab enamikes keskkondades lineaarsetest segamudelitest täpsema tulemuse ning ei ole üheski märgatavalt halvem. Seega kui tuleks valida edaspidiseks kasutamiseks üks mudeliklass, tasuks eelistada juhuslikku segametsa.

Tabel 5. Genotüübil põhinevate mudeliklasside võrdlus testandmetel. Lineaarne segamudel (LMM), mitmemõõtmeline lineaarne segamudel (MT LMM), juhuslik segamets (RMF). (+) tähistab aditiivseid efekte ning (*) interaktsiooniefekte.

	glycine	isoleuc.	NaCl	rapam.	galactose	hydroxyurea	phleom.
LMM(+)	0.68	0.69	0.76	0.75	0.79	0.67	0.75
LMM(*)	0.71	0.76	0.78	0.77	0.92	0.77	0.82
MT LMM(+)	0.68	0.68	0.76	0.75	0.80	0.67	0.75
MT LMM(*)	0.71	0.76	0.78	0.76	0.92	0.74	0.81
MRF	0.74	0.78	0.79	0.80	0.92	0.76	0.85

Võrreldes tinglikke mudeleid, kus oleme lisaks genotüübile arvesse võtnud ka kõiki teisi fenotüüpe (tabel 6), on mõnes keskkonnas juhuslik segamets täpsem kui MT LMM ning mõnes vastupidi (näiteks *glycine* korral on prognooside täpsused vastavalt 0.85 ja 0.81, aga *galactose* korral 0.94 ja 0.96). Üldiselt on kõik näitajad on väga lähedased eelmistes peatükkides nähtuile.

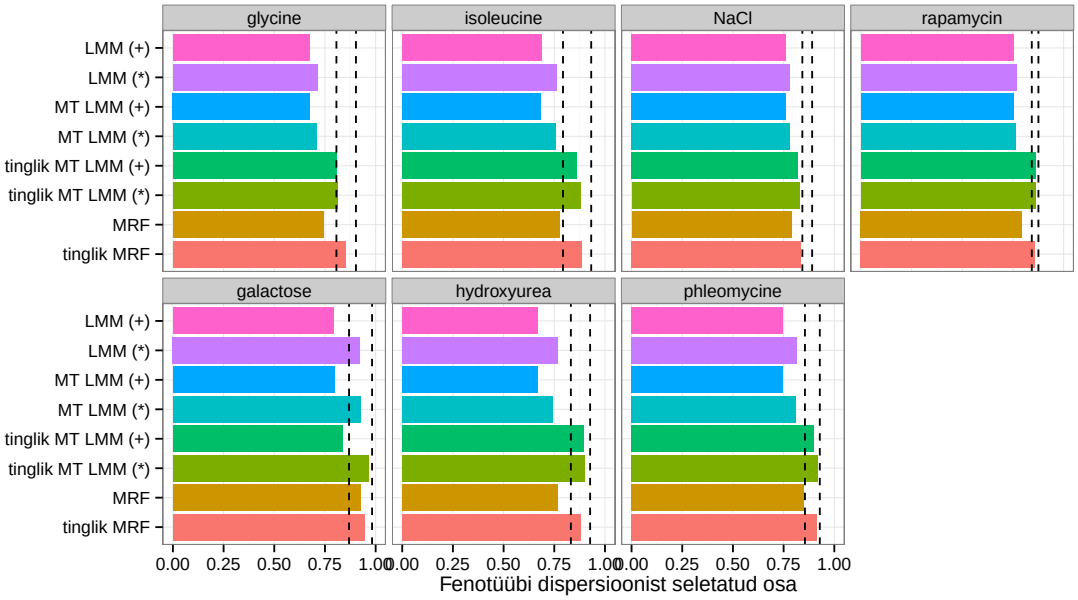
Tabel 6. Tinglike mudelite prognoositäpsuste võrdlus testandmetel. Aditiivsete (+) ja interaktsiooniefektidega (*) mitmemõõtmeline segamudel (MT LMM) ning juhuslik segamets (MRF).

	glycine	isoleuc.	NaCl	rapam.	galactose	hydroxyurea	phleom.
t MT LMM (+)	0.81	0.86	0.82	0.86	0.84	0.89	0.90
t MT LMM (*)	0.81	0.88	0.83	0.86	0.96	0.90	0.92
t MRF	0.85	0.89	0.84	0.86	0.94	0.88	0.92

Kõigi testandmetel sobitatud mudelite tulemused koos kitsa ja laia pärilikkuse hinnangutega on näidatud joonisel 15. Kitsa pärilikkuse piiri võiksid idee poolest ületada nii interaktsioonidega mudelid kui ka teistel fenotüüpidel põhinevad mudelid. Keskkonnas *galactose* õnnestus see piir ületada koosmõjudega segamudeli abil.

Kasutades mudelis teiste fenotüüpide väärtuseid, suureneb prognooside täpsust märgatavalt: tinglike mudelite abil ületasime kitsa pärilikkuse piiri kõigis keskkondades peale *NaCl*. Arvestades, et laiast pärilikkusest üle jääva variatsiooni allikas on katsepõhine mõõtmismüra, pole sellest näitajast paremini võimalik prognoosida.

Saavutatud prognoositäpsust illustreerib ka joonis 16. Võttes arvesse ainult genotüüpi, saavutame kõigis keskkondades 92% kitsast pärilikkusest (tabel 7). Kui kasutame prognoosimisel ka teiste fenotüüpide väärtuseid, saavutame kõigis keskkondades vähemalt



Joonis 15. Igast mudeliklassist valitud parima mudeli prognooside täpsus testandmetel. Lineaarne segamudel (LMM), mitmemõõtmeline lineaarne segamudel (MT LMM), juhuslik segamets (RMF). (+) tähistab aditiivseid efekte ning (*) interaktsiooniefekte. Märksõna "tinglik" tähendab, et prognoosimisel kasutati kõiki teisi fenotüüpe.

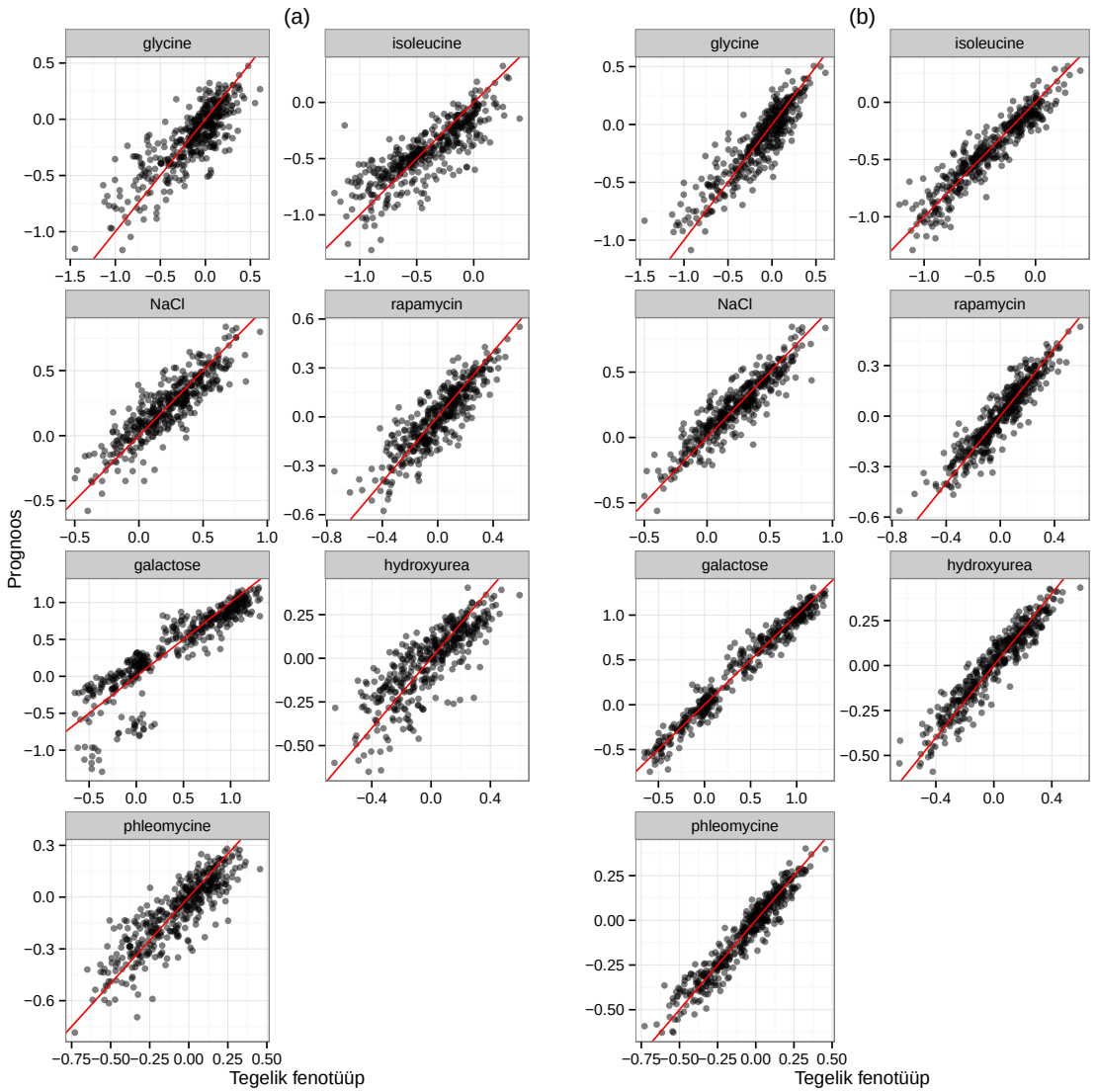
Tabel 7. Kitsast pärilikkusest parima aditiivse (+) või interaktsioonidega (*) mudeli poolt seletatud osa.

	glycine	isoleucine	NaCl	rapamycin	galactose	hydroxyurea	phleomycine
(+)	0.84	0.87	0.91	0.89	0.92	0.80	0.88
(*)	0.92	0.98	0.94	0.94	1.06	0.92	0.99

Tabel 8. Laiast pärilikkusest parima tingliku mudeli poolt seletatud osa.

glycine	isoleucine	NaCl	rapamycin	galactose	hydroxyurea	phleomycine
0.94	0.95	0.94	0.98	0.98	0.97	0.99

94% laiast pärilikkusest (tabel 8). Selline prognoositäpsus on silmapaistev: näiteks artiklis [21] suudeti prognooside abil seletada 72% kuni 100% kitsast pärilikkusest, aga seda piiri ei ületatud.



Joonis 16. Tegeliku fenotüübi (x-teljel) ja prognoositud väärtuste (y-teljel) hajuvusdiagrammid: Mudelid (a) LMM (+) ning (b) tinglik MT LMM (*) on valitud jooniselt 15. Ideaalsed prognoosid asuksid punasel joonel $y = x$.

3.5 Juhusliku segametsa edasiarendus

Oleme näinud kahte erinevat lähenemist prognoosimisele: lineaarne segamudel on efektiivne aditiivsete efektide arvessevõtmisel, aga juhusliku segametsa algoritmis modelleeritakse puude abil mittelineaarseid sõltuvusi. Kuna andmetes võib esineda nii üht kui ka teist tüüpi seoseid, tekib küsimus, kas neid kahte lähenemist ei saaks kombineerida. Juhuslik segamets on iseenesest paindlik, kuna võimaldab komplekssete koosmõjude modelleerimist, aga lineaarse efekti jaoks pole binaarne kaheksjagamine optimaalne. Seetõttu pakume välja juhusliku segametsa algoritmi täiustatud versiooni.

3.5.1 Meetodi kirjeldus

Täiustatud juhuslik segamets kasutab analoogilisi otsustuspuid, nagu peatükis 1.4 kirjeldatud, ainult et puu tippudes lubame binaarse hargnemise asemel ka lineaarset efekti. Mõneti üllatuslikult ei too selline muudatus kaasa põhimõttelisi muutuseid otsustuspuu koostamise protseduuris, sest selle protsessi käigus sobitasime juba niigi lineaarseid segamudeleid.

Vaatleme näiteks peatükis 1.4 tutvustatud otsustuspuu esimese hargnemise konstrueerimise kriteeriumit, mis seisnes lineaarse segamudeli (10) optimaalse argumendi $\mathbf{z}_j$ valimises. Siiani oleme binaarse tunnuse $\mathbf{z}_j$ rollis kasutanud indikaatortunnuseid $(\mathbf{x}_j > \mathbf{0})$ ja $(\mathbf{x}_j < \mathbf{2})$. See tähendab, et oleme sobitanud mudelid

$$\begin{aligned} \mathbf{y} &\sim \mathcal{N}(\beta_0\mathbf{1} + \beta_1(\mathbf{x}_j > \mathbf{0}), \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I}) \\ \mathbf{y} &\sim \mathcal{N}(\beta_0\mathbf{1} + \beta_1(\mathbf{x}_j < \mathbf{2}), \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I}) \end{aligned} \tag{17}$$

kõigi SNPide $j = 1, \ldots, M$ jaoks ning valinud välja sellise mudeli ning indeksi j, mille korral tõepära on suurim. Nüüd on esimese hargnemise tulemusena prognoosideks ühes grupis β_0 ja teises grupis $\beta_0 + \beta_1$, nii et prognooside vektori saab kirja panna kui $\beta_0\mathbf{1} + \beta_1\mathbf{z}_j$.

Arvestades, et tegu on lineaarse segamudeli sobitamisega, võime kandidaattunnuste sekka lisada ka mittebinaarse tunnuse $\mathbf{x}_j$ ning sellise "hargnemise" korral oleks vastavaks prognoosiks $\beta_0\mathbf{1} + \beta_1\mathbf{x}_j$ (rangelt võttes pole enam tegu hargnemisega, kuna me ei jaga valimit kahte ossa). Eelnevale arutelule tuginedes tuleb meil modifitseeritud juhusliku segametsa esimese hargnemise konstrueerimiseks sobitada lisaks mudelitele (17) veel järgnev jaotus

$$\mathbf{y} \sim \mathcal{N}(\beta_0\mathbf{1} + \beta_1\,\mathbf{x}_j,\ \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I})$$

kõigi SNPide korral ning valida välja mudel, mille tõepära on suurim.

Detailsem ülevaade algoritmist on toodud pseudokoodis 1. Kui binaarse hargnemise korral jagatakse valim kaheks alamhulgaks, siis lineaarse efekti korral kaheksjagamist ei toimu. Konstrueerides otsustuspuud, uuendame puu igas tipus mudelimaatriksit, lisades uuele tunnusele vastava veeru, ning samuti hindame uue kordajate vektori $\hat{\boldsymbol{\beta}}$. Nii saame lõpuks iga viimase taseme tipu jaoks omaette kordajate hinnangud, sest arvatavasti on erinevates tippudes osutunud valituks erinevad argumendid. Teisisõnu, igale tipule vastab omaette lineaarne segamudel, mida saame kasutada uute indiviidide fenotüübi

Pseudokood 1 Juhusliku segametsa edasiarendus

1: Funktsiooni sisenditeks on: fenotüübi vektori $\mathbf{y}$, argumentide maatriks $\mathbf{X}$, geneetilise sarnasuse maatriks $\mathbf{K}$ (vt ptk 1.1), eelnevalt valmis konstrueeritud mudelimaatriks $\mathbf{X}_M$ ning antud tippu kuuluvate indiviidide indeksite hulk $\mathcal{S}$

2: Rekursiivse funktsiooni esimeseks väljakutsumiseks (st esimese hargnemise konstrueerimiseks) peab $\mathbf{X}_M$ sisaldama ühtede veergu ning $\mathcal{S}$ kõigi valimi elementide indekseid, st $\mathbf{X}_M := \mathbf{1}$ ning $\mathcal{S} := \{1, \ldots, N\}$.

3: **function** KONSTRUEERI_HARGNEMINE($\mathbf{X}, \mathbf{y}, \mathbf{K}, \mathbf{X}_M, \mathcal{S}$)

4: **for all** $j \in \{1, \ldots, M\}$ **do**

5: Defineerime $\tilde{\mathbf{x}}_j$ nii, et hulgas $\mathcal{S}$ on tema väärtus $\mathbf{x}_j$ ning muidu $\mathbf{0}$, st

$$\tilde{x}_{ij} = \begin{cases} x_{ij}, & \text{kui } i \in \mathcal{S} \\ 0, & \text{kui } i \notin \mathcal{S} \end{cases}$$

6: Hindame parameetrid $\boldsymbol{\beta}, \beta_j, \sigma_g^2, \sigma_e^2$ ning arvutame normaaljaotuste tõepära

$$\begin{aligned} L_{0,j} &\leftarrow L(\mathbf{y} | \mathbf{X}_M\boldsymbol{\beta} + \beta_j\tilde{\mathbf{x}}_j, \qquad \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I}) \\ L_{1,j} &\leftarrow L(\mathbf{y} | \mathbf{X}_M\boldsymbol{\beta} + \beta_j(\tilde{\mathbf{x}}_j > \mathbf{0}), \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I}) \\ L_{2,j} &\leftarrow L(\mathbf{y} | \mathbf{X}_M\boldsymbol{\beta} + \beta_j(\tilde{\mathbf{x}}_j < \mathbf{2}), \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I}) \end{aligned}$$

7: **end for**

8: $i, j \leftarrow \arg\max\{L_{ij} : i \in \{0, 1, 2\}, j \in \{1, \ldots, M\}\}$

9: **if** $i = 0$ **then** ▷ Kas valituks osutus lineaarne efekt?

10: $\mathbf{X}_M \leftarrow [\mathbf{X}_M, \tilde{\mathbf{x}}_j]$ ▷ Lisame mudelimaatriksisse veeru

11: **if** puu_kõrgus < max_puu_kõrgus **then**

12: **return** KONSTRUEERI_HARGNEMINE($\mathbf{X}, \mathbf{y}, \mathbf{K}, \mathbf{X}_M, \mathcal{S}$)

13: **end if**

14: **else**

15: **if** $i = 1$ **then** ▷ Kas valituks osutus tunnus $(\mathbf{x}_j > \mathbf{0})$?

16: $\mathbf{X}_M \leftarrow [\mathbf{X}_M, (\tilde{\mathbf{x}}_j > \mathbf{0})]$ ▷ Lisame mudelimaatriksisse binaarse veeru

17: $\mathcal{S}_0 \leftarrow \mathcal{S} \cap \{i : \tilde{x}_{ij} > 0\}$ ▷ Jagame valimi indeksid $\mathcal{S}$ kaheks osaks

18: $\mathcal{S}_1 \leftarrow \mathcal{S} \cap \{i : \tilde{x}_{ij} = 0\}$

19: **else if** $i = 2$ **then** ▷ Kas valituks osutus tunnus $(\mathbf{x}_j < \mathbf{2})$?

20: $\mathbf{X}_M \leftarrow [\mathbf{X}_M, (\tilde{\mathbf{x}}_j < \mathbf{2})]$ ▷ Lisame mudelimaatriksisse binaarse veeru

21: $\mathcal{S}_0 \leftarrow \mathcal{S} \cap \{i : \tilde{x}_{ij} < 2\}$ ▷ Jagame valimi indeksid $\mathcal{S}$ kaheks osaks

22: $\mathcal{S}_1 \leftarrow \mathcal{S} \cap \{i : \tilde{x}_{ij} = 2\}$

23: **end if**

24: **if** puu_kõrgus < max_puu_kõrgus **then**

25: tulemus$_0 \leftarrow$ KONSTRUEERI_HARGNEMINE($\mathbf{X}, \mathbf{y}, \mathbf{K}, \mathbf{X}_M, \mathcal{S}_0$)

26: tulemus$_1 \leftarrow$ KONSTRUEERI_HARGNEMINE($\mathbf{X}, \mathbf{y}, \mathbf{K}, \mathbf{X}_M, \mathcal{S}_1$)

27: **return** [tulemus$_0$, tulemus$_1$]

28: **end if**

29: **end if**

30: **return** $[\hat{\boldsymbol{\beta}}, \hat{\beta}_j]$ ▷ Lisaks on vaja tagastada puu struktuur

31: **end function**

prognoosimiseks: kõigepealt tuleb leida, millises puu tipus uus indiviid paikneb ning seejärel rakendada ennustamiseks antud tipule vastavat segamudelit.

Tegu on arvutuslikult mahuka algoritmiga, sest ainuüksi esimese hargnemise jaoks on vaja sobitada $3 \cdot M$ lineaarset segamudelit (meie genotüübi andmetes $M \approx 10000$). Kui puu kõrgus on näiteks 5, on tippude arv kokku $\sum_{n=0}^{5} 2^n = 63$. Kui metsas on 100 puud, tuleb meil kokku sobitada $18900 \cdot M$ lineaarset segamudelit.

Implementeerisime modifitseeritud juhusliku segametsa algoritmi Pythonis. Järgnevalt demonstreerime enda meetodi headust genereeritud andmestikel. Et algoritmi saaks pärisandmetel mõistliku arvutusajaga rakendada, tuleks programmikoodi efektiivsemaks muuta, näiteks implementeerida selle põhilised osad keeles C++.

3.5.2 Meetodi valideerimine simuleeritud andmetel

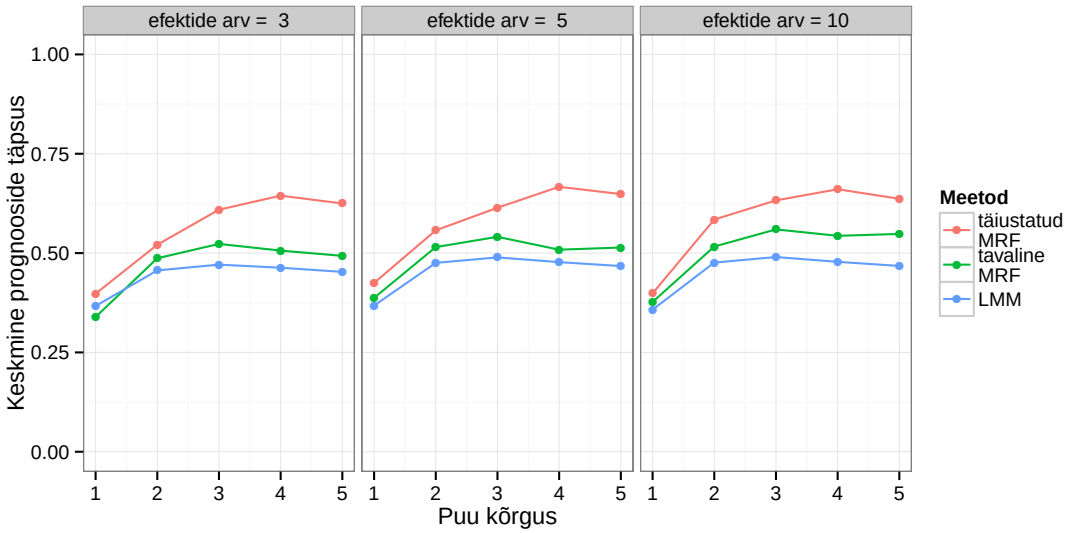
Veendumaks, et eelnevalt kirjeldatud lähenemine töötab hästi ning uurimaks, kas see võib olla efektiivsem kui lineaarne segamudel või juhuslik segamets, viisime läbi simulatsiooni. Genereerisime andmeid, mis sisaldaks lisaks populatsioonistruktuurile nii lineaarseid kui ka mittelineaarseid sõltuvusi y-tunnuse ning argumentide vahel. Rakendasime pseudokoodis 2 toodud skeemi selliste andmestike genereerimiseks, kus oleks tegelike efektide arv 3, 5 või 10. Kuna efektisuurused on genereeritud jaotusest $5 \cdot \text{Beta}(2, 2)$, siis need ulatuvad 0st 5ni, keskväärtusega 2.5. Võrdluseks, lisatud juhusliku müra dispersioon on $\sigma_e^2 = 1$.

Iga efektide arvu jaoks genereerisime 30 andmestikku (igaühes 500 indiviidi fenotüübi ja genotüübi väärtused, SNPide koguarv oli 500). Genereeritud andmetest eraldasime treening- ja testhulga (80% ja 20%) ning võrdlesime prognooside täpsust testandmetel. Rakendasime kolme meetodit: lineaarne segamudel edaspidise mudelivalikuga, juhuslik segamets ning modifitseeritud juhuslik segamets. Et mõistliku arvutusajaga tulemusi saada, kasutasime metsas 5 puud ning vaatlesime puid kõrgusega 1 kuni 5, analoogiliselt kasutasime segamudelis 1 kuni 5 argumenttunnnust. Genereeritud efektide arvud 3, 5, 10 on valitud nii, et 5 liiget sisaldavate mudelite potentsiaalne keerukus oleks vastavalt suurem, võrdne ja väiksem kui läheb ideaalis vaja genereeritud efektide modelleerimiseks.

Tulemused (keskmine R^2 väärtus testandmetel, joonis 17) näitavad, et meie täiustatud juhusliku segametsa algoritm võimaldab suurima prognoositäpsuse. Kuna lineaarse ning mittelineaarse efekti genereerimise tõenäosus oli võrdne (vt skeem 2), võib eeldada, et mõnedel andmestikel töötab paremini lineaarne segamudel ning teistel juhuslik sega-

Pseudokood 2 Andmete genereerimise skeem

Tähistagu $U\{a_1, \ldots, a_n\}$ ühtlast diskreetset jaotust üle hulga $\{a_1, \ldots, a_n\}$
$N \leftarrow 500$ ▷ indiviidide arv
$M \leftarrow 500$ ▷ SNPide arv
Genereeri $\mathbf{X} \in \{0, 1, 2\}^{N \times M}$ nii, et $x_{ij} \leftarrow U\{0, 1, 2\}$ ▷ Genotüübi maatriks
Arvutame $\mathbf{X}$ põhjal suguluse maatriksi $\mathbf{K}$ (vt ptk 1.1)
$\sigma_g^2 \leftarrow 1$
$\sigma_e^2 \leftarrow 1$
$\mathbf{y} \leftarrow \mathcal{N}(\mathbf{0},\ \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I})$ ▷ Populatsioonistruktuur ja mõõtmismüra
function GENEREERI_REKURSIIVSELT($\mathbf{y}, \mathbf{X}$, loendur)
    $\beta \leftarrow 5 \cdot \text{Beta}(2, 2)$ ▷ Efektisuurus beeta-jaotusest
    märk $\leftarrow U\{1, -1\}$ ▷ Kas β on pluss- või miinusmärgiga
    $j \leftarrow U\{1, \ldots M\}$ ▷ Valime indeksi j
    efekti_tüüp $\leftarrow U\{$"lineaarne", "dominantne"$\}$ ▷ Juhuslik efektitüübi valik
    if efekti_tüüp $=$ "lineaarne" **then**
        $\mathbf{y} \leftarrow \mathbf{y} +$ märk $\cdot\, \beta \cdot \mathbf{x}_j$ ▷ Lisa tunnuse $\mathbf{x}_j$ lineaarne efekt
        if loendur < efektide_arv **then**
            return GENEREERI_REKURSIIVSELT($\mathbf{y}, \mathbf{X}$, loendur+1)
        end if
    else
        $\mathbf{y} \leftarrow \mathbf{y} +$ märk $\cdot\, \beta \cdot (\mathbf{x}_j > \mathbf{0})$ ▷ Lisa tunnuse $\mathbf{x}_j$ dominantsuse efekt
        if loendur < efektide_arv **then**
            $\mathbf{y}_0 \leftarrow \mathbf{y}[(\mathbf{x}_j > \mathbf{0})]$ ▷ Jaga $\mathbf{y}$ kaheks osaks $\mathbf{y}_0$ ja $\mathbf{y}_1$
            $\mathbf{y}_1 \leftarrow \mathbf{y}[(\mathbf{x}_j = \mathbf{0})]$ ▷ tingimuse $(\mathbf{x}_j > \mathbf{0})$ järgi
            $\mathbf{X}_0 \leftarrow \mathbf{X}[(\mathbf{x}_j > \mathbf{0}),]$ ▷ Analoogiliselt jaga $\mathbf{X}$ kaheks
            $\mathbf{X}_1 \leftarrow \mathbf{X}[(\mathbf{x}_j = \mathbf{0}),]$
            $\text{tulemus}_0 \leftarrow$ GENEREERI_REKURSIIVSELT($\mathbf{y}_0, \mathbf{X}_0$, loendur+1)
            $\text{tulemus}_1 \leftarrow$ GENEREERI_REKURSIIVSELT($\mathbf{y}_1, \mathbf{X}_1$, loendur+1)
            return $[\text{tulemus}_0, \text{tulemus}_1]$
        end if
    end if
    return $\mathbf{y}$
end function
GENEREERI_REKURSIIVSELT($\mathbf{y}, \mathbf{X}$, loendur=0)



Joonis 17. Lineaarse segamudeli (LMM), juhusliku segametsa (tavaline MRF) ning selle täiustatud versiooni (täiustatud MRF) prognooside täpsuse võrdlus simuleeritud andmetel. Näidatud on keskmine R^2 väärtus üle 30 genereeritud andmestiku.

mets. Viimase täiustatud versioon kombineerib mõlema algoritmi tugevad küljed ning saavutab täpsemad prognoosid kui kumbki mudel eraldiseisvalt.

Kokkuvõte

Töö laiemaks eesmärgiks oli uurida, kui täpselt saab indiviidi pärilike tunnuste väärtuseid ennustada. Selleks käsitlesime käesolevas töös fenotüüpide prognoosimist ühe pärmipopulatsiooni näitel. Kuna selles esines tugev populatsioonistruktuur, st klassikaline *i.i.d.* eeldus ei kehtinud, oli modelleerimisel vaja arvesse võtta indiviididevahelisi sõltuvusi. Mudelitena kasutasime nii lineaarset segamudelit (*Linear Mixed Model*) kui ka juhuslikku segametsa (*Mixed Random Forest*), kusjuures mitmemõõtmelise lineaarse segamudeli (*Multi-Trait Linear Mixed Model*) abil modelleerisime ka fenotüüpidevahelist sõltuvust. Kõigi mudelite prognoosivõimet hindasime eraldiseisval testandmestikul.

Kõik käsitletud mudelid kasutasid populatsioonistruktuuri modelleerimiseks juhuslikke efekte, mis toetusid indiviididevahelisele sugulusele ehk geneetilisele sarnasusele. Selle intuitiivne seletus on, et mida suurem osa genoomist on kahe indiviidi vahel identne, seda sarnasemad nad fenotüüpide poolest on. Nägime, et ainuüksi geneetilisel sarnasusel põhinevad segamudelid andsid suhteliselt täpseid prognoose ning konkreetsete SNPide mudelisse lisamine omas võrdlemisi väikest mõju, mõnel juhul ei pruukinud see üldse ennustustäpsust suurendada. Sellest järeldame, et tugeva populatsioonistruktuuri korral on prognoosimisel suureks abiks teada indiviidide üldist geneetilist sarnasust.

Modelleerides pärilikke tunnuseid genotüübi põhjal, seab pärilikkus prognoositäpsusele ülemise piiri: aditiivsete efektide modelleerimisel on tõkkeks kitsas pärilikkus, dominantsuse ja koosmõjude lisamisel lai pärilikkus. Ühe fenotüübi korral õnnestus meil koosmõjusid sisaldava mudeli abil ületada kitsa pärilikkuse piir, ülejäänud tunnuste korral seletasid parimate mudelite prognoosid vähemalt 92% kitsast pärilikkusest.

Et prognooside täpsust veelgi suurendada, kasutasime uuritava fenotüübi ennustamisel lisaks genotüübile ka indiviidi teisi fenotüüpe. Sellega imiteerisime olukorda, kus meil pole võimalik mõõta huvipakkuvat fenotüüpi või selle mõõtmine on kallis, aga meil on indiviidi kohta teada nii genotüüp kui ka teiste tunnuste väärtused. Nägime, et teiste tunnuste abil saame täpsemad prognoosid, aga ainult nendest ei piisa: parima prognoositäpsuse saamiseks tuleb arvesse võtta nii genotüübi kui fenotüübi informatsiooni. Kuigi sel juhul pole enam tegu ainult genoomi informatsioonil põhineva prognoosiga, ületasime selliste mudelite abil kitsa pärilikkuse piiri peaaegu kõigi fenotüüpide korral ning jõudsime lähedale laia pärilikkuse piirile. Meile teadaolevalt pole varem keegi nii täpselt fenotüüpi ennustanud.

Töös kasutatud kaks mudeliklassi, lineaarsed segamudelid ning juhuslikud segametsad,

kasutavad mudeli ehitamisel kaht erinevat tüüpi lähenemist: modelleeritakse vastavalt lineaarseid ja mittelineaarseid sõltuvusi argumenttunnustest. Seetõttu pakkusime prognoosimiseks välja uue meetodi, mis kombineeriks endas mõlema mudeliklassi häid külgi ning oleks seeläbi universaalsem. Tegu on juhusliku segametsa algoritmi edasiarendusega, mis lubab puu struktuuris ka lineaarsete sõltuvuste arvessevõtmist. Meetodi headuse testimiseks viisime läbi simulatsiooni. Genereerisime andmeid, kus leidub nii lineaarseid kui mittelineaarseid seoseid, ning rakendasime neil lineaarseid segamudeleid, juhuslikke segametsi ja enda täiustatud meetodit. Tulemusena selgus, et väljapakutud meetod võimaldab saada parema prognoositäpsuse kui kumbki alternatiiv.

Tööd saame jätkata mitmes suunas. Esiteks, kaalume väljapakutud juhusliku segametsa edasiarenduse efektiivsemat implementeerimist, et see muutuks kasutatavaks suurte andmemahtude juures mõistliku arvutusajaga. Teiseks, saame meetodit veel edasi arendada, otsides puu igas tipus optimaalset fenotüübi transformatsiooni, mis tagaks parema kooskõla sobitatava normaaljaotusega ning seeläbi täpsemad prognoosid.

Kuna kõnealune pärmi eksperiment on veel pooleli, suureneb järk-järgult andmete kogus. Suurema arvu fenotüüpide korral loodame suuremat võitu nende ühisjaotuse modelleerimisest ning teiste fenotüüpide järgi tinglikustades saada veelgi täpsemaid prognoose. Kuna lisaks 7396 indiviidile on andmed ka mõnede nende kaugemate sugulast kohta, kel esineb näiteks uusi mutatsioone, saame uurida, kuidas üldistub prognoosivõime uutele indiviididele.

Viited

[1] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang, "Five years of gwas discovery", *The American Journal of Human Genetics*, vol. 90, no. 1, pp. 7–24, 2012.

[2] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, *et al.*, "Finding the missing heritability of complex diseases", *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[3] O. Zuk, E. Hechter, S. R. Sunyaev, and E. S. Lander, "The mystery of missing heritability: Genetic interactions create phantom heritability", *Proceedings of the National Academy of Sciences*, vol. 109, no. 4, pp. 1193–1198, 2012.

[4] H. L. Allen, K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, F. Rivadeneira, C. J. Willer, A. U. Jackson, S. Vedantam, S. Raychaudhuri, *et al.*, "Hundreds of variants clustered in genomic loci and biological pathways affect human height", *Nature*, vol. 467, no. 7317, pp. 832–838, 2010.

[5] T. F. Mackay, E. A. Stone, and J. F. Ayroles, "The genetics of quantitative traits: Challenges and prospects", *Nature Reviews Genetics*, vol. 10, no. 8, pp. 565–577, 2009.

[6] D. B. Goldstein, "Common genetic variation and human traits", *New England Journal of Medicine*, vol. 360, no. 17, p. 1696, 2009.

[7] B. Rakitsch, C. Lippert, O. Stegle, and K. Borgwardt, "A lasso multi-marker mixed model for association mapping with population structure correction", *Bioinformatics*, vol. 29, no. 2, pp. 206–214, 2013.

[8] X. Zhou and M. Stephens, "Efficient multivariate linear mixed model algorithms for genome-wide association studies", *Nature methods*, vol. 11, no. 4, pp. 407–409, 2014.

[9] C. Lippert, F. P. Casale, B. Rakitsch, and O. Stegle, "Limix: Genetic analysis of multiple traits", *BioRxiv*, p. 003 905, 2014.

[10] G. Moser, S. H. Lee, B. J. Hayes, M. E. Goddard, N. R. Wray, and P. M. Visscher, "Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model", 2015.

[11] C. Lippert, J. Listgarten, Y. Liu, C. M. Kadie, R. I. Davidson, and D. Heckerman, "Fast linear mixed models for genome-wide association studies", *Nature Methods*, vol. 8, no. 10, pp. 833–835, 2011.

[12] B. J. Hayes, P. M. Visscher, and M. E. Goddard, "Increased accuracy of artificial selection by using the realized relationship matrix", *Genetics Research*, vol. 91, no. 01, pp. 47–60, 2009.

[13] G. de los Campos, A. I. Vazquez, R. Fernando, Y. C. Klimentidis, and D. Sorensen, "Prediction of complex human traits using the genomic best linear unbiased predictor", *PLoS genetics*, vol. 9, no. 7, e1003608, 2013.

[14] O. Stegle, C. Lippert, J. M. Mooij, N. D. Lawrence, and K. M. Borgwardt, "Efficient inference in matrix-variate gaussian models with iid observation noise", in *Advances in neural information processing systems*, 2011, pp. 630–638.

[15] T. Hastie, R. Tibshirani, J. Friedman, T Hastie, J Friedman, and R Tibshirani, *The elements of statistical learning*, 1. Springer, 2009, vol. 2.

[16] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning", *Foundations and Trends® in Computer Graphics and Vision*, no. 7, pp. 81–227, 2011.

[17] J. Stephan, O. Stegle, and B. Andreas, "Mixed random forest: An efficient approach to capture additive and epistatic genetic effects in the presence of population structure", *Nature Communications (accepted)*, 2015.

[18] L. Parts, "Genome-wide mapping of cellular traits using yeast", *Yeast*, vol. 31, no. 6, pp. 197–205, 2014.

[19] F. A. Cubillos, L. Parts, F. Salinas, A. Bergström, E. Scovacricchi, A. Zia, C. J. Illingworth, V. Mustonen, S. Ibstedt, J. Warringer, *et al.*, "High-resolution mapping of complex traits with a four-parent advanced intercross yeast population", *Genetics*, vol. 195, no. 3, pp. 1141–1155, 2013.

[20] R. Jelier, J. I. Semple, R. Garcia-Verdugo, and B. Lehner, "Predicting phenotypic variation in yeast from individual genome sequences", *Nature genetics*, vol. 43, no. 12, pp. 1270–1274, 2011.

[21] J. S. Bloom, I. M. Ehrenreich, W. T. Loo, T.-L. V. Lite, and L. Kruglyak, "Finding the sources of missing heritability in a yeast cross", *Nature*, vol. 494, no. 7436, pp. 234–237, 2013.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Kaspar Märtens (sünnikuupäev 23. november 1990),

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose **"Pärilike tunnuste täpne prognoosimine DNA põhjal"**, mille juhendaja on Leopold Parts,

    1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

    1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus 13. mail 2015