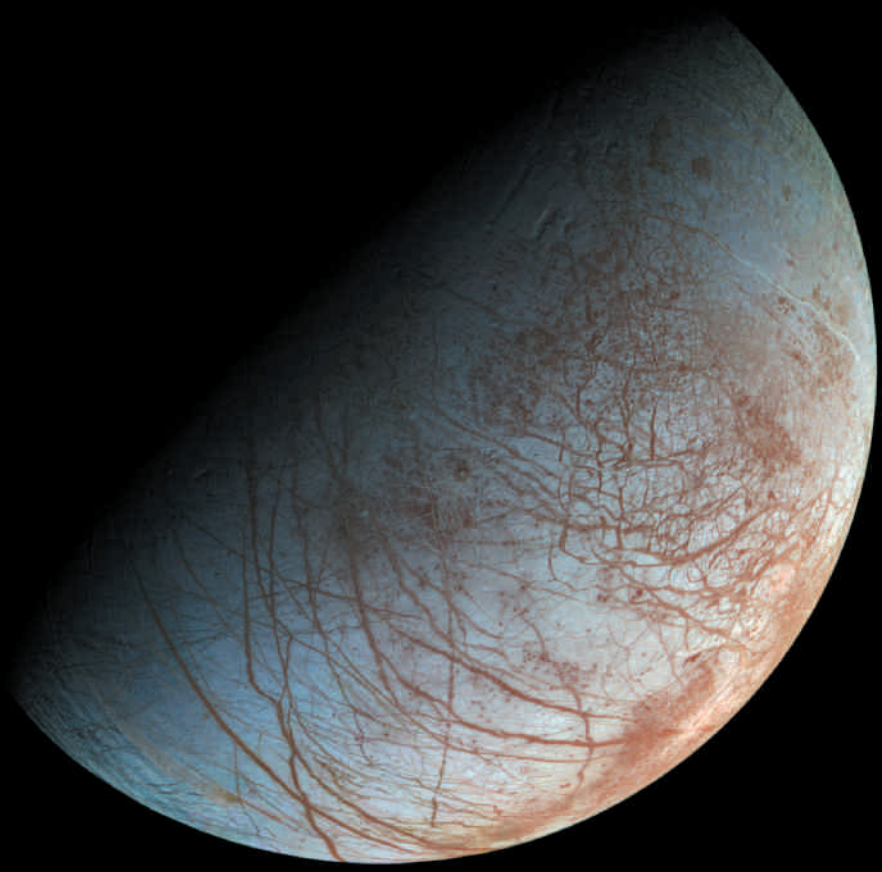
RONALD DE VLAMING

Linear Mixed Models in Statistical Genetics



Linear Mixed Models in Statistical Genetics

Linear Mixed Models in Statistical Genetics

Lineaire gemengde modellen in de statistische genetica

THESIS

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. H.A.P. Pols

and in accordance with the decision of the Doctorate Board.

The public defence shall be held on
Thursday July 6, 2017 at 15:30 hours

by

RONALD DE VLAMING
born in Nieuwerkerk aan den IJssel.

Erasmus University Rotterdam

Doctoral Committee

Promotor:	Prof.dr. A.R. Thurik
	Prof.dr. P.J.F. Groenen
	Prof.dr. P.D. Koellinger
Other members:	Prof.dr. P.H.C. Eilers
	Prof.dr. R. Paap
	Prof.dr. D. Posthuma

Erasmus Research Institute of Management – ERIM
The joint research institute of the Rotterdam School of Management (RSM)
and the Erasmus School of Economics (ESE) at the Erasmus University Rotterdam.
Internet: http://www.erim.eur.nl.

ERIM Electronic Series Portal: http://hdl.handle.net/1765/1

ERIM PhD Series in Research in Management, 416
ERIM reference number: EPS-2017-416-S&E
ISBN 978-90-5892-481-0

Design: PanArt, www.panart.nl.
Cover photo: courtesy NASA/JPL-Caltech
(source: https://photojournal.jpl.nasa.gov/catalog/PIA19048; accessed: Jan. 16, 2017).

This publication (cover and interior) is printed by Tuijtel on recycled paper, BalanceSilk®.
The ink used is produced from renewable resources and alcohol free fountain solution.
Certifications for the paper and the printing production process: Recycle, EU Ecolabel,
FSC®, ISO14001.
More information: www.tuijtel.com.

Contents

List of Figures

List of Tables

Voorwoord

Hoewel men – naar mijn mening – nooit kan zeggen dat een proefschrift écht helemaal af is, treft u bij dezen de versie waarvan ik zeg: zo is het goed genoeg. Dit boek bestaat uit twee delen, met in ieder deel drie hoofdstukken. Het eerste deel is vooral empirisch van aard, terwijl het tweede deel hoofdzakelijk theoretisch is. Een samenvatting van mijn onderzoek kunt u terugvinden aan het einde van dit proefschrift. Daarnaast kan de lezer die wat dichter bij de materie wil komen het inleidende hoofdstuk lezen dat voorafgaat aan de twee eerder genoemde delen.

De rest van het voorwoord gaat niet zozeer over de inhoud als wel over het proces. Hoe ben ik gekomen waar ik nu ben? En – nog belangrijker – wie hebben mij in de afgelopen jaren geholpen om op dit punt te komen? Om deze vragen te beantwoorden moet ik teruggaan naar het moment waarop ik mijn bachelor 'Econometrie en Operationele Research' net had afgerond, medio 2010.

In september 2010 begon de eenjarige master 'Econometrics and Management Science' waarvoor ik mij had ingeschreven. Op voorhand was ik vooral bezig met de vraag hoe ik zonder problemen door het programma heen zou komen. Aandacht voor de vervolgstap had ik nog niet. Een jaar lijkt immers lang. Maar toen de boel eenmaal in beweging was, schrok ik na twee maanden flink; het eerste blok was voorbij gevlógen, terwijl ik nog geen flauw benul had over wat ik eigenlijk wilde gaan doen als het eenmaal voorbij zou zijn.

Ik wist wel dat ik voor het einde van mijn studie nog graag een tijd stage wilde lopen ergens buiten de EU. Die wens viel binnen dat jaar alleen te verenigen met mijn scriptie. Anderzijds zag ik het ook wel zitten om bij *De Nederlandsche Bank* (DNB) aan de slag te gaan. Maar de enige springplank die ik naar DNB zag was een onderzoeksstage waaruit een scriptie zou moeten voortvloeien. Dus moest ik een keuze maken tussen die twee.

Na een aantal slapeloze nachten en – achteraf bezien – naïeve pogingen om een stageplek te bemachtigen bij DNB, zat ik tegen het

einde van februari 2011 met mijn handen in het haar. De betreffende organisatie leek nog niet echt warm te lopen voor het bieden van een stageplek en een goede plek in het buitenland had ik ook nog niet voorbij zien komen. In deze periode van veel hoofdbrekens heb ik onbeschrijflijk veel steun gevonden bij dr. Niels van der Bijl. Ondanks mijn bij vlagen ietwat weerbarstige karakter is hij een steun en toeverlaat geweest in het vaststellen van prioriteiten en het nemen van doelmatige beslissingen.

Uiteindelijk kwam ik toch iets tegen wat mij wel interessant leek; eind februari brachten prof. Dennis Fok en dr. Andreas Pick de masterstudenten op de hoogte van een mogelijke onderzoeksstage in Sint-Petersburg, Rusland, voor twee à drie maanden, met als doel een voorspelmodel te bouwen voor een online adverteerder van het aantal weergaves van hun advertenties via verschillende 'kanalen'. Daarmee was de kogel door kerk: het sluitstuk van het masterprogramma zou een stage in het buitenland worden met als beoogde uitkomst een stevige scriptie. De vraag wat ik na mijn master zou gaan doen was weer even op de lange baan geschoven.

Vanaf dat moment was ik in de ban van 'big data'. De dataset waarmee ik aan de slag kon was fors: één maand aan data, van uur tot uur, voor circa 28.000 verschillende kanalen. Het was flink aanpoten om een goed model te krijgen. Uiteindelijk was het beste voorspelmodel één waar 126 verschillende 'eenvoudige' voorspelmodellen aan ten grondslag lagen, waar al die voorspellingen werden geaggregeerd door een gewogen gemiddelde te nemen, waarbij de model-, kanaal- en tijdspecifieke gewichten dynamisch werden gekozen op basis van de historische voorspelkwaliteit van een gegeven model in een gegeven kanaal. Enerzijds duizelde het mij, maar tegelijkertijd vond ik het razend interessant!

Voor de stage had ik nagenoeg geen aandacht besteed aan een carrière in de wetenschap; ik kom uit een familie zonder academici. De wetenschap was simpelweg een andere – ogenschijnlijk zelfs ietwat saaie – wereld voor mij. Maar tijdens de stage merkte ik dat het mij eigenlijk wel beviel om te worstelen met data, systematisch te zoeken naar een optimum, nieuwe modellen toe te passen en ondertussen goed rekening te houden met allerlei zaken die om statistische redenen roet in het eten kunnen gooien.

Naarmate mijn stage verder vorderde was ik er wel uit: ik wilde de wetenschap in—ik wilde stoeien met grote datasets en voorspelmodellen, en daar papers over schrijven. Maar goed, ik had geen flauw benul hoe een promotietraject eruitzag, laat staan de stappen die eraan vooraf zouden gaan. Dus toen ik vol frisse moed terug was in Nederland dacht ik dat ik met een bijna afgeronde master econometrie wel eventjes iets zou kunnen vinden. Natuurlijk bleek niets minder

waar; het vinden van een project dat bij mij paste was lastig en de weg van sollicitatie naar goedkeuring door de *Vaste Commissie voor de Wetenschapsbeoefening* was lang. Het pad dat ik zo graag wilde bewandelen werd nauwer met iedere maand die verstreek. Wel kwam mij het research-masterprogramma van het *Tinbergen Instituut* (TI) ter ore. Na een half jaar aan deze master te hebben deelgenomen werd het echter duidelijk dat dit voor mij niet de juiste weg voorwaarts was.

Dus was ik dolgelukkig toen ik begin 2012, met hulp van dr. Jan Brinkhuis, de ideale plek vond. Een plek waar ik, onder leiding van professoren Roy Thurik, Patrick Groenen en Philipp Koellinger, zou kunnen werken met big data, maar dan op veel grotere schaal en op een fundamenteler niveau dan ik gewend was: genetische data van tienduizenden mensen voor miljoenen genetische 'markers'. Een latente interesse in de biologie – mede ooit teweeggebracht door een bevlogen biologieleraar op de middelbare school, Hans van Zuylen – werd weer aangewakkerd.

De exercitie waar professoren Thurik, Groenen en Koellinger zich aan waagden was het gebruiken van moleculaire genetische data om variatie in sociaaleconomische uitkomsten en individuele voorkeuren te kunnen verklaren, en om deze verklaarde genetische variantie te herleiden tot specifieke gebieden in het DNA. Het was in 2012 nog een vergezicht en bovenal een waagstuk. Maar ondanks de ogenschijnlijke risico's gingen deze hoogleraren en hun toenmalige promovendi, dr. Matthijs van der Loos en dr. Niels Rietveld, onverdroten voort; erfelijkheidsstudies en berekeningen van statistische kracht lieten zien dat er wat te halen viel. Mede dankzij de hulp van dr. Jan Brinkhuis, dr. Adriana Gabor, dr. Christiaan Heij, dr. Andreas Pick en Peter Post slaagde de sollicitatie, en waren professoren Thurik, Groenen en Koellinger bereid mij in dienst te nemen als promovendus. Na een initiële aanstelling als onderzoeksassistent begon het avontuur in januari 2013 dan echt!

Zoals eerder gezegd ben ik beslist geen telg uit een wetenschappelijke familie. Daarnaast was en is de wetenschap, met het oog op mijn karakter, wellicht niet de meest voor de hand liggende optie. Spelenderwijs ideeën ontwikkelen, werken met statistische modellen, met wiskundige afleidingen stoeien, programmeren, dat beviel allemaal prima. Maar het ging er soms rusteloos aan toe; het ene idee was nog niet uitgewerkt of mijn aandacht was alweer op een volgend probleem gericht. Naast de haperingen in mijn focus had ik ook nog eens drie begeleiders, ieder met zijn eigen ideeën. Om nog maar niet te spreken over samenwerkingsverbanden met vele ijverige Amerikanen, ambitieuze Europeanen en wat dies meer zij. Enfin, halverwege het tweede jaar zag ik door de bomen het bos niet meer.

Maar door veel te praten met andere jonge onderzoekers – in het

bijzonder dr. Sophie Bruinsma, Ekaterina Isakina, Richard Karlsson Linnér, Fleur Meddens, Gertjan van den Burg en dr. Peter van der Zwan – kwam ik erachter dat dit mij niet per definitie tot een slechte wetenschapper zou maken; dat soort strubbelingen horen er gewoon bij. Men is niet van nature een wetenschapper, men wordt het zogezegd door schade en schande. Rond het einde van mijn tweede jaar kreeg ik langzaam maar zeker de smaak te pakken. Mijn werkwijze werd systematischer en lopende projecten kwamen nader tot volbrenging.

Als promovendus heb ik mogen samenwerken met begenadigde wetenschappers, zoals dr. Jonathan Beauchamp, dr. Daniel Benjamin en dr. David Cesarini, en professoren Peter Visscher, Naomi Wray en Jian Yang. In het bijzonder ben ik professoren Visscher en Wray zeer dankbaar voor het faciliteren van een onderzoeksstage binnen hun team aan de *University of Queensland* in de zomer en het najaar van 2016. Hun vermogen baanbrekend onderzoek uit te voeren en tegelijkertijd oog te houden voor menselijke aspecten verdient alle lof. Ik ben hun uiterst dankbaar voor de kans op plezierige wijze met hen te hebben mogen samenwerken en ik hoop dan ook in de toekomst te kunnen blijven samenwerken. Voorts wil ik dr. Beben Benyamin, dr. Fleur Garton, dr. Matt Keller, Robert Maier, Adriaan van der Graaf, dr. Anna Vinkhuyzen, dr. Loic Yengo, dr. Jian Zeng, dr. Zhihong Zhu en de vele anderen die ik heb leren kennen tijdens mijn onderzoeksstage bij de *University of Queensland* hartelijk danken voor hun gastvrijheid en de zeer leerzame tijd.

Graag richt ik nog een dankwoord aan dr. Benjamin, dr. Cesarini en professor Koellinger omwille van de ruimte die zij creëren voor onderzoek naar de genetische architectuur van sociaaleconomische uitkomsten en voorkeuren, onder andere middels het oprichten van het – internationaal actieve – *Social Science Genetic Association Consortium* (SSGAC). Voorts ben ik professoren Koellinger, Groenen en Thurik alsmede dr. Rietveld dankbaar voor hun hulp en begeleiding en voor de wijze waarop zij gestalte hebben gegeven aan onderzoek op het raakvlak van de biologie en de economie, bijvoorbeeld door de oprichting van het *Erasmus University Rotterdam Institute for Behavior and Biology* (EURIBEB). Ook wil ik professoren Bert Hofman en André Uitterlinden en – in brede zin – de *Rotterdam Study* bedanken voor het faciliteren van spannend onderzoek en hun nuttige adviezen.

Voor het slagen van mijn onderzoek is veel rekenkracht noodzakelijk geweest. Daarom bedank ik graag de stichting *SURFsara* voor haar diensten, zoals toegang tot de supercomputer *Lisa* en later ook *Cartesius*. Ik ben met name Wim Rijks en Markus van Dijk erg dankbaar voor al hun hulp de afgelopen jaren.

Daarnaast ben ik het *Erasmus Research Institute of Management* (ERIM), het TI, de afdeling *Toegepaste Economie* binnen de *Erasmus*

School of Economics (ESE) en allen die betrokken zijn bij deze organisaties ten zeerste dankbaar voor de ondersteuning die ik door de jaren heb mogen ontvangen. In het bijzonder wil ik Kim Beerentemfel-Laarman, Gerda de Rave, Manuela Ettekoven, Nita Ramsaransing en Judith van Kronenburg bedanken. Ook ben ik mijn mede-promovendi, waaronder Indy Bernoster, Casper Burik, dr. Pourya Darnihamedani, Ekaterina Isakina, Richard Karlsson Linnér, Plato Leung, Fleur Meddens, dr. Aysu Okbay, dr. Wim Rietdijk, dr. Niels Rietveld, Eric Slob, Gertjan van den Burg, dr. Matthijs van der Loos en Caroline Witte, alsmede vele andere collega's, waaronder dr. Jolanda Hessels, dr. Peter van der Zwan en dr. André van Stel, en vele anderen erg dankbaar voor de mooie tijd bij de ESE. Ik ben blij uiterst prettig met jullie te hebben mogen samenwerken.

Voorlaatst wil ik de leden van de commissie, professoren Paul Eilers, Richard Paap en Danielle Posthuma alsmede dr. Jonathan Beauchamp, dr. Matt Keller en dr. Katrijn Van Deun, bedanken voor hun bereidwilligheid zitting te nemen in de commissie en voor de daaruit voortvloeiende inspanningen. Graag wil ik Gertjan van den Burg nog bedanken voor het mooie LaTeX-proefschriftsjabloon. Daarnaast wil ik dr. Peter van der Zwan bedanken voor zijn aanmerkingen op dit voorwoord en de samenvatting van dit proefschrift. Ook wil ik mijn paranimfen, Vincent Okhuyzen en Fleur Meddens, bedanken voor al hun hulp en voor het onbevangen enthousiasme waarmee zij deze verantwoordelijkheid hebben aanvaard.

Tot slot richt ik graag een zeer groot dankwoord aan een aantal vrienden, mijn familie en mijn partner. Beste Alain, Bertine, Debby, Fatma, Gidius, Gies, Jan, Kirstin, Léon, Niels van der Bijl, Niels Vriethoff, Peter, Ruud, Sarah, Sophie, Stefan, Steven en Vincent, ik koester mijn vriendschappen met jullie. De vele leuke momenten die ik met jullie heb doorgebracht zijn onontbeerlijk geweest de afgelopen jaren. Beste Marijke en Paul, met veel liefde en tederheid heb ik mogen aanschouwen hoe jullie de afgelopen jaren twee prachtige kinderen op de wereld hebben gezet. Loes en Sven herinneren mij er dikwijls aan dat er zo veel meer is in het leven dan werken. Ik ben onwijs trots hun oom te mogen zijn. Lieve Hans en Willie, met veel geduld hebben jullie mij zo nu en dan zien stoeien om een plek te vinden waar ik tot mijn recht kom. De afgelopen vier jaren zullen jullie bij vlagen bezorgd zijn geweest. Desondanks is mijn indruk al die tijd geweest dat jullie een rotsvast vertrouwen hadden dat het mij zou lukken om het proefschrift tot een goed einde te brengen. Dank voor het vertrouwen en alle hulp. Lieve Jordi, veel dank voor jouw onvoorwaardelijke liefde, steun, geduld en – wanneer het nodig was – ook ongeduld. Jij bent instrumenteel geweest in het vinden van de juiste koers in het woelige proces dat promoveren heet.

Welnu, voordat u de inhoud van dit proefschrift gaat bestuderen zou ik graag nog een paar woorden willen wijden aan de kaft van mijn proefschrift. De omslag toont een foto van een van de manen van Jupiter, genaamd Europa (met dank aan NASA/JPL-Caltech). Daar Europa vijf keer zo ver van de zon verwijderd is als de Aarde, is deze maan aan de oppervlakte een ijzige woestenij. Desondanks is het een van de weinige bekende plekken waar buitenaards leven mogelijk wordt geacht, doch niet aan de oppervlakte maar in ondergrondse oceanen. Een gangbare hypothese luidt dat deze oceanen voldoende door getijdenfrictie worden verhit om ze deels vloeibaar te houden. Hiermee is leven op deze maan – in ieder geval in theorie – mogelijk. Mocht er ooit een spoor van leven worden aangetroffen op de maan Europa dan onderschrijft dit de uiterst taaie aard van het leven.

Maar deze foto staat voor mij symbool voor meer dan alleen het feit dat leven vernuftig is. Allereerst, sterkt het feit dat we deze maan – in onze zoektocht naar buitenaards leven – willen doorgronden mij in de overtuiging dat de wetenschappelijke gemeenschap bereid is hypotheses op grond van wetenschappelijke merites te overwegen, zelfs als ze op het eerste gezicht vergezocht lijken. Ten tweede, vind ik dat er een schoonheid uitgaat van het feit dat we op een astronomische afstand de mogelijke aanwezigheid van leven beschouwen – leven dat wellicht weinig meer voorstelt dan een dans van zelfreplicerende moleculen – op een maan die miljoenen kilometers van onze planeet verwijderd is, ver van onze maatschappij en ver van onze biologie.

De wijze waarop het microscopische en het macroscopische ondanks de enorme verschillen in schaal lijken samen te hangen, heeft mij al van jongs af aan versteld doen staan. Ik ben evenzeer verwonderd – hoewel in een wat aardsere zin – wanneer ik de maatschappij beschouw die wij hebben opgebouwd als soort en ik mij realiseer dat wij op dit moment bezig zijn ons ingewikkelde gedrag, onze politiek, onze cultuur, onze nieuwsgierigheid en onze voorkeuren deels te herleiden tot de moleculaire bouwstenen waaraan wij ontspruiten.

Hoewel u als lezer wellicht een beetje overrompeld bent door al deze metaforen is er toch nog een laatste zienswijze die ik wil delen. Wanneer men velden zoals kwantitatieve genetica, gedragsgenetica of geno-economie beschouwt heeft de maan een schaduwzijde; ik vermoed dat er menig wetenschapper te vinden valt die zal beamen dat we nog maar een klein deel van het verhaal kennen. We hebben weliswaar een basaal begrip van de wijze waarop de machinerie des levens werkt. Maar we hebben tevens zaken zoals de ontbrekende erfelijkheid. Ook hebben we voor veel belangrijke uitkomsten nog maar een handvol genen of zelfs nog geen genen te pakken die robuust geassocieerd zijn met een gegeven uitkomst. Nieuwe biologische mechanismes die ons leven schapen worden nog steeds op een gestaag tempo ontdekt.

De vragen en de verwondering duren dus voort. Afbeeldingen zoals deze foto van de maan Europa inspireren mij om de onbekende zaken te blijven overpeinzen en vervullen mij met de wens antwoorden op deze vragen te blijven zoeken en daarmee nieuw licht te werpen op de ontbrekende puzzelstukjes in onze collectieve kennis *ad infinitum*.

Amsterdam, 16 januari 2017 Ronald de Vlaming

Preface

Although one can never really say a Ph.D. dissertation is completely finished – after all, there is always room for improvement – I deem this version of my thesis to be sufficiently ready. This book consists of two parts, where each part consists of three chapters. The first part comprises empirical work, whereas the second part is more theoretical in nature. A summary of my research can be found at the end of this thesis. For those who desire a more in-depth understanding of this thesis, I recommend reading the introductory chapter which precedes the aforementioned two parts.

The remainder of this preface is not so much about the content, rather it is about the process. My aim here is to ponder how I came to be a PhD candidate in the first place, how I managed to finish this considerably arduous process, and – most importantly – who have helped me during these past few years to get through it. In order to answer these questions, I need to go back to the time when I had just finished my bachelor program in 'Econometrics and Operations Research', in the middle of 2010.

After finishing the program, I immediately enrolled for the master program in 'Econometrics and Management Science', which started in September. At the outset of this program, I was primarily concerned with questions like: how am I going to get through this program, with above-average results, without suffering some kind of breakdown? At that time, I was not bothered in the slightest by subsequent steps in my career; a whole year seemed like a sufficiently long time to figure something out. Yet, once the program had commenced I was in for a rude awakening; the first two months passed by in what felt like the blink of an eye, and – by that time – I still had no idea at all about what to do next.

On the one hand, I knew that I wanted to spend some time abroad as an intern, preferably somewhere outside the EU, before finishing the master program. Such an internship seemed feasible only if it could be combined with writing my master's thesis. On the other hand,

the idea of a research internship at the Dutch central bank (DNB) in conjunction with writing my thesis over there was quite tempting. So I had to make a choice between the two.

After several sleepless nights and a few somewhat half-hearted attempts at securing an internship at DNB, I was forced to reconsider by the end of February 2011; they did not seem terribly excited about offering me an internship position and, so far, I had not seen a really interesting position abroad yet either.

In this period of considerable uncertainty, I have had tremendous support from dr. Niels van der Bijl. I owe him my sincere gratitude, as he played a pivotal role in helping me to get my priorities straight and to implement those priorities through persistent and consistent action.

After some time, I finally encountered an internship that seemed interesting. Professor Dennis Fok and dr. Andreas Pick invited master students to apply for a position in Saint Petersburg, Russia, for a period of two to three months, with the aim of developing a model for an online advertiser, with the purpose of predicting the number of ads that will be displayed in different 'channels'. I decided that this project would form the basis of my master's thesis. The question what I would want to do after finishing the master program was pushed aside for some more time.

From that moment on, I was captivated by 'big data'. I was working with a large dataset: one month's worth of data, hour-by-hour, for roughly 28,000 different channels. Finding the right model was quite laborious. In the end, my best prediction model consisted of 126 underlying 'simple' models, where the forecasts from the simple models were aggregated by taking a weighted average, where model-, channel-, and time-specific weights were chosen and updated dynamically, based on the past predictive accuracy of a given model in a given channel. Although I was at times bamboozled by the complexities involved, I also found the line of work exciting and intellectually stimulating.

Before the internship, I had hardly considered a career in science at all; there are no academics in my family. Science was simply a different – seemingly even somewhat boring – world to me. Yet, during the internship, I noticed that I actually liked playing around with data, searching systematically for some kind of optimum, developing and applying new models taking into account all kinds of pitfalls that can confound results for various statistical reasons.

As my time in Saint Petersburg wore on, it became increasingly clear to me that I wanted to pursue a career in science—I wanted to play around with large-scale data and prediction models, and to write papers about such work. But I still was blissfully unaware of what a PhD track would actually look like and how difficult it would be to find a good position. So when I got back to the Netherlands, I was supremely

confident that with a master's degree in econometrics I would quickly find a suitable PhD position. Of course, nothing could have been further from the truth; finding a research project that matched my interests and abilities was difficult, and the road from application to approval by the standing committee for research (*Vaste Commissie voor de Wetenschapsbeoefening*) was long and laborsome.

The road I wanted to take became increasingly winding; I even enrolled in a research-master program offered by the *Tinbergen Instituut* (TI). After trying that route for half a year it became clear to me that I was simply on the wrong track. Hence, I could not have been happier when in early 2012, with the help of dr. Jan Brinkhuis, I finally found the ideal position: a PhD project, under the supervision of professors Roy Thurik, Patrick Groenen, and Philipp Koellinger, in which I would be working with big data, but on an even larger scale and on a far more fundamental level than anything I had gotten accustomed to during my time as an intern; I would be working with genetic data from tens of thousands of people, for millions of genetic 'markers'. A latent interest in biology – instilled by a passionate and energetic biology teacher in high school, Hans van Zuylen – got stirred.

Professors Thurik, Groenen, and Koellinger wanted to use molecular genetic data to explain variation in socioeconomic outcomes and differences in individual preferences, and to relate this explained genetic variance to specific regions in the human genome. When I became interested in joining their research group, in the middle of 2012, this whole endeavour was still quite a moonshot. Yet, in spite of the seeming risks – in terms of foregoing a safe career in a well-established field – these professors and their erstwhile PhD candidates, dr. Matthijs van der Loos and dr. Niels Rietveld, pushed this effort forward without wavering; studies of heritability and calculations of statistical power had shown that so-called genome-wide significant hits were within reach. Thanks to the support of dr. Jan Brinkhuis, dr. Adriana Gabor, dr. Christiaan Heij, dr. Andreas Pick, and Peter Post, I was able to secure the job; professors Thurik, Groenen, and Koellinger were willing to hire me as a PhD candidate. After an initial appoint as a research assistant the adventure really started in January 2013!

As I already wrote, I am not quite a scion from a scientific family. Perhaps one could even argue that, in light of my character, a career in science is not necessarily the most obvious choice for me. Playing around with ideas, working with statistical models and methods, scribbling down mathematical derivations, programming, these were all things I felt fairly comfortable doing. But at times I was quite restless; one idea would not even be a fully-fledged project to work on, when my attention typically had already shifted to some other intellectually challenging problem. In addition to lapses in focus, I also had

three supervisors, each supervisor – of course – with his own ideas. Not to mention the international collaborations with assiduous Americans, ambitious Europeans, and goodness knows what more. Halfway through the second year of my PhD track I was simply at a loss; I did not know how to best proceed, which projects to prioritize, which things to work on most diligently, and so on.

But by talking with many other young researchers – in particular dr. Sophie Bruinsma, Ekaterina Isakina, Richard Karlsson Linnér, Fleur Meddens, Gertjan van den Burg, and dr. Peter van der Zwan – I gradually learned that this did not necessarily mean that I would be a lousy scientist; those struggles are simply a part of learning to navigate one's way in science. One is not a scientist simply by nature; gaining experience is vital to shaping and honing the right skill set for a career in science. By the end of the second year, I slowly but surely started finding my way. The manner in which I started tackling research questions became more systematic and focussed, and, hence, projects finally started nearing completion.

As a candidate, I have had the pleasure of working with highly skilled scientists, like dr. Jonathan Beauchamp, dr. Daniel Benjamin, and dr. David Cesarini, and professors Peter Visscher, Naomi Wray, and Jian Yang. In particular, I am very much indebted to professors Visscher and Wray for facilitating a research visit to their group at the *University of Queensland* in 2016. Their ability to commit to ground-breaking research, whilst also keeping an eye on social aspects, deserves praise. I am very grateful for having had the chance to collaborate with them, and I hope to be able to continue collaborating with them in the future. Furthermore, I would like to sincerely thank dr. Beben Benyamin, dr. Fleur Garton, dr. Matt Keller, Robert Maier, Adriaan van der Graaf, dr. Anna Vinkhuyzen, dr. Loic Yengo, dr. Jian Zeng, dr. Zhihong Zhu, and the many others that I got to know during my visit to the *University of Queensland* for their hospitality and the many things that I have learned during my stay.

This preface would be incomplete without thanking dr. Benjamin, dr. Cesarini, and professor Koellinger for their consistent efforts to create permanent scope for research on the genetic architecture of socioeconomic outcomes and preferences by founding the *Social Science Genetic Association Consortium* (SSGAC) and by many other efforts. In addition, I would like to thank professors Koellinger, Groenen, and Thurik as well as dr. Rietveld for their support and supervision, and for the way in which they have promoted and advanced research at the intersection of biology and economics, for instance, by setting up the *Erasmus University Rotterdam Institute for Behavior and Biology* (EURIBEB). I also want to thank professors Bert Hofman and André Uitterlinden, and all those involved in the *Rotterdam Study* for

valuable advice and enabling this exciting line of research.

Access to exceptional computational resources is vital to my research. Hence, I would like to thank *SURFsara* for super-computer services such as *Lisa* and *Cartesius*. In particular, I am extremely thankful to Wim Rijks and Markus van Dijk for all their help during the past four years.

Moreover, I am immensely thankful to the *Erasmus Research Institute of Management* (ERIM), TI, the *Department of Applied Economics* within the *Erasmus School of Economics* (ESE) and all involved in these organisations, for the help and support I have received over the years. In particular, I would like to thank Kim Beerentemfel-Laarman, Gerda de Rave, Manuela Ettekoven, Nita Ramsaransing, and Judith van Kronenburg. In addition, I want to thank my fellow PhD candidates, Indy Bernoster, Casper Burik, dr. Pourya Darnihamedani, Ekaterina Isakina, Richard Karlsson Linnér, Plato Leung, Fleur Meddens, dr. Aysu Okbay, dr. Wim Rietdijk, dr. Niels Rietveld, Eric Slob, Gertjan van den Burg, dr. Matthijs van der Loos, and Caroline Witte, as well as colleagues, dr. Jolanda Hessels, dr. Peter van der Zwan, and dr. André van Stel, and many others. It has been a delight working with you.

I want to thank the members of my PhD committee, professors Paul Eilers, Richard Paap, and Danielle Posthuma as well as dr. Jonathan Beauchamp, dr. Matt Keller, and dr. Katrijn Van Deun, for their willingness to join the committee, and for the efforts and responsibilities involved therein. I would like to thank Gertjan van den Burg for the neat LaTeX template for my thesis. In addition, I would like to thank dr. Peter van der Zwan for his comments on this preface and on the summary of this dissertation. Also, I want to thank my 'paranimfen', Vincent Okhuyzen and Fleur Meddens, for their enthusiasm and for their tremendous help and support.

Finally, I would like to sincerely thank a number of friends, my family, and my partner. Dear Alain, Bertine, Debby, Fatma, Gidius, Gies, Jan, Kirstin, Léon, Niels van der Bijl, Niels Vriethoff, Peter, Ruud, Sarah, Sophie, Stefan, Steven, and Vincent, I cherish our bonds of friendship. The many moments of shared joy have been of great importance in keeping a healthy work-life balance. My dear Marijke and Paul, with much affection I have borne witness to the love and care with which you are raising a family. Loes and Sven often remind me that there is so much more to life than just work. I am very proud to be their uncle. My dear Hans and Willie, you have always supported me in my efforts to find a place where I can flourish. The past four years you have been worried at times about how this type of work would pan out for me. Nevertheless, I have always been under the impression that – as time passed by – you have kept the steadfast belief that, in the end, I would be able to write my thesis and graduate successfully. Thank you

for your kind concerns and your great trust. My dear Jordi, to you I owe the biggest thanks of all, for your unwavering love, support, patience, and – when needed – some impatience. You have been absolutely instrumental in making me see things in ways that helped me to keep my head cool, and in finding the right way forward during the – at times challenging – past four years.

Before moving on to the actual scientific content, I would like to say a few words about the cover of my dissertation, which shows a photo of one of the moons of Jupiter, called Europa (courtesy NASA/JPL-Caltech). Owing to the fact that Europa is about five times as far away from the sun as Earth is, its surface is frigid. Yet it is one of the very few known extraterrestrial places where scientist deem life possible, though not on the surface, but rather in subsurface oceans; these oceans are hypothesized to be heated sufficiently by tidal friction to remain liquid. Should a trace of life ever be found on moon Europa it would illustrate the truly hardy nature of life.

For me personally, however, this picture symbolizes more than the fact that life is versatile. First, our efforts to probe this moon, both by intellect and experiment, tells me that the scientific community is willing to actively consider hypotheses that may seem outlandish at first glance, but which may still hold scientific merit. Second, I find there is tremendous beauty to the idea that we consider, from afar, the potential presence of a microbial form of life – perhaps little more than an intricate dance of a few self-replicating molecules – on a moon which is many millions of miles away from our planet, far away from our society and from our biology.

This connection between the very small and the very large has captivated me for as long as I can remember. I feel the same sense of awe – albeit in a much more humble and earthly sense – when I consider the society that we have built as a species, and realize that we are trying to trace our intricate behaviors, our politics, our culture, our curiosity, our preferences, and so on, all the way back – at least in part – to the molecular building blocks from which we arise.

Although by this point all these metaphorical interpretations may have befuddled you as a reader, there is one last point of view that I want to share. When considering fields such as quantitative genetics, behavior genetics, and genoeconomics or – more broadly – the social sciences, the moon has a proverbial dark side. I suspect plenty of scientists in these fields would willingly confirm that what we know is only the tip of the iceberg; we have a basic understanding of the molecular machinery of life, yet we also have things like missing heritability and only a smattering of independent genome-wide significant loci for many important traits. New aspects of the biological machinery shaping life as we know it are still being discovered at a steady pace.

Hence, both the questions and the awe persist, and images such as this photo of Europa inspire me to keep pondering these unknowns, instilling the desire to explore, probe, and shed new light on the missing pieces in our collective knowledge *ad infinitum*.

Amsterdam, January 16, 2017 Ronald de Vlaming

1

Introduction and Conclusions

ABSTRACT

The recent ascent of large-scale genotyping efforts, combined with the increased collection of data on social-scientific outcomes in such genotyping efforts, has paved the way for studying such outcomes from a genetic perspective. This development has given rise to fields such as genoeconomics, in which the genetic architecture of socioeconomic outcomes and preferences is studied. There are several difficulties in studying the genetic building blocks of such traits. First, social-scientific outcomes tend to be highly polygenic (i.e., affected by many genetic variants), making the contribution of each variant small and, therefore, hard to detect. Second, behavioral traits and preferences can be rather difficult to measure in an objective and homogeneous manner across individuals and studies. Finally, many social-scientific outcomes tend to be interrelated. Therefore, in this thesis, I focus on the use of linear mixed models to elucidate the genetic architecture of such polygenic, heterogeneous, and related traits. In terms of trait complexity and heterogeneity, by developing a versatile calculator of the statistical power of a genome-wide association study, I show that even for highly polygenic traits with substantial heterogeneity across studies, we are now entering an era in which sample sizes are large enough to overcome the problem of both small effect sizes and heterogeneity. Hence, the detection of many of the genetic variants affecting biologically-distal traits is within reach. Regarding correlated traits, I illustrate how the aforementioned linear mixed models can be generalized to deal with multiple traits without becoming numerically infeasible. This thesis (i) underlines the feasibility of uncovering the genetic architecture of social-scientific traits and (ii) provides methodological insights for geneticists interested in the various uses of linear mixed models. This introductory chapter provides an overview of this thesis and its findings. For those who have little or no training in genetics, a non-technical glossary can be found at the end of this chapter.

1.1. MOTIVATION AND CONTRIBUTIONS

The advent of large-scale dense molecular genetic data on so-called single-nucleotide polymorphisms (SNPs) has paved the way for using SNP data to understand the biological nature of complex human traits, to assist in prediction of such traits, and to diagnose and treat complex diseases (Pharoah et al., 2002, Donnelly, 2008, Visscher et al., 2012).

A central concept in the field of genetics is that of heritability, which is defined as the proportion of phenotypic variation that can be explained by genetic variation. Decades of twin studies have revealed that almost any conceivable human trait is to some extent heritable (Polderman et al., 2015). This observation has ramifications not only for fields such as medicine, but also for the social sciences; if one aims to understand why individuals differ, or – as economists would put it – why agents are heterogeneous (e.g., in terms of preferences and behaviors) an understanding of the genetic drivers of such differences is required.

The conjunction of the increased availability of SNP data and the observation that even social behaviors and individual preferences are considerably heritable (Martin et al., 1986, Alford et al., 2005, Fowler et al., 2008, Cesarini et al., 2009, Benjamin et al., 2012b) has given rise to fields such as genoeconomics (Benjamin et al., 2012a), genopolitics (Fowler and Dawes, 2013), and sociogenomics (Robinson et al., 2005).

This thesis lies at the intersection of genoeconomics and statistical genetics, with a strong emphasis on theory and methodology. The contributions of my thesis are fourfold. The first two contributions are presented in Part I and the last two contributions in Part II of this thesis. Importantly, all chapters can be read independently. Moreover, a non-technical glossary can be found at the end of this section.

The first contribution of this thesis can be found in Chapters 3 and 4, where the genetic architecture of reproductive behavior and educational attainment are investigated. These traits are of great importance to the social sciences, including economics (Becker, 1962, Mincer, 1974, Mills and Tropf, 2016). For educational attainment, SNPs are grouped into several functional categories (e.g., coding for proteins and non-coding), after which the proportion of phenotypic variance explained by the respective categories is estimated. Thereby, such an effort helps to understand the biological etiology of educational

attainment at a global level. For reproductive behavior, a genome-wide association study (GWAS) is carried out in order to find SNPs that are robustly associated with different measures of reproductive success. More specifically, in line with the common practice in genetics, GWAS results from different studies are meta-analyzed.

The second contribution of this thesis pertains to the aforementioned widespread practice of meta-analyzing GWAS results. This approach often assumes, implicitly, that the trait of interest has a homogeneous genetic architecture within and across studies. However, in fields such as genetic epidemiology, quantitative genetics, and genoeconomics one often works with heterogeneously measured biologically-distal outcomes. One must, therefore, acknowledge that the environment has ample opportunity to moderate genetic effects, even when traits are homogeneously measured. Consider, for instance, the genetic architecture of years of education under a compulsory-education policy; by putting a lower bound on the years of education one receives, such a policy may dampen the genetic effects, thereby, potentially reducing the heritability of years of education. On this premise, I study in detail how heterogeneity at the level of studies attenuates (i) the statistical power to detect associated SNPs using a meta-analysis of GWAS results and (ii) the predictive accuracy of polygenic predictors constructed from these meta-analysis results (Chapter 2).

Third, a central concept in quantitative genetics is that of genetic covariance between traits, which is defined as the covariance that can be explained by genetic factors. The advent of genome-wide SNP data has facilitated the estimation of genetic variance and covariance in unrelated individuals by means of two forms of restricted maximum likelihood (REML), *viz*., univariate and bivariate genetic-relationship-matrix (GRM) restricted maximum likelihood (GREML), where the elements of the GRM comprise a SNP-based measure of genetic similarity between individuals. However, little attention has yet been paid to joint estimation of genetic and environment covariance matrices for multiple traits using a multivariate GREML approach in lieu of a pairwise bivariate approach. Therefore, I study the theory of multivariate GREML estimation. In this effort, I pay considerable attention to the numerical feasibility and efficiency (Chapter 6).

The fourth contribution of my thesis is to illustrate and establish equivalence principles between different methods in quantitative ge-

netics. Chapter 5 considers five methods used to estimate SNP effects, *viz*., a classical GWAS, multiple regression, ridge regression, best linear unbiased prediction, and maximum *a posteriori* estimation. Based on the existing literature, I illustrate that these methods may be considered as different forms of one general method. Finally, in the last chapter, I establish an equivalence between a measure of population stratification that is estimated using summary statistics from a GWAS and a measure of stratification that is estimated using individual-level data (Chapter 7).

Although the current chapter is written from a first-person singular perspective, subsequent chapters are based on co-authored work. In fact, parts of the work reported in this thesis carry with it long lists of co-authors. There are two main reasons for the multitude of co-authors. The first reason being the 'authorship culture'. In the natural sciences this culture leans more strongly towards substantially co-authored works than in the social sciences (The Economist, 2016a); some large-scale projects in physics have thousands of co-authors (The Economist, 2016b). A field such as genetic epidemiology lies between the extremes of physics and economics; many works in genetic epidemiology have several dozen co-authors. The second reason is that in this line of work one often uses data for which there are legitimate privacy concerns. Therefore, many cohorts carry out relevant analyses themselves (e.g., a GWAS) and only share results (e.g., for meta-analytic purposes)—they typically do not share the genotypes and phenotypes. Consequently, many participating studies contribute in terms of the analyses. In addition to such analyses, many of the participating studies also provide indispensable resources, leadership, and research infrastructures. Hence, investigators and analysts of those studies are typically listed as co-authors. To ensure credit is awarded fairly among those involved in the works presented in this thesis, the contributions of the authors are discussed in considerable detail in this chapter.

The remainder of this introductory chapter is organized as follows. In Section 1.2 I formulate the research questions considered in this thesis. The main results are presented in Section 1.3 and the implications of these results are discussed in Section 1.4. Author contributions are presented in Section 1.5. In Section 1.6 the publication status of each chapter is discussed. Finally, a non-technical glossary can be found in Section 1.7.

1.2. RESEARCH QUESTIONS

Part I – The Architecture of Complex Traits

In this part of the thesis, I first focus on detecting specific SNPs contributing to variation in heterogeneously measured social-scientific traits that are genetically complex. That is, traits that are affected by many genetic variants, where the effects of these variants may be moderated by the environment (i.e., gene–environment interactions), and which may be difficult to measure in an objective fashion across different samples.

CHAPTER 2 – HIDING HERITABILITY AND CROSS-STUDY GENETIC OVERLAP

Since environments tend to differ, not only between individuals, but also between different regions and time periods, and, therefore, between studies, the aforementioned gene–environment interactions that can arise for complex traits can lead to heterogeneity in the genetic architecture across studies. Moreover, biologically-distal traits of relevance to social sciences can be rather difficult to measure homogeneously across studies. Classical calculations of statistical power, however, rely on the assumption of cross-study homogeneity in the phenotypic measure and in its genetic architecture across the studies included in the meta-analysis of GWAS results. Consequently, classical power calculations may be overoptimistic. Similarly, when considering a linear combination of SNPs with SNP-weights based on such a meta-analysis, called a polygenic score (PGS), classical calculations of the predictive accuracy of such a PGS also assume cross-study homogeneity. Hence, such calculations may also be overoptimistic in terms of PGS accuracy. Therefore, the following question is considered in this chapter:

> *What is the exact attenuation (i) of the statistical power to detect associated SNPs in a meta-analysis of GWAS results and (ii) of the predictive accuracy of a PGS constructed using such meta-analysis results under the presence of cross-study heterogeneity in genetic architecture?*

CHAPTER 3 – GWAS ON HUMAN REPRODUCTIVE BEHAVIOR

In the third chapter, two complex traits that have been shown to be strongly affected by environmental factors are considered, *viz*., 'age at first birth' and the 'number of children ever born'. Although previous efforts have already shown the ability of meta-analyses of GWAS results to robustly detect trait-associated SNPs for biologically distal and complex behavioral traits such as educational attainment (Rietveld et al., 2013a, Okbay et al., 2016b) and subjective well-being (Okbay et al., 2016a), this approach has not yet been applied on the same scale to traits reflecting reproductive behavior. Bearing in mind that reproductive success can be considered as a measure of fitness, studying the genetic architecture of reproductive behavior is particularly poignant, not only because of the potential heterogeneity in its genetic architecture but also in light of the obviously strong selection pressure against variants that are deleterious in terms of reproductive success. Hence, the following question is considered in this chapter:

> *Which SNPs associated with differences in age at first birth and the number of children ever born can be discovered by a meta-analysis of GWAS results from over 300,000 individuals?*

CHAPTER 4 – PARTITIONING EDUCATIONAL-ATTAINMENT HERITABILITY

In addition to the discovery of trait-associated SNPs by means of a GWAS, one can also use individual-level genome-wide SNP data to estimate the total proportion of phenotypic variation that can be explained when considering all SNPs jointly. Moreover, since a set of SNPs can be partitioned (e.g., in coding versus non-coding variants), one can assess the contribution of different SNP categories to phenotypic variation. In this chapter, the following question is raised:

> *Can a partitioning of genome-wide SNPs be used in order to disentangle the contribution of different biological mechanisms to variation in a biologically distal and complex trait, viz., educational attainment?*

Part II – Advanced Methods for Individual-Level Data

The first part of this dissertation relies strongly on a so-called linear mixed model (LMM) in which the effects of standardized SNPs are assumed to be random and the effects of certain confounding covariates to be fixed. These random SNP effects are considered to be independent homoskedastic draws from a distribution with mean zero.

In Chapter 2, such an LMM is used in the derivations of the statistical power and predictive accuracy of meta-analyses of GWAS results (Appendix A). In those derivations, it is assumed that the effects of a given SNP are correlated across studies – albeit potentially imperfectly correlated (i.e., less than one) – whilst there is no correlation between the effects of different SNPs.

Chapter 3 relies on such an LMM indirectly, as the theory from Chapter 2 can be used to show that meta-analyses of GWAS results for the traits of interest are sufficiently well-powered to detect several trait-associated SNPs, even under considerable cross-study heterogeneity in the genetic architecture of these traits. For instance, using the theory from Chapter 2, one can show for age at first birth (with a SNP heritability of 15% for women according to Tropf et al. 2015) that (i) when the cross-study genetic correlation is as low as 0.5 and (ii) when there are as many as 20k associated SNPs, a meta-analysis of GWAS results from 62 studies with a pooled sample size of 250k individuals has 85% statistical power to detect at least one associated locus and is expected to find two independent associated loci.

Finally, Chapter 4 employs an LMM where heteroskedasticity in SNP effects across the SNP categories is permitted (preserving the assumption of homoskedasticity within the categories). Finally, LMMs can also be used to jointly estimate SNP effects, predict breeding values, and to estimate genetic covariance between traits and samples.

Hence, given the ubiquitous use of LMMs in current research in statistical genetics, the second part of this thesis considers several of the more theoretical aspects of such models in detail.

CHAPTER 5 – A REVIEW OF RIDGE REGRESSION IN QUANTITATIVE GENETICS

In the statistical-genetics literature much attention is paid to the best linear unbiased prediction of both SNP effects and breeding values,

introduced by Henderson (1975). The best linear unbiased prediction of SNP effects – in essence – is a joint estimate of the effects, even when there are far more SNPs than observations (i.e. $M \gg N$, where M denotes the number of SNPs or makers and N the number of observations). Similarly, in the econometric literature, a method called ridge regression is suitable for joint estimation of the effects of M regressors observed in N observations, which – opposed to multiple regression – yields unique estimates even when $M \gg N$ and/or when there is high multicollinearity amongst the regressors. Finally, so-called maximum *a posteriori* estimation, seen in the Bayesian literature, can also be used to estimate the parameters of a model with more regressors than observations.

In spite of the existence of these sophisticated methods, a GWAS typically involves repeated simple regressions (i.e., performing a regression for each SNP in a dense set of genome-wide SNPs, where in each regression only the confounders and the SNP of interest are included as regressors). This approach completely ignores linkage disequilibrium (LD; i.e., correlation) between SNPs.

These seemingly different approaches can, however, be unified in one framework. Hence, I ponder two questions in this review chapter:

> *First, what is the relation between (i) the best linear unbiased prediction of SNP effects, (ii) maximum a posteriori estimation under a normal prior on SNP effects, (iii) ridge regression, (iv) multiple regression, and (v) the classical GWAS approach of repeated simple regressions? Second, what are the future uses of ridge-regression-type methods in quantitative genetics?*

CHAPTER 6 – MULTIVARIATE AVERAGE-INFORMATION CONSTRAINED GREML

LMMs can be used to estimate genetic and environment covariances between traits using bivariate methods. Such methods typically employ GREML in order to estimate so-called variance and covariance components (e.g., Yang et al. 2011a, Lee et al. 2012).

A natural extension of these bivariate methods is a multivariate approach. However, instead of estimating multivariate LMMs, researchers often estimate pairwise bivariate LMMs for each combination

of two phenotypes from a set of multiple phenotypes. When combining such pairwise bivariate estimates for multiple traits, the resulting genetic and environment covariance matrices may not be covariance matrices at all (i.e., they are not necessarily positive (semi)-definite). In addition, when considering a balanced dataset of N individuals for whom P phenotypes are observed, the computational complexity of a naïve approach is of the order $(NP)^3$ regardless of whether a multivariate or pairwise bivariate approach is used, rendering the estimation problem computationally infeasible in case both N and P grow large. Therefore, I raise the following two questions:

> *First, can a multivariate LMM and its estimation procedure be formulated in such a way (i) that the resulting estimates yield valid genetic and environment covariance matrices (i.e. positive (semi)-definite) and (ii) that the procedure is computationally feasible? Second, what is the statistical efficiency of such a multivariate approach compared to a pairwise bivariate approach?*

CHAPTER 7 – LD-SCORE-REGRESSION INTERCEPT IN INDIVIDUAL-LEVEL DATA

Finally, a recently developed method, called LD-score regression (Bulik-Sullivan et al., 2015b), enables users to disentangle the contribution of polygenic signal and population stratification to the observed inflation in the χ^2-statistics (i.e., the squared Wald-test or t-test statistics) from a GWAS. As before, this work rests on the assumption of random SNP effects in the underlying data-generating process. LD-score regression reports both the estimated SNP heritability and an intercept. Bulik-Sullivan et al. (2015b) show theoretically that this intercept is expected to be greater than one under the presence of confounding stratification.

LD-score regression uses summary statistics from a GWAS as input, whereas REML uses individual-level data. Hence, given the difference in data used by summary-statistics methods and individual-level-data methods, these two classes of methods are often considered to be profoundly different from one another even though both can be used to estimate SNP heritability. This idea is strengthened by the absence of an equivalent quantity for the LD-score-regression intercept – reflecting the amount of confounding population stratification – in methods

using individual-level data. Nevertheless, as these methods are identical in terms of the underlying data-generating process, I posit that these methods are approximately equivalent; both the SNP heritability and the LD-score-regression intercept can be estimated using summary statistics as well as individual-level data. Hence, in this chapter, the following question is considered:

> *How can individual-level data be used to estimate the LD-score-regression intercept directly?*

1.3. RESULTS

CHAPTER 2 – HIDING HERITABILITY AND CROSS-STUDY GENETIC OVERLAP

In order to answer the question how cross-study heterogeneity affects statistical power and predictive accuracy of a GWAS, I develop the online meta-GWAS accuracy and power (MetaGAP) calculator, which takes cross-study genetic correlations into account. This calculator infers the statistical power to detect associated SNPs and the predictive accuracy of a PGS in a meta-analysis of GWAS results from genetically and phenotypically heterogeneous studies, and quantifies the loss in power and predictive accuracy incurred by this cross-study heterogeneity. Using simulations, I show that the MetaGAP calculator is accurate under a wide range of genetic architectures, even when the assumptions of the calculator are strongly violated.

In an empirical application, I use GREML to estimate the SNP-based heritability and cross-study genetic correlation of several polygenic traits across three distinct studies: the Rotterdam Study (RS), the Swedish Twin Registry (STR), and the Health and Retirement Study (HRS). For self-rated health, years of education, body-mass index, and height I obtain point estimates of cross-study genetic correlation between 0.47 and 0.97. Based on these estimates of SNP-based heritability and cross-study genetic correlation, the MetaGAP calculator is used to quantify the expected number of hits and predictive accuracy of the PGS in recent GWAS efforts for these traits. The theoretical predictions align with empirical observations.

For height, under an estimated cross-study genetic correlation of 0.97, the expected loss in the number of genome-wide significant hits due to the imperfect cross-study genetic correlation is 8–9%, whereas for years of education, under an estimated cross-study genetic correlation of 0.78, I expect a loss of 51–62% in the number of hits. Moreover, I find that the relative loss in PGS R^2 is expected to be 6–7% for height and 36–38% for years of education.

CHAPTER 3 – GWAS ON HUMAN REPRODUCTIVE BEHAVIOR

Human reproductive behavior is an important topic of research across the medical, social, and biological sciences (Mills and Tropf, 2016). In this chapter, I report the design and outcomes of the largest GWAS to date on human reproductive behavior, measured by age at first birth and number of children ever born.

The GWAS of reproductive behavior includes 251,151 individuals for age at first birth and 343,072 for number of children ever born. Ten novel associated loci are identified and two recently identified loci are confirmed. These loci harbor genes that are likely to play a role – either directly or by affecting non-local gene expression – in human reproduction and fertility.

CHAPTER 4 – PARTITIONING EDUCATIONAL-ATTAINMENT HERITABILITY

By applying GREML estimation to pooled data from the HRS, the RS, and the STR, I partition the SNP-based heritability of years of education between (i) coding and non-coding regions of the genome and (ii) regions of the genome that are DNase I hypersensitive regions in different cell types.

Partitioned heritability estimates indicate that years-of-education-associated SNPs enrich nonsynonymous sites and regions that are DNase I hypersensitive in both blood cells and the brain. Only the enrichment of regions that are DNase I hypersensitive in blood, however, is statistically significant. The lack of statistical significance is likely to be driven by the poor representation of causal SNPs in enriched regions by the subset of SNPs used for these analyses.

CHAPTER 5 – A REVIEW OF RIDGE REGRESSION IN QUANTITATIVE GENETICS

I investigate the theory underlying ridge regression, best linear unbiased prediction, maximum *a posteriori* estimation, multiple regression, and a classical GWAS for the purpose of performing association analyses and constructing PGSs. As existing literature shows, these five methods can be perceived as a regression of the phenotype on all SNPs jointly, where these methods account for LD between SNPs to different degrees. In fact, a classical GWAS is on one side of the spectrum, giving no weight to LD at all, whereas a multiple regression is on the other side of the spectrum, attempting to fully account for LD. The other methods mentioned, such as ridge regression, lie in between those two extremes. Provided the weight given to the LD is fixed at some non-zero quantity, ridge regression, best linear unbiased prediction, and maximum *a posteriori* estimation are equivalent – up to a scalar – in terms of both the estimated SNP effects and resulting PGSs.

Based on a simulation study, I gauge the current and future potential of ridge-regression-type methods for prediction of human traits using genome-wide SNP data. I conclude that for outcomes with a relatively simple genetic architecture, given current sample sizes in most cohorts (i.e., $N < 10$k) the predictive accuracy of ridge regression is only slightly higher than the classical GWAS approach, which ignores LD. Moreover, both types of methods only capture only a small proportion of the heritability. Based on extrapolations from the simulation results, I posit that in large-scale initiatives sample sizes can be attained where ridge regression improves predictive accuracy substantially when compared to the classical GWAS approach.

CHAPTER 6 – MULTIVARIATE AVERAGE-INFORMATION CONSTRAINED GREML

In this chapter, I develop a multivariate average-information constrained GREML (MacGREML) estimation method. This method consists of an iterative procedure, based on a Newton-Raphson algorithm, to obtain unbiased estimates of the parameters of a multivariate SNP-based LMM for balanced data on P phenotypes observed for N individuals. The LMM is parametrized such that the $(NP) \times (NP)$ covariance matrices are positive (semi)-definite, irrespective of starting values

and updates of the estimates throughout the iterations. I rewrite the log-likelihood, the gradient, and the average-information matrix in terms of the eigendecomposition of an $N \times N$ GRM and transformations of $P \times P$ matrices of parameters. In doing so, I am able to reduce the computational complexity of MacGREML estimation from the order $(NP)^3$ to an order of N^3. In addition, this parametrization is such that two basic factor restrictions can be imposed: (i) a restriction where all traits have a perfect genetic correlation and (ii) a restriction where the traits have no genetic correlation. The significance of the additional fit of the saturated model I employ, compared to the two restricted versions of the model, can be tested easily using a likelihood-ratio test.

CHAPTER 7 – LD-SCORE-REGRESSION INTERCEPT IN INDIVIDUAL-LEVEL DATA

I show that, in an admixed sample drawing from two discrete populations, the theoretical LD-score-regression intercept – measuring the amount of confounding population stratification in the GWAS χ^2-test statistics – can also be estimated using individual-level data. More specifically, I show that the LD-score-regression intercept can be inferred from individual-level data directly by (i) performing ordinary least squares (OLS), where the phenotype is regressed on the leading principal component from the GRM, and (ii) appropriately transforming the resulting OLS estimate. Using simulations, I show that this estimator of the LD-score-regression intercept is approximately unbiased. Moreover, I posit the conjecture that under more complex forms of stratification (i.e., with $P > 2$ discrete populations) an equivalence principle also holds for the LD-score-regression intercept and a transformation of the estimates of a regression using individual-level data.

1.4. CONCLUSIONS AND DISCUSSION

In this dissertation, I considered empirical applications of LMMs in the pursuit of unraveling the genetic architecture of complex and heterogeneous traits, such as educational attainment and reproductive behavior. In addition, I investigated several theoretical aspects of LMMs.

In Chapter 2, I found that cross-study heterogeneity considerably attenuates the statistical power and predictive accuracy of meta-analyses

of GWAS results. This finding has important ramifications for research efforts considering meta-analyses of GWAS results under heterogeneity. First, these results show that such heterogeneity ought to be reckoned with when considering the statistical power and predictive accuracy of a meta-GWAS. Therefore, I believe that the online MetaGAP calculator will prove to be an important tool for assessing whether an intended meta-analysis of GWAS results from different studies is likely to yield meaningful outcomes. Second, these findings stress the importance of considering the use of more sophisticated GWAS meta-analysis methods that account for cross-study heterogeneity (Lebrec et al., 2010, Han and Eskin, 2011, Bhattacharjee et al., 2012, Wen and Stephens, 2014, Shi and Lee, 2016). Finally, the results show that cross-study heterogeneity may contribute to the so-called hiding heritability (Wray et al., 2013b, Witte et al., 2014, Wray and Maier, 2014), which is defined as the difference between the SNP-based heritability estimate (Yang et al., 2010) and the proportion of phenotypic variation explained by genetic variants that reach genome-wide significance in a GWAS.

In Chapter 3, using a large-scale GWAS, twelve independent loci were identified, of which ten are novel, that are robustly associated with age at first birth and/or number of children ever born. These loci harbor genes that are likely to play a role – either directly or by affecting non-local gene expression – in human reproduction and infertility, thereby increasing the understanding of these complex traits. These findings are anticipated to (i) lead to insights into how postponing reproduction may be more detrimental for some – based on their genetic make-up – than others, (ii) fuel experiments to determine "*how late can you wait?*" (Menken, 1985), and (iii) stimulate reproductive awareness. This study is the first to examine the genetics of reproductive behavior in both men and women, and the first that is adequately well-powered to identify loci both in women and men. While effect sizes of the identified common variants are small, there are examples of GWAS-identified loci with small effects that end up leading to important biological insights (Manolio et al., 2008, Hindorff et al., 2009).

In Chapter 4, I followed the method developed by Gusev et al. (2014) and estimated the extent to which the heritable variance of years of education enriches coding SNPs and also SNPs residing in regions that are DNase I hypersensitive in particular cell types. Partitioning heritability in this way can help to elucidate the biological mechanisms

through which genetic variation affects the phenotype of interest. Partitioned heritability estimates suggest that SNPs associated with years of education enrich nonsynonymous sites and regions that are DNase I hypersensitive in both blood cells and the brain. Despite the suggestive findings, only the enrichment for SNPs located in regions that are DNase I hypersensitive in blood cells is statistically significant. The lack of statistical significance is likely to be driven by the poor representation of causal SNPs in enriched regions by the HapMap 3 SNPs (Altshuler et al., 2010) used in these analyses; when attempting to partition a fixed SNP-based heritability with a reduced subset of all common SNPs, the true heritability contributed by a SNP that bears a particular annotation but is missing from the panel must be captured by other SNPs in LD, and these proxy SNPs will often fall in other functional categories. This LD-induced misattribution will tend to reduce the estimated heritability accounted for by SNPs in enriched regions and to increase the estimated heritability accounted for by SNPs in impoverished regions. Gusev et al. (2014) noted that DNase I hypersensitive regions are especially prone to a misallocation of their SNP-based heritability to other regions when panels of SNPs smaller than 1000 Genomes (McVean et al., 2012) are used.

In Chapter 5, I investigated the use of ridge regression for performing a GWAS. Ridge regression can be perceived as method that partially accounts for LD between markers. On the one hand, for a sufficiently low penalty the method fully accounts for LD and is, therefore, equivalent to the OLS estimator of the multiple regression problem using all SNPs jointly. On the other hand, for a sufficiently high penalty ridge regression ignores LD and is, therefore, equivalent – in terms of prediction – to the approach of a simple regression per SNP, which is common in a GWAS. When the amount of weight given to LD is fixed at a non-zero quantity, ridge regression is equivalent to so-called best linear unbiased prediction and maximum *a posteriori* estimation of SNP effects. Using a suite of simulations, I assessed the predictive accuracy of PGSs resulting from ridge regression and from the classical GWAS approach. I found for prediction of complex traits, using training data with sample sizes far below 100k individuals, that PGSs have little predictive accuracy regardless of whether one applies the classical GWAS approach or ridge regression. However, as sample sizes increase to what one typically sees in large-scale efforts (e.g.,

UK Biobank; Ollier et al. 2005), I expect the PGS based on a classical GWAS approach to be able to explain a substantial proportion of the genetic variance. Moreover, under this scenario, prediction using ridge regression is likely to outperform the accuracy of PGSs constructed using classical GWAS results. Therefore, ridge regression in the near future may be able to make a substantial contribution to the prediction of complex traits. This expected gain in predictive accuracy by ridge-regression-type methods has been confirmed in more recent work by Vilhjálmsson et al. (2015), where they show an increase in predictive accuracy of methods accounting for LD in several large-scale samples.

In Chapter 6, I presented a multivariate average-information constrained GREML estimation method. This method consists of an iterative procedure, based on a Newton-Raphson algorithm, to obtain unbiased estimates of the parameters of a multivariate SNP-based LMM for balanced data on P phenotypes observed for N individuals. The LMM has been parametrized such that the $(NP)\times(NP)$ phenotypic covariance matrix is positive definite, irrespective of starting values and updates of the estimates throughout the iterations. By combining various computationally efficient expressions, I am able to provide a method which – in terms of computational complexity – is of the order N^3. This order does not depend on the number of phenotypes being considered. Therefore, the MacGREML estimation method I propose can – in theory – be applied in a large set of phenotypes, provided the data are balanced and only one GRM is considered. Despite these theoretical advances, analyses using simulations and real data are still needed in order to assess the overall empirical merits of this method. Hence, the question whether the statistical efficiency of joint estimation for P traits is considerably higher than the efficiency attained by estimating $P(P-1)/2$ pairwise bivariate models still goes unanswered.

In Chapter 7, I considered the question whether the so-called intercept from LD-score regression, reflecting confounding stratification permeating GWAS summary statistics, can also be estimated directly from individual-level data. I show that, in an admixed sample drawing from two discrete populations, this intercept can indeed be estimated using individual-level data. This theoretical finding illustrates the fact that even though LD-score regression uses summary statistics and, therefore, seems to be profoundly different from LMMs or other methods using individual-level data, an equivalence between these two

types of methods exists. Upon closer inspection this equivalence makes sense, since both the LMM and LD-score regression assume the data-generating process of the phenotype to follow a model with random SNP effects. Using simulations, I show that the individual-level-data estimator of the LD-score-regression intercept is approximately unbiased. Whether this equivalence also holds in empirical data is a question that still needs to be answered.

As this thesis shows – both theoretically and empirically – studies aimed at discovering genetic variants associated with polygenic and heterogeneous social-scientific outcomes are – statistically speaking – increasingly well-powered and are increasingly likely to yield PGSs with sufficient predictive accuracy to be of direct relevance to the social sciences as a whole. Regarding methodology, this thesis provides several examples of the key uses of LMMs. As is shown, the use of LMMs (i) is indispensable for *a priori* inferences on statistical power and predictive accuracy of an intended meta-GWAS, (ii) enables estimating the SNP heritability and genetic covariance of multiple traits jointly, (iii) allows users to partition SNP heritability according to (biological) function, thereby, providing insight into the etiology of traits, (iv) helps improving the accuracy of PGSs, and (v) reveals that parameters estimated from summary statistics may also be estimated from individual-level data directly.

1.5. INDIVIDUAL CONTRIBUTIONS

Chapter 2 was conceptualized by me and Philipp D. Koellinger. I developed the MetaGAP calculator as well as the underlying theory and methodology. Aysu Okbay and I validated this method and carried out the analyses. Resources to carry out this project were made available by Philipp D. Koellinger and A. Roy Thurik. Empirical data were made available by Magnus Johannesson, Patrik K. E. Magnusson, André G. Uitterlinden, Frank J. A. van Rooij, and Albert Hofman. Quality control of the data was performed by me and Aysu Okbay. The original draft was written by me, Aysu Okbay, and Philipp D. Koellinger. Further reviewing and editing was done by me, Philipp D. Koellinger, Aysu Okbay, A. Roy Thurik, Patrick J. F. Groenen, and Cornelius A. Rietveld.

Chapter 3 is based on parts of the work by Barban et al. (2016). The full manuscript comprises a wide range of analyses, such as an

investigation of population stratification, polygenic prediction, and the biological annotation of the GWAS results. Hence, the manuscript in its entirety entails far more work and consists of far more results than the parts reported in my thesis. Regarding the full study, Melinda C. Mills, Harold Snieder, and Marcel den Hoed led the design. In addition, Philipp D. Koellinger, Daniel J. Benjamin, David Cesarini, and Nicola Barban contributed to the design of the study. The study was managed by Melinda C. Mills and Nicola Barban. Nicola Barban and Felix C. Tropf investigated population stratification; I contributed to this investigation in terms of the analyses using LD-score regression. Analyses regarding genetic correlations and PGSs were carried out by Nicola Barban. Meta-analysis and quality control (both at the sex-specific and at the pooled level) were performed by Nicola Barban and me, as well Jornt J. Mandemakers and Ilja M. Nolte. Analyses for the biological annotation were carried out by Rick Jansen, Marcel den Hoed, and Ahmad Vaez. Sex-specific genetic effects were investigated by Nicola Barban and Felix C. Tropf. Bivariate and conditional analyses of the two fertility traits were carried out by Xia Shen, James F. Wilson, and Daniel I. Chasman. Gene-based analysis were performed by Vinicius Tragante and Sander W. van der Laan. Regarding the writing of the parts included in my thesis, I did contribute directly to the write-up of Sections 3.2–3.6. The abstract as well as Sections 3.7 and 3.8 are based on excerpts from the full manuscript by Barban et al. (2016) and the corresponding supplementary information, with minor changes in wording to align with the contents of this thesis. Section 3.1, also apart from minor changes in wording, is based directly on parts of the supplementary information that were written by Melinda C. Mills. Finally, this work, as it is reported in this introductory chapter, is based directly on the content of Chapter 3 in this dissertation, which in turn draws from the full manuscript by Barban et al. (2016) in the aforementioned manners.

Chapter 4 is based on one section of the supplementary information to the work by Okbay et al. (2016b). This manuscript has resulted from efforts and resources provided by hundreds of co-authors. I do not claim credit for any other piece of this manuscript than the supplementary-information section on partitioning the heritability of years of education using GREML. Consequently, I will here only report the contributions of others to this specific section. This work was conceptualized by

Philipp D. Koellinger, Daniel J. Benjamin, and Jonathan P. Beauchamp. Empirical data were made available by Magnus Johannesson, Patrik K. E. Magnusson, André G. Uitterlinden, Frank J. A. van Rooij, and Albert Hofman. Analyses were carried out by me. Quality control of the data was performed by me and Aysu Okbay. The original draft was written by James J. Lee, Jonathan P. Beauchamp, and me. Further reviewing and editing was done by me, James J. Lee, Jonathan P. Beauchamp, Daniel J. Benjamin, and Philipp D. Koellinger.

Chapter 5 was conceptualized by Patrick J. F. Groenen and me. I carried out the review of the literature. Derivations were carried out by me and checked by Patrick J. F. Groenen. I conceptualized and carried out the simulation study. The original draft was written by me. Further reviewing and editing was done by Patrick J. F. Groenen and me.

Chapter 6 was conceptualized by me and Patrick J. F. Groenen. I carried out the derivations and wrote the original draft. Reviewing and editing was done by Patrick J. F. Groenen and me.

Chapter 7 was conceptualized by Peter M. Visscher and me. I carried out the derivations and wrote the original draft. Reviewing and editing was done by Peter M. Visscher and me.

1.6. PUBLICATION STATUS

Chapter 2 has been published in *PLOS Genetics* in January 2017 (De Vlaming et al., 2017). Chapter 5 has been published in its entirety in *BioMed Research International* in 2015 (De Vlaming and Groenen, 2015).

As indicated, Chapter 3 is based on the broader work by Barban et al. (2016), which has been published in *Nature Genetics* in 2016. Similarly, Chapter 4 is based on a small part of the work by Okbay et al. (2016b), which has been published in *Nature* in 2016.

Finally, Chapters 6 and 7 are based on manuscripts which are still in a theoretical stage. Once adequate empirical applications have been found, the aim is – of course – to publish the resulting manuscripts in peer-reviewed journals.

1.7. GLOSSARY

Allele – A genetic variant observed within a population at a given locus.

Allosome – See sex chromosome.

Assortative mating – A pattern where pairs of individuals with similar phenotypes tend to mate more frequently than individuals with dissimilar phenotypes.

Autosomal biallelic SNP – A SNP located on an autosomal chromosome with two alleles occurring within the population. This type of SNP can be recoded in terms of the number of minor alleles occurring on the coding strand of a SNP. *Example*: a biallelic SNP with alleles A-C and T-G, where the first letter denotes the abbreviated nucleotide on the coding strand and the second letter the nucleotide on the alternative strand. Across the chromosomes in the autosomal pair, an individual can have the following genotypes: (i) A-C, A-C, (ii) A-C, T-G, and (iii) T-G, T-G (where the two subsequent pairs of abbreviated nucleotides indicate the observed genotypes across the two autosomal chromosomes). If A on the coding strand is the minor allele, we have the following minor allele counts (MAC) for the three possible genotypes: (i) MAC = 0, (ii) MAC = 1, and (iii) MAC = 2.

Autosome – A type of chromosome that comes in pairs, where both autosomes in a pair have the same form, one coming from the father and one from the mother, with the only major differences between the pairs being in terms of the specific alleles present at common-variant loci.

Base pair – Complementary nucleotides on the two strands in the DNA at a given locus. The human genome consists of approximately 3 billion base pairs (Collins et al., 2004).

Best linear unbiased prediction (ABBREV. BLUP) – Posterior estimates of random effects in an LMM.

Biallelic – Having two alleles within a population.

Breeding value – The genetic contribution to a phenotype, when written as $y = g + e$, where y is the phenotype, g the genetic contribution aggregated across all variants, and e the contribution of the environment. The breeding value can be conceptualized as a polygenic score, where estimated effects are replaced by 'true' effects.

Broad-sense heritability – Specific definition of heritability, where this definition includes the phenotypic variation explained by dominance and epistasis.

Chromosome – A structure consisting of one long DNA molecule.

Coding region – A region in the DNA that codes directly for a protein. Variants in this region may affect protein shape and functioning.

Coding SNP – A SNP in a coding region. Coding SNPs that have alleles which leave the protein being coded for unaltered are referred to as synonymous SNPs, whereas coding SNPs for which the different alleles lead to different proteins are referred to as nonsynonymous SNPs.

Collinearity – See multicollinearity.

Common SNP – A (biallelic) SNP with a considerable MAF (e.g., MAF > 1%).

Complex trait – A trait which is typically (i) not fully heritable (i.e., heritability less than one) and (ii) polygenic, and which may be shaped by both gene–gene and gene–environment interactions. Lander and Schork (1994) describe complex traits as an all-inclusive category of traits, for which a one-to-one "[...] *correspondence between genotype and phenotype breaks down, either because the same genotype can result in different phenotypes* [...] *or different genotypes can result in the same phenotype*".

Confounder – A variable associated with both regressor and regressand, inducing a spurious association between regressor and regressand when omitted as control variable.

Covariance component – The covariance between two phenotypes that can be explained by a certain similarity metric. *Example:* the covariance between two traits across individuals that coincides with genetic similarity across individuals.

Cross-study genetic correlation – The genetic correlation of a trait in one study with a trait in another study. In bivariate GREML estimation, this parameter is identified by genetic 'chance' similarity between individuals across studies. Under a non-zero cross-study genetic correlation, phenotypic similarity between pairs of individuals across studies is expected to be affected by such chance similarities in genotypes between individuals from the different studies.

Deoxyribonucleic acid (ABBREV. DNA) – A molecule consisting of two strands coiled around each other, forming a double helix. Each strand consists of a series different nucleotides, each nucleotide being either cytosine (C), guanine (G), adenine (A), or thymine (T). These nucleotides encode information for 'building' organisms using molecular biological machinery. The two strands in a DNA molecule are complementary, in the sense that A on one strand matches with T on the other strand, and C on one strand matches with G on the other strand. A pair of complementary nucleotides is called a base pair. Between the members of a given species, the majority of the information encoded by the base pairs is identical.

DNase I hypersensitive site (ABBREV. DHS) – A region in the DNA that is physically exposed at the molecular level when the DNA is folded in a given manner. Importantly, the folding of DNA varies between cell types,

causing DHSs to differ across cell types. A DHS often constitutes an important regulatory pathway, affecting the expression of coding regions.

Dominance – A violation of additive linear effects of alleles such that, in case of a biallelic locus, the expected phenotypic value of the heterozygote is not the midpoint of the expected phenotypic values of the two homozygotes. Can be represented statistically as an interaction of a locus with itself. *Example:* the expected phenotypic values for the genotypes (i) A-T, A-T, (ii) A-T, G-C, and (iii) G-C, G-C are given by 0, 10, and 100, respectively, instead of 0, 50, and 100.

Endophenotype – An intermediate phenotype (i.e., a mediating factor on the path from genotype to phenotype).

Enrichment – A pattern where a region in the DNA explains a disproportionally large share of the heritability. In case of SNPs: when a subset of M SNPs, from a set of P genome-wide SNPs, explains significantly more than fraction M/P of the SNP-based heritability inferred from the P genome-wide SNPs.

Epistasis – Interaction effects between different loci.

Gene–environment interaction (ABBREV. $\mathbf{G} \times \mathbf{E}$) – Genetic effects that are moderated by environment factors and *vice versa*.

Gene–gene interaction (ABBREV. $\mathbf{G} \times \mathbf{G}$) – See epistasis and dominance.

Genetic-relatedness matrix – See genomic-relatedness matrix.

Genetic architecture – A description of how the trait of interest is shaped by genetics. At the global level, the genetic architecture of a trait is described by aspects such as the polygenicity, the number of trait-affecting rare and common variants, the relation between the frequency of trait-affecting alleles and effect size, whether the effects of genetic variants are additive, nonlinear (i.e., dominance effects and epistasis), and/or moderated by the environment (i.e., gene–environment interaction), and the degree of genetic heterogeneity across environments and populations. At a lower level, the genetic architecture can be described by the specific variants that are associated with the trait, and the biological pathways and endophenotypes via which these variants affect the trait.

Genetic correlation – The ratio of the genetic covariance of two traits and the square root of the product of the genetic variance of the two traits.

Genetic risk score – See polygenic score.

Genetic value – See breeding value.

Genome-wide association study (ABBREV. GWAS) – A massive association analysis, where a phenotype of interest is regressed on a set of control variables and a dense set of M SNPs, one SNP at a time. Hence, a GWAS typically consists of running M regression analyses, and assessing the genome-wide significance of the estimated effect of the given SNP.

Genome-wide significance – Since a GWAS typically considers $M \gg 10^5$ SNPs, a Bonferroni correction based on the number of independent test is used to keep the false-positive rate low. Common practice is a corrected significance level of $\alpha = 5 \cdot 10^{-8}$.

Genome-wide significant hit – A SNP that reaches genome-wide significance in a GWAS.

Genotype – The inherited genetic makeup of an individual.

Genotyped SNP – A SNP that is directly measured using a genotyping array.

Genotyping – The process of assessing an individual's genotype.

Genomic-relatedness matrix (ABBREV. GRM) – A matrix that consists of estimates of relatedness based on a standardized $N \times M$ matrix $\mathbf{X}$ consisting of M SNPs observed in N individuals. The GRM is then defined as $\mathbf{A} = M^{-1}\mathbf{X}\mathbf{X}^\top$.

Genomic-relatedness-matrix restricted maximum likelihood (ABBREV. GREML) – Univariate GREML estimates the additive genetic variance and environment variance components using a SNP-based GRM. Bivariate GREML estimates genetic and environment variance and covariance. For bivariate GREML the environment covariance is only identified when the samples for the two phenotypes overlap at least partially.

Genomic-relationship matrix – See genomic-relatedness matrix.

GWAS hit – See genome-wide significant hit.

Hardy-Weinberg equilibrium (ABBREV. HWE) – This equilibrium holds when the minor allele count of an autosomal biallelic SNP follows a Binom$(2, f)$ distribution, where f denotes the frequency of the minor allele on the coded strand. Under HWE, the expected frequency across the population for minor allele counts (MACs) equal to zero is $(1-f)^2$, for MACs equal to one is $2f(1-f)$, and for MACs equal to two is f^2. An exact χ^2-test can be used to test deviations from the expected frequencies. This test is referred to as an HWE test. Under the absence of assortative mating (i.e., random mating) and no differences in allele frequencies for a SNP of interest across admixed populations, HWE is generally considered to hold.

Heritability – The proportion of phenotypic variation that can be explained by genetic variants.

Heteroskedastic – Not having the same variance.

Heterozygote – An individual who, at a given autosomal locus, has different genotypes on both chromosomes in the autosomal pair. *Example:* an individual with genotype T-A, C-G at a given autosomal locus.

Hiding heritability – The gap between SNP-based heritability estimates and the proportion of phenotypic variance explained by GWAS hits to date.

Homoskedastic – Having the same variance.

Homozygote – An individual who, at a given autosomal locus, has the same genotype on both chromosomes in the autosomal pair. *Example:* an individual with genotype T-A, T-A at a given autosomal locus.

Human genome – A genome that comprises 22 different types of autosomes, where each autosome has two copies (i.e., one from the father and one from the mother) and two allosomes (i.e., one X chromosome from the mother and, in case of a female, also an X chromosome from the father, and, in case of a male, a Y chromosome from the father).

Impoverishment – The opposite of enrichment.

Imputed SNP – SNPs that are not directly genotyped can be imputed by combining the genotyped SNPs from the data of interest with a dense set of genotyped SNPs in a reference sample (e.g., HapMap; Bentley et al. 2003, or 1000Genomes; McVean et al. 2012). The more common a SNP is and the higher its LD with genotyped SNPs, the higher the accuracy of the imputation. The primary use of imputed SNPs is that it improves the overlap between GWAS results from studies that have used different genotyping platforms.

Linear mixed model (ABBREV. LMM) – A model where a subset of regressors are assumed to have random effects and a subset of the regressors fixed effects. In quantitative genetics an LMM that is frequently used assumes that SNPs have random effects and potential confounders fixed effects. When (i) standardized SNP are assumed to have homoskedastic independent normal effects and (ii) the residuals are assumed to be independent homoskedastic draws from a normal distribution (i.e., independent environment effects), these two sets of random effects induce a phenotypic covariance matrix that is shaped by a linear combination of the GRM and the identity matrix, weighted according to a genetic variance component and an environment variance component, where the total phenotypic variance is given by the sum of these two components.

Linkage disequilibrium (ABBREV. LD) – The correlation structure between SNPs. Typically SNPs in close proximity to each other on the DNA are highly correlated in terms of genotypes. This correlation decreases with distance. However, long-distance LD has been known to occur, and may point to epistatic effects and/or population stratification.

Locus (PLURAL loci) – A specific location in the genome.

Major allele – The allele that is most common within the population at the locus of interest.

Marker – A locus in the DNA where variation occurs within a population. In this thesis, SNPs are the only markers considered.

Maximum *a posteriori* (ABBREV. MAP) – MAP estimation obtains estimates of the parameters of a model by finding the maximum (i.e., mode) of the posterior density function (i.e., likelihood).

Meta-GWAS – A meta-analysis of GWAS results from different studies.

Minor allele – The allele that is least common within the population at the locus of interest.

Minor allele frequency (ABBREV. MAF) – The frequency of the allele that is least common within the population at the locus of interest.

Missing heritability – The gap between heritability estimates from twin- and family-based studies and the proportion of phenotypic variance explained by GWAS hits to date. Colloquially referred to as the dark matter of genetics.

Mixed linear model (ABBREV. MLM) – See linear mixed model.

Molecular genetics – The field of genetics aimed at measuring and inferring genetic structures and mechanisms, at a molecular level.

Monogenic trait – A trait that is shaped by a single genetic variant.

Multicollinearity – A property of a set of variables, such that there exists at least one non-trivial linear combination of these variables (i.e., a linear combination where at least one of the variables receives a non-zero weight) which is equal to a vector of zeros.

Narrow-sense heritability – Specific definition of heritability, considering the proportion of phenotypic variation explained when considering additive contributions of variants at different loci. This measure of heritability is typically inferred from a twin study, where the similarity of dizygotic twins is compared to the similarity of monozygotic twins.

Noncoding DNA – Regions in the DNA that do not directly code for a protein, but which may be involved in regulatory pathways (e.g., the regulation of the expression of coding regions).

Nonsynonymous SNP – See coding SNP.

Ordinary least squares (ABBREV. OLS) – A method that estimates parameters by minimizing the sum of squared differences between an outcome variable and a linear combination of predictor variables.

Phenotype – Measurable characteristics of an individual, including biologically-distal outcomes such as behavior and products of behavior (e.g., the highest degree of education attained or the number of children ever born). In genetic epidemiology and quantitative genetics the phenotype is typically treated as regressand.

Pleiotropy – A gene affecting multiple traits is called pleiotropic. Pleiotropy can induce correlations between traits. An estimate of genetic correlation significantly different from zero is indicative of underlying genes being pleiotropic. However, a genetic correlation equal to zero does not mean there is no pleiotropy, in the same way that, in probability theory, no correlation does not equal independence.

Polygenic – Shaped by variants at many loci in the DNA.

Polygenic score (ABBREV. PGS) – A linear combination of SNP data with the aim of predicting (polygenic) traits. Main approaches for constructing such a score are counting the number 'risk' alleles (i.e., trait-increasing alleles) a person has across (a subset of the) SNPs, or weighting SNPs according to the SNP-effect estimates from a GWAS.

Polygenic risk score – See polygenic score.

Polygenicity – The degree to which a trait is polygenic.

Population admixture – See population stratification.

Population stratification – A pattern where allele frequencies differ significantly across two or more subpopulations, within one large population. If differences in allele frequencies coincide with differences in phenotypic mean, not accounting for population stratification can lead to false GWAS results. The common way to control for population stratification in a GWAS, is to include the top principal components from the GRM as covariates.

Quantitative genetics – A branch of genetics that is concerned with the variation in continuous traits – typically normally distributed – that is attributable to genetic variation.

Random mating – No discernible relation between mating probabilities and phenotypic similarity.

Rare SNP – A (biallelic) SNP with a low MAF (e.g., MAF $\leq 1\%$).

Regressor – Explanatory variable in a regression.

Regressand – Outcome variable in a regression.

Restricted maximum likelihood (ABBREV. REML) – REML estimation is a maximum likelihood method that provides unbiased estimates of variance components (e.g., phenotypic variance in the simplest case) in case fixed-effect covariates play a role.

Ridge regression – A regularized form of ordinary least squares, enforcing a unique solution for the effect estimates of regressors, even under perfect collinearity and/or more regressors than observations.

Sex chromosome – A type of chromosome that comes in pairs, not necessarily of the same form, with either an X or a Y chromosome coming from the father, and an X chromosome from the mother. Females have two X chromosomes and males have one X and one Y chromosome.

Single-nucleotide polymorphism (ABBREV. SNP, IPA PRON. /snɪp/, PLURAL SNPs, IPA PRON. /snɪps/) – A specific base pair where the complementary nucleotides vary across members of a species (e.g., a base pair where across the population we observe the following alleles: A-C, C-A, G-T, and T-G; where the first letter denotes the abbreviated nucleotide on the coding strand and the

second letter the nucleotide on the alternative strand). Most SNPs are biallelic, meaning that only two combinations of complementary base pairs are observed in the population at the given locus. *Example*: at a given locus A-C is observed in 90% of the population (population of genotypes at the given locus, in this case), T-G is observed in the 10%, and other genotypes are not observed.

Statistical genetics – The field concerned with drawing inferences from genetic data by developing new statistical methods and applying existing methods.

Still-missing heritability – The gap between heritability estimates from twin- and family-based studies on the one hand, and SNP-based heritability estimates on the other hand.

SNP-based heritability (ABBREV. h^2_{SNP}) – Idem to narrow-sense heritability, yet only considering the proportion of phenotypic variation explained by the additive effects stemming from a dense set of SNPs.

SNP heritability – See SNP-based heritability.

Trait – See phenotype.

Variance component – The phenotypic variance that can be explained by a certain similarity metric. *Example:* the variance captured by the GRM, which may be referred to as a genetic variance component.

I

The Architecture of Complex Traits

2

Hiding Heritability and Cross-Study Genetic Overlap

Based on De Vlaming et al. (2017)

ABSTRACT

Large-scale genome-wide association results are typically obtained from a fixed-effects meta-analysis of GWAS summary statistics from multiple studies spanning different regions and/or time periods. This approach averages the estimated effects of genetic variants across studies. In case genetic effects are heterogeneous across studies, the statistical power of a GWAS and the predictive accuracy of polygenic scores are attenuated, contributing to the so-called 'missing heritability'. Here, we describe the online Meta-GWAS Accuracy and Power (MetaGAP) calculator (available at www.devlaming.eu) which quantifies this attenuation based on a novel multi-study framework. By means of simulation studies, we show that under a wide range of genetic architectures, the statistical power and predictive accuracy provided by this calculator are accurate. We compare the predictions from the MetaGAP calculator with actual results obtained in the GWAS literature. Specifically, we use genomic-relatedness-matrix restricted maximum likelihood to estimate the SNP heritability and cross-study genetic correlation of height, BMI, years of education, and self-rated health in three large samples. These estimates are used as input parameters for the MetaGAP calculator. Results from the calculator suggest that cross-study heterogeneity has led to attenuation of statistical power and predictive accuracy in recent large-scale GWAS efforts on these traits (e.g., for years of education we estimate a relative loss of 51–62% in the number of genome-wide significant loci and a relative loss in polygenic score R^2 of 36–38%). Hence, cross-study heterogeneity contributes to the missing heritability.

2.1. INTRODUCTION

Large-scale GWAS efforts are rapidly elucidating the genetic architecture of polygenic traits, including anthropometrics (Wood et al., 2014, Locke et al., 2015) and diseases (Eeles et al., 2009, Ehret et al., 2011, Ripke et al., 2014), as well as behavioral and psychological outcomes (Rietveld et al., 2013a, Okbay et al., 2016b,a). These efforts have led to new biological insights, therapeutic targets, and polygenic scores (PGS), and help to understand the complex interplay between genes and environments in shaping individual outcomes (Okbay et al., 2016b, Visscher et al., 2012, Benjamin et al., 2012a). However, GWAS results do not yet account for a large part of the estimated heritability (Wood et al., 2014, Locke et al., 2015, Okbay et al., 2016b,a). This dissonance, which is referred to as the 'missing heritability', has received broad attention (Maher, 2008, Manolio et al., 2009, Eichler et al., 2010, Zuk et al., 2012, Wray et al., 2013b, Witte et al., 2014, Wray and Maier, 2014).

Differences across strata (e.g., studies and populations), in genetic effects, phenotype measurement, and phenotype accuracy, lead to loss of signal (Evangelou et al., 2011, Wray et al., 2012, 2013a). Hence, such forms of heterogeneity attenuate the statistical power of a GWAS (Evangelou et al., 2011, Wray and Maier, 2014, Lee et al., 2013, Sham and Purcell, 2014) and the predictive accuracy of a PGS in a hold-out sample (Dudbridge, 2013), and, thereby, contribute to the missing heritability. Since large-scale GWAS results are typically obtained from a meta-analysis of GWAS results from many different studies, we focus on the attenuation resulting from heterogeneity at the level of studies included in such a meta-analysis. Given the importance of discovering trait-affecting variants and obtaining accurate polygenic predictions, it is vital to understand to which extent cross-study heterogeneity attenuates the statistical power and predictive accuracy of GWAS efforts. By considering cross-study differences in genetic effects and heritability, we can quantify this attenuation.

Despite empirical evidence of transethnic genetic heterogeneity in diseases (Brown et al., 2016) and the fact that cross-study heterogeneity has been found to decrease the chances of a study to yield meaningful results (Sham and Purcell, 2014, Wray et al., 2007), a theoretical multi-study framework that quantifies the effect of cross-study heterogeneity

on statistical power and predictive accuracy is still absent. We bridge this gap by developing a Meta-GWAS Accuracy and Power (MetaGAP) calculator (available at www.devlaming.eu) that accounts for the cross-study genetic correlation (CGR). This calculator infers the statistical power to detect associated SNPs and the predictive accuracy of the PGS in a meta-analysis of GWAS results from genetically and phenotypically heterogeneous studies, and quantifies the loss in power and predictive accuracy incurred by this cross-study heterogeneity. Using simulations, we show that the MetaGAP calculator is accurate under a wide range of genetic architectures, even when the assumptions of the calculator are violated.

Although meta-analysis methods accounting for heterogeneity exist (Lebrec et al., 2010, Han and Eskin, 2011, Morris, 2011, Bhattacharjee et al., 2012, Wen and Stephens, 2014, Shi and Lee, 2016), large-scale GWAS results are typically still obtained from fixed-effects meta-analysis methods (Evangelou and Ioannidis, 2013, Nalls et al., 2014) such as implemented in `METAL` (Willer et al., 2010). Therefore, the MetaGAP calculator assumes the use of a fixed-effects meta-analysis method. Thus, the calculator will help researchers to assess the merits of an intended fixed-effects meta-analysis of GWAS results and to gauge whether it is more appropriate to apply a meta-analysis method that accounts for heterogeneity.

In an empirical application, we use genomic-relatedness-matrix restricted maximum likelihood (GREML) to estimate the SNP-based heritability (h^2_{SNP}) and CGR of several polygenic traits across three distinct studies: the Rotterdam Study (RS), the Swedish Twin Registry (STR), and the Health and Retirement Study (HRS). For self-rated health, years of education, BMI, and height, we obtain point estimates of CGR between 0.47 and 0.97. Based on these estimates of h^2_{SNP} and CGR, we use the MetaGAP calculator to quantify the expected number of hits and predictive accuracy of the PGS in recent GWAS efforts for these traits. Our theoretical predictions align with empirical observations.

For height, under an estimated CGR of 0.97, the expected relative loss in the number of genome-wide significant hits is 8–9%, whereas for years of education, under an estimated CGR of 0.78, we expect a relative loss of 51–62% in the number of hits. Moreover, we find that the relative loss in PGS R^2 is expected to be 6–7% for height

and 36–38% for years of education. Hence, our findings show that cross-study heterogeneity attenuates the statistical power and PGS accuracy considerably, thus, contributing substantially to the missing heritability, and, more specifically, to the 'hiding heritability' (Wray et al., 2013b, Witte et al., 2014, Wray and Maier, 2014) – defined as the difference between the SNP-based heritability estimate (Yang et al., 2010) and the proportion of phenotypic variation explained by genetic variants that reach genome-wide significance in a GWAS.

2.2. MATERIALS AND METHODS

2.2.1. *Definitions and assumptions*

The MetaGAP calculator is based on theoretical expressions for statistical power and PGS accuracy, derived in Appendices A.1 and A.2. In these expressions, within-study estimates of SNP heritability (e.g., inferred using `GCTA`; Yang et al. 2011a) are required input parameters. Estimates of CGR (e.g., inferred as genetic correlations across studies using pairwise bivariate methods as implemented in `GCTA` (Lee et al., 2012) and LD-score regression (Bulik-Sullivan et al., 2015b,a), or as genetic-impact correlation from summary statistics (Brown et al., 2016)) also play a central role in those expressions. As we show in Appendix A.3, such estimates of CGR are affected by the cross-study overlap in trait-affecting loci as well as the cross-study correlation in the effects of these overlapping loci. In our derivations of statistical power and predictive accuracy, we assume, however, that the set of trait-affecting loci is the same across all studies and that CGRs are, consequently, shaped solely by cross-study correlations in the effects. Using simulation studies, discussed in Appendix A.4, we assess how violations of this assumption affect our results.

In addition, genetic correlations as inferred using `GCTA` (Lee et al., 2012) or LD-score regression (Bulik-Sullivan et al., 2015a) effectively estimate the cross-trait and/or cross-study correlation in the effects of standardized SNPs. This correlation has been referred to as the genetic-impact correlation (Brown et al., 2016). The scale of rare variants is inflated most by standardization (i.e., genotypes are scaled by $1/\sqrt{2f(1-f)}$, where f denotes the allele frequency of the SNP of interest). Therefore, the scale of the effects of these variants is de-

creased most by standardization of SNPs (i.e., when standardizing a SNP, the effect is scaled by $\sqrt{2f(1-f)}$). Hence, the genetic-impact correlation emphasizes the contribution of common variants (Brown et al., 2016). If rare alleles tend to have larger effects than common alleles, as assumed in GCTA (Yang et al., 2011a) and LD-score regression (Bulik-Sullivan et al., 2015b), these two opposing forces may cancel each other out; the effects of rare alleles are then bigger, but also scaled downwards more strongly by considering standardized SNPs. Alternatively, one can also consider the correlation in the effect of non-standardized SNPs, referred to as the genetic-effect correlation (Brown et al., 2016). This genetic-effect correlation gives rare and common variants equal weight in theory. However, in case rare alleles have larger effects than common alleles, this genetic-effect correlation, in practice, gives a disproportional weight to rare variants.

A clear definition of genetic correlation can be further complicated by the presence of allele frequency differences across samples. Whereas GCTA assumes fixed allele frequencies across the samples included in the analysis (Yang et al., 2011a), there also exist methods which allow for differences in allele frequencies. Ideally, estimates of cross-study genetic-impact correlation accounting for allele frequency differences (Brown et al., 2016) should be used in the MetaGAP calculator as input for CGR. However, provided the genetic drift is small, whether to account for allele frequency differences across samples or not, will – in all likelihood – hardly affect the CGR estimates. Therefore, under little genetic drift, estimates of CGR obtained by methods ignoring cross-study differences in allele frequencies (e.g., bivariate GREML; Lee et al. 2012), suffice as input for the MetaGAP calculator.

In line with other work, we define the effective number of SNPs, S, as the number of haplotype blocks (i.e., independent chromosome segments; Daetwyler et al. 2008), where variation in each block is tagged by precisely one genotyped SNP. By genotyped SNPs we also mean imputed SNPs. Hence, in our framework, there are S SNPs contributing to the polygenic score. Due to linkage disequilibrium (LD) this number is likely to be substantially lower than the total number of SNPs in the genome (Li et al., 2012), and is inferred to lie between as little as 60k (Wray et al., 2013b) and as much as 5 million (Li et al., 2012).

In terms of trait-affecting variants, we consider a subset of M

SNPs from the set of S SNPs. Each SNP in this subset tags variation in a segment that bears a causal influence on the phenotype. We refer to M as the associated number of SNPs. We assume that the M associated SNPs jointly capture the full SNP-based heritability for the trait of interest and, moreover, that each associated SNP has the same theoretical R^2 with respect to the phenotype. In the simulation studies, we also assess the impact of violations of this 'equal-R^2' assumption.

By considering only independent genotyped SNPs that are assumed to fully tag the causal variants, we can ignore LD among genotyped variants and between the causal variant and the genotyped variants. Thereby, we can greatly reduce the theoretical and numerical complexity of the MetaGAP calculator. However, a genotyped tag SNP does not necessarily capture the full variation of the causal variant present in that independent segment. Nevertheless, the inputs for SNP heritability used in the MetaGAP calculator are within-study GREML estimates of heritability, based on the available SNPs. Therefore, if these genotyped SNPs are in imperfect LD with the causal variants, this will lead to a downward bias in the SNP-based heritability estimates (Yang et al., 2015a). Hence, the imperfect tagging of the causal variants is likely to be absorbed by a downward bias in the SNP-based heritability estimates.

2.2.2. *Statistical power of a GWAS meta-analysis*

The theoretical distribution of the Z statistic, resulting from a meta-analysis of GWAS results under imperfect CGRs, can be found in Appendix A.1. These expressions allow for differences in sample size, h^2_{SNP}, and CGR across (pairs of) studies. For intuition, we here present the specific case of a meta-analysis of results from two studies with CGR $\rho_{\mathbf{G}}$, with equal SNP-based heritability h^2_{SNP}, and equal sample sizes (i.e., N in Study 1 and N in Study 2). Under this scenario, we find that under high polygenicity, the Z statistic of an associated SNP k is normally distributed with mean zero and the following variance:

$$\text{Var}(Z_k) = \mathbb{E}\left[Z_k^2\right] \approx 1 + \frac{h^2_{\text{SNP}}}{M} N\left(1 + \rho_{\mathbf{G}}\right). \tag{2.1}$$

We incorporate cross-study genetic heterogeneity by assuming that the data-generating process follows a random-effects model, where

cross-study correlations in SNP effects shape the inferred CGRs. When one has random effects, under the null hypothesis a SNP effect follows a degenerate distribution with all probability mass at zero, whereas under the alternative hypothesis a SNP effect follows a distribution with mean zero and a finite non-zero variance. Bearing in mind that we can write a meta-analysis Z statistic as a weighted average of true effects across studies and noise terms, the null hypothesis leads to a Z statistic with a mean equal to zero and a variance equal to one, whereas the alternative hypothesis does not lead to a non-zero mean in the Z statistic, but rather to excess variation (i.e., a variance larger than one).

The larger the variance in the Z statistic, the higher the probability of rejecting the null. The ratio of h^2_{SNP} and M can be regarded as the theoretical R^2 of each associated SNP with respect to the phenotype. Equation 2.1 reveals that (i) when sample size increases, power increases, (ii) when h^2_{SNP} increases, the R^2 per associated SNP increases and therefore power increases, (iii) when the number of associated SNPs increases, the R^2 per associated SNP decreases and therefore power decreases, (iv) when the CGR is zero the power of the meta-analysis is identical to the power obtained in each of the two studies when analyzed separately, yielding no strict advantage to meta-analyzing, and (v) when the CGR is positive one, the additional variance in the Z statistic – compared to the variance under the null – is twice the additional variance one would have when analyzing the studies separately, yielding a strong advantage to meta-analyzing.

Notably, our expression for $\mathbb{E}\left[Z_k^2\right]$ bears a great resemblance to expressions for the expected value of the squared Z statistic when accounting for LD, population stratification, and polygenicity (Yang et al., 2011b, 2014, Bulik-Sullivan et al., 2015b). Consider the scenario where the CGR between two samples of equal size is positive one. Based of Equation 2.1, we then have that $\mathbb{E}\left[Z_k^2\right] \approx 1 + \frac{h^2_{\text{SNP}}}{M} N_T$ for a trait-affecting haplotype block, where $N_T = 2N$ denotes the total sample size. This expression is equivalent to the expected squared Z statistic from the linear regression analysis for a trait-affecting variant reported in Section 4.2 of the Supplementary Note to Yang et al. (2014) as well as the first equation in Bulik-Sullivan et al. (2015b) when assuming that confounding biases and LD are absent.

In order to compute statistical power in a multi-study setting, we

first use the generic expression for the variance of the GWAS Z statistic derived in Appendix A.1 to characterize the distribution of the Z statistic under the alternative hypothesis. Given a genome-wide significance threshold (denoted by α; usually $\alpha = 5 \cdot 10^{-8}$), we use the normal cumulative distribution function under the alternative hypothesis to quantify the probability of attaining genome-wide significance for an associated SNP. This probability we refer to as the 'power per associated SNP' (denoted here by β). Given that we use SNPs tagging independent haplotype blocks, we can calculate the probability of rejecting the null for at least one SNP and the expected number of hits, true positives, false positives, false negatives, and positive negatives, as functions of α, β, the number of truly associated SNPs (denoted by M), and the number of non-associated SNPs (denoted by $S - M$). Letting '#' denote the number of elements in a set, we have that

$$\begin{aligned}
\mathbb{P}[\#\text{ true positives} \geq 1] &= 1 - \left(1-\beta\right)^{M}, \\
\mathbb{P}[\#\text{ hits} \geq 1] &= 1 - \left[\left(1-\beta\right)^{M}(1-\alpha)^{S-M}\right], \\
\mathbb{E}[\#\text{ hits}] &= \beta M + \alpha(S-M), \\
\mathbb{E}[\#\text{ true positives}] &= \beta M, \\
\mathbb{E}[\#\text{ false positives}] &= \alpha(S-M), \\
\mathbb{E}[\#\text{ false negatives}] &= \left(1-\beta\right)M, \text{ and} \\
\mathbb{E}[\#\text{ true negatives}] &= (1-\alpha)(S-M).
\end{aligned}$$

2.2.3. *Predictive accuracy of a polygenic score based on estimates from a GWAS meta-analysis*

In Appendix A.2 we derive a generic expression for the theoretical R^2 of a PGS in a hold-out sample, with SNP weights based on a meta-analysis of GWAS results under imperfect CGRs. We consider a PGS that includes all the SNPs that tag independent haplotype blocks (i.e., there is no SNP selection).

For intuition, we here present an approximation for prediction in a hold-out sample, with SNP weights based on a GWAS in a single discovery study with sample size N, where both studies have SNP heritability h^2_{SNP}, and with CGR $\rho_{\mathbf{G}}$, between the studies. Under high polygenicity, the R^2 of the PGS in the hold-out sample is then given by

the following expression:

$$R^2 \approx h^2_{\text{SNP}} \rho^2_{\mathbf{G}} \frac{h^2_{\text{SNP}}}{\frac{S}{N} + h^2_{\text{SNP}}}. \tag{2.2}$$

In case the CGR is one, and we consider the R^2 between the PGS and the genetic value (i.e., the genetic component of the phenotype) instead of the phenotype itself, the first two terms in Equation 2.2 disappear, yielding an expression equivalent to the first equation in Daetwyler et al. (2008). Assuming a CGR of one and that all SNPs are associated, Equation 2.2 is equivalent to the expression in Dudbridge (2013) for the R^2 between the PGS and the phenotype in the hold-out sample.

From Equation 2.2, we deduce that (i) as the effective number of SNPs S increases, the R^2 of the PGS deteriorates (since every SNP-effect estimate contains noise, owing to imperfect inferences in finite samples), (ii) given the effective number of SNPs, under a polygenic architecture, the precise fraction of effective SNPs that is associated does not affect the R^2, (iii) R^2 is quadratically proportional to $\rho_{\mathbf{G}}$, implying a strong sensitivity to CGR, and (iv) as the sample size of the discovery study grows, the upper limit of the R^2 is given by $h^2_{\text{SNP}} \rho^2_{\mathbf{G}}$, implying that the full SNP heritability in the hold-out sample cannot be entirely captured as long as CGR is imperfect.

2.2.4. *Online power and R^2 calculator*

An online version of the MetaGAP calculator can be found at www.devlaming.eu. This calculator computes the theoretical power per trait-affecting haplotype block, the power to detect at least one of these blocks, and the expected number of (a) independent hits, (b) true positives, (c) false positives, (d) false negatives, and (e) true negatives, for a meta-analysis of GWAS results from C studies. In addition, it provides the expected R^2 of a PGS for a hold-out sample, including all GWAS SNPs, with SNP weights based on the meta-analysis of the GWAS results from C studies. Calculations are based on the generic expressions for GWAS power derived in Appendix A.1 and PGS R^2 derived in Appendix A.2.

The calculator assumes a quantitative trait. Users need to specify either the average sample size per study or the sample size of each study separately. In addition, users need to specify either the average

within-study SNP heritability or the SNP heritability per study. The SNP heritability in the hold-out sample also needs to be provided. Users are required to enter the effective number of causal SNPs and the effective number of SNPs in total. The calculator assumes a fixed CGR between all pairs of studies included in the meta-analysis and a fixed CGR between the hold-out sample and each study in the meta-analysis. Hence, one needs to specify two CGR values: one for the CGR within the set of meta-analysis studies and one to specify the genetic overlap between the hold-out sample and the meta-analysis studies.

Finally, a more general version of the MetaGAP calculator is provided in the form of `MATLAB` code (www.mathworks.com), also available at www.devlaming.eu. This code can be used in case one desires to specify a more versatile genetic-correlation matrix, where the CGR can differ between all pairs of studies. Therefore, this implementation requires the user to specify a full (C+1)-by-(C+1) correlation matrix. Calculations in this code are also fully in line with the generic expressions in Appendices A.1 and A.2.

2.2.5. *Assessing validity of theoretical power and R^2*

We simulate data for a wide range of genetic architectures in order to assess the validity of our theoretical framework. As we show in Appendix A.4, the theoretical expressions we derive for power and R^2 are accurate, even for data generating processes substantially different from the process we assume in our derivations. Our strongest assumptions are that all truly associated SNPs have equal R^2 with respect to the phenotype, regardless of allele frequency, and that genome-wide CGRs are shaped solely by the cross-study correlations in the effects of causal SNPs. When we simulate data where the former assumption fails and where – in addition – allele frequencies are non-uniformly distributed and different across studies, the root-mean-square prediction error of statistical power lies below 3% and that of PGS R^2 below 2%. Moreover, when we simulate data where the CGR is shaped by both non-overlapping causal loci across studies and the correlation of the effects of the overlapping loci, the RMSE is less than 2% for both statistical power and PGS R^2.

2.2.6. *Estimating SNP heritability and CGR*

Using 1000 Genomes-imputed (1kG) data from the RS, STR, and HRS, we estimate SNP-based heritability and CGR, respectively, by means of univariate and bivariate GREML (Yang et al., 2011a, Lee et al., 2012) as implemented in `GCTA` (Yang et al., 2011a). In our analyses we consider the subset of HapMap3 SNPs available in the 1kG data. In Appendix A.5 we report details on the genotype and phenotype data, as well as our quality control (QC) procedure. After QC we have a dataset, consisting of $\approx$ 1 million SNPs and $\approx$ 20k individuals, from which we infer h^2_{SNP} and CGR. In Appendix A.6 we provide details on the specifications of the models used for GREML estimation.

2.3. RESULTS

2.3.1. *Determinants of GWAS power and PGS R^2*

Using the MetaGAP calculator, we assessed the theoretical power of a meta-analysis of GWAS results from genetically heterogeneous studies and the theoretical R^2 of the resulting PGS in a hold-out sample, for various numbers of studies and sample sizes, and different values of CGR and h^2_{SNP}.

SAMPLE SIZE AND CGR

Figure 2.1 shows contour plots for the power per truly associated SNP and R^2, for a setting with 50 studies, for a trait with $h^2_{\text{SNP}} = 50\%$, for various combinations of total sample size and CGR. Increasing total sample size enhances both power and R^2. When the CGR is perfect, power and R^2 (relative to SNP heritability) have a near-identical response to sample size. This similarity in response gets distorted when the CGR decreases. For instance, in the scenario of 100k SNPs of which a subset of 1k SNPs is causal with $h^2_{\text{SNP}} = 50\%$, in a sample of 50 studies with a total sample size of 10 million individuals, a CGR of one yields 94% power per causal SNP and an R^2 of 49%, which is 98% of the SNP heritability, whereas for a CGR of 0.2 the power is still 87% per SNP, while the R^2 of the PGS is 8.5%, which is only 17% of h^2_{SNP}. Thus, R^2 is far more sensitive to an imperfect CGR than the meta-analytic power is. This finding is also supported by the approximations of power

in Equation 2.1 and of PGS R^2 in Equation 2.2; these expressions show that, for two discovery studies, the CGR has a linear effect on the variance of the meta-analysis Z statistic, whereas for one discovery and one hold-out sample, the PGS R^2 is quadratically proportional to the CGR.

SNP HERITABILITY AND CGR

Figure 2.2 shows contour plots for the power per truly associated SNP and R^2 for a setting with 50 studies, with a total sample of 250k individuals, for 1k causal SNPs and 100k SNPs in total, for various combinations of h^2_{SNP} and CGR. The figure shows a symmetric response of both power and R^2 to CGR and h^2_{SNP}. For instance, when $h^2_{\text{SNP}} = 25\%$ and CGR = 0.5 across all studies, the power is expected to be around 34% and the R^2 3.0%. When these numbers are interchanged (i.e., $h^2_{\text{SNP}} = 50\%$ and CGR = 0.25), similarly, the power is expected to be 35% and the R^2 2.9%. Hence, in terms of both R^2 and power, a low heritability can be compensated by a high CGR (e.g., by means of homogeneous measures across studies) and a low CGR can be compensated by high heritability. When either CGR or heritability is equal to zero, both power and R^2 are decimated in the multi-study setting. However, when both are moderately low but still substantially greater than zero, neither power nor R^2 are completely diminished.

NUMBER OF STUDIES AND CGR

Figure 2.3 shows contour plots for the power per truly associated SNP and R^2 for a trait with $h^2_{\text{SNP}} = 50\%$, 1k causal SNPs, 100k SNPs in total, and a fixed total sample size of 250k individuals. In this figure, various combinations of the CGR and the number of studies are considered. Logically, when there is just one study for discovery, CGR does not affect power. However, even for two studies, the effect of CGR on power is quite pronounced. For instance, when CGR is a half, the power per causal SNP is 63% for one study, 58% for two studies, 51% for ten studies, and 50% for 100 studies. Thus, when the number of studies is low, increasing the number of studies makes the effect of CGR on power more pronounced rapidly. When the number of studies is large, further increases in the number of studies hardly make the effect of CGR on power more pronounced.

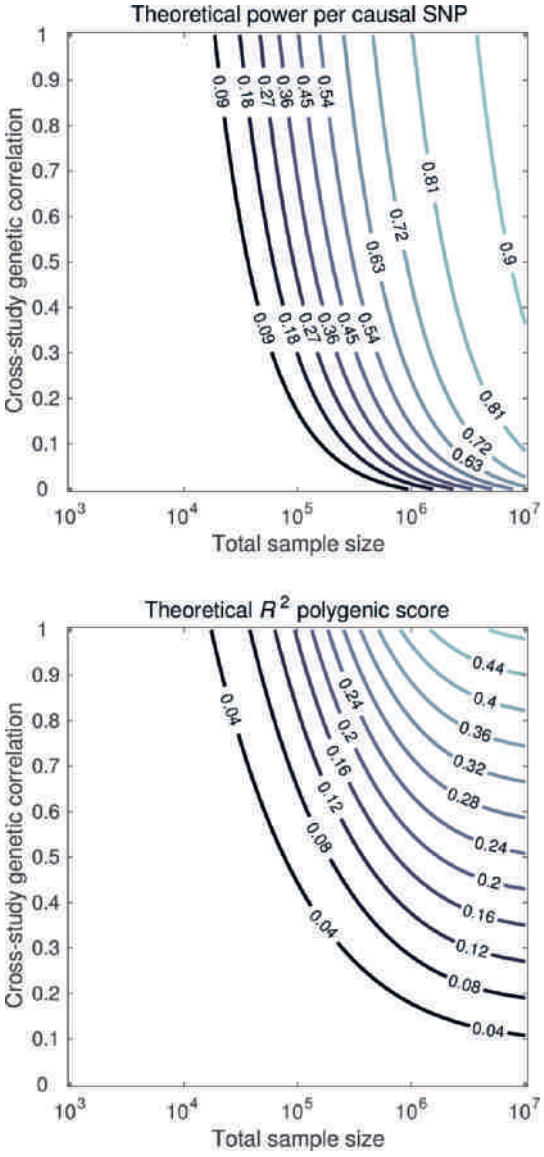


Figure 2.1: *Contour plots of the theoretical statistical power per causal SNP (upper panel) and out-of-sample polygenic-score R^2 (lower panel), resulting from a meta-analysis of GWAS results, for various combinations of total sample size (x-axis) and cross-study genetic correlation (y-axis). Calculations assume a meta-analysis of GWAS results from 50 studies of equal sample size, yielding a given total sample size (x-axis), for 100k independent SNPs, and a SNP heritability of 50% arising from a subset of 1k causal SNPs.*

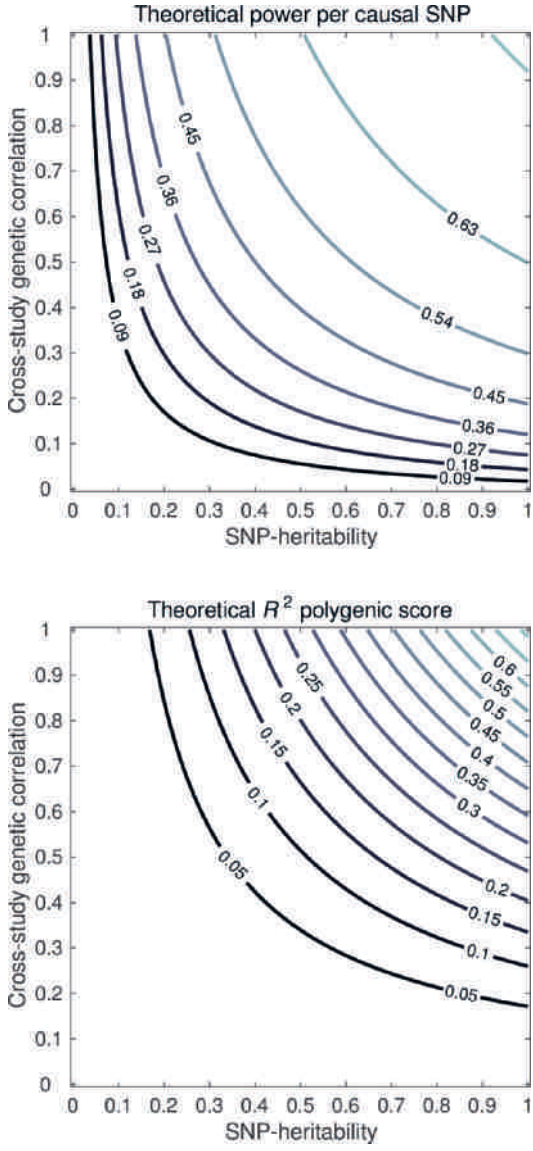


Figure 2.2: *Contour plots of the theoretical statistical power per causal SNP (uper panel) and out-of-sample polygenic-score R^2 (lower panel), resulting from a meta-analysis of GWAS results, for various combinations of SNP heritability (x-axis) and cross-study genetic correlation (y-axis). Calculations assume a meta-analysis of GWAS results from 50 studies, with a sample size of 5k individuals per study, for 100k independent SNPs, and a given SNP heritability (x-axis) arising from a subset of 1k causal SNPs.*

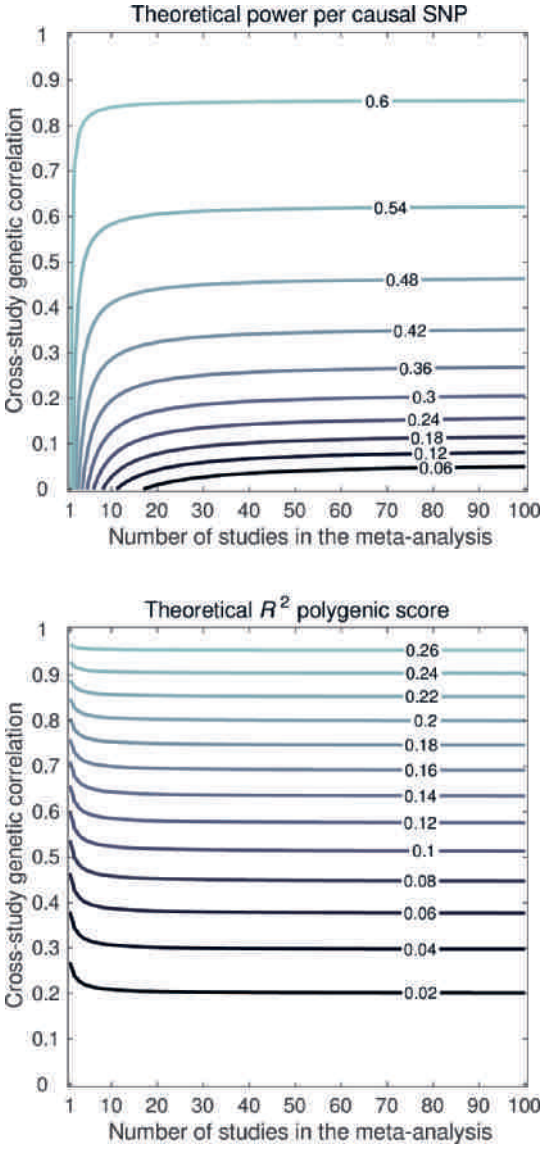


Figure 2.3: *Contour plots of the theoretical statistical power per causal SNP (upper panel) and out-of-sample polygenic-score R^2 (lower panel), resulting from a meta-analysis of GWAS results, for various combinations of the number of studies in the meta-analysis (x-axis) and cross-study genetic correlation (y-axis). Calculations assume a meta-analysis of GWAS results from a given number of studies (x-axis) of equal sample size, yielding a total sample size of 250k individuals, for 100k independent SNPs, and a SNP heritability of 50% arising from a subset of 1k causal SNPs.*

For a given number of studies, we observed that the effect CGR has on R^2 is stronger than the effect it has on power. This observation is in line with the approximated theoretical R^2 in Equation 2.2, indicating that R^2 is quadratically proportional to CGR. However, an interesting observation is that this quadratic relation lessens as the number of studies grows large, despite the total sample size being fixed. For instance, at a CGR of a half, the R^2 in the hold-out sample is expected to be 6.9% when there is only one discovery study. However, the expected R^2 is 8.1% for two discovery studies, 9.3% for ten discovery studies, and 9.6% for 100 discovery studies. A likely reason for this pattern is that, in case of one discovery study, the PGS is influenced relatively strongly by the study-specific component of the genetic effects. This idiosyncrasy is not of relevance for the hold-out sample. As the number of studies increases – even though each study brings its own idiosyncratic contribution – each study consistently conveys information about the part of the genetic architecture which is common across the studies. Since the idiosyncratic contributions from the studies are independent, they tend to average each other out, whereas the common underlying architecture gets more pronounced as the number of studies in the discovery increases, even if the total sample size is fixed.

SNP HERITABILITY IN THE HOLD-OUT SAMPLE

Figure 2.4 shows a contour plot for the PGS R^2 based on a meta-analysis of 50 studies with a total sample size of 250k individuals, with 1k causal SNPs and 100k SNPs in total, and a CGR of 0.8 between both the discovery studies and the hold-out sample. In the plot, various combinations of h^2_{SNP} in the discovery samples and h^2_{SNP} in the hold-out sample are considered. The response of PGS R^2 to heritability in the discovery sample and the hold-out sample is quite symmetric, in the sense that a low h^2_{SNP} in the discovery samples and a high h^2_{SNP} in the hold-out sample yield a similar R^2 as a high h^2_{SNP} in the discovery sample and a low h^2_{SNP} in the hold-out sample. However, R^2 is slightly more sensitive to h^2_{SNP} in the hold-out sample than in the discovery samples. For instance, when SNP heritability in the discovery samples is 50% and 25% in the hold-out sample, the expected R^2 is 10%, whereas in case the SNP heritability is 25% in the discovery samples and 50% in the hold-out sample, the expected R^2 is 13%.

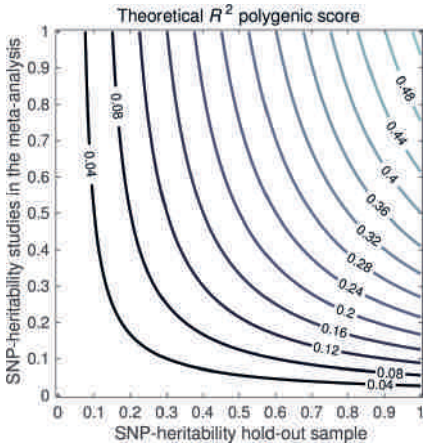


Figure 2.4: *Contour plot of the theoretical out-of-sample polygenic-score R^2, resulting from a meta-analysis of GWAS results, for various combinations of SNP heritability in the studies included in the meta-analysis (y-axis) and SNP heritability in the hold-out sample (x-axis). Calculations assume a meta-analysis of GWAS results from 50 studies, with a sample size of 5k individuals per study and a cross-study genetic correlation of 0.8, for 100k independent SNPs, where a subset of 1k SNPs is causal.*

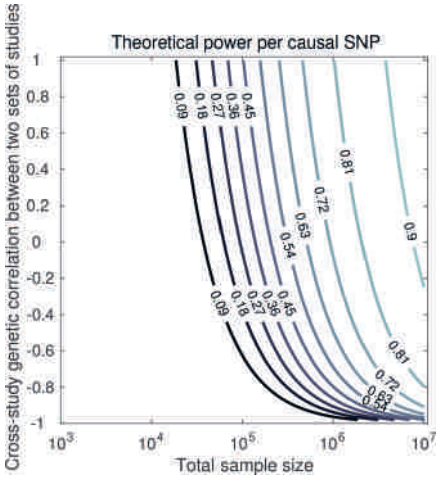


Figure 2.5: *Contour plot of the theoretical statistical power per causal SNP, resulting from a meta-analysis of GWAS results from two sets of studies, for various combinations of total sample size (x-axis) and cross-study genetic correlation between the sets of studies (y-axis). Calculations assume a meta-analysis of GWAS results from two sets of 50 equally sized studies, a within-set cross-study genetic correlation equal to one, for 100k independent SNPs, and a trait with a SNP heritability of 50% arising from a subset of 1k causal SNPs.*

CGR BETWEEN SETS OF STUDIES

Figure 2.5 shows a contour plot for the power per truly associated SNP in a setting where there are two sets consisting of 50 studies each. Within each set, the CGR is equal to one, whereas between sets the CGR is imperfect. Consider, for example, a scenario where one wants to meta-analyze GWAS results for height from a combination of two sets of studies; one set of studies consisting primarily of individuals of European ancestry and one set of studies with mostly people of Asian ancestry in it. Now, one would expect CGRs close to one between studies consisting primarily of individuals of European ancestry and the same for the CGRs between studies consisting primarily of people of Asian ancestry. However, the CGRs between those two sets of studies may be less than one.

As is shown in Appendix A.1, in case the CGR between the two sets of studies, $\mathscr{C}_1$ and $\mathscr{C}_2$, is zero, meta-analyzing the two sets jointly yields power $\beta_{\mathscr{C}_1 \cup \mathscr{C}_2} \leq \max\{\beta_{\mathscr{C}_1}, \beta_{\mathscr{C}_2}\}$ and $\beta_{\mathscr{C}_1 \cup \mathscr{C}_2} \geq \min\{\beta_{\mathscr{C}_1}, \beta_{\mathscr{C}_2}\}$, where $\beta_{\mathscr{A}}$ denotes the power in set of studies $\mathscr{A}$. In particular, when $\beta_{\mathscr{C}_1} = \beta_{\mathscr{C}_2}$ we have under a CGR of zero between the sets, that $\beta_{\mathscr{C}_1 \cup \mathscr{C}_2} = \beta_{\mathscr{C}_1} = \beta_{\mathscr{C}_2}$. Since in Figure 2.5 we considered two equally well-powered sets, the power of a meta-analysis using both sets, under zero CGR between sets, is identical to the power obtained when meta-analyzing, for instance, only the first set. However, as CGR between sets increases, so does power. For instance, when a total sample size of 250k individuals is spread across 2 clusters, each cluster consisting of 50 studies (i.e., sample size of 125k individuals per cluster and 2,500 individuals per study), under $h^2_{\text{SNP}} = 50\%$ due to 1k causal SNPs, a CGR of one within each cluster, and CGR of zero between clusters, the power is expected to be 49%, which is identical to the power of a meta-analysis of either the first or the second cluster. However, if the CGR between clusters is 0.5 instead of zero, the power goes up to 58%. In terms of the expected number of hits, this cross-ancestry meta-analysis yields an expected 82 additional hits, compared to a meta-analysis considering only one ancestry.

Alternatively, one could carry out a meta-analysis in each set of studies and pool the hits across these sets. However, this would imply more independent tests being carried out, and, hence, the need for a more stringent genome-wide significance threshold, in order to keep the false-positive rate fixed. Therefore, this route may yield less statistical

power than a meta-analysis of merely one of the two sets or a joint analysis of both. Ideally, in the scenario where between-population heterogeneity is likely, one should apply a meta-analysis method that accounts for the heterogeneity (e.g., Lebrec et al. 2010, Han and Eskin 2011, Bhattacharjee et al. 2012, Wen and Stephens 2014, Shi and Lee 2016). By applying such a method, one can consider all GWAS results from different ancestry groups in one analysis.

2.3.2. *Empirical results for SNP heritability and CGR*

In Table 2.1 we report univariate GREML estimates of SNP heritability and bivariate GREML estimates of genetic correlation for traits that attained a pooled sample size of at least 18k individuals, which gave us at least 50% power to detect a genetic correlation near one for a trait that has a SNP heritability of 10% or more (Visscher et al., 2014). The smallest total sample size is N_T = 19,184 for self-rated health. Details per phenotype (i.e., sample size, univariate estimates of SNP heritability, and bivariate estimates of genetic correlation, stratified across studies and sexes, as well as cross-study and cross-sex averages) are provided in Appendix A.7.

The univariate estimates of SNP heritability based on the pooled data assume perfect CGRs. Therefore, such estimates of SNP heritability are downwards biased when based on data from multiple studies with imperfect CGRs. To circumvent this bias, we estimated SNP heritability in each study separately, and focused on the sample-size-weighted cross-study average estimate of SNP heritability.

For both height and BMI, we observed genetic correlations close to one across pairs of studies and between females and males. For years of schooling (*EduYears*) we found a CGR around 0.8 when averaged across pairs of studies. Similarly, the genetic correlation for *EduYears* in females and males lies around 0.8. The CGR of self-rated health is substantially below one across the pairs of studies, whilst the genetic correlation between females and males seems to lie around one. The reason for this difference in the genetic correlation of self-rated health between pairs of studies and between females and males may be due to the difference in the questionnaire across studies, discussed in Appendix A.5. The questionnaire differences can yield a low CGR, while not precluding the remaining genetic overlap for this measure across the three studies, to be highly similar for females and males.

Table 2.1: *GREML estimates of SNP heritability and genetic correlation across studies and sexes.*

Phenotype	***N***	**Estimates SNP heritability**[1]			**Estimates genetic correlation**[1,2]			
		pooled[3]	**study**[4]	**sexes**[5]	**RS–STR**	**RS–HRS**	**STR–HRS**	**Females–Males**
Height	20,458	43.3% (1.8%) ***	44.9%	44.0%	0.976 (0.102) ***	0.954 (0.095) ***	0.967 (0.106) ***	0.981 (0.067) ***
BMI	20,449	20.9% (1.7%) ***	21.9%	22.8%	1.000 (0.269) ***	0.914 (0.172) ***	0.847 (0.246) ***	0.794 (0.122) *** †
EduYears	20,619	16.4% (1.7%) ***	18.2%	18.4%	0.690 (0.233) ***	0.659 (0.224) *** †	1.000 (0.263) ***	0.832 (0.162) ***
CurrCigt	20,686	18.2% (4.0%) ***	19.1%	24.2%	1.000 (0.643) ***	0.611 (0.448) *	1.000 (0.607) ***	0.543 (0.257) *** †
CurrDrinkFreq	20,072	7.0% (2.6%) ***	10.3%	8.3%	1.000 (0.666) ***	0.298 (0.670)	-0.056 (0.647)	1.000 (2.068) *
Self-rated health	19,184	10.3% (1.8%) ***	15.7%	9.5%	0.626 (0.439) **	0.363 (0.223) ** ††	0.447 (0.278) **	1.000 (0.349) ***

[1] Standard errors between parentheses.
[2] Significance of deviations from one only tested for genetic correlations.
[3] Univariate estimates from pooled data.
[4] Sample-size weighted averages of univariate estimates across studies.
[5] Sample-size weighted averages of univariate estimates across sexes.

* > 0 at 10% sign. † < 1 at 10% sign.
** > 0 at 5% sign. †† < 1 at 5% sign.
*** > 0 at 1% sign. ††† < 1 at 1% sign.

For *CurrCigt* and *CurrDrinkFreq*, the estimates of CGR and of genetic correlation between females and males are non-informative. For these two traits the standard errors of the genetic correlations estimates are large, mostly greater than 0.5. In addition, for *CurrDrinkFreq* there is strong volatility in the CGR estimate across pairs of studies.

2.3.3. *Attenuation in power and R^2 due to imperfect CGR*

Considering only the traits for which we obtained accurate estimates of CGR and SNP heritability (i.e., with low standard errors), we used the MetaGAP calculator to predict the number of hits in a set of discovery samples and the PGS R^2 in a hold-out sample, in prominent GWAS efforts for these traits. Details and notes on the results from existing studies, used as input for the MetaGAP calculations, can be found in Appendix A.8. Importantly, as reported in Table A.7, for the traits under consideration here, large-scale GWAS results to date have been obtained using fixed-effects meta-analyses.

Since we only had accurate estimates for height, BMI, *EduYears*, and self-rated health, we focused on these four phenotypes. For these traits, we computed sample-size-weighted average CGR estimates across the pairs of studies. Table 2.2 shows the number of hits and PGS R^2 reported in the most comprehensive GWAS efforts to date for the traits of interest, together with predictions from the MetaGAP calculator. We tried several values for the number of independent haplotype blocks (i.e., 100k, 150k, 200k, 250k) and for the number of trait-associated blocks (i.e., 10k, 15k, 20k, 25k). Overall, 250k blocks of which 20k trait-affecting yielded theoretical predictions in best agreement with the empirical observations; we acknowledge the potential for some overfitting (i.e., two free parameters set on the basis of 17 data points; 10 data points for the reported number of hits and 7 for PGS R^2).

For height – the trait with the lowest standard error in the estimates of h^2_{SNP} and CGR – the predictions of the number of hits and PGS R^2 for the two largest GWAS efforts are much in line with theoretical predictions. For the smaller GWAS of 13,665 individuals (Weedon et al., 2008), our estimates seem slightly conservative; 0 hits expected versus the 7 reported. However, in our framework, we assumed that each causal SNP has the same R^2. Provided there are some differences in R^2 between causal SNPs, the first SNPs that are likely to reach

Table 2.2: *Predicted and observed number of genome-wide-significant hits and polygenic-score R^2, for large-scale GWAS meta-analysis efforts to date for height, BMI, EduYears, and self-rated health, assuming 250k effective SNPs (i.e., independent haplotype blocks) of which 20k trait-affecting, using averaged GREML estimates from Table 2.1 for setting SNP heritability and CGR. Notes on the sources for the large-scale GWAS efforts are listed in Appendix Table A.8.*

Phenotype	**Main studies**			**Architecture**		**Number of hits**				**PGS R^2 using all SNPs**			
	Study	***N***	***C*****	**h^2_{SNP}**	**CGR**	**Study**	**Theory|CGR <1**	**Theory|CGR =1**	**Attenuation***	**Study**	**Theory|CGR <1**	**Theory|CGR =1**	**Attenuation***
Height	Wood et al. (2014)	253,288	79	44.9%	0.965	697	647.26	700.24	8%	13.5%	13.2%	14.0%	6%
	Lango Allen et al. (2010)	183,727	61	44.9%	0.965	180	292.03	320.77	9%	10.0%	10.5%	11.1%	6%
	Weedon et al. (2008)	13,665	5	44.9%	0.965	7	0.00	0.00	*n.a.*	2.9%	1.0%	1.1%	7%
BMI	Locke et al. (2015)	339,224	125	21.9%	0.917	97	188.52	241.07	22%	6.5%	4.3%	5.0%	14%
	Speliotes et al. (2010)	123,865	46	21.9%	0.917	19	5.48	7.64	28%	2.5%	1.8%	2.1%	15%
	Willer et al. (2008)	32,387	15	21.9%	0.917	1	0.01	0.02	65%	*n.a.*	0.5%	0.6%	16%
EduYears	Okbay et al. (2016b)	405,072	65	18.2%	0.783	162	115.28	235.90	51%	*n.a.*	2.7%	4.1%	36%
	Okbay et al. (2016b)	293,723	64	18.2%	0.783	74	39.30	88.93	56%	3.9%	2.0%	3.2%	36%
	Rietveld et al. (2013b)	101,069	42	18.2%	0.783	1	0.63	1.64	62%	2.5%	0.8%	1.2%	38%
Self-rated health	Harris et al. (2016)	111,749	1	15.7%	0.468	13	1.35	1.35	0%	*n.a.*	0.2%	1.0%	78%

* Attenuation measures the relatively loss in expected power and R^2 due to a CGR in accordance with averaged GREML estimates from Table 2.1.
** C denotes the number of studies in the meta-analysis.

genome-wide significance in relatively small samples, are the ones with a comparatively large R^2. This view is supported by the fact that a PGS based on merely 20 SNPs already explains 2.9% of the variation in height. Hence, for relatively small samples our theoretical predictions of power and R^2 may be somewhat conservative. In addition, the 10k SNPs with the lowest meta-analysis p-values can explain about 60% of the SNP heritability (Wood et al., 2014). If the SNPs tagging the remaining 40% each have similar predictive power as the SNPs tagging the first 60%, then the number of SNPs needed to capture the full h^2_{SNP} would lie around 10k/0.6=17k, which is somewhat lower than the 20k which yields the most accurate theoretical predictions. However, as indicated before, the SNPs which appear most prominent in a GWAS are likely to be the ones with a greater than average predictive power. Therefore, the remaining 40% of h^2_{SNP} is likely to be stemming for SNPs with somewhat lower predictive power. Hence, 20k associated independent SNPs is not an unreasonable number for height.

The notion of a GWAS first picking up the SNPs with a relatively high R^2 is also supported by the predicted and observed number of hits for the reported self-rated-health GWAS (Harris et al., 2016); given a SNP heritability estimate between 10% (Harris et al., 2016) and 16% (Table 2.2), according to our theoretical predictions, a GWAS in a sample of around 110k individuals is unlikely to yield even a single genome-wide significant hit. Nevertheless, this GWAS has yielded 13 independent hits. This finding supports the idea that for various traits, some SNPs with a relatively high R^2 are present. However, there is uncertainty in the number of truly associated loci. More accurate estimates of this number may improve the accuracy of our theoretical predictions.

For BMI our predictions of PGS R^2 were quite in line with empirical results. However, for the number of hits, our predictions for the largest efforts seemed overly optimistic. We therefore suspect that the number of independent SNPs associated with BMI is higher than 20k; a higher number of associated SNPs would reduce the GWAS power, while preserving PGS R^2, yielding good agreement with empirical observation. Nevertheless, given the limited number of data points, this strategy of setting the number of causal SNPs would increase the chance of overfitting.

For *EduYears* we observed that the reported number of hits is

in between the expected number of hits when the CGR is set to the averaged GREML estimate of 0.783 and when the CGR is set to one. Given the standard errors in the CGR estimates for *EduYears*, the CGR might very well be somewhat greater than 0.783, which would yield a good fit with the reported number of hits. However, as with the number of truly associated SNPs for BMI, in light of the risk of overfitting, we can make no strong claims about a slightly higher CGR of *EduYears*.

Overall, our theoretical predictions of the number of hits and PGS R^2 are in moderate agreement with empirical observations, especially when bearing in mind that we are looking at a limited number of data points, making chance perturbations from expectation likely. In addition, regarding the number of hits, the listed studies are not identical in terms of the procedure to obtain the independent hits. Therefore, the numbers could have been slightly different, had the same pruning procedure been used across all reported studies.

Regarding attenuation, we observed a substantial spread in the predicted number of hits and PGS R^2 when assuming either a CGR equal to one, or a CGR in accordance with empirical estimates, with traits with lower CGR suffering from stronger attenuation in power and predictive accuracy. In line with theory, R^2 falls approximately quadratically with CGR. For instance, for self-rated health, the estimated CGR of about 0.5, would yield a PGS that retains approximately 0.5^2=25% of the R^2 it would have had under a CGR of one. Hence the approximated attenuation is 75%. This approximation is corroborated by the theoretical relative attenuation of 78%.

Given our CGR estimates, the theoretical relative loss in PGS R^2 is 6% for height, 14% for BMI, 36% for *EduYears*, and 78% for self-rated health, when compared to the R^2 of PGSs under perfect CGRs (Table 2.2). These losses in R^2 are unlikely to be reduced by larger sample sizes and denser genotyping.

Somewhat contrary to expectation, the number of hits seems to respond even more strongly to CGR than PGS R^2. However, since in each study under consideration the average power per associated SNP is quite small, a small decrease in power per SNP in absolute terms can constitute a substantial decrease in relative terms. For instance, when one has 2% power per truly associated SNP, an absolute decrease of 1% – leaving 1% power – constitutes a relative decrease of 50% of power per causal SNP, and thereby a 50% decrease in the expected

number of hits. This strong response shows, for example, in the case of *EduYears*, where the expected number of hits drop by about 37% when going from a CGR of one down to a CGR of 0.783.

2.4. DISCUSSION

We have shown that imperfect CGRs are likely to contribute to the gap between the phenotypic variation accounted for by all SNPs jointly and by the leading GWAS efforts to date. We arrived at this conclusion in five steps. First, we developed a Meta-GWAS Accuracy and Power (MetaGAP) calculator that accounts for the CGR. This online calculator relates the statistical power to detect associated SNPs and the R^2 of the polygenic score (PGS) in a hold-out sample to the number of studies, sample size and SNP heritability per study, and the CGR. The underlying theory shows that there is a quadratic response of the PGS R^2 to CGR. Moreover, we showed that the power per associated SNP is also affected by CGR.

Second, we used simulations to demonstrate that our theory is robust to several violations of the assumptions about the underlying data-generating process, regarding the relation between allele frequency and effect size, the distribution of allele frequencies, and the factors contributing to CGR. Further research needs to assess whether our theoretical predictions are also accurate under an even broader set of scenarios (e.g., when studying a binary trait).

Third, we used a sample of unrelated individuals from the Rotterdam Study, the Swedish Twin Registry, and the Health and Retirement Study, to estimate SNP-based heritability as well as the CGR for traits such as height and BMI. Although our CGR estimates have considerable standard errors, the estimates make it likely that for many polygenic traits the CGR is positive, albeit smaller than one.

Fourth, based on these empirical estimates of SNP heritability and CGR for height, BMI, years of education, and self-rated health, we used the MetaGAP calculator to predict the number of expected hits and the expected PGS R^2 for the most prominent studies to date for these traits. We found that our predictions are in moderate agreement with empirical observations. Our theory seems slightly conservative for smaller GWAS samples. For large-scale GWAS efforts our predictions were in line with the outcomes of these efforts. More accurate estimates

of the number of truly associated loci may further improve the accuracy of our theoretical predictions.

Fifth, we used our theoretical model to assess statistical power and predictive accuracy for these GWAS efforts, had the CGR been equal to one for the traits under consideration. Our estimates of power and predictive accuracy in this scenario indicated a strong decrease in the PGS R^2 and the expected number of hits, due to imperfect CGRs. Though these observations are in line with expectation for predictive accuracy, for statistical power the effect was larger than we anticipated. This finding can be explained, however, by the fact that though the absolute decrease in power per SNP is small, the relative decrease is large, since the statistical power per associated SNP is often low to begin with.

Overall, our study affirms that although PGS accuracy improves substantially with further increasing sample sizes, in the end PGS R^2 will continue to fall short of the full SNP-based heritability. Hence, this study contributes to the understanding of the hiding heritability reported in the GWAS literature.

Regarding the etiology of imperfect CGRs, the likely reasons are heterogeneous phenotype measures across studies, gene–environment interactions with underlying environmental factors differing across studies, and gene–gene interactions where the average effects differ across studies due to differences in allele frequencies. Our study is not able to disentangle these different causes; by estimating the CGR for different traits we merely quantify the joint effect these three candidates have on the respective traits.

However, in certain situations it may be possible to disentangle the etiology of imperfect CGRs to some extent. For instance, in case one considers a specific phenotype that is usually studied by means of a commonly available but relatively heterogeneous and/or noisy measure, while there also exists a less readily available but more accurate and homogeneous measure. If one has access to both these measures in several studies, one can compare the CGR estimates for the more accurate measure and the CGR estimates for the less accurate but more commonly available measure. Such a comparison would help to disentangle the contribution of phenotypic heterogeneity and genetic heterogeneity to the CGR of the more commonly available measure.

In considering how to properly address imperfect CGRs, it is impor-

tant to note that having a small set of large studies, rather than a large set of small studies, does not necessarily abate the problem of imperfect genetic correlations. Despite the fact that having fewer studies can help to reduce the effects of heterogeneous phenotype measures, larger studies are more likely to sample individuals from different environments. If gene–environment interactions do play a role, strong differences in environment between subsets of individuals in a study can lead to imperfect genetic correlations within that study. The attenuation in power and accuracy resulting from such within-study heterogeneity may be harder to address than cross-study heterogeneity.

Our findings stress the importance of considering the use of more sophisticated meta-analysis methods that account for cross-study heterogeneity (Lebrec et al., 2010, Han and Eskin, 2011, Bhattacharjee et al., 2012, Wen and Stephens, 2014, Shi and Lee, 2016). We believe that the online MetaGAP calculator will prove to be an important tool for assessing whether an intended fixed-effects meta-analysis of GWAS results from different studies is likely to yield meaningful outcomes.

3

GWAS on Human Reproductive Behavior

Based on parts of Barban et al. (2016)

ABSTRACT

The genetic architecture of human reproductive behavior, as measured by age at first birth and number of children ever born, has a strong relationship with fitness, human development, infertility, and risk of neuropsychiatric disorders. However, very few genetic loci have been identified and the underlying mechanisms of age at first birth and number of children ever born are poorly understood. We report the design and outcomes of the largest genome-wide association study to date of both sexes including 251,151 individuals for age at first birth and 343,072 for number of children ever born. We identify twelve independent loci – of which ten novel – that are significantly associated with age at first birth and/or number of children ever born in a SNP-based genome-wide association study. These loci harbor genes that are likely to play a role – either directly or by affecting non-local gene expression – in human reproduction and infertility, thereby, increasing our understanding of these complex traits.

3.1. INTRODUCTION

Human reproductive behavior – measured by age at first birth (AFB) and number of children ever born (NEB) – is a core topic of research across the medical, social and biological sciences (Mills and Tropf, 2016). Two central indicators are the tempo of childbearing of age at first birth (AFB) and the quantum or number of children ever born (NEB). NEB is also often referred to in biological research as life-time reproductive success (Byars et al., 2010), number of offspring (Zietsch et al., 2014), or as 'fitness' in evolutionary studies, which is the function of the number of children of a person in relation to the number of children of peers of the same birth cohort (Kirk et al., 2001, Stearns et al., 2010). Due to improvements in hygiene and the reduction in prenatal, infant and child mortality in industrialized societies, NEB has emerged as the gold standard to measure lifetime reproductive success indicating biological fitness (Stearns et al., 2010). AFB and NEB are complex phenotypes related not only to biological fecundity, but also behavior, in the sense that they are driven by the reproductive choice of individuals and their partners, and shaped by the social, cultural, economic and historical environment. Genetic factors influence the first two factors of biological fecundity and choice, with the social and historical environment filtering the types of behavior that are possible (e.g., via contraceptive legislation, social norms).

Although interrelated, AFB and NEB, but also childlessness, are distinct phenotypes. Late AFB, low NEB or remaining childless is not only due to 'involuntary' infertility or factors outside of the individual's control (e.g., inability to find a partner), but also 'voluntary' choices to remain 'childfree' (Tanturri and Mencarini, 2008). In the past four decades there has been a rapid postponement by around 4–5 years in the AFB to advanced ages in many industrialized societies (Mills et al., 2011) and a growth in childlessness, with around 20% of women born from 1965–1969 in Southern and Western European countries having no children (OECD, 2011). The biological ability to conceive a child starts to steeply decline for some women as of age 25, with almost 50% of women sterile by the age of 40 (Leridon, 2008). Birth postponement and a lower number of children has been largely attributed to social, economic and cultural environmental factors (i.e., individual and partner characteristics, socioeconomic status; Mills et al.

2011, Balbo et al. 2013). Not surprisingly, this delay has led to an unprecedented growth in infertility (i.e., involuntary childlessness), which impacts between 10–15% of couples in Western countries, with men and women affected equally (OECD, 2011). An estimated 48 million couples worldwide are infertile (Mascarenhas et al., 2012), with a large part of subfertility, particularly in men, remaining unexplained (Boivin et al., 2007). Although therapeutic options for infertility in the form of Assisted Reproductive Technology (ART) are available, they are highly ineffective at later ages and older mothers have considerably more problems during gestation and delivery, also associated with low birth weight and preterm delivery (Messerlian et al., 2013, Jolly et al., 2000, Tarín et al., 1998). Recent studies have also linked advanced maternal age to a higher risk of schizophrenia in offspring (Mehta et al., 2016).

Childless individuals (and those with a low NEB) are a heterogeneous group consisting of the involuntary childless (e.g., infertility, sterility) and voluntarily childless or 'childfree' (e.g., out of choice). Although primarily related to biological fecundity, involuntary childlessness may also be due to circumstantial socio-environmental reasons outside of the individual's control, including a lack of ability to find a stable partner (Berrington, 2004), divorce and lack of housing, employment or material resources to start a family (Mills et al., 2011). Those who are voluntarily childless are generally considered to have made an active choice or to be endowed with an underlying preference (Hakim, 2003) or personality traits that pull individuals towards or away from parenthood (Avison and Furnham, 2015). It is difficult, however, to disentangle the voluntary from the involuntary, since fertility intentions can be adjusted in relation to circumstances (Jeffries and Konnert, 2002) and these modifications are age-related (Koropeckyj-Cox and Call, 2007).

A better understanding of the genetic architecture of human reproductive behavior and its relation to the environment would enable the discovery of predictors of infertility, which would in turn greatly improve family planning but also reduce costly and invasive ART treatments. Examination of AFB and NEB may also produce a better understanding of the biology of human reproduction, which in turn may give insight into fundamental biological mechanisms and could have ramifications for the study of many health outcomes, especially the etiology of diseases related to the reproductive tract. Furthermore, it is

important to understand whether and which proportion of these traits are driven by genetic, behavioral and environmental factors. Relatively little is known about the relationship between indicators of women's reproductive lifespan (menarche, menopause) and reproductive success. A smaller and recent study has produced some evidence of the link between age at first sexual intercourse (AFS) with AFB and NEB, with a focus on puberty and development (Day et al., 2016).

By systematically investigating the relationship with genetic variants for a multitude of phenotypes related to human reproduction we can establish to what extent diseases related to the reproductive tract play a role in human reproduction and *vice versa*, and begin to chart the complex biological and related mechanisms that drive human reproduction. It is therefore crucial to examine not only genetic determinants of more biologically proximate phenotypes (e.g. age at menarche, endometriosis, and polycystic ovary syndrome) but also human reproductive behavior and success. AFB and NEB represent more accurate and concrete measures of observed reproductive success in comparison with proxies which capture the reproductive life span (e.g., age at menarche, menopause) or infertility measures (e.g., endometriosis, polycystic ovary syndrome).

3.2. STUDY

To our knowledge, the current study is the largest meta-GWAS effort on human reproductive behavior. We launched this study in early 2012. As mentioned previously, a recently published smaller and related study of cohorts also involved in our study focused on AFS, also linking it to AFB and NEB, among other traits (Day et al., 2016). Several studies have shown promising results for fertility-related outcomes related to both infertility and the reproductive life span. Previous research has uncovered a genetic component to reproduction with over 70 genome-wide association studies (GWAS) published for 32 traits and diseases associated with reproduction (Montgomery et al., 2014). This includes identification of genes such as those related to age at menarche (Sulem et al., 2009, Elks et al., 2010, Day et al., 2015a), menopause (Snieder et al., 1998, Stolk et al., 2009, 2012, Perry et al., 2013, He et al., 2009), endometriosis (Painter et al., 2011, Rahmioglu et al., 2014, Albertsen et al., 2013, Dhawan et al., 2004) and polycystic ovary syndrome (Day

et al., 2015b). This study is the first step towards understanding the pathways between genes and the complex relationship between reproduction and other phenotypes and the environment.

The current study measures human reproductive choice by the age at first birth (AFB) and the number of children ever born (NEB). AFB is the self-reported age when subjects had their first child. In most cohorts this was asked directly (e.g. "How old were you when you had your first child?"). Alternatively, it could also be calculated based on several survey questions (such as the date of birth of the subject and date of birth of the first child). Often these questions were part of a medical questionnaire about women's reproductive health. In a large number of cohorts, this means that only women were asked this question. For this reason, the sample size for AFB for women is considerably larger than for men. Note that only people who have had at least one child (parous) are eligible to be included for the analysis of the AFB phenotype.

Number of children ever born (NEB) is the self-reported number of children. This phenotype was either asked directly (e.g. "How many children do you have?" or "How many natural (biological) children have you ever had, that is, all children who were born alive?", or "How many children have you had - not counting any step, adopted, or foster children, or any who were stillborn?") or it was calculated based on several survey questions (such as pregnancy histories and outcomes, number of deliveries). In most cases it was possible to distinguish between biological (live born or stillborn) and adopted or step-children. When it was possible to distinguish between cases, we used the number of live born biological children. We included cases for NEB if they finished their reproductive career (aged at least 45 for women and 55 for men at time of study) and were thus unlikely to have future biological children.

3.3. OVERVIEW OF ANALYSES

The instructions given to cohorts who agreed to participate in our study are described in detail in the original Analysis Plan that was posted on the Open Science Framework preregistration site, uploaded December 9, 2013 at https://osf.io/53tea/. Due to the fact that we started our study in 2012, before 1000-Genomes imputation (McVean et al., 2012), our

analysis plan recommended using resulted imputed using the HapMap 2 CEU (r22.b36) reference sample (Bentley et al., 2003, McVean et al., 2007). For ease of analysis, we advised that AFB should be treated as a continuous measure. When possible, we asked analysts to use the more direct question: "How old were you when you had your first child?" Another variant of this question is: "What is the date of birth of your first child?" In the case of the latter, we advised analysts to create the AFB variable by subtracting the date of birth of the first child from the date of birth of the subject.

Analysts then normalized the raw measure of the age at first birth for sex and/or birth-cohort specific means and standard deviations. In other words, we asked them to compute a mean and standard deviation separately for men and women by birth cohort category (generally ten-year intervals), subtract the mean value for that group from the respondent's value, and then divide the result by the appropriate standard deviation. This standardized measure was used as the final AFB variable, measured in sex/cohort specific Z-scores, and is our regressand.

Analysts were asked to include birth year of the respondent (represented by birth year – 1900) linearly, squared and cubed, to control for nonlinear birth cohort effects. Combined analyses that included both men and women also needed to include interactions of birth year and its polynomials with sex. Some cohorts only used birth year and not its polynomials because of multi-collinearity issues/convergence of the genome-wide association analyses.

The principal investigator of each cohort confirmed that the results on these analyses were approved by the local Research Ethics Committee and/or the relevant Institutional Review Board. All participants fell under the written informed consent protocol of each participating study. The entire project was also approved by the local Research Ethics Committee of the principal investigator.

Cohorts with acceptable measures of AFB and/or NEB were eligible to participate. Some measured one or both of the phenotypes, and there was also variation by whether the question was asked to women and/or also men. Most participating cohorts were included in the meta-analysis (i.e., 62 cohorts are included the meta-analyses, constituting 26 files for AFB men, 50 for AFB women, 64 for AFB pooled, 47 for NEB men, 60 for NEB women, and 91 for NEB pooled). Table 3.1 provides

an overview of the studies included in the meta-analyses of GWAS results. Cohorts of unrelated individuals uploaded separate results for men and women. In addition to sex-specific association results, family-based cohorts uploaded pooled results. In addition to results from these 62 cohorts, GWAS results from the Australian Breast Cancer Family Study (ABCFS) and the Longenity study were also available. However, as discussed in Section 3.5, these results did not meet our QC requirements, and were, therefore, excluded from further analyses.

The genome-wide association study (GWAS) of human reproductive behavior is based on the summary statistics that were uploaded to a central server by cohort-level analysts. Our analysis includes the two phenotypes of age at first birth (AFB) and number of children ever born (NEB), with analysts producing association results for women, men, and combined analyses of both sexes, also including birth cohort as a covariate. The summary statistics were subsequently quality-controlled and meta-analyzed by two independent centers at the University of Oxford and Erasmus University Rotterdam.

We followed the QC protocol of the GIANT consortium's study of human height (Wood et al., 2014) and employed the software packages `QCGWAS` (Van Der Most et al., 2014) and `EasyQC` (Winkler et al., 2014), which allowed us to harmonize the files and identify possible sources of errors in association results. This procedure entailed that diagnostic graphs and statistics were generated for each set of GWAS results (i.e., for each file). In the case where apparent errors could not be amended by stringent QC, cohorts were excluded from the meta-analysis.

We first circulated three documents to interested cohorts at the end of April 2012, which included: (a) Rationale for a GWAS of Fertility Behavior, (b) GWAS Fertility Behavior Analysis Plan, and (c) Collaboration Agreement for Fertility GWAS Meta-analyses. These documents were circulated after a meeting and approval from the REPROGEN working group of the CHARGE consortium, on December 9, 2011, that we were not competing with or unduly replicating existing efforts. Preliminary results were presented at various CHARGE meetings between the years 2012 and 2015. This study was initially set up as a two-stage GWAS with a large discovery and smaller replication phase. Due to an increasing influx of new data, we opened the participation to cohorts that had genome-wide data, but also to cohorts that had Metabochip (Voight et al., 2012) data. We also included a list of 15 independent

Table 3.1: *Cohorts included in the meta-analyses of GWAS results on reproductive behavior.*

Study name	**Design**	$N \approx$
1958BC - Type 1 Diabetes Genetics Consortium	Prospective birth cohort	2,530
1958BC - Wellcome Trust Case Control Consortium	Prospective birth cohort	2,703
23andMe	Population-based	24,609
Avon Longitudinal Study of Parents and Children	Prospective pregnancy cohort	14,541
Amish Study	Community-based	1,457
Austrian Stroke Prevention Study	Population-based	2,008
Blue Mountains Eye Study	Population-based	4,828
Chicago Health and Aging Project	Population-based	624
Cilento	Population based	1,147
Lausanne Cohort	Population-based	6,188
FINRISK (Corogene)	Case-control	2,066
Croatia Korcula	Population-based	899
Croatia Split	Population-based	499
Croatia Vis	Population-based	924
deCODE genetics	Population-based	98,712
DESIR	Case-control	731
Dortmund Health Study	Population-based	1,021
Estonian Genome Center	Population-based	51,000
EPIC-Norfolk	Prospective cohort	1852
Erasmus Rucphen Family Study	Family-based	3,437
Finnish Twin Cohort	Cohort-based	688
Genetic Epidemiology Network of Arteriopathy	Family-based	1,464
Helsinki Birth Cohort Study	Birth Cohort	2,003
Health 2000	Case-control	2,123
HPFS Subcohort 1	Cohort-based	8,323
HPFS Subcohort 2	Cohort-based	8,323
Health, Aging, and Body Composition Study	Cohort-based	3,075
Health and Retirement Study	Population-based	12,507
Oxford Family Blood Pressure Study	Family-based	1,265
Genetic Park of Carlantino Project	Population-based	400
Genetic Park of Friuli Venezia Giulia Project	Population-based	1,254
Val Borbera Isolated Population Project	Population-based	1,664
KORA - F3	Population-based	1,643
KORA - S4	Population-based	1,814
RPGEH / GERA	Population-based	43,939
Lothian Birth Cohort 1936	Narrow-range birth cohort	1,005
Lothian Birth Cohort 1921	Narrow-range birth cohort	517
LifeLines Cohort Study	Population-based	6089
Minnesota Center for Twin and Family Research	Family-based	7,702
Multi-Ethnic Study of Atherosclerosis study	Population-based	1,167
Mother and Child Cohort of NIPH	Population-based	883
Osteoporotic Fractures in Men (MrOS) Sweden	Population-based	940
Netherlands Epidemiology of Obesity Study	Population-based	6,624
Netherlands Study of Depression and Anxiety	Case-control	974
Nurses' Health Study	Cohort-based	12,446
Netherlands Twin Register	Twin-family population	175,000
Ogliastra Genetic Park	Case-control	400
Ogliastra Genetic Park-Talana	Population-based	1208
Orkney Complex Disease Study	Population-based	899
Queensland Institute of Medical Research	Population-based	20,217
Rotterdam Study Baseline	Cohort-based	7,983
Rotterdam Study Extension of Baseline	Cohort-based	3,011
Rotterdam Study Young	Cohort-based	3,932
SardiNIA Study of Aging	Family-based	1829
Study of Health in Pomerania	Population-based	4,308
Sorbs cohort	Population-based	1,046
Swedish Twin Registry	Population-based	10,917
THISEAS	Case-control	1,097
St Thomas' UK Adult Twin Registry	Twin-family population	5,638
UK Biobank	Population-based	112,338
Women's Genome Health Study	Population-based	23,294
Young Finns Study	Cohort-based	2,442

SNPs with associated p-value below 10^{-6} for cohorts that did not have genome-wide data available, but could perform de-novo replication on a limited number of SNPs.

Agreements at a later stage included data from RPGEH/GERA (Kaiser Permanente Research Program on Genes, Environment, and Health; N(AFB women)=31,898, N(NEB women)=39,576), deCODE (N(AFB pooled)=60,602, N(NEB pooled)=65,228), and UK Biobank (N(AFB women)=40,082, N(NEB pooled)=88,094). Given the resulting well-powered total sample size of $N\approx$250k for AFB and $N\approx$340k for NEB, we chose to merge the discovery and replication cohorts into a single large discovery phase, as in other recent well-powered GWAS efforts (Wood et al., 2014, Locke et al., 2015, Okbay et al., 2016b). We also opted to include only cohorts with genome-wide data in the meta-analysis, leaving the remaining cohorts that performed de-novo replication for follow-up analysis.

3.4. GENOME-WIDE ASSOCIATION ANALYSES

Cohorts were asked to only include participants of European ancestry, with no missing values on all relevant covariates (sex, birth year, and cohort specific covariates), who were successfully genotyped genome-wide (e.g., genotyping rate greater than 95%), and who passed cohort-specific quality controls (e.g., no genetic outliers).

Cohorts used the fully imputed set of HapMap Phase 2 autosomal SNPs (Bentley et al., 2003, McVean et al., 2007), and estimated an additive linear model, including top principal components from the SNP data to control for population stratification and cohort specific covariates if appropriate. They were specifically instructed to control for population stratification using principal components of the genotype data, with reference to Price et al. (2006). In addition, cohorts were requested to include the birth year of the respondent (represented by birth year – 1900) linearly, squared, and cubed, to control for nonlinear age and birth-cohort effects. Analyses pooling data across sexes also needed to include interactions of birth year and its polynomials with sex. Some cohorts only used birth year and not its polynomials because of multi-collinearity issues/convergence of the GWA analysis. Omission of

these nonlinear birth year effects is unlikely to lead to biased inferences, since genotypes are not usually considered as truly associated with birth year. However, inferences might be less accurate (i.e., have larger standard errors), since omission of nonlinear birth year effects can lead to larger residual variation.

3.5. QUALITY-CONTROL PROCEDURE

In this section, we summarize the main steps and diagnostic tests of the Quality Control (QC) procedure. The QC was conducted in two independent analysis centers at the University of Oxford and at Erasmus University Rotterdam. Once data were submitted, each study was independently subjected to QC in the two centers according to standard protocols. We followed the QC protocol of the GIANT consortium's recent study of human height (Wood et al., 2014) and the SSGAC studies of educational attainment (Rietveld et al., 2013a, Okbay et al., 2016b).

Since this study began, QC procedures have become more stringent. Recently, a comprehensive set of guidelines for GWAS QC was released (Winkler et al., 2014). For the cohorts initially included in the study a first round of QC was performed using the R package `QCGWAS` (Van Der Most et al., 2014). We updated the QC protocol based on the GIANT consortium's and SSGAC's protocols. The updated QC protocol was applied to all cohorts using the R package `EasyQC` (Winkler et al., 2014). Findings of the first round of QC were used as a starting point for the updated QC.

In the QC procedure, diagnostic graphs and statistics were generated for each set of GWAS results (i.e., for each result file uploaded by the cohort analysts). Most errors (e.g., coded allele reported as other allele and *vice versa*) could be easily addressed. When apparent errors could not be amended by combining stringent QC with file-specific inspections and corrections, cohorts were excluded from the meta-analysis.

3.5.1. *Filters*

We harmonized base-pair positions of the markers across files using NCBI build 37. For each result file, a given marker was excluded in

case:

1. The combination of chromosome and base-pair position could not be uniquely linked to the HapMap Phase II CEU panel (McVean et al., 2007).

2. The marker had missing or incorrect values. Specifically,

    - the effect allele and other allele were missing,
    - the association p-value was missing or outside the unit interval,
    - the effect estimate was missing or reported to have infinite magnitude,
    - the standard error (SE) of the effect estimate was missing, negative, or infinite,
    - the allele frequency was missing or outside the unit interval,
    - the sample size was not reported, or zero or below,
    - the reported callrate (i.e., fraction of genotypes that is non-missing) was outside unit interval,
    - the reported imputation quality was negative, and
    - the reported imputed dummy was not binary.

3. The marker was not a SNP, not biallelic, non-autosomal, and/or monomorphic.

4. The sample size was below 30. This filter is to guard against spurious associations due to overfitting of the model.

5. The minor allele count was 6 or below. This filter is to guard against spurious associations with low-frequency SNPs in small samples. The risk of spurious associations has shown to be particularly high for SNPs that are extremely rare (Winkler et al., 2014).

6. Minor allele frequency (MAF) was below 1%. For all the cohorts, we dropped SNPs with a MAF below 1%. For small cohorts we applied more stringent filters based on diagnostic tests and figures.

7. The standard error (SE) of the effect estimate was greater than $100/\sqrt{N}$. Based on the approximation to the expected standard error by Winkler et al. (2014), we calculated that an SE greater than $100/\sqrt{N}$ is at least 40% greater than the expected SE of the estimated effect of a SNP with a MAF of 1% for a trait with standard deviation of 10. Since in our analyses we only consider SNPs with MAF $\geq$ 1% and traits with a standard deviation below 10, an effect estimate with an SE greater than $100/\sqrt{N}$ can be considered to be unreasonably large.

8. The R^2 of the marker with respect to the phenotype was greater than 10%. We excluded SNPs for which the estimated R^2, based on the approximation by Rietveld et al. (2013a), was greater than 10% because such an R^2 would defy all upper bounds on reasonable effect sizes of SNPs.

9. The marker was imputed while imputation quality was missing.

10. The marker was imputed while imputation quality was below 0.4. For all the cohorts, we dropped imputed SNPs with an imputation quality below 0.4. For several cohorts we apply more stringent filters based on diagnostic tests and figures.

11. The callrate of the SNP was below 95%.

12. The SNP was genotyped and not in Hardy-Weinberg Equilibrium (HWE). We excluded genotyped SNPs if they fail the HWE χ^2-test. Violations of HWE will lead to lower χ^2-test p-values as sample size increases. The p-value threshold is therefore sample-size dependent. We applied an HWE p-value threshold of 10^{-3} in case N <1k, 10^{-4} in case 1k $\leq N <$ 2k, 10^{-5} in case 2k $\leq N <$ 10k, and no filter in case $N \geq$ 10k.

3.5.2. *Diagnostic Checks*

For the SNPs remaining, after applying Filters 1–12, we generated four key diagnostic graphs:

1. Allele frequency (AF) plots—to identify errors in allele frequencies and strand orientations. The AF plot shows the expected AF (based on the HapMap II CEU2 reference panel or the 1000

Genomes Phase 1 European panel in case of 1000-Genomes imputed data) versus the reported AF.

2. Reported p-values versus p-values of the Z-scores (PZ) plots—to assess the consistency of the reported p-values with respect to those implied by the effect estimates and the corresponding standard errors.

3. Quantile-Quantile (Q-Q) plots—to check for evidence of unaccounted population stratification.

4. Reported Standard Error versus Expected Standard Error (SE) plots—to assess whether the reported standard errors behave in line with the approximation of the expected standard errors provided by Winkler et al. (2014), implemented as a QC step by Okbay et al. (2016a)

These diagnostic plots were examined by two independent analysts. If problems were detected which could not be resolved by more stringent thresholds, we applied the following *ad hoc* filters (descending order in terms of frequency used).

1. MAF filters more stringent than the generic MAF filter (e.g., 5% instead of 1%).

2. Imputation quality filters more stringent than the generic filter (e.g., 0.8 instead of 0.4).

3. Filter on the absolute difference between expected (based on the HapMap II CEU2 reference panel or the 1000 Genomes Phase 1 European panel in case of 1000-Genomes imputed data) and reported allele frequencies. This filter helps to remove clear outliers in the AF-plots (e.g., strand-ambiguous SNPs that are likely to have been reverse-coded).

4. Filter on the absolute difference between the reported log(p-value) and the log(p-value) derived from the reported Z-score. This filter helps to remove clear outliers in the PZ-plots. Such outliers can arise when software such as `SNPTEST` (Marchini and Howie, 2010) switches to another estimation method, for reasons such as poor convergence of the estimates.

Table 3.2: *Cohort-specific quality-control filters, in deviation of generic filters, applied prior to the meta-analyses of GWAS results on reproductive behavior.*

Cohort	Trait	Sex	Common filters[1]		Other filters[2]	
			MAF[3]	ImpQ[4]	DAF[5]	DLP[6]
1958BC.T1	AFB	men	0.03			
1958BC.T1	AFB	women	0.03			
1958BC.T1	NEB	men	0.03	0.6		
1958BC.T1	NEB	women	0.03			
1958BC.WT	AFB	men	0.03	0.8		
1958BC.WT	AFB	women	0.03	0.8		
1958BC.WT	NEB	men	0.03	0.8		
1958BC.WT	NEB	women	0.03	0.8		
AMISH	AFB	women	0.05			
AMISH	NEB	women	0.05			
ASPS	AFB	women	0.03			
CHAP	NEB	women	0.05			
Cilento	NEB	men	0.03			
CoLaus	AFB	men			0.3	
CoLaus	AFB	women			0.3	
CoLaus	NEB	men			0.3	
CoLaus	NEB	women			0.3	
COROGENE	AFB	women	0.03			
COROGENE	NEB	men	0.03			
COROGENE	NEB	women	0.03			
CROATIA.Korcula	NEB	men	0.03			
CROATIA.Korcula	NEB	women	0.03			
CROATIA.Split	NEB	men	0.05			
CROATIA.Split	NEB	women	0.05			
CROATIA.Vis	NEB	men	0.03			
CROATIA.Vis	NEB	pooled	0.03			
CROATIA.Vis	NEB	women	0.03			
DESIR	AFB	women	0.05	0.6		
DESIR	NEB	men	0.05	0.6		
DESIR	NEB	women	0.03	0.6		
EGCUT.cohort1	AFB	men	0.03			
EGCUT.cohort1	AFB	women	0.03			
EGCUT.cohort1	NEB	men		0.6		
EGCUT.cohort2	AFB	men		0.8		
EGCUT.cohort2	AFB	women	0.05	0.7		
EGCUT.cohort2	NEB	men	0.05	0.7		
EGCUT.cohort2	NEB	women		0.7		
EPIC cases	AFB	women		0.6		
FinnTwinCohort	AFB	men	0.03			
FinnTwinCohort	NEB	women	0.03			

Table 3.2: *(continued)*

Cohort	Trait	Sex	Common filters[1]		Other filters[2]	
			MAF[3]	ImpQ[4]	DAF[5]	DLP[6]
GENOA	NEB	women	0.03			
HBCS	NEB	women	0.03			
HEALTH2000	AFB	women	0.05			
HEALTH2000	NEB	men	0.05			
HEALTH2000	NEB	women	0.03			
INGI.CARL	AFB	women	0.05	0.8		
INGI.CARL	NEB	women	0.05	0.8		
INGI.FVG	AFB	men	0.03			
INGI.FVG	AFB	pooled	0.03			
INGI.FVG	AFB	women	0.03			
INGI.FVG	NEB	men	0.03			
INGI.FVG	NEB	pooled	0.03			
INGI.FVG	NEB	women	0.03			
MCTFR	AFB	men			0.3	
MCTFR	AFB	pooled			0.3	
MCTFR	AFB	women			0.3	
MCTFR	NEB	women			0.3	
MESA	AFB	women			0.2	0.1
MESA	NEB	women	0.03		0.2	
MOBA	AFB	women			0.3	
MrOS Sweden	NEB	men	0.03			
NESDA	AFB	men	0.03	0.8		0.1
NESDA	AFB	pooled		0.8		0.1
NESDA	AFB	women		0.8		0.1
NESDA	NEB	men		0.8		0.1
NESDA	NEB	pooled		0.8		0.1
NESDA	NEB	women	0.03	0.8		0.1
NTR	AFB	men	0.03			
NTR	NEB	men	0.03			
OGP.Talana	AFB	men	0.05			
OGP.Talana	AFB	pooled	0.05			
OGP.Talana	AFB	women	0.05			
OGP.Talana	NEB	men	0.05			
OGP.Talana	NEB	pooled	0.05			
OGP.Talana	NEB	women	0.05			
SORBS	NEB	men			0.5	
SORBS	NEB	women	0.03		0.5	
THISEAS	NEB	men	0.05		0.3	
THISEAS	NEB	women	0.05		0.3	
TwinsUK	AFB	women	0.05	0.8		
TwinsUK	NEB	women	0.05	0.8		
VB	AFB	men	0.03			
VB	AFB	pooled	0.03			

Table 3.2: *(continued)*

Cohort	Trait	Sex	Common filters[1]		Other filters[2]	
			MAF[3]	ImpQ[4]	DAF[5]	DLP[6]
VB	AFB	women	0.03			
VB	NEB	men	0.03			
VB	NEB	pooled	0.03			
VB	NEB	women	0.03			

[1] For minor allele frequency (resp. imputation quality) a standard filter of 0.01 (0.4) is applied unless otherwise stated.
[2] These filter are applied to remove outliers in AF-plots and in PZ-plots.
[3] MAF: Minor Allele Frequency (default = 0.01).
[4] ImpQ: Imputation Quality (default = 0.4).
[5] DAF: absolute Difference Allele Frequency reported in the cohort and in the reference sample (default = no filter).
[6] DLP: absolute Difference of the Logarithm of the reported p-value and and the logarithm of the p-value of the Z-score (default = no filter).

Non-standard filters (e.g., MAF filter of 5% instead of 1%), per cohort, per association file, are reported in Table 3.2.

In each results file where the AF plot revealed an anti-diagonal (i.e., where reported allele frequencies in a subset of the SNPs had a strongly negative correlation with the expected allele frequencies), we investigated whether this anti-diagonal persisted for SNPs that had been used as a basis for imputation. Such an anti-diagonal could imply that reverse-coded SNPs had been used for imputation, thereby yielding unreliable imputed SNP data. Therefore, such files were excluded from further analyses. In case the anti-diagonal was only observed for imputed SNPs, and not for the genotyped SNPs used as basis for imputation, we used the third *ad hoc* filter on allele frequency differences to excluded the aberrant SNPs.

Only the AF plots for ABCFS ($N = 410$ for both AFB and NEB), in Figure 3.1, show a strong anti-diagonal that persists when considering only genotyped markers that have been used for imputation of the ABCFS data. Consequently, we excluded the ABCFS result files from the meta-analyses.

In addition, for Longenity ($N = 285$ for AFB and $N = 352$ for NEB) many SNPs have far greater standard errors for the effect estimates than expected based on the approximation of the SE by Winkler et al. (2014). Moreover, many SNPs in Longenity have callrates substantially below 95%. When applying QC to Longenity, only several hundreds of SNPs remain. Consequently, we also exclude Longenity results from the meta-analyses of GWAS results.

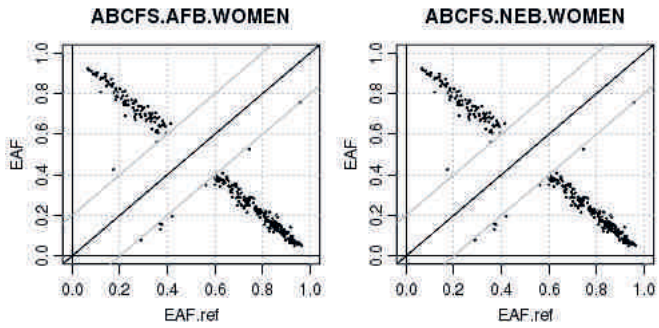


Figure 3.1: *Example of allele-frequency plots revealing many likely reverse-coded genotyped SNPs that have been used as basis for imputation. Plots of allele frequencies of genotyped SNPs used for imputation of the Australian Breast Cancer Family Study (ABCFS), with frequencies in the reference sample on the x-axis of both plots and frequencies in the female (resp. male) ABCFS sample on the y-axis of the left (right) plot. SNPs with an absolute difference in allele frequency, between ABCFS and the reference sample, below 0.2 (i.e., in between the two gray lines parallel to the main diagonal) are not displayed.*

3.6. META-ANALYSES

Cohort-specific association results (after applying the QC filters) were combined, by means of sample-size weighted meta-analyses with genomic control (GC) correction (Devlin et al., 2001) within each study (i.e., single genomic control), using `METAL` (Willer et al., 2010). Sample-size weighting is based on Z-scores and can account for different phenotypic measurements among cohorts (Evangelou and Ioannidis, 2013). Only SNPs that were observed in at least 50% of the participants across cohorts, for a given phenotype-sex combination, were passed to the meta-analysis. SNPs were considered genome-wide significant at p-values smaller than $5 \cdot 10^{-8}$ ($\alpha = 5\%$, Bonferroni-corrected for a million tests). The meta-analyses were carried out by two independent analysts. Comparisons were made to ensure concordance of the identified signals between the two analysis centers. The `PLINK` clumping function (Purcell et al., 2007, Chang et al., 2015) was used to identify the most significant SNPs in associated regions (termed “lead SNPs”).

3.6.1. *Pooled meta-analyses*

The total sample size of the meta-analysis is $N = 251,151$ for AFB pooled (i.e., pooled across sexes) and $N = 343,072$ for NEB pooled. To understand the magnitude of the estimated effects, we used an approximation method to compute unstandardized regression coefficients based on the Z-scores of `METAL` (Willer et al., 2010) output obtained from the sample-size-weighted meta-analyses, allele frequencies, and phenotypic standard deviations. Further details of the approximation procedure are available in the Supplementary Information of Rietveld et al. (2013a).

Figure 3.2 shows Manhattan plots for the AFB and NEB meta-analysis results and Figure 3.3 the corresponding quantile-quantile (Q-Q) plots. The Q-Q plots show substantial inflation due to either polygenic signal or population stratification. As we discuss in Section 3.7, the inflation – in all likelihood – stems at least partially from true polygenic signal. The Manhattan plots reveal nine genome-wide significant lead SNPs from the meta-analysis of AFB and two from the analysis of NEB, both pooled across sexes.

3.6.2. *Sex-specific meta-analyses*

In addition to the meta-analyses of pooled results, we also performed sex-specific GWAS meta-analyses for AFB and NEB, with the following sample size per analysis:

- AFB women: $N = 189,656$,
- AFB men: $N = 48,408$,
- NEB women: $N = 225,230$, and
- NEB men: $N = 103,909$.

Our results indicate six genome-wide significant (p-values $< 5 \cdot 10^{-8}$) independent SNPs for AFB women and one genome-wide significant independent SNP for NEB men. We do not find any genome-wide significant loci for AFB men and NEB women. Among the six hits for AFB women, five were also significant in the AFB pooled analysis, while one hit (rs2721195; chromosome 8, base-pair position 145,677,011) is specific for women. The single independent hit for NEB men (rs13161115;

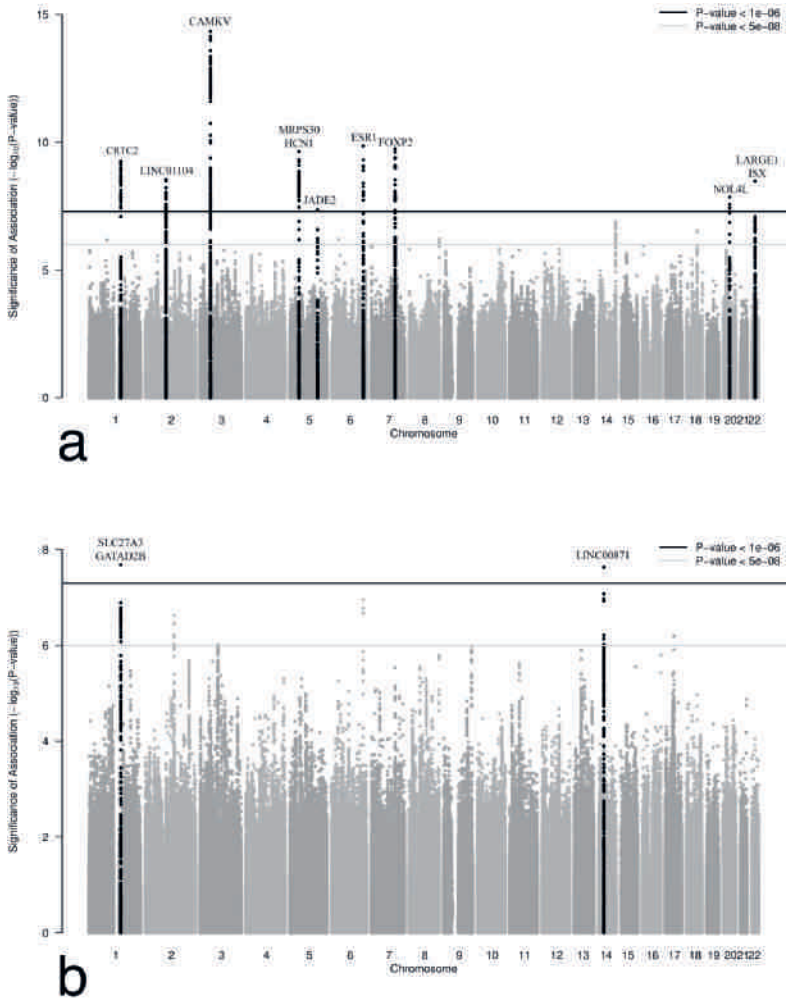


Figure 3.2: *Manhattan plots of SNPs for age at first birth (AFB; panel a) and number of children ever born (NEB; panel b), resulting from single-genomic-control meta-analyses of GWAS results. SNPs are plotted according to their position on each chromosome (x-axis) and the* $-log_{10}(p\text{-}value)$ *from the association (y-axis). The horizontal black line indicates the threshold for genome-wide significance (p-value* $< 5 \cdot 10^{-8}$*) and the horizontal gray line the threshold for suggestive hits (p-value* $< 5 \cdot 10^{-6}$*). Black points indicate SNPs in a* ± 100 *kB region around genome-wide significant hits. Gene labels are annotated as the nearby genes to the significant SNPs.*

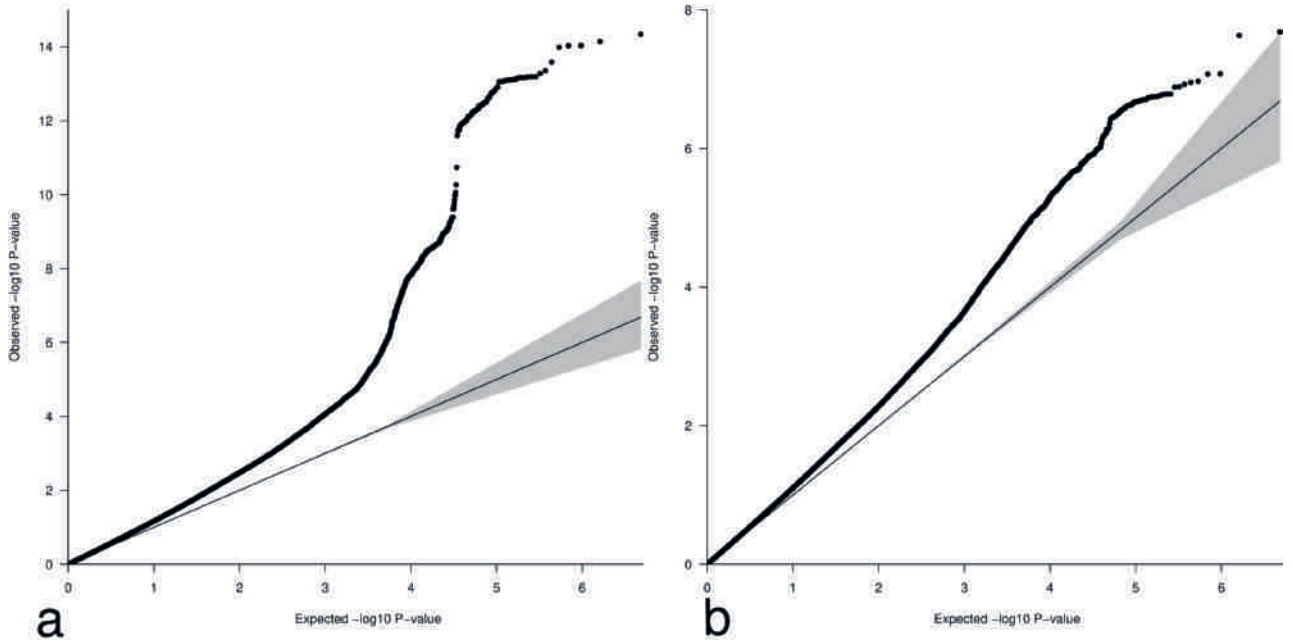


Figure 3.3: *Quantile-quantile (Q-Q) plots of SNPs for age at first birth (AFB; panel a) and number of children ever born (NEB; panel b), resulting from single-genomic-control meta-analyses of GWAS results. The gray shaded areas in the Q-Q plots represent the 95% confidence bands around the p-values under the null hypothesis.*

chromosome 5, base-pair position 107,050,002) is not significant in the NEB pooled analysis. Figure 3.4 shows the Miami plots for AFB and NEB sex-specific analyses. Figure 3.5 depicts the Q-Q plots of men and women's meta-analyses for AFB and NEB. The figure shows a noteworthy departure from the null hypothesis of no statistical association, in particular for the analysis of AFB women.

Table 3.3 shows both the sex-specific and pooled-analyses signals, respectively, for AFB and NEB. The effects of all significant hits in AFB have the same direction for both men and women. The single locus found in NEB men (rs13161115) has an opposite effect on NEB for women, although the p-value associated with its effect size in NEB for women does not reach statistical significance.

3.7. POPULATION STRATIFICATION

Population stratification can severely bias GWAS estimates for causal variants and lead to false positives. Such a bias can occur if a particular variant of a SNP is more common in a particular subpopulation and if there are mean differences in the phenotype of interest between subpopulations due to factors that do not involve that SNP. As described in Section 3.4, all cohorts in the GWAS of AFB and NEB included the top principal components from the genotypes (Price et al., 2006) in their analyses, to account for population stratification. Despite this inclusion, residual stratification could still remain and affect the results. To test the extent of this problem, we used LD-score regression (Bulik-Sullivan et al., 2015b) to estimate the 'intercept', which reflects inflation in the GWAS χ^2-test statistics due to confounding stratification.

The LD-score-intercept test uses GWAS summary statistics for all measured SNPs. LD-score regression is a method that can disentangle inflation in the χ^2-test statistics that is due to a true polygenic signal throughout the genome from inflation that is due to confounding biases such as cryptic relatedness and population stratification. The inflation due to a true polygenic signal impacts the slope of the LD-score regression, whereas inflation due to population stratification, and other confounding biases, affects the intercept of the regression.

We used the LD-score-regression software (Bulik-Sullivan et al., 2015b,a) to estimate the intercept from the summary statistics of our GWAS results for AFB and NEB, in both pooled and sex-stratified

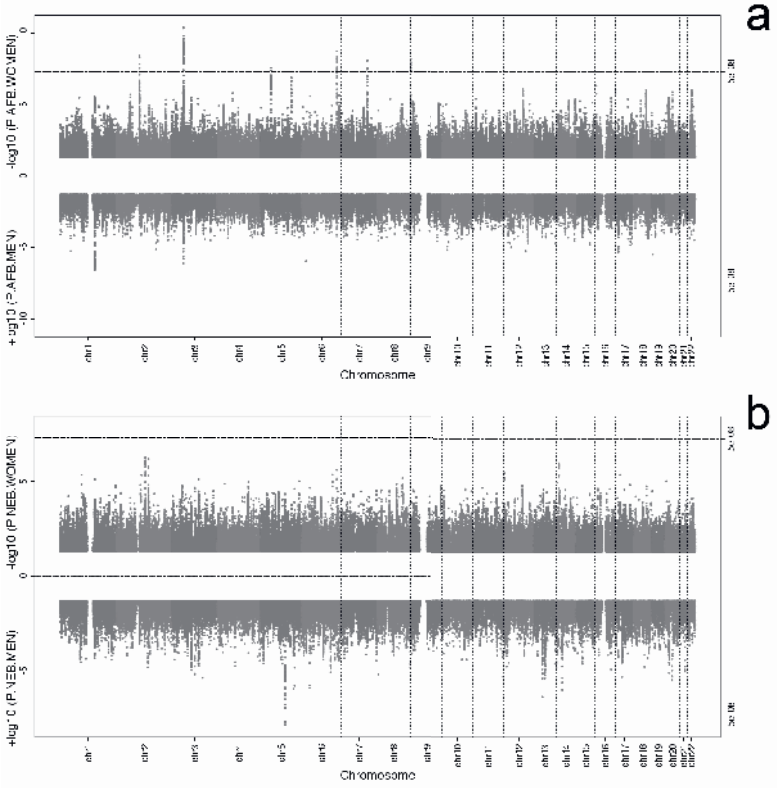


Figure 3.4: *Miami plots of SNPs for age at first birth (AFB; panel a) and number of children ever born (NEB; panel a), resulting from sex-specific single-genomic-control meta-analyses of GWAS results. SNPs are plotted according to their position on each chromosome (x-axis) and the* $-log_{10}(p\text{-value})$ *from the association (y-axis). The upper half (resp. lower half) of both panels indicates the strength of associations in the female (male) sample. The dashed horizontal black lines indicate the threshold for genome-wide significance (i.e., p-value* $< 5 \cdot 10^{-8}$*).*

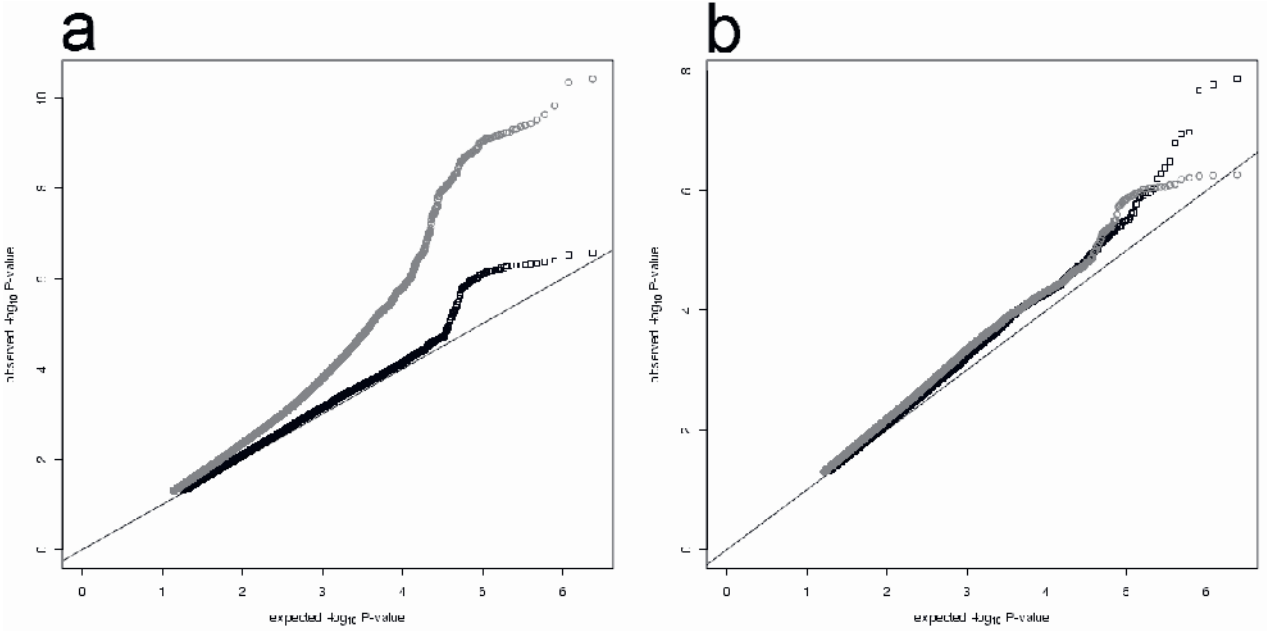


Figure 3.5: *Quantile-quantile plots of SNPs for age at first birth (AFB; panel a) and number of children ever born (NEB; panel a), resulting from sex-specific single-genomic-control meta-analyses of GWAS results. The gray circles (resp. black squares) indicate the quantiles of GWAS meta-analyses results in the female (male) sample.*

Table 3.3: *Independent SNPs reaching genome-wide significance (i.e., association p-value* $< 5 \cdot 10^{-8}$*) in meta-analyses of GWAS results for age at first birth and number of children ever born, in sex-stratified and/or pooled analyses.*

rsID	Chr**	Base-pair position	Nearest genes	Effect/ other allele	*N*	Effect allele frequency	Pooled		Men		Women	
							$\widehat{\beta}$	*p*-value	$\widehat{\beta}$	*p*-value	$\widehat{\beta}$	*p*-value
Age at first birth (AFB)												
rs10908557	1	153927052	CRTC2	C/G	249,025	0.695	0.091	$5.59{\cdot}10^{-10}$*	0.185	$2.98{\cdot}10^{-07}$	0.076	$5.38{\cdot}10^{-06}$
rs1160544	2	100832218	LINC01104	A/C	250,330	0.395	−0.082	$2.90{\cdot}10^{-09}$*	−0.042	$2.12{\cdot}10^{-01}$	−0.092	$5.00{\cdot}10^{-09}$*
rs2777888	3	49898000	CAMKV	A/G	250,941	0.507	0.106	$4.58{\cdot}10^{-15}$*	0.155	$2.40{\cdot}10^{-06}$	0.095	$6.07{\cdot}10^{-10}$*
rs6885307	5	45094503	MRPS30, HCN1	A/C	248,999	0.799	−0.107	$2.32{\cdot}10^{-10}$*	−0.131	$2.07{\cdot}10^{-03}$	−0.104	$3.90{\cdot}10^{-08}$*
rs10056247	5	133898136	JADE2	T/C	249,429	0.289	0.082	$4.37{\cdot}10^{-08}$*	0.050	$1.68{\cdot}10^{-01}$	0.089	$1.28{\cdot}10^{-07}$
rs2347867	6	152229850	ESR1	A/G	248,039	0.649	0.091	$1.38{\cdot}10^{-10}$*	0.098	$4.69{\cdot}10^{-03}$	0.097	$1.80{\cdot}10^{-09}$*
rs10953766	7	114313218	FOXP2	A/G	248,039	0.429	0.087	$1.82{\cdot}10^{-10}$*	0.106	$1.31{\cdot}10^{-03}$	0.089	$8.41{\cdot}10^{-09}$*
rs2721195	8	145677011	CYHR1	T/C	250,493	0.469	−0.073	$6.25{\cdot}10^{-07}$	−0.014	$6.85{\cdot}10^{-01}$	−0.099	$6.13{\cdot}10^{-09}$*
rs293566	20	31097877	NOL4L	T/C	245,995	0.650	0.081	$1.41{\cdot}10^{-08}$*	0.110	$1.47{\cdot}10^{-03}$	0.079	$1.31{\cdot}10^{-06}$
rs242997	22	34503059	LARGE1, ISX	A/G	238,002	0.613	−0.084	$3.38{\cdot}10^{-09}$*	−0.139	$8.51{\cdot}10^{-05}$	−0.076	$1.82{\cdot}10^{-06}$
Number of children ever born (NEB)												
rs10908474	1	153753725	SLC27A3, GATAD2B	A/C	342,340	0.384	0.020	$2.08{\cdot}10^{-08}$*	0.021	$8.10{\cdot}10^{-04}$	0.020	$7.89{\cdot}10^{-06}$
rs13161115	5	107050002	EFNA5, FBXL17	C/G	341,737	0.234	−0.041	$1.34{\cdot}10^{-02}$	−0.041	$1.37{\cdot}10^{-08}$*	0.005	$3.29{\cdot}10^{-01}$
rs2415984	14	46873776	LINC00871	A/G	315,167	0.470	−0.020	$2.34{\cdot}10^{-08}$*	−0.029	$2.41{\cdot}10^{-06}$	−0.016	$3.71{\cdot}10^{-04}$

* Independent signals reaching *p*-value $< 5 \cdot 10^{-8}$ in the meta-analysis.

** Chr: chromosome.

samples. More precisely, we performed a separate LD-score regression for meta-analysis results of (i) AFB in the pooled sample, (ii) AFB for women, (iii) AFB for men, (iv) NEB in the pooled sample, (v) NEB for women, and (vi) NEB for men. For each phenotype, we used the "`eur_w_ld_chr/`" files of LD scores computed by Finucane et al. (2015). These LD Scores were computed with genotypes from the European-ancestry samples in the 1000 Genomes Project (McVean et al., 2012) using only HapMap3 SNPs (Altshuler et al., 2010). Only HapMap3 SNPs with MAF > 1% were included in the LD-score regression.

Because GC will tend to bias the intercept of the LD-score regression downward, we did not apply GC to the summary statistics we used to estimate the LD-score regression. Furthermore, we excluded the deCODE cohort from the data for the estimation of the LD-score intercept for AFB and NEB, since the cohort-level regression estimates for deCODE did not directly correct for the high level of relatedness in the sample (their standard procedure is to apply GC). Our intercept estimates from the LD-score regressions are, thus, unbiased measures of the amount of stratification there is in the data (excluding deCODE) that we used for the GWAS of each phenotype.

For AFB, the estimated LD-score intercept is 1.022 (SE=0.008) and for NEB the estimated intercept is 1.009 (SE = 0.006). All intercept estimates, including those based on sex-stratified GWAS results, are not significantly different from one. By comparison, the mean of the χ^2 statistics for all the SNPs in the LD-score regressions is 1.239 for AFB and 1.141 for NEB. Under the null hypothesis that there is no confounding bias and that the SNPs have no causal effects on the phenotypes, the mean χ^2 statistics would be one. Thus, the mean χ^2 statistics being significantly greater than one, combined with intercept estimates not significantly different from one, indicate that some SNPs are associated with the phenotypes. The estimates we obtained, imply that about 9% of the observed inflation in the mean χ^2 statistics for AFB and about 6% of the inflation for NEB is accounted for by confounding bias (e.g., due to relatedness) rather than a polygenic signal.

As described in Section 3.6, we applied the standard single GC correction to obtain our main estimates. Once single GC is applied, the LD-score-regression estimates indicate no confounding bias due to population stratification. The LD-score-regression-intercept estimate after single GC is 0.9618 (SE= 0.0077) for AFB and 0.9763 (SE=0.0068)

for NEB. We can, therefore, conclude that the amount of inflation in our final results due to confounding biases is likely to be negligible.

3.8. CONCLUSIONS

This GWAS is the largest genetic epidemiological discovery effort for human reproduction to date, with critical implications for population fitness and clear physiological mechanisms linking hypothesized genes and observed phenotypes. Related studies previously focused on reproductive life span (Day et al., 2015c, Perry et al., 2013), age at first sexual intercourse (Day et al., 2016), and more proximal infertility phenotypes (Perry et al., 2014, Rahmioglu et al., 2014, Day et al., 2015b), largely overlooking AFB and NEB. The rapid postponement of AFB and increased infertility and involuntary childlessness in many societies (Mills et al., 2011) makes it important to uncover the genetic and biological architecture of reproduction.

We identify ten novel and confirm two recently identified genetic loci that are robustly associated with AFB and NEB. Our findings are anticipated to lead to insights into how postponing reproduction may be more detrimental for some – based on their genetic make-up – than others, fuel experiments to determine "*how late can you wait?*" (Menken, 1985) and stimulate reproductive awareness. Causal genes in the loci we identified may serve as novel drug targets, to prevent or delay age-related declines in fertility and sperm quality, and increase the efficiency of assisted reproductive technology. Our study is the first to examine the genetics of reproductive behavior in both men and women, and the first that is adequately powered to identify loci both in women and men. While effect sizes of the identified common variants are small, there are examples of GWAS-identified loci of a small effect that end up leading to important biological insights (Manolio et al., 2008, Hindorff et al., 2009).

4

Partitioning Educational-Attainment Heritability

Based on parts of Okbay et al. (2016b)

ABSTRACT

We partition the SNP-based heritability of years of education between (i) coding and non-coding regions of the genome and (ii) regions of the genome that are DNase I hypersensitive regions in different cell types by applying genomic-relatedness-matrix restricted-maximum-likelihood estimation to pooled data from the Health and Retirement Study, the Rotterdam Study, and the Swedish Twin Registry. Partitioned heritability estimates indicate that years-of-education-associated SNPs enrich nonsynonymous sites and regions that are DNase I hypersensitive in both blood cells and the brain. Only the enrichment of regions that are DNase I hypersensitive in blood, however, is statistically significant. A likely explanation for the typical failure of enrichment to reach significance is that the available SNPs in our analysis poorly represent nonsynonymous sites and DNase I hypersensitive regions and, thus, lead to biased partitioned heritability estimates.

4.1. BACKGROUND

Explanations of genomic-relatedness-matrix restricted maximum likelihood (GREML), at various levels of formality, have been given in previous publications (Yang et al., 2010, Visscher et al., 2010, Rietveld et al., 2013a, Lee and Chow, 2014). We followed the method developed by Gusev et al. (2014) and estimated the extent to which the heritable variance of years of education (*EduYears*) enriches coding SNPs and also SNPs residing in regions that are DNase I hypersensitive in particular cell types. Partitioning heritability in this way can help to elucidate the biological mechanisms through which genetic variation affects the phenotype of interest.

4.2. DATA AND METHODS

The investigators of the Rotterdam Study (RS) I and II genotyped their samples with the Illumina-550K chip; RS III, the Illumina-610K-Quad; the Swedish Twin Registry (STR), the HumanOmniExpress-12v1-A; and the Health and Retirement Study (HRS), the Illumina-Omni2.5-Beadchip. In all cohorts, the worldwide 1000 Genomes (1000G) phase I reference sample was used for imputation.

From the 1000G SNPs we selected the subset of available autosomal HapMap3 SNPs with an imputation R^2 above 70%. We rounded the dosages to best-guess genotypes. In each cohort we performed quality control (QC) on the best-guess genotypes, excluding all SNPs meeting any of the following criteria: minor allele frequency (MAF) < 0.01, Hardy-Weinberg-Equilibrium (HWE)-test p-value < 0.01, and missingness (i.e., the fraction of genotypes that is missing) > 0.05. We also excluded individuals missing more than 5 percent of their genotype calls. After QC we merged the five cohorts. This procedure yielded a merged dataset consisting of 1,062,589 SNPs available in all cohorts. The total number of individuals in the merged set was 29,765.

In each cohort we corrected *EduYears* for age, squared age, and sex. The resulting residuals were standardized within cohort to have sample mean zero and unit sample variance. In the merged dataset we selected individuals with non-missing measurements of the control and outcome variables. In addition, from each twin pairship in the pooled data, we selected at most one twin. The sample size remaining after

these steps was 26,180.

We applied a second round of QC to the merged data, with the same thresholds applied at the cohort level, leading to a dataset comprising 1,052,745 SNPs. In this final set of markers and individuals, we used GCTA (Yang et al., 2011a) to construct the genomic-relatedness-matrix (GRM) and calculate the eigendecomposition of the GRM. From this decomposition we retained the first 20 principal components. Finally, we included cohort dummies as additional controls. From each pair of individuals with a genetic relatedness greater than 0.025, one individual was excluded using the pruning function in GCTA. This relatedness cutoff led to a final sample of 20,450 individuals.

In the taxonomy of Gusev et al. (2014), SNPs are assigned to six different categories (i.e., nonsynonymous, UTR, promoter, DNase I hypersensitive regions, intronic, and intergenic). We adopted the data sources of Gusev et al. (2014), but for simplicity collapsed all SNPs in the five noncoding categories. This classification yielded 16,565 coding and 1,036,180 noncoding SNPs. For each of the two categories, we constructed a GRM. In essence, we modeled elements of the matrix of phenotypic products as a linear combination of corresponding elements of the GRMs. The variance components that give weights to the GRMs correspond to the SNP-based additive genetic variances attributable to the different classes of SNPs (i.e., coding or non-coding).

We carried out another partitioning analysis by constructing three (partially overlapping) subsets, containing SNPs located in regions that are DNase I hypersensitive in blood cells, brain cells, and other cell types, respectively. For each subset, we constructed a GRM based on the SNPs in the subset, a GRM based on SNPs located in regions that are DNase I hypersensitive region in some cells but not in the cell type under consideration, and a GRM based on SNPs outside any region ever observed to be DNase I hypersensitive. We, subsequently, used GREML with three genetic variance components to estimate the respective contributions to heritability made by three types of SNPs.

We note that the GREML procedure we employ can produce downward biased estimates if the SNPs with non-zero partial regression coefficients are not representative of the entire category with respect to linkage disequilibrium (LD; Speed et al. 2012), but the magnitude of any such bias is likely to be small (Lee and Chow, 2014). Moreover, in partitioning heritability, we have no *a priori* reason to suspect any

specific functional category is more prone to a downwards bias than the other categories (i.e., we assume the average LD in the different functional categories to be even across categories). If this assumption holds, although the estimated component of each functional category can be biased, the contribution relative to the total estimated genetic variance will on average not be biased, as the bias in numerator and denominator tend to cancel out.

4.3. PARTITIONED HERITABILITY RESULTS

Table 4.1 shows the results of our GREML partitions. Turning first to the partition between coding and non-coding SNPs, one can see that noncoding SNPs explain the bulk of the genetic variation. This is not surprising since coding SNPs are outnumbered by a factor ≈ 60. The enrichment statistic – defined as the proportion of heritability captured by a set of SNPs divided by the proportion of SNPs in that set – suggests that coding SNPs are enriched by $\approx$ 3-fold; however, using a likelihood-ratio test, we found that this statistic is not significantly greater than one.

Similarly, in our partitions between SNPs in regions that are DNase I hypersensitive in a particular cell type and other SNPs, there appears to be enrichment of regions that are DNase I hypersensitive regions in blood and the brain, but only the $\approx$ 2-fold enrichment of blood is statistically significant.

The lack of statistical significance is likely to be driven by the poor representation of causal SNPs in enriched regions by our subset of HapMap3 SNPs. In more detail, we must consider that the SNP-based heritability captured by genotyping chips is already near the asymptote once the number of SNPs is about 400k (Yang et al., 2010, Vattikuti et al., 2012), which is a small subset of the roughly 8 million SNPs in European populations where both alleles are common. Therefore, when attempting to partition a fixed SNP-based heritability with a reduced subset of all common SNPs, the true heritability contributed by a SNP that bears a particular annotation but is missing from the panel must be captured by other SNPs in LD, and these proxy SNPs will often fall in other functional categories; this will tend to reduce the estimated

Table 4.1: *Functional partition of the SNP heritability of years of education using GREML estimation. DHS denotes DNase I hypersensitive site.*

Annotation	Number of SNPs	Proportion of SNPs	Heritability (SE)	Enrichment
Synonymous				
Coding	16,565	0.016	0.008 (0.007)	3.212
Non-coding	1,036,180	0.984	0.148 (0.018)	0.959
DNase I hypersensitive in blood				
DHS in blood	145,361	0.138	0.047 (0.021)	2.054*
DHS in non-blood tissues	118,742	0.113	0.000 (0.022)	0.000
not DHS in any tissue	788,642	0.749	0.119 (0.027)	0.956
DNase I hypersensitive in the brain				
DHS in brain	101,653	0.097	0.023 (0.019)	1.499
DHS in non-brain tissues	162,450	0.154	0.014 (0.023)	0.558
not DHS in any tissue	788,642	0.749	0.121 (0.027)	1.027
DNase I hypersensitive in other tissues				
DHS in tissues other than blood or brain	233,872	0.222	0.021 (0.024)	0.603
DHS only in blood or brain	30,231	0.029	0.018 (0.012)	4.060
not DHS in any tissue	788,642	0.749	0.118 (0.027)	1.001

* P-value < 0.01 in one-sided likelihood-ratio test whether enrichment is greater than that of non-DHS SNPs.

heritability accounted for by SNPs in enriched regions (and to increase the estimated heritability accounted for by SNPs in impoverished regions). For instance, the very numerous classes of SNPs that are not DNase I hypersensitive in the brain will appear to capture more of the fixed SNP-based heritability than these classes actually contribute, because many of their SNPs tag DNase I hypersensitive regions that are not well-represented in the panel. Gusev et al. (2014) noted that DNase I hypersensitive regions are especially prone to a misallocation of their SNP-based heritability to other regions when panels of SNPs smaller than 1000G are used. Rather than attempting to remedy this limitation, in the full manuscript by Okbay et al. (2016b), we turned to stratified LD-score regression, a novel method for partitioning heritability that is not constrained in this manner.

II

Advanced Methods for Individual-Level Data

5

A Review of Ridge Regression in Quantitative Genetics

Based on De Vlaming and Groenen (2015)

ABSTRACT

In recent years, there has been a considerable amount of research into the use of regularization methods for inference and prediction in quantitative genetics. Such research mostly focuses on selection of markers and shrinkage of their effects. In this review chapter, the use of *ridge regression* for prediction in quantitative genetics using *single-nucleotide polymorphism* data is discussed. In particular, we consider (i) the theoretical foundations of ridge regression, (ii) its link to commonly used methods in animal breeding, (iii) the computational feasibility, and (iv) the scope for constructing prediction models with nonlinear effects (e.g., *dominance* and *epistasis*). Based on a simulation study we gauge the current and future potential of ridge regression for prediction of human traits using genome-wide SNP data. We conclude that for outcomes with a relatively simple genetic architecture, given current sample sizes in most cohorts (i.e., $N < 10$k) the predictive accuracy of ridge regression is slightly higher than the classical *genome-wide association study* approach of *repeated simple regression* (i.e., one regression per SNP). However, both capture only a small proportion of the heritability. Nevertheless, we find evidence that for large-scale initiatives, such as biobanks, sample sizes can be achieved where ridge regression improves predictive accuracy substantially when compared to the classical approach.

5.1. INTRODUCTION

The advent of large-scale molecular genetic data has paved the way for using these data to help predict, diagnose, and treat complex human diseases (Pharoah et al., 2002). In recent years, the use of such data for the prediction of polygenic diseases and traits has become increasingly popular (Meigs et al., 2008, Purcell et al., 2009, Smoller et al., 2013). This venue has proved successful even for traits such as educational attainment and cognitive performance (Rietveld et al., 2013a, 2014). The vast majority of research into the genetic architecture of human traits and diseases is exploratory, and considers the effects of at least hundreds of thousands of *single-nucleotide polymorphisms* (SNPs) on the outcome of interest (Purcell et al., 2007).

Predictions based on molecular genetic data are typically constructed as a weighted linear combination of the available SNPs. This yields a so-called *polygenic risk score* (or *polygenic score*, *genetic risk score*, *genome-wide score*; Purcell et al. 2009, Evans et al. 2009). *Multiple regression* (*ordinary least squares*, OLS) is a natural technique for estimating the weights of the predictors (SNPs) in this context, but cannot be applied here: in general, the number of samples (N) available is far lower than the number of SNPs (P); typically, $N <$ 10k and $P >$ 100k. OLS would yield a perfect in-sample prediction without any predictive value out-of-sample, and would not allow to draw inferences on the weights of the SNPs, as they are nonunique. A commonly accepted solution to this problem is to carry out a *genome-wide association study* (GWAS), where one regresses the outcome of interest on each SNP separately. In this paper, we call this the *repeated simple regression* (RSR) approach.

Polygenic scores are typically constructed as the weighted sum of the SNPs with weights resulting from a GWAS using RSR. We raise four points of critique regarding this method. The first problem with this approach is that, in contrast to multiple regression, there is no search for the best linear combination over all SNPs jointly for predicting the outcome. A second, related, problem is that highly correlated SNPs (i.e., SNPs in strong *linkage disequilibrium*) repeatedly contribute very similar information, thereby distorting the risk score. For example, consider a set of ten perfectly correlated SNPs. In the RSR, they receive exactly the same weight. As the polygenic risk score

is a weighted linear sum of the SNPs with the weights coming from RSR, these perfectly correlated SNPs contribute a factor ten stronger to the risk score, than a single SNP capturing all information from that region does. This factor ten does not depend on the predictive power of the information in that region. A third problem is that the polygenic risk score can theoretically be correlated with *confounding variables* (*confounders*, *control variables*, *controls*). For instance, SNPs can be correlated with the population structure. Therefore, the polygenic risk – being a linear combination of SNPs – can be correlated with the confounders. Usually, confounders, such as age and gender, are included as regressors in order to control for spurious relations through these covariates. However, we find that often in empirical work researchers do not control properly for the confounders in at least one of the many steps that lead from phenotype and genotype data, to evaluation of the out-of-sample predictive accuracy of the polygenic risk score. A fourth problem is that the RSR approach is not able to handle even two-way interactions between the SNPs, as it would lead to a number of weights to be estimated that is quadratic in the number of SNPs, which is clearly computationally infeasible.

In this paper, we review the use of *ridge regression* (RR; Hoerl and Kennard 1970) to tackle the four problems discussed above. The purpose of this paper is threefold. First, we discuss how prediction using RR can address the aforementioned four points of critique pertaining to a typical polygenic score. That is, how RR can be used to search for the best linear combination of SNPs jointly, to address the multicollinearity of SNPs (Malo et al., 2008, Abraham et al., 2013), to account for the presence of confounding variables and of nonlinear SNP effects (González-Recio et al., 2008, Gianola and Van Kaam, 2008, De Los Campos et al., 2009, Crossa et al., 2010, Endelman, 2011, Morota and Gianola, 2014). Second, we review relevant work on ridge regression both in and outside the field genetics. Third, we assess the merits of prediction using ridge regression in the new domain of biobanks. That is, we predict the expected accuracy of ridge regression in large scale initiatives with over a 100k observations.

An important property of RR is that it cannot select a subset of predictors (e.g., SNPs). Other regularization methods related to RR are able to select a subset of predictors from a large set of predictors. Examples of such methods are the *least absolute shrinkage and selec-*

tion operator (LASSO), group LASSO (Yuan and Lin, 2006), adaptive LASSO (Zou, 2006), and the elastic net (Zou and Hastie, 2005).

In a GWAS, SNP selection is a desirable property when trying to find regions in the DNA that bear a causal influence on the outcome. However, there is mixed evidence for the claim that selection techniques in general improve the overall predictive accuracy of the polygenic score. Some studies suggest that preselection of markers (e.g., SNPs), either based on linkage disequilibrium or (in-sample) univariate association results, is detrimental to predictive accuracy (Purcell et al., 2009, Evans et al., 2009, Abraham et al., 2013, Benner et al., 2010). Moreover, there is no conclusive evidence on the relative performance of RR-type methods and LASSO-type methods. For instance, using a simulation study, Ogutu et al. (2012) find that LASSO-type methods outperform classic RR, whereas other studies find that RR outperforms LASSO and similar variable selection methods (Frank and Friedman, 1993, Bøvelstad et al., 2007, Van Wieringen et al., 2009). A reasonable proposition is that the relative performance of RR and LASSO depends on trait architecture (Benner et al., 2010, Usai et al., 2009). In particular, a low number of causal SNPs favors LASSO-type methods, whereas an intermediate or high number of causal variants favors RR-type methods. Regularization methods performing selection are computationally more involved and less amenable to incorporate nonlinear SNP effects than RR. For the above reasons, as well as our aim to provide a clear overview of RR, we focus in this paper primarily on RR.

The remainder of this paper is organized as follows. In Section 5.2, we present the theory underlying RR. In Section 5.3, we show that RR can be perceived as a method between OLS and RSR, leveraging the advantages of these two methods. Subsequently, in Section 5.4, we discuss the relation between RR and the *best linear unbiased prediction* used in animal breeding, and the relation between RR and LASSO-type methods. In Section 5.5, we pay special attention to the effect standardization of SNP data has on the implicit assumptions about the genetic architecture of traits. As indicated, the feasibility of RR depends critically on the use of computationally efficient approaches. These will be discussed in Section 5.6. Related to this, in Section 5.7, we will discuss methods to tune the penalty parameter of RR. Following, in Section 5.8, advanced RR techniques will be discussed, such as

modelling nonlinear effects using RR, weighting SNPs differently, and incorporating information from earlier studies.

In order to assess the current and future use of ridge regression for prediction in quantitative genetics, we run a suite of simulations. The design of the simulations and the results are presented in Section 5.9. Based on these results we will estimate the effect sample size, the number of SNPs, the number of causal SNPs, and trait heritability have on the predictive accuracy of RR and the classical RSR approach. Using these estimates we will extrapolate how RR and RSR are expected to perform relative to each other in large scale studies (e.g., N >100k). Finally, in Section 5.10, we summarize the most important aspects of RR in the context of prediction in quantitative genetics, and discuss our expectations for its future uses.

5.2. RIDGE REGRESSION

Using *ridge regression* (RR) for prediction in quantitative genetics was first proposed by Whittaker et al. (2000). RR can be understood as follows. Like regular *least-squares* methods RR minimizes a loss function that includes the sum of squared regression residuals. However, opposed to least squares, the loss function also includes a term consisting of positive penalty parameter λ times the model complexity, measured by the sum of squared regression weights (Hoerl and Kennard, 1970). This penalty prevents overfitting by shrinking the weights towards zero, ensuring that even in case of multicollinearity and $P \gg N$, the estimator has a solution. The RR estimator has a simple analytical solution.

More formally, given a set of N individuals, P SNPs, and K confounders, a linear model for quantitative outcome vector $\mathbf{y}$ ($N \times 1$), with a matrix of SNP data $\mathbf{X}$ ($N \times P$), and a matrix of confounders $\mathbf{Z}$ ($N \times K$) as predictors, is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{5.1}$$

where $\boldsymbol{\beta}$ is the vector of SNP effects, $\boldsymbol{\gamma}$ the vector of effects of the confounders, and $\boldsymbol{\varepsilon}$ the phenotype noise.

In this particular case, we consider a large set of SNPs and a small set of potential confounders. Since one of our aims is to prevent any

spurious relations via the confounders, we use a loss function that does not apply shrinkage to these. Therefore, the RR estimator minimizes

$$\mathscr{L}_{\text{RR}}(\boldsymbol{\beta},\boldsymbol{\gamma}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\gamma})^{\top}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\gamma}) + \lambda\boldsymbol{\beta}^{\top}\boldsymbol{\beta}. \tag{5.2}$$

Under this loss function, the RR estimator of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}}_{\text{RR}} = (\mathbf{X}^{\top}\mathbf{M}_{\mathbf{Z}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\top}\mathbf{M}_{\mathbf{Z}}\mathbf{y}, \tag{5.3}$$

where $\mathbf{M}_{\mathbf{Z}} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}^{\top}\mathbf{Z})^{-1}\mathbf{Z}^{\top}$ is the projection matrix, removing the effects of the confounding variables. The larger λ, the more shrinkage is applied. When $\lambda = 0$, RR corresponds to OLS. The OLS estimator only exists if $\text{rank}(\mathbf{X}^{\top}\mathbf{M}_{\mathbf{Z}}\mathbf{X}) = P$, meaning that there is no perfect collinearity amongst the SNPs and that $P \leq N$. However, in a GWAS, almost invariably $P \gg N$. Therefore, OLS cannot be applied in this context. However, the RR estimator has a solution for any $\lambda > 0$, even if $P \gg N$.

Heteroskedastic ridge regression (HRR) is a generalization of RR, where each SNP p receives a different amount of shrinkage, $\lambda_p \geq 0$. The loss function of HRR is given by

$$\mathscr{L}_{\text{HRR}}(\boldsymbol{\beta},\boldsymbol{\gamma}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\gamma})^{\top}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\gamma}) + \lambda\boldsymbol{\beta}^{\top}\boldsymbol{\Lambda}\boldsymbol{\beta},$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1,\ldots,\lambda_P)$. The corresponding estimator is given by

$$\widehat{\boldsymbol{\beta}}_{\text{HRR}} = (\mathbf{X}^{\top}\mathbf{M}_{\mathbf{Z}}\mathbf{X} + \lambda\boldsymbol{\Lambda})^{-1}\mathbf{X}^{\top}\mathbf{M}_{\mathbf{Z}}\mathbf{y}. \tag{5.4}$$

The $P \times P$ matrix $\mathbf{X}^{\top}\mathbf{M}_{\mathbf{Z}}\mathbf{X}$ in Equations 5.3 and 5.4 can be regarded as a map of the estimated correlation (linkage disequilibrium) between markers. OLS takes this linkage disequilibrium fully into account at the expense of overfitting the data, whereas RSR completely ignores it. For this reason, when constructing a polygenic score, RSR is often used in combination with a heuristic procedure, known as linkage disequilibrium pruning, which selects SNPs that are not too strongly correlated. As is shown in the next section, RR leverages the two extremes of OLS and RSR. Therefore, opposed to RSR, RR does not require the *a priori* selection of SNPs; RR is able to handle linkage disequilibrium between markers (Malo et al., 2008, Abraham et al., 2013).

RR is expected to perform particularly well under a scenario where a substantial proportion of the SNPs is expected to contribute to the phenotype, and where each contribution is small.

5.3. LIMITING CASES

Varying the penalty weight, λ, allows specifying special cases of RR. Prediction by RR can be perceived as a method that lies between prediction based on OLS estimates considering all SNPs jointly and OLS estimates considering each SNP separately. By definition of RR (Hoerl and Kennard, 1970), for sufficiently low shrinkage the RR estimates converge to the multiple regression estimates (Malo et al., 2008), provided these are unique. For sufficiently high shrinkage a RR prediction score is equivalent to an RSR prediction score, in terms of the proportion of variance accounted for by the respective scores. For ease of notation, we assume in this section that there are no confounders $\mathbf{Z}$.

To establish aforementioned relations, two conditions are needed. First, the measure of predictive accuracy is independent of scale. That is, given an out-of-sample quantitative outcome vector ($\mathbf{y}_2$) and its prediction ($\widehat{\mathbf{y}}_2$), the accuracy measure should be such that for any coefficient $b > 0$, the accuracy of prediction $\widehat{\mathbf{y}}_2$ is identical to that of prediction $\widehat{\mathbf{y}}_2^* = b\widehat{\mathbf{y}}_2$. An example of such a measure is the R^2 of an outcome and its prediction. The second condition is that SNP data are standardized, such that each SNPs p has mean zero ($\mathbf{x}_p^\top \boldsymbol{\iota} = 0$, where $\boldsymbol{\iota}^\top = (1,\ldots,1)$) and equal standard deviation ($\mathbf{x}_p^\top \mathbf{x}_p = c$, where c is a scalar).

Consider the prediction of $\mathbf{y}_2$ based on $N_2 \times P$ out-of-sample genotype matrix $\mathbf{X}_2$, using in-sample RR estimates $\widehat{\boldsymbol{\beta}}_{\text{RR}}$. This prediction is given by $\widehat{\mathbf{y}}_2 = \mathbf{X}_2\widehat{\boldsymbol{\beta}}_{\text{RR}}$. Based on the first condition, we can multiply the prediction $\widehat{\mathbf{y}}_2$ by $b = (1+\lambda)$. This is equivalent to inflating the RR estimates by $(1+\lambda)$ instead of inflating the predictions. Thus, we can take $\widehat{\boldsymbol{\beta}}^*_{\text{RR}} = (1+\lambda)\widehat{\boldsymbol{\beta}}_{\text{RR}}$. This yields

$$\widehat{\boldsymbol{\beta}}^*_{\text{RR}} = \left(\alpha\mathbf{I} + (1-\alpha)\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y},$$

where $\alpha = (1+\lambda)^{-1}\lambda \in (0,1)$. The OLS estimator considering all SNPs jointly is given by $\widehat{\boldsymbol{\beta}}_{\text{OLS}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$. Thus, it follows that when α goes to zero (i.e., λ goes to zero), the RR estimator goes to the OLS

estimator. Moreover, as α goes to one (i.e., λ becomes sufficiently large), the inflated RR estimator goes to $\mathbf{X}^\top\mathbf{y}$.

Using the condition of having standardized SNPs, we can rewrite the RSR for SNP p as $\widehat{\beta}_p = \mathbf{x}_p^\top\mathbf{y}$, where $\mathbf{x}_p$ is the standardized genotype vector of SNP p. This expression can vectorized over all SNPs as $\widehat{\boldsymbol{\beta}}_{\text{RSR}} = \mathbf{X}^\top\mathbf{y}$. From this, it follows that the inflated RR estimates approach the RSR estimates as λ becomes sufficiently large.

5.4. RELATED METHODS

Prediction using RR is related to the predictions that arise under a widely used *linear mixed model* (LMM), commonly referred to as the *animal model*. In such a model, expected genetic relatedness is mapped to phenotypic relatedness. Usually pedigree information is used to infer genetic relatedness. However, with the advent of genome-wide molecular data, LMMs that use SNPs to estimate genetic relatedness have been proposed (e.g., Yang et al. 2011a). In most LMMs using SNPs, the prior assumption is that SNP effects are normally distributed with mean zero and variance $\sigma^2_{\boldsymbol{\beta}}$, and the error terms in the phenotype are also normally distributed with variance $\sigma^2_{\boldsymbol{\varepsilon}}$.

To understand the relation between RR and mixed models, consider the following LMM,

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},\\ \boldsymbol{\beta} &\sim \mathcal{N}\left(\mathbf{0}, \sigma^2_{\boldsymbol{\beta}}\mathbf{I}_P\right),\\ \boldsymbol{\varepsilon} &\sim \mathcal{N}\left(\mathbf{0}, \sigma^2_{\boldsymbol{\varepsilon}}\mathbf{I}_N\right),\end{aligned} \tag{5.5}$$

where $\sigma^2_{\boldsymbol{\beta}}$ is the SNP effect variance and $\sigma^2_{\boldsymbol{\varepsilon}}$ the noise variance. In this model the effects of the confounders, $\mathbf{Z}$, are assumed to be fixed. For the remainder of this section we ignore the confounders for ease of notation. The parameters $\sigma^2_{\boldsymbol{\varepsilon}}$ and $\sigma^2_{\boldsymbol{\beta}}$ can be estimated using, for instance, *maximum likelihood*, *restricted maximum likelihood* (Patterson and Thompson, 1971), or *expectation maximization* (Dempster et al., 1977). Alternatively, these parameters can be fixed by using prior information from other data sets, see, for instance, Hofheinz et al. (2012).

Consider conditional expectations $\mathbb{E}\left[\boldsymbol{\beta} \,\middle|\, \mathbf{y}\right]$ and $\mathbb{E}\left[\mathbf{y}_2 \,\middle|\, \mathbf{y}\right]$. In an LMM such expectations are known as the *best linear unbiased prediction*

(BLUP; Henderson 1950, 1953, 1963, 1975, 1985). BLUP was first proposed by Henderson (1950) in order to obtain estimates of the so-called *breeding values*, that is, the part of the phenotype that can be attributed to genetic variation.

Provided that the RR penalty $\lambda = \sigma^2_{\boldsymbol{\varepsilon}}/\sigma^2_{\boldsymbol{\beta}}$, the BLUP of SNP effects (Yang et al., 2011a, Meuwissen et al., 2001, Schaeffer, 2006) is equivalent to the RR estimator. Under that same condition, the BLUP of the SNP-based breeding values is equivalent to RR prediction. Such *genomic estimated breeding values* (Schaeffer, 2006) contain the part of the phenotype that can be attributed to the genetic variation in the genotyped markers.

To understand this equivalence, first we rewrite the RR estimator in Equation 5.3. By applying the *Sherman-Morrison-Woodbury formula* (Sherman and Morrison, 1950, Woodbury, 1950) to the $P \times P$ inverse of $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$, we obtain

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\text{RR}} &= \frac{1}{\lambda}\left[\mathbf{I}_P - \mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N\right)^{-1}\mathbf{X}\right]\mathbf{X}^\top\mathbf{y} \\
&= \frac{1}{\lambda}\mathbf{X}^\top\left[\mathbf{I}_N - \left(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N\right)^{-1}\mathbf{X}\mathbf{X}^\top\right]\mathbf{y} \\
&= \frac{1}{\lambda}\mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N\right)^{-1}\left[\left(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N\right) - \mathbf{X}\mathbf{X}^\top\right]\mathbf{y} \\
&= \mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N\right)^{-1}\mathbf{y}. \qquad (5.6)
\end{aligned}$$

Second, by rewriting Equation 5.5 in terms of the joint distribution of $\mathbf{y}$ and $\boldsymbol{\beta}$,

$$\begin{pmatrix} \mathbf{y} \\ \boldsymbol{\beta} \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma^2_{\boldsymbol{\beta}}\mathbf{X}\mathbf{X}^\top + \sigma^2_{\boldsymbol{\varepsilon}}\mathbf{I}_N & \sigma^2_{\boldsymbol{\beta}}\mathbf{X} \\ \sigma^2_{\boldsymbol{\beta}}\mathbf{X}^\top & \sigma^2_{\boldsymbol{\beta}}\mathbf{I}_P \end{bmatrix}\right),$$

the BLUP of $\boldsymbol{\beta}$ is given by the expectation of $\boldsymbol{\beta}$ conditional on $\mathbf{y}$ (Morota and Gianola, 2014). This yields,

$$\widehat{\boldsymbol{\beta}}_{\text{BLUP}} = \sigma^2_{\boldsymbol{\beta}}\mathbf{X}^\top\left(\sigma^2_{\boldsymbol{\beta}}\mathbf{X}\mathbf{X}^\top + \sigma^2_{\boldsymbol{\varepsilon}}\mathbf{I}_N\right)^{-1}\mathbf{y} = \mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top + \frac{\sigma^2_{\boldsymbol{\varepsilon}}}{\sigma^2_{\boldsymbol{\beta}}}\mathbf{I}_N\right)^{-1}\mathbf{y}.$$

Clearly, when $\lambda = \sigma^2_{\boldsymbol{\varepsilon}}/\sigma^2_{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\beta}}_{\text{RR}} = \widehat{\boldsymbol{\beta}}_{\text{BLUP}}$.

In addition, from a Bayesian perspective the posterior mode of the distribution of SNP effects (i.e., the mode of the distribution conditional on a training set) can also be used as point estimator. Estimation using

the posterior mode is known as *maximum a posteriori* (MAP) estimation. However, due to the normality of $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ the mode coincides with the conditional expectation $\mathbb{E}[\boldsymbol{\beta} \mid \mathbf{y}]$. Therefore, MAP estimation of $\boldsymbol{\beta}$ in Equation 5.5 is equivalent to BLUP.

Consequently, there exists a λ such that the RR estimator of SNP effects is equivalent to its BLUP (Endelman, 2011), and by extension to the MAP estimator. The diagram in Figure 5.1 summarizes the relations between RR, BLUP, and MAP.

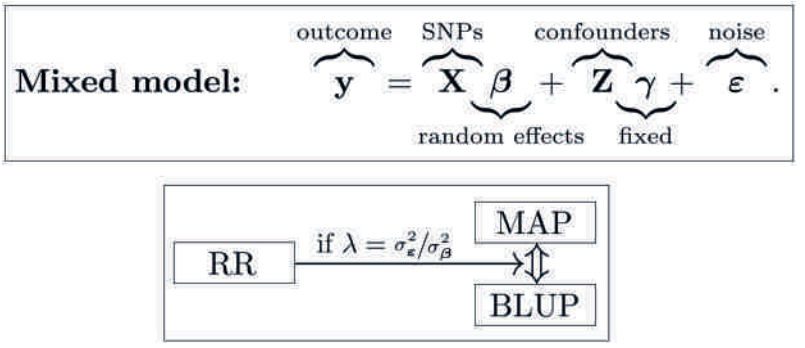


Figure 5.1: *Diagram (lower panel) showing the relation between estimation of SNP effects and prediction using ridge regression (RR), best linear unbiased prediction (BLUP), and maximum a posteriori (MAP) estimation, under the specified linear mixed model (upper panel), where* $\sigma^2_{\boldsymbol{\varepsilon}}$ *denotes the variance of noise* $\boldsymbol{\varepsilon}$ *and* $\sigma^2_{\boldsymbol{\beta}}$ *the variance in the random SNP effects* $\boldsymbol{\beta}$.

5.4.1. *LASSO-type methods*

An important feature that RR lacks is the selection of SNPs. LASSO-type methods, such as the LASSO, group LASSO, adaptive LASSO, and the elastic net, are able to select SNPs. The key to achieving SNP selection is to include an L_1 penalty, that is, adding a penalty consisting of a penalty parameter, λ, times $||\boldsymbol{\beta}||_1 = |\beta_1| + \ldots + |\beta_P|$. The loss function of the LASSO is given by

$$\mathscr{L}_{\text{LASSO}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma}) + \lambda ||\boldsymbol{\beta}||_1.$$

This function is highly similar to the RR loss function in Equation 5.2. The most important property of the LASSO is that it performs variable selection, that is, for a sufficiently large λ many of the SNP

coefficients β_p will be zero. The higher the λ, the fewer non-zero SNP effects are obtained by the LASSO. Moreover, this method also shrinks the non-zero coefficients, that is, the estimated effects of the selected SNPs.

The loss function of the elastic net (Zou and Hastie, 2005) is obtained by taking a convex combination of $\boldsymbol{\beta}^\top\boldsymbol{\beta}$ and $||\boldsymbol{\beta}||_1$ as penalty, that is,

$$\mathscr{L}_{\text{net}}(\boldsymbol{\beta},\boldsymbol{\gamma}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\gamma})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\gamma}) + \lambda\left(\alpha\boldsymbol{\beta}^\top\boldsymbol{\beta} + (1-\alpha)||\boldsymbol{\beta}||_1\right),$$

with $\lambda \geq 0$, and $\alpha \in [0,1]$. The elastic-net method preserves SNP selection, while allowing more than N of P SNPs to be selected. Taking a convex combination of the two norms hardly increases the computational costs of solving this problem, when compared to solving the LASSO problem (Zou and Hastie, 2005). Typically, the LASSO solution is obtained by means of the least-angle regression algorithm (Efron et al., 2004). This algorithm entails an iterative procedure, where at most one SNP can enter the model at a time. Therefore, LASSO-type methods are computationally far more involved than RR-type methods.

Finally, the group LASSO (Yuan and Lin, 2006) splits the P predictors in G mutually disjoint groups, with p_g predictors in group g, and associated effects $\boldsymbol{\beta}_g$, for groups $g = 1,\ldots,G$. The group LASSO minimizes

$$\mathscr{L}_{\text{group}}(\boldsymbol{\beta},\boldsymbol{\gamma}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\gamma})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\gamma}) + \lambda\sum_{g=1}^{G}\sqrt{\boldsymbol{\beta}_g^\top\boldsymbol{\beta}_g}.$$

Each group can be chosen, for instance, to represent a single gene in terms of its SNPs. The group LASSO induces sparsity at the group level (e.g., a gene is either included as a whole or wholly excluded), whereas within a group the individual regressors receive an L_2 penalty. To our best knowledge, Sabourin et al. (2015) provide the first, and so far only, application of a (modified) group LASSO using SNP data to construct polygenic scores. In this study, each SNP is considered as a group, with two effects: an additive and a dominance effect. In a simulation with mild to strong dominance, this method improves accuracy, compared to an RSR-type approach (Sabourin et al., 2015). For a detailed comparison of LASSO-type methods and RR, we refer to Hastie et al. (2009).

5.5. STANDARDIZING SNPS

In the preceding sections, we have only considered SNP standardization as a tool to show that RR can be perceived as a method between the classical GWAS approach and the OLS approach considering all SNPs jointly. However, SNP standardization is often used in the LMM in Equation 5.5.

The reason for this is that standardization has a profound effect on the implicit assumptions about the effect sizes of SNPs. We show in this section that the standardization we use is equivalent to HRR applied to raw genetic data, where SNPs measuring rare variants receive less shrinkage than SNPs measuring common variants.

More specifically, let $\mathbf{G}$ (resp. $\mathbf{G}_2$) denote raw SNP data in-sample (out-of-sample), that has already been mean-centered, but not yet standardized to have the same variance. The standardized data $\mathbf{X}$ in Section 5.3 can now be obtained by postmultiplying $\mathbf{G}$ by a diagonal matrix $\mathbf{D}$. That is, $\mathbf{X} = \mathbf{G}\mathbf{D}$, where

$$\mathbf{D} = \text{diag}\left(\left\{\sqrt{\frac{N-1}{\mathbf{x}_p^\top \mathbf{x}_p}}\right\}_{p=1,\ldots,P}\right).$$

Under the reasonable assumption that only SNPs are considered for which in-sample variation occurs, this matrix $\mathbf{D}$ is invertible.

By applying this transformation in both the training and test set, RR prediction based on standardized data is given by

$$\begin{aligned}\widehat{\mathbf{y}}_2 &= \mathbf{X}_2\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y} \\ &= \mathbf{G}_2\mathbf{D}\left(\mathbf{D}\mathbf{G}^\top\mathbf{G}\mathbf{D} + \lambda\mathbf{I}\right)^{-1}\mathbf{D}\mathbf{G}^\top\mathbf{y} = \mathbf{G}_2\left(\mathbf{G}^\top\mathbf{G} + \lambda\mathbf{D}^{-2}\right)^{-1}\mathbf{G}^\top\mathbf{y}.\end{aligned}$$

This shows that RR applied to standardized SNP data is equivalent to HRR, with $\mathbf{\Lambda} = \mathbf{D}^{-2}$, applied to raw genotype data. Here, the SNP-specific shrinkage depends on the amount of SNP variation. This type of shrinkage implicitly assumes that the standardized SNPs have homoskedastic effects, whereas the underlying raw genotypes (i.e., the count data) have effects of which the variance decreases with minor allele frequency. That is, rare alleles are assumed to have larger effects on average than common variants. For a qualitative treatment of the relation between allele frequency and expected effect sizes, see, for

instance, Manolio et al. (2009).

To be more precise, this type of shrinkage corresponds to the implicit assumption that the variance of the effect of raw SNP p, denoted by $\sigma^2_{\boldsymbol{\beta}_p}$, with allele frequency f_p, is proportional to $(2f_p(1-f_p))^{-1}$. This assumption implies that when f_p is close to one or zero, the variance of the effect size is expected to be large, whereas for f_p close to 50% the variance of the effect size attains its minimum.

Naturally, raw SNP effect variances responding differently to allele frequency can be conceived. As indicated by Manolio et al. (2009) such relations depend on the effect trait under consideration has on fitness. Therefore, a natural extension would be to consider HRR with $\boldsymbol{\Lambda} = \mathbf{D}^{\alpha}$. Here $\alpha = 0$ corresponds to a trait for which allele frequency is independent of effect size, $\alpha = -2$ corresponds to the relation described before. Moreover, $-2 < \alpha < 0$ describes a trait for which there is a slight relation between allele frequency and effect size. It is interesting to note that $\alpha > 0$ corresponds to a trait where diversity is an asset; that is, a trait in which variants causing phenotypic divergence between individuals, tend to become common. Finally, $\alpha < -2$ would correspond to a trait for which there has been strong selection pressure causing convergence; only very rare variants are expected to have a large effect. Thus, in future work α can be considered as an additional hyperparameter which might boost predictive accuracy and of which the estimate would reveal something about the selection pressure regarding the trait under consideration. The same type of transformation has been proposed by Speed et al. (2012) for improving estimation of SNP-based heritability in an LMM.

5.6. COMPUTATIONAL COSTS

The main hurdle in computing RR predictions is estimating the P parameters, when $P \gg N$. In particular, a naïve approach requires solving a system with P unknowns. However, RR can be implemented in a computationally efficient way. When $P > N$, using dimensionality reduction techniques the complexity of RR can be reduced from $\mathcal{O}(P^3)$ to $\mathcal{O}(PN^2)$ in case one is interested in the estimated effects (Hastie and Tibshirani, 2004).

Moreover, if the focus lies solely on obtaining predictions, a non-parametric representation of RR reveals that a dual formulation exists,

which can be perceived as solving a linear model with N unknowns (Kimeldorf and Wahba, 1970). Solving such a system has a complexity slightly less than $\mathcal{O}(N^3)$. Building on this computationally efficient approach, RR can also efficiently control for confounders, both in-sample and out-of-sample.

Finally, when considering a wide array of values of λ, RR can be reformulated to generate predictions for all values of λ jointly by exploiting the properties of the eigendecomposition of an $N \times N$ matrix, thereby yielding a complexity of $\mathcal{O}(N^3)$.

To understand these reductions in computational costs, consider the RR estimator in Equation 5.6, used to show equivalence of RR and the BLUP. Premultiplying this expression by $\mathbf{X}_2$, the out-of-sample prediction is given by

$$\widehat{\mathbf{y}}_2 = \mathbf{X}_2\mathbf{X}^\top \left(\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_N\right)^{-1}\mathbf{y}. \tag{5.7}$$

As discussed, accounting for confounding variables is important. Let $\mathbf{Z}$ be the in-sample $N \times K$ matrix of confounders and $\mathbf{Z}_2$ the out-of-sample $N_2 \times K$ matrix of confounders. By replacing $\mathbf{X}$ by $\mathbf{X}^* = \mathbf{M}_\mathbf{Z}\mathbf{X}$ and $\mathbf{X}_2$ by $\mathbf{X}_2^* = \mathbf{M}_{\mathbf{Z}_2}\mathbf{X}_2$, where $\mathbf{M}_\mathbf{C}$ is the projection matrix removing the effects of $\mathbf{C}$, we find that

$$\widehat{\mathbf{y}}_2 = \mathbf{A}_{21}^* \left(\mathbf{A}^* + \lambda_{\text{GRM}}\mathbf{I}_N\right)^{-1}\mathbf{y}, \tag{5.8}$$

where $\mathbf{A}^* = \mathbf{M}_\mathbf{Z}\mathbf{A}\mathbf{M}_\mathbf{Z}$ and $\mathbf{A}_{21}^* = \mathbf{M}_{\mathbf{Z}_2}\mathbf{A}_{21}\mathbf{M}_\mathbf{Z}$, $\mathbf{A} = P^{-1}\mathbf{X}\mathbf{X}^\top$ and $\mathbf{A}_{21} = P^{-1}\mathbf{X}_2\mathbf{X}^\top$, and $\lambda_{\text{GRM}} = P^{-1}\lambda$. Therefore, one can correct for covariates by simply pre- and postmultiplying $N_{(2)} \times N$ matrices, by appropriate projection matrices.

Matrices $\mathbf{A}$ and $\mathbf{A}_{21}$ both have the interpretation of a SNP-based *Genetic Relationship Matrix* (GRM; Yang et al. 2011a), measuring the genetic similarity of individuals in the space of additive SNP effects.

Given the eigendecomposition

$$\mathbf{A}^* = \mathbf{Q}\text{diag}\left(\{\theta_i\}_{i=1,\ldots,N}\right)\mathbf{Q}^\top, \tag{5.9}$$

RR prediction can be written as

$$\widehat{\mathbf{y}}_2 = \mathbf{A}_{21}^*\mathbf{Q}\text{diag}\left(\left\{\frac{1}{\theta_i + \lambda_{\text{GRM}}}\right\}_{i=1,\ldots,N}\right)\mathbf{Q}^\top\mathbf{y}. \tag{5.10}$$

If $P \gg N$, this approach is far more efficient than the naïve approach to RR prediction. GRMs can be computed efficiently in packages such as `PLINK` 1.9 (Chang et al., 2015) and `GCTA` (Yang et al., 2011a). The most involved step in the prediction procedure, is finding the eigendecomposition of $\mathbf{A}^*$.

5.7. TUNING AND INTERPRETING λ

So far, it was assumed that the penalty strength parameter λ is given. However, in most applications of RR the optimal λ is not known in advance. Here, we discuss three ways for choosing λ.

The dominant approach in the machine learning literature for tuning λ is by maximizing out-of-sample predictive accuracy of RR using *cross-validation* (CV). In CV one considers a fine grid $\mathscr{L}$ of potential values of λ. The data are randomly split in a (small) test set (e.g., 10% of the sample) and CV set (90%). To the CV set one applies K-fold CV (e.g., $K = 10$), meaning that one splits the CV sample randomly in K blocks of (approximately) equal size. In each fold $K-1$ blocks are considered as CV training set and the remaining block as CV test set. Using RR for all values of $\lambda \in \mathscr{L}$, predictions in the CV test set are generated. Each block is the CV test set precisely once. After the K folds, the predictive accuracy over all CV test sets is evaluated for all $\lambda \in \mathscr{L}$. Now, $\widehat{\lambda}$ is set to maximize the cross-validation accuracy. Finally, using $\widehat{\lambda}$ the predictive accuracy in the final test is considered, using the full CV set as training data. For a more detailed treatment of CV, see, for instance, Hastie et al. (2009).

Nested cross-validation (NCV) is a natural extension of CV, where the sample is randomly split in S "super"-blocks of approximately equal size (e.g., $S = 10$) and where there are S "super"-folds. In each super-fold, one block is considered as final test set and $S-1$ other blocks as CV set. To this CV set and test set one applies regular K-fold CV. Each super-block is the final test precisely once.

Classical CV is used to fit the model and to assess its predictive accuracy; one can judge the merits of a set of values of the hyperparameter by means of the CV procedure and apply the optimal value to a new part of the sample which has not yet been considered. Using NCV one can test whether the hyperparameter and accuracy that result from classical CV are robust; NCV can show the amount of variation in

either of these over the "super"-folds.

CV requires a computationally efficient strategy since a different set of RR predictions will result for each different value of λ. However, a large set of different values of λ can be evaluated in one step at nearly the same costs of evaluating a single value of λ. This approach avoids computing a full RR solution for each λ separately. To see this, the formulation of RR prediction in Equation 5.10 is highly relevant. In this equation, the eigendecomposition of $\mathbf{A}^*$ is independent of λ. Thus, predictions for each $\lambda \in \{\lambda_1, \ldots, \lambda_L\}$ can be obtained by the following equation.

$$\widehat{\mathbf{Y}}_2 = \mathbf{A}^*_{2,1}\mathbf{Q}\left[\begin{pmatrix} (\theta_1+\lambda_1)^{-1} & \ldots & (\theta_1+\lambda_L)^{-1} \\ \vdots & \ddots & \vdots \\ (\theta_N+\lambda_1)^{-1} & \ldots & (\theta_N+\lambda_L)^{-1} \end{pmatrix} \circ \left(\left(\mathbf{Q}^\top \mathbf{y}\right)\boldsymbol{\iota}^\top\right)\right], \quad (5.11)$$

where $\boldsymbol{\iota}^\top = (1, \ldots, 1)$, and "$\circ$" denotes the element-wise (Hadamard) product. The computation of the eigendecomposition of $\mathbf{A}^*$ has a computational complexity of $\mathcal{O}(N^3)$. Given this decomposition, the prediction consists of $(N_2+3)NL + (L+1)N^2$ simple operations such as multiplication and addition of scalars.

To illustrate the differences in the respective approaches to RR, Figure 5.2 shows the CPU time for (i) the naïve approach in Equation 5.3 which involves solving P unknowns, (ii) the dual formulation in Equation 5.8 which requires solving L systems with N unknowns each, and (iii) the dual formulation in Equation 5.11 solving for all values of λ jointly. These results are obtained by applying the approaches to simulated data, with baseline settings $N = 100$, $N_2 = 10$, $P = 1000$, and $L = 100$, and by varying the levels of the factors N and L, one factor at a time. In order to ensure no approach has an advantage in terms of preprocessing of the data (e.g., constructing $P^{-1}\mathbf{X}\mathbf{X}^\top$ and its eigendecomposition) all reported CPU times include these preprocessing steps.

In the upper panel of Figure 5.2, we see that as the number of SNPs P increases the time required by the naïve approach keeps growing at a fixed rate, whereas the time required by the dual approaches remains unchanged. Moreover, the approach considering all values of λ jointly outperforms the dual approach solving L separate systems. When sample size N is relatively large compared to P the dual formulations

lose their advantage compared to the naïve approach. This is not surprising; when $N > P$ the dual formulation requires solving more unknowns than the naïve approach. Concordantly, when faced with data in which $N \leq P$ one can apply the dual approach, and when $N > P$ one can use the classical approach to RR. The lower panel of Figure 5.2 shows that for a very small set of λ's the dual formulation solving L systems with N unknowns is faster than the formulation solving for all values of λ jointly. However, the CPU time required by the former approach increases continuously with L, whereas the CPU time of the method considering all λ's jointly hardly changes. When $L \geq 10$ the latter method attains a better CPU time than the former method does.

The second method for setting λ is based on the LMM in Equation 5.5. In this model, the optimal hyperparameter is a function of $\sigma^2_{\boldsymbol{\varepsilon}}$ and $\sigma^2_{\boldsymbol{\beta}}$. Therefore, one can estimate the LMM using methods such as (restricted) maximum likelihood (Yang et al., 2011a, Patterson and Thompson, 1971) and take $\lambda = \sigma^2_{\boldsymbol{\varepsilon}}/\sigma^2_{\boldsymbol{\beta}}$.

Finally, one can use an existing heritability estimate of the trait under consideration. Given the following definition of SNP-based heritability

$$h^2_{\text{SNP}} = \frac{P\sigma^2_{\boldsymbol{\beta}}}{P\sigma^2_{\boldsymbol{\beta}} + \sigma^2_{\boldsymbol{\varepsilon}}},$$

provided the SNP data are standardized as Z-scores, it is shown by Hofheinz et al. (2012) that the RR shrinkage parameter λ can be written as a function of the SNP-based heritability. Specifically, simple algebra shows that under the above definition of SNP-based heritability,

$$\lambda = P\left(\frac{1}{h^2_{\text{SNP}}} - 1\right). \tag{5.12}$$

This implies that heritability estimates can be used to set λ (Hofheinz et al., 2012). When using a GRM ($P^{-1}\mathbf{X}\mathbf{X}^\top$) to carry out RR prediction, the corresponding shrinkage parameter $\lambda_{\text{GRM}} = P^{-1}\lambda$. This implies the relation between λ_{GRM} and h^2_{SNP} is given by $\lambda_{\text{GRM}} = \left(h^2_{\text{SNP}}\right)^{-1} - 1$.

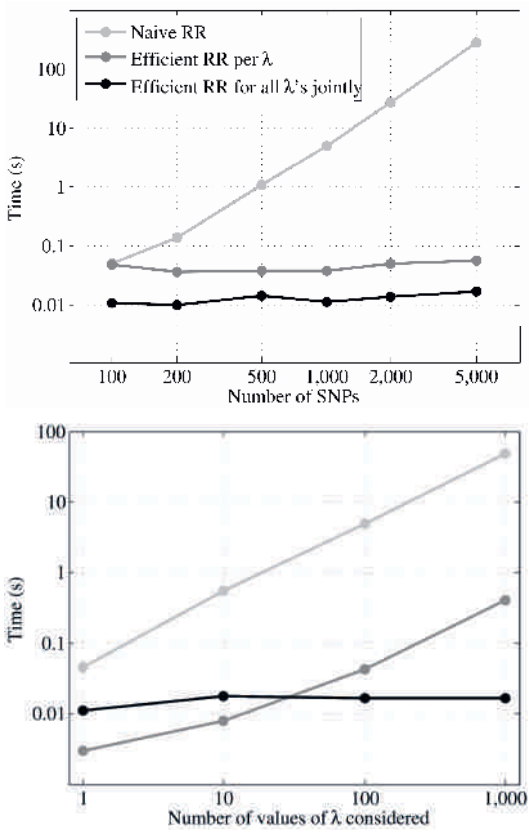


Figure 5.2: *CPU time (seconds) of prediction based on ridge regression using a naïve approach (light gray line), an efficient approach for each λ separately (dark gray line), and an efficient approach considering all values of λ jointly (black line), for various combinations of the number of SNPs in the prediction model (upper panel, with $N = 100$, $N_2 = 10$, $L = 100$) and the number of values of the penalty parameter being considered (lower panel, with $N = 100$, $N_2 = 10$, $P = 1,000$).*

5.8. ADVANCED METHODS

5.8.1. *Heteroskedastic ridge regression*

A point of critique regarding the use of RR is the lack of SNP selection. However, for highly polygenic traits, given current sample sizes, there is evidence that SNP selection is sometimes detrimental to predictive accuracy (e.g., Purcell et al. 2009, Evans et al. 2009, Benner et al. 2010). Nevertheless, since RR can be used for inference just as well as RSR, the approach of selecting SNPs that attain a p-value below some threshold τ in the GWAS, can also be extended to RR.

In a spirit similar to that of SNP selection, one can argue in favor of a *heteroskedastic ridge regression* (HRR), where each SNP receives a different amount of shrinkage (Shen et al., 2013, Hofheinz and Frisch, 2014). As with homoskedastic shrinkage, this SNP-specific shrinkage might either be based on results from the training set or prior information from different data sets. Depending on the size of SNP-specific shrinkage, this method can leverage between SNP selection and full inclusion. Based on prior evidence or in-sample evidence the weight assigned to a SNP can be made arbitrarily small or arbitrarily large given the amount of evidence for association with the outcome. SNP-specific shrinkage opens up the door for a whole array of HRR methods (e.g., Shen et al. 2013, Hofheinz and Frisch 2014).

The HRR estimator in Equation 5.4 and resulting predictions can be rewritten as

$$\widehat{\boldsymbol{\beta}}_{\text{HRR}} = \boldsymbol{\Lambda}^{-1}\mathbf{X}^{\top}\left(\mathbf{X}\boldsymbol{\Lambda}^{-1}\mathbf{X}^{\top} + \lambda\mathbf{I}\right)^{-1}\mathbf{y}, \tag{5.13}$$

$$\widehat{\mathbf{y}}_2 = \mathbf{X}_2\boldsymbol{\Lambda}^{-1}\mathbf{X}^{\top}\left(\mathbf{X}\boldsymbol{\Lambda}^{-1}\mathbf{X}^{\top} + \lambda\mathbf{I}\right)^{-1}\mathbf{y}, \tag{5.14}$$

where $\boldsymbol{\Lambda} = \text{diag}\left(\{\lambda_p\}_{p=1,\ldots,P}\right)$ is a diagonal matrix with SNP specific shrinkage effects.

It is implied by Equations 5.13 and 5.14 that HRR can be carried out using the same machinery as homoskedastic RR, by first weighting the SNPs appropriately. More specifically, take $\mathbf{X}^* = \mathbf{X}\boldsymbol{\Lambda}^{1/2}$ and $\mathbf{X}_2^* = \mathbf{X}_2\boldsymbol{\Lambda}^{1/2}$ and construct corresponding weighted GRMs by taking

$$\mathbf{A}^* = \mathbf{M}_{\mathbf{Z}}\left(\frac{1}{P}\mathbf{X}^*\mathbf{X}^{*\top}\right)\mathbf{M}_{\mathbf{Z}}, \tag{5.15}$$

$$\mathbf{A}_{21}^{*} = \mathbf{M}_{\mathbf{Z}_2}\left(\frac{1}{P}\mathbf{X}_2^{*}\mathbf{X}^{*\top}\right)\mathbf{M}_{\mathbf{Z}}.$$

Now, using the eigendecomposition defined in Equation 5.9 of the weighted GRM defined in Equation 5.15 and by subsequently applying Equation 5.11 to resulting eigenvectors in $\mathbf{Q}$ and eigenvalues, $\{\theta_i\}_{i=1,\ldots,N}$, we obtain efficient out-of-sample HRR predictions.

5.8.2. *Incorporating information from earlier studies*

Using HRR prediction it is possible to include results from a GWAS in other samples as prior information. Consider SNP-specific shrinkage, given by $\lambda_p = \sigma_{\boldsymbol{\epsilon}}^2/\sigma_{\beta_p}^2$, and a set of GWAS t-test statistics from another study without the presence of confounding variables. Given that $\widehat{\sigma}_{\boldsymbol{\epsilon}}$ is approximately constant over the SNPs in the GWAS, the t-test statistic of SNP p can be written as

$$t_p \approx \frac{1}{\widehat{\sigma}_{\boldsymbol{\epsilon}}}\left(\frac{\mathbf{x}_p^{\top}}{\sqrt{\mathbf{x}_p^{\top}\mathbf{x}_p}}\right)\mathbf{y} = \frac{1}{\widehat{\sigma}_{\boldsymbol{\epsilon}}}\mathbf{x}_p^{*\top}\mathbf{y} = \frac{1}{\widehat{\sigma}_{\boldsymbol{\epsilon}}}\widehat{\beta}_p$$

where $\mathbf{x}_p^{*}$ denotes SNP p standardized to unit length and $\widehat{\beta}_p$ the estimated effect of the standardized SNP. It follows from this equation that these statistics are proportional to the estimated effects of standardized SNPs. Therefore, the square t-test statistics are approximately proportional to the square standardized GWAS estimates. Now, under the prior that $\beta_p \sim \mathcal{N}\left(0, \sigma_{\beta_p}^2\right)$ we have that $\widehat{\beta}_p^2$ is a consistent estimator of $\sigma_{\beta_p}^2$. Correspondingly, the square t-test statistics are proportional to this estimator of the SNP-specific effect variance. Therefore, for a suitable choice of λ a consistent estimator of λ_p is given by $\lambda t_p^{-2} = \sigma_{\boldsymbol{\epsilon}}^2/\widehat{\beta}_p^2$. In the framework of HRR, this entails setting $\widehat{\boldsymbol{\Lambda}} = \text{diag}(\left\{t_p^{-2}\right\}_{p=1,\ldots,P})$. This definition of $\widehat{\boldsymbol{\Lambda}}$ implies that SNPs are weighted according to t_p. From these weighted SNPs we can construct the weighted GRM and apply Equation 5.14 to obtain out-of-sample HRR predictions which incorporate information from a GWAS in another dataset.

5.8.3. *Nonlinear prediction methods*

An important question in genetics is how nonlinear effects (e.g., *dominance* and *epistasis*) contribute to the variation of complex traits. RR

can efficiently implement such nonlinear SNP effects using the *kernel trick* from machine learning. Resulting *kernel ridge regression* (KRR) extends the non-parametric approach to RR, where genetic "similarities" in the space of additive effects are replaced by genetic "similarities" in a larger (potentially infinite) feature space, for instance, including two or three-way interactions.

The efficient RR predictions in Equation 5.7 are in essence a weighted average of the observed phenotypes in the training set. Weights are based on the genetic similarity of individuals in the test set and the training set. The more genetically similar two individuals are in the test and training set, the more weight will be given to the phenotype of the similar individual in the training set.

Classical RR measures genetic similarity of individuals in the space of additive effects and assigns weights accordingly. KRR however, can measure genetic similarity in the space of more than just additive effects. This extended space can include, for instance, d-way interactions between SNPs. Now, a GWAS estimating all potential d-way interactions between SNPs is not feasible. However, with KRR, rather than having to estimate all coefficient of all nonlinear combinations of regressors, one can instead obtain the measure of genetic similarity in this higher-dimensional space by applying a simple kernel function $k\left(\mathbf{x}_i,\mathbf{x}_j\right)$ to any two genotype vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ corresponding to individuals i and j.

In this context, classical RR corresponds to $k\left(\mathbf{x}_i,\mathbf{x}_j\right)=\mathbf{x}_i^\top\mathbf{x}_j$. Similarly, a function measuring similarity in the space consisting only of two-way linear interactions is given by

$$k\left(\mathbf{x}_i,\mathbf{x}_j\right)=\left(\mathbf{x}_i^\top\mathbf{x}_j\right)^2. \tag{5.16}$$

To see why this is so, consider expanding Equation 5.16. We then have

$$\begin{aligned} k\left(\mathbf{x}_i,\mathbf{x}_j\right)&=\left(\sum_{p=1}^{P}x_{ip}x_{jp}\right)^2=\sum_{p=1}^{P}\sum_{q=1}^{P}x_{ip}x_{jp}x_{iq}x_{jq}\\ &=\sum_{p=1}^{P}\sum_{q=1}^{P}(x_{ip}x_{iq})(x_{jp}x_{jq})=\boldsymbol{\phi}(\mathbf{x}_i)^\top\boldsymbol{\phi}(\mathbf{x}_j),\\ \text{where }\boldsymbol{\phi}(\mathbf{x}_i)^\top&=\left(\left\{\left\{x_{ip}x_{iq}\right\}_{q=1,\ldots,P}\right\}_{p=1,\ldots,P}\right). \end{aligned}$$

Thus, $\boldsymbol{\phi}(\mathbf{x}_i)$ is a vector that contains all possible two-way interactions

between the P markers. Kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ represents the genetic similarity of individuals i and j in this space of all two-way interactions between SNPs.

The essence of KRR is the so-called *kernel trick* that allows one to efficiently compute the higher-dimensional similarity measure by applying a simple *kernel function* $k(\mathbf{x}_i, \mathbf{x}_j)$ to any two input vectors for individuals i and j (Aizerman et al., 1964). Provided the kernel is positive definite it constitutes the reproducing kernel of a unique *reproducing kernel Hilbert space* (RKHS; Aronszajn 1950). KRR then is equivalent to a so-called RKHS regression.

In the case of d-way interactions the associated kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ can be evaluated for all pairs of individuals by raising each element of the GRM, $P^{-1}\mathbf{X}\mathbf{X}^\top$, to the power d. An alternative is the nonhomogeneous polynomial kernel of degree d, given by $k(\mathbf{x}_i, \mathbf{x}_j) = (c + \mathbf{x}_i^\top \mathbf{x}_j)^d$. This kernel, similar to the regular polynomial kernel of degree d, includes d-way interactions but also lower-order interaction terms including the "one-way interactions", that is, simple additive linear effects.

The preceding example of the polynomial kernel of degree two shows how KRR can include dominance and epistasis in the prediction model. For frequently used kernels, such as the Gaussian (radial basis function) kernel, there exists a representation in which classical RR is applied to a model with infinitely many predictors, nevertheless yielding finite predictions. Obtaining the weights for infinitely many predictors is not possible. Hence, rather than aiming to obtain point estimates of $\boldsymbol{\beta}$, KRR only aims to obtain predictions.

BLUP and, by extension, RR are special cases of prediction using KRR (e.g., Harville 1983, Speed 1991). There has been a substantial amount of work in plant and animal breeding, aiming to improve predictive accuracy using KRR (e.g., González-Recio et al. 2008, Crossa et al. 2010, Endelman 2011). A generally used kernel is the aforementioned Gaussian kernel, defined as

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left[-d^2(\mathbf{x}_i, \mathbf{x}_j)/\eta\right],$$

where $d^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top (\mathbf{x}_i - \mathbf{x}_j)$ and hyperparameter $\eta > 0$. This type of kernel includes all conceivable linear interactions between the P SNPs and with themselves. Endelman (2011) finds that the Gaussian

kernel outperforms accuracy of RR and a Bayesian approach to LASSO, used to predict wheat and maize traits in samples, typically with about 300 observations and 3000 SNPs. Similarly, using a Bayesian approach, Crossa et al. (2010) find in samples of about 250 observations, with 1100 SNPs, that both the Gaussian kernel and the LASSO outperform predictive accuracy of RR for grain yield and maize flowering traits. However, when comparing the LASSO with the Gaussian KRR, which of two the methods is better depends on the trait. An efficient implementation of KRR based on maximum likelihood, using the Gaussian kernel, is available in the R package `rrBLUP` (Endelman, 2011).

Morota and Gianola (2014) compare a wide range of kernels, such as the exponential (González-Recio et al. 2008, Endelman 2011, Piepho 2009), Matérn, diffusion (e.g., Morota et al. 2013), and t kernel (Tusell et al., 2014), for the purpose of obtaining genomic estimated breeding values (Schaeffer, 2006). Though it is argued that selecting a suitable kernel is the most precarious step (e.g., De Los Campos et al. 2009), current evidence suggests that most considered kernels attain a predictive accuracy similar to that of the Gaussian kernels (Morota and Gianola, 2014). Thus, it appears that the Gaussian KRR is a robust prediction method for quantitative traits, able to handle nonlinear genetic architectures. Moreover, Endelman (2011) finds little evidence supporting the hypothesis that a Gaussian kernel is likely to overfit the data (Piepho, 2009).

Given the current evidence, KRR using an appropriate kernel (e.g., the Gaussian kernel) is a promising prediction technique, especially for traits where epistatic effects and dominance are expected to contribute to trait variation. De Los Campos et al. (2009) suggest an interesting venue for further research on the use of KRR for prediction in quantitative genetics, by combining multiple kernels in a single model, each kernel representing a single variance component (e.g., additive, dominance, or epistasis). For a more detailed treatment of KRR and its uses in quantitative genetics, see Morota and Gianola (2014).

Regarding the computation of predictions using KRR, let $\mathbf{K}$ denote the matrix of similarities in the higher-dimensional feature space in the training set, such that an element of this matrix k_{ij} is given by $k(\mathbf{x}_i, \mathbf{x}_j)$ and let $\mathbf{K}_{21}$ be defined similarly for individuals in the test set versus individuals in the training set.

Now, KRR prediction without confounders is given by

$\widehat{\mathbf{y}}_2 = \mathbf{K}_{21}(\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{y}$ and with confounders by

$$\widehat{\mathbf{y}}_2 = \mathbf{M}_{\mathbf{Z}_2}\mathbf{K}_{21}\mathbf{M}_{\mathbf{Z}}(\mathbf{M}_{\mathbf{Z}}\mathbf{K}\mathbf{M}_{\mathbf{Z}} + \lambda \mathbf{I})^{-1}\mathbf{y},$$

where, as before, $\mathbf{M}_{\mathbf{C}}$ is the projection matrix removing the effects of $\mathbf{C}$.

In the case of the nonhomogeneous polynomial kernel of degree d, given the GRMs, $P^{-1}\mathbf{X}\mathbf{X}^\top$ and $P^{-1}\mathbf{X}_2\mathbf{X}^\top$, the matrices $\mathbf{K}$ and $\mathbf{K}_{21}$ can be obtained efficiently by adding a constant c to each element of the GRMs and by raising each resulting element to the power d. When $c > 0$ and $d \in \{1,2,\ldots\}$ are not fixed, these are additional hyperparameters which can be tuned via (N)CV.

5.9. SIMULATION STUDY

An important question is under what circumstances we can expect RR to yield more accurate predictions than RSR. The answer to this question can help us assess the merits of RR in quantitative genetics. As discussed, prediction using RR is intimately related to the BLUP of the phenotype under an LMM in which SNP effects are assumed to be all drawn from a normal distribution. This corresponds to idea of each SNP making a tiny contribution to phenotype. Therefore, it is reasonable to assume that RR will perform well when the SNP effects are such. However, given that not all SNPs in existence are causally affecting the outcome, an open question is how RR performs when only a subset of SNPs affects the outcome.

Moreover, an important factor influencing predictive accuracy of a classical polygenic score is the training sample size. Therefore, RR is likely also to be very sensitive to the sample size. Finally, the more heritable a trait is the easier it should be to detect the effects of SNPs. Thus, an additional question is how RR and RSR perform under different levels of heritability.

In short, we want to know the relative predictive accuracy of RR and RSR (i) for a wide range of trait architectures and (ii) under particular combinations of sample size and the number of genotyped SNPs. To answer this question we run a suite of simulations. In these analyses, we vary sample size of the training set (N), the number of genotyped SNPs (P), the fraction of SNPs exerting a causal influence (f_C), and the SNP-based heritability (h^2_{SNP}).

Table 5.1: *Settings of the data-generating processes considered in the simulation study of the predictive accuracy of ridge regression and repeated simple regression. N denotes the sample size of the training set, P the number of SNPs, f_C the fraction of SNPs that is causal, and h^2_{SNP} the SNP heritability.*

Factor	**Levels**	**#Levels**
N	$\{200;500;1k;\ldots;10k;20k\}$	7
P	$\{100;200;500;\ldots;100k;200k;500k\}$	12
f_C (%)	$\{0.1;\ldots,100\}$ (linear increases on log scale)	37
h^2_{SNP} (%)	$\{5;10;15;\ldots;100\}$	20

Table 5.1 shows the levels we consider for these factors. In addition, a range of values for λ on the interval $[10^{-6};10^{9}]$ is considered. Each unique combination of levels of these factors constitutes a scenario. The total number of scenarios is $S = 7 \times 12 \times 37 \times 20 = 62,160$. We consider $R = 21$ runs, yielding $S \times R = 1,305,360$ combinations of levels and runs.

For a combination of sample size, the number of SNPs, trait heritability, and a fraction of causal SNPs chosen from the levels shown in Table 5.1, let C be the corresponding number of causal SNPs. Now, the data generating process for this combination of levels is given by

$$
\begin{aligned}
y_i &= \sum_{p=1}^{C} x_{ip}\beta_p + \varepsilon_i, \text{ for } i = 1,\ldots,N_{\text{total}},\\
x_{ip} &= \frac{g_{ip} - 2f_p}{\sqrt{2f_p(1-f_p)}}, \text{ for } i = 1,\ldots,N_{\text{total}} \text{ and } p = 1,\ldots,P,\\
g_{ip} &\sim \binom{2}{f_p}, \text{ for } i = 1,\ldots,N_{\text{total}} \text{ and } p = 1,\ldots,P,\\
f_p &\sim \mathcal{U}(0.05, 0.95), \text{ for } p = 1,\ldots,P,\\
\beta_p &\sim \mathcal{N}\left(0, \sigma^2_{\boldsymbol{\beta}}\right), \text{ for } p = 1,\ldots,P, \text{ and}\\
\varepsilon_i &\sim \mathcal{N}\left(0, \sigma^2_{\boldsymbol{\varepsilon}}\right), \text{ for } i = 1,\ldots,N_{\text{total}},
\end{aligned}
$$

where $\binom{a}{b}$ denotes the binomial distribution with a draws each with probability of success b, and $\mathcal{U}(a,b)$ denotes the uniform distribution

on the interval (a, b). This data generating process corresponds to a quantitative trait which is normally distributed and has only additive genetic variation to which common variants contribute (i.e., minor allele frequency above 5%).

The total number of observations N_{total} includes the individuals in the test set. The size of the test set is 10% of the size of the training set. Hence, yielding $N_{\text{total}} = \lfloor 1.1N \rfloor$. Here, $\lfloor x \rfloor$ denotes the nearest smaller integer.

In order not to be dependent on a single generated dataset, the entire simulation consists of $R = 21$ independent runs (replications). In each run we simulate only one set of genotype data for $N_{max} = 22$k individuals and $P_{max} = 500$k SNPs. Given any combination of N and P listed in Table 5.1 we can take an appropriate submatrix of the genotype matrix. To this submatrix we apply a set of P weights of which some are zero, such that we attain the desired fraction of SNPs being causal. Moreover, by scaling these weights and the noise vector $\boldsymbol{\epsilon}$ appropriately we can attain any specified heritability. The result is a four-dimensional phenotype array with individuals along the first dimension and the factors P, f_C, and h^2 along the remaining dimensions.

When computing the out-of-sample predictions based on RR and RSR, the available genotype matrix only depends on N and P, not on h^2 nor on f_C. Therefore, given N and P, when $N \leq P$ the eigendecomposition of the $N \times N$ GRM, $P^{-1}\mathbf{X}\mathbf{X}^\top$, can be reused for all combinations of h^2 and f_C. Moreover, the approach has already been amended to reuse the eigendecomposition for different values of λ. Similarly, when $N > P$ the eigendecomposition of $P \times P$ matrix $P^{-1}\mathbf{X}^\top\mathbf{X}$ can be reused. Since there only are 7 unique levels of N and 12 unique levels of P, RR prediction for the 62,160 scenarios per replication reduces to (i) computing $7 \times 12 = 84$ eigendecompositions and (ii) for each scenario carrying out the matrix multiplications seen in Equation 5.11.

In a typical run it takes 4.5 hours to predict using RR on a machine with 16 cores at 2.60 GHz per core with 64GB RAM. The RSR predictions are generated alongside at virtually no costs in terms of CPU and memory. The computing time includes computation of the GRM, $P^{-1}\mathbf{X}\mathbf{X}^\top$, when $N \leq P$ and $P^{-1}\mathbf{X}^\top\mathbf{X}$ when $N > P$. Given N and P, failure to exploit (i) the constancy of the GRM and of $P^{-1}\mathbf{X}^\top\mathbf{X}$ over the 20×37 different combinations of h^2 and f_C and (ii) the properties of the eigendecomposition which enable the joint evaluation of the 151

values of λ we consider, dramatically increases the CPU time of RR. In fact, we infer that the less efficient approach yields a CPU time that is at most a factor $20 \times 37 \times 151 = 111,740$ larger than the 4.5 hours we attain (i.e., about 57 years per run). Even worse, when the naïve RR approach is applied, also when $P \gg N$, RR predictions cannot be obtained for datasets with more than 50k SNPs on the machine we use. Thus, using the efficient approach based on the GRM when $N \leq P$ and based on $P^{-1}\mathbf{X}^\top\mathbf{X}$ when $N > P$, combined with the smart use of eigen-decompositions and constancy of GRMs over different combinations of f_C and h^2 we are able to reduce CPU times from several decades to several hours.

In each run, for each combination of levels we compute the R^2 of the RSR prediction with the outcome and the R^2 of the RR prediction with the outcome. R^2 is measured by the squared sample correlation coefficient between the polygenic score and the outcome in the test set. Our aim is to assess predictive accuracy of RSR and see whether it differs significantly from zero for a wide range on configurations. Moreover, we want to test whether RR provides a significant improvement compared to RSR. Therefore, the performance of RR is measured relative to RSR. That is, we take the log-ratio of the two, given by $\log\left(R^2_{\text{RR}}/R^2_{\text{RSR}}\right)$. This measure is continuously distributed over $(-\infty, +\infty)$.

We measure the absolute performance of RSR by the logit transformation of $R^2_{\text{RSR}}/h^2_{\text{SNP}}$, that is,

$$\text{logit}\left(\frac{R^2_{\text{RSR}}}{h^2_{\text{SNP}}}\right) = \log\left(\frac{\frac{R^2_{\text{RSR}}}{h^2_{\text{SNP}}}}{1 - \frac{R^2_{\text{RSR}}}{h^2_{\text{SNP}}}}\right).$$

This measure is also distributed over $(-\infty, +\infty)$. The reason for dividing R^2_{RSR} by h^2_{SNP} is that we want to know what part of the genetic variation the polygenic score captures. If h^2_{SNP} is low, for instance 5%, we consider a polygenic score that attains an R^2 of 4% to be more impressive than a risk score that explains 10% of the variation in a highly heritable trait (e.g., $h^2_{\text{SNP}} = 50\%$). Note that we exclude observations with $R^2_{\text{RSR}} > h^2_{\text{SNP}}$ as these are aberrant observations; a polygenic score that "explains" more genetic variation than there actually is, is simply wrong.

Regarding the RR penalty, let $R^2_{\text{RR}}(\lambda, r)$ denote the accuracy of RR

in run r, given penalty λ, conditional on some N, P, f_C, and h^2. Now, let

$$R^2_{\text{RR,med}}(\lambda) = \text{median}\left(\left\{R^2_{\text{RR}}(\lambda, r)\right\}_{r=1,\ldots,R}\right),$$

denote the median of the RR performance over the runs for a specific value of λ. Now, for this combination of N, P, f_C, and h^2 we take

$$\widehat{\lambda} = \underset{\lambda \in \{\lambda_1,\ldots,\lambda_L\}}{\arg\max}\, R^2_{\text{RR,med}}(\lambda).$$

Thus, for a given combination of levels of factors λ is tuned by setting it such that it maximizes the median R^2 of RR over the runs for the given combination of levels. Based on this procedure, the optimal R^2 of RR in run r is given $R^2_{\text{RR}}(\widehat{\lambda}, r)$. This yields a single measure of accuracy of RR per replication and per combination of levels. This procedure results in a value of λ that performs well in 21 independent samples. Hence, it is similar to a value that would result from CV; there is little scope for overfitting. Moreover, since the median is less sensitive to outliers than, for instance, the mean, we make our measure more robust by taking the median over the runs. The reason that we choose for this approach instead of CV is to reduce the computational complexity of the simulation procedure at the expense of having a slightly less elegant approach.

5.9.1. *Simulation results*

Table 5.2 shows the summary statistics of the measure $\log\left(R^2_{\text{RR}}/R^2_{\text{RSR}}\right)$ and of R^2_{RSR}. As can be seen, over all the combinations of levels and runs RR seems to outperform RSR on average by about 6%. However, there is much variation in the log-ratio. The lowest log-ratio is -22.2 and the highest $+20.7$. Since this ratio is on a log scale this implies a tremendous difference in R^2. The reason for this is that when either the nominator or denominator of $R^2_{\text{RR}}/R^2_{\text{RSR}}$ gets close to zero, the log-ratio can attain a large value (in absolute terms). For this reason we excluded log-ratios outside the interval $(-1, +1)$. This leads to a drop in the variance from about 0.4 to 0.04, only at the expense of losing 3.9% of the observed combinations of levels and runs. Moreover, the mean log-ratio hardly changes by removing the outliers. This reduction in variance allows us to display the results in a more insightful manner

Table 5.2: *Summary statistics of (1) the predictive accuracy of ridge regression (RR) compared to repeated simple regression (RSR), measured by* $\log\left(R^2_{RR}/R^2_{RSR}\right)$*, across simulations and simulation designs, for the full set of observed log-ratios and for the subset excluding log-ratios outside* $(-1,+1)$*, and (2) the predictive accuracy of RSR, measured by* R^2_{RSR}*, across simulations and simulation designs, for the full set and for the subset excluding observations for which* $R^2_{RSR} \geq h^2_{SNP}$*.*

Outcome	**Count**	**(% total)**	**Mean**	**Var**	**Min**	**Max**
$\log\left(\frac{R^2_{\text{RR}}}{R^2_{\text{RSR}}}\right)$	1,305,360	(100.0%)	0.065	0.403	-22.2	20.7
$\log\left(\frac{R^2_{\text{RR}}}{R^2_{\text{RSR}}}\right) \in (-1,+1)$	1,254,168	(96.1%)	0.060	0.041	-1.00	1.00
R^2_{RSR}	1,305,360	(100.0%)	0.177	0.058	0.000	0.997
$R^2_{\text{RSR}} < h^2_{\text{SNP}}$	1,239,721	(95.0%)	0.160	0.051	0.000	0.997

and ensures further inferences about the relation between our factors (e.g., sample size) and predictive accuracy are not influenced by aberrant observations. For R^2_{RSR} we see that the average R^2 of about 17% is significantly greater than zero.

The upper panel in Figure 5.3 shows the histogram of $\log\left(R^2_{\text{RR}}/R^2_{\text{RSR}}\right)$ over the combinations of runs and levels inside the range $(-1,+1)$. This histogram confirms that there are long and thin tails. Most mass centers around zero. However, the empirical distribution is slightly skewed to the right, giving rise to the positive average log-ratio. The figure shows that RR often performs better than RSR. Given the fact that RR lies between RSR and OLS, this is not surprising. Using the penalty parameter λ, RR tries to find the optimum between these two extremes. The lower panel in Figure 5.3 shows the histogram of $\text{logit}\left(R^2_{\text{RR}}/h^2_{\text{SNP}}\right)$ excluding observations for which $R^2_{\text{RR}} > h^2_{\text{SNP}}$. The observations are smoothly distributed. A value of zero, corresponds to an R^2 equal to half the heritability. Thus, in a substantial proportion of the cases RSR captures more than half of the genetic variation.

Figure 5.4 shows the log-ratio of the median R^2 of ridge regression and of RSR, with values outside the interval $(-1,+1)$ truncated to corresponding extremes of this interval. This truncation is necessary in order for the figure not to be dominated by the outliers. For $N \ll P$ (see the lower right block in Figure 5.4), the performance of RR and RSR is volatile. The more so, since we consider the median over the

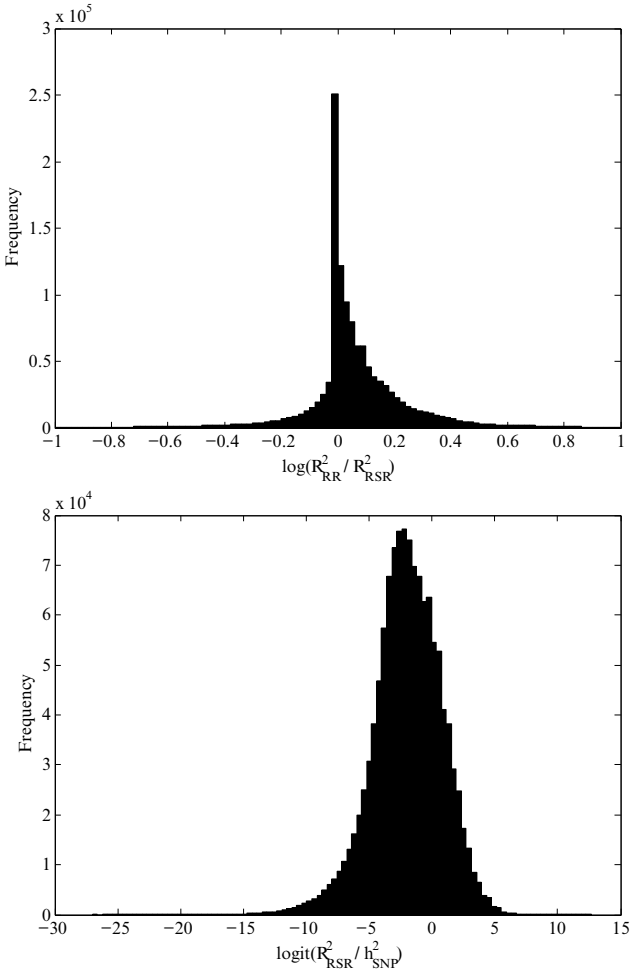


Figure 5.3: *Histograms of the relative predictive accuracy of ridge regression (RR) and repeated simple regression (RSR), across 21 runs of simulations for different combinations of sample size, number of SNPs, fraction of SNPs causal, and SNP heritability, specified in Table 5.1. Upper panel: the logarithm of the ratio of the R^2 of RR and the R^2 of RSR, where RR penalty parameter (λ) is chosen to maximize median R^2 of RR and where values outside the interval $(-1, +1)$ are excluded. Lower panel: the logit transformation of the ratio of the R^2 of RSR and the SNP heritability, where observations for which the R^2 of RSR exceeds the SNP heritability are excluded.*

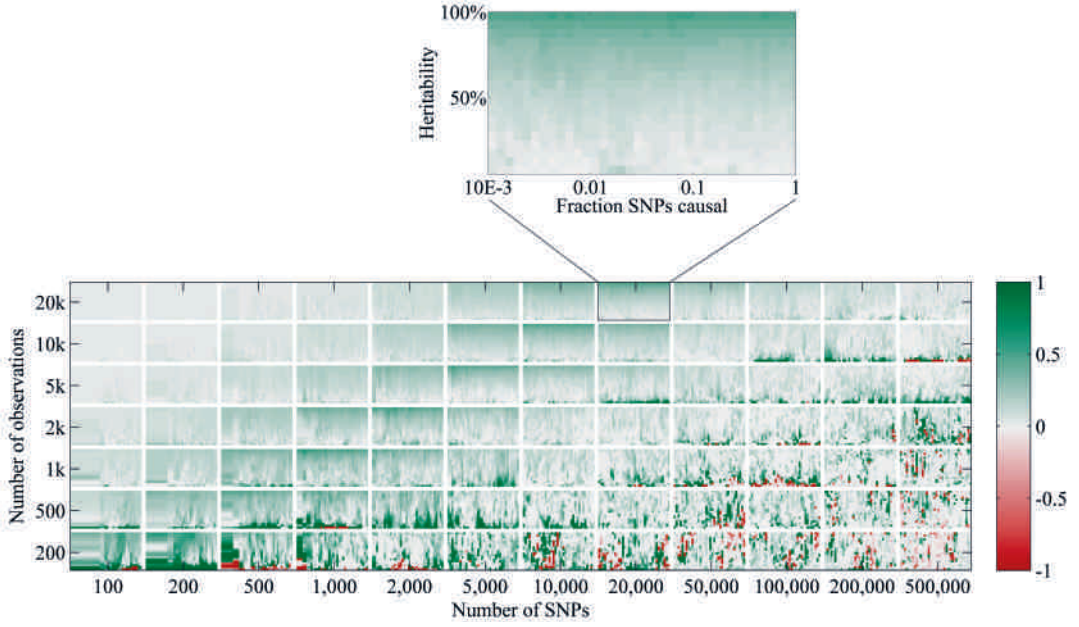


Figure 5.4: *Heat maps of the ratio of the median of the R^2 over 21 simulations, attained by ridge regression (RR) and by the classical GWAS approach of repeated simple regression (RSR), for various combinations of training sample size (y-axis across heat maps), the number of SNPs (x-axis across heat maps), the fraction of SNPs that are causal (x-axis within each heat map), and the SNP heritability (y-axis within each heat map). The RR penalty parameter (λ) is chosen to maximize the median R^2 of RR. Values are truncated to lie between minus one (red) and plus one (green).*

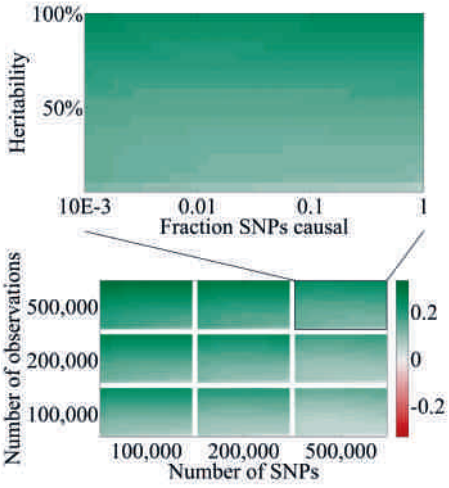


Figure 5.5: *Heat maps of the predicted logarithm of the ratio of the R^2 attained by ridge regression (RR) and repeated simple regression (RSR), for various combinations of training sample size (y-axis across heat maps), the number of SNPs (x-axis across heat maps), the fraction of SNPs that are causal (x-axis within each heat map), and the SNP heritability (y-axis within each heat map). Predictions are based on a model fitted to simulation results for this measure.*

runs here. Sometimes, RR strongly outperforms RSR and sometimes it is the other way round. However, on average RR seems to outperform RSR. As N approaches P (see the lower left and upper right blocks in Figure 5.4) RR starts to outperform RSR. There are large regions, where the log of the gain in accuracy is consistently between zero and a half. This corresponds to a relative increases between zero and 65%. For example, for $N = P = 20$k, $h^2_{\text{SNP}} = 50\%$, and 200 causal SNPs RSR attain a median R^2 of 17% and RR 20%, constituting a relative increase of 16%. This gain in accuracy peaks when $N \approx P$.

When $N \gg P$ (see the upper left block in Figure 5.4), the gain in accuracy drops to zero. However, it is unlikely that this pattern, where the gain of RR dies out as N keeps increasing, replicates empirically. The reason for this is that the patterns is probably an artefact of the design of the simulation; all SNPs are simulated independent from each other. Even though empirical correlations between SNPs can arise in the simulations, asymptotically there are none. Thus, for sufficiently

Table 5.3: *The median R^2 of repeated simple regression (RSR) and the median R^2 of ridge regression (RR) relative to the median R^2 of RSR across simulations, for various combinations of sample size (N), number of SNPs (P), and SNP heritability (h^2_{SNP}), where 1% of the SNPs is causal in each simulation. The RR penalty parameter (λ) is chosen to maximize the median R^2 of RR.*

N	P	h^2_{SNP}	median R^2_{RSR}	$\frac{\text{median } R^2_{\text{RR}}}{\text{median } R^2_{\text{RSR}}}$
5k	500k	0.50	0.003	1.078
10k	500k	0.50	0.005	1.029
20k	500k	0.50	0.009	1.038
10k	100k	0.50	0.027	1.079
10k	200k	0.50	0.011	1.000
10k	500k	0.50	0.005	1.029
10k	500k	0.25	0.001	1.000
10k	500k	0.50	0.005	1.029
10k	500k	0.75	0.011	1.011

large N (compared to P) the standardized simulated SNP data are such that $\mathbf{X}^\top\mathbf{X}$ approaches the identity matrix and RR becomes equivalent to RSR (see Section 5.3). Therefore, the accuracy of RR and RSR does not differ for such extremely large values of N. How the performance differs in these large samples when there is linkage disequilibrium in the data remains to be seen.

Table 5.3 shows the median of the R^2 of RSR and that of RR relative to RSR for combinations of sample size and the number of genotyped SNPs that are typically seen in a GWAS (e.g., $N = 10$k, $P = 500$k). We see that for these data dimensions a trait with a heritability of 50% has a classical polygenic score which on average only explains 0.5% of the total phenotypic variation. Moreover, RR yields a relative increase of just 2.9%. This increase gives an absolute R^2 of 0.51% for RR. This observations clearly illustrates that the so-called missing heritability (Manolio et al., 2009) is hard to find, even under a very simple data generating process, that is, a process for which we are sure that both RSR and RR should asymptotically capture all genetic variation.

5.9.2. *Modelling the simulation results*

To understand the relation between the various factors in the simulation study and the gain in predictive accuracy by RR we fit a linear model to the logarithm of the ratio $R^2_{\text{RR}}/R^2_{\text{RSR}}$ for all replications and for all considered levels of factors, such as sample size. Moreover, in order to obtain the R^2 of RSR as a benchmark we also fit a linear model to the logit transformation of R^2_{RSR} relative to the SNP heritability.

The results in the previous section indicate that the relation between sample size N and the performance is nonlinear. The relation seems to exhibit an inverted U-shape. For this purpose, we include $\log(N)$ and its square as regressors. Moreover, the location of the peak depends on the number of SNPs, implying that the parameters of regressors related to sample size depend on P. Consequently, interactions between P and N are added to the model. By symmetry of Figure 5.4, similar arguments hold for the performance as function of P. Based on this argument we consider up to three-way interactions between the regressors.

In addition, we see in many subplots of Figure 5.4 that the gain in predictive accuracy differs systematically between low, intermediate, and high heritabilities. Therefore, heritability is included as regressor. Finally, although the effect of the fraction of causal SNPs is hard to judge from Figure 5.4, we include this factor as regressor as well.

Both outcomes are modelled as a linear function of the aforementioned basic regressors. These regressors are reported in Table 5.4. We consider models ranging from merely an intercept, up to all 3-way interactions between the explanatory variables. We choose the model that minimizes the Bayesian information criterion (BIC; Schwarz 1978).

Table 5.5 reports the BIC values of the respective models. On the basis of these values we find that a model including all three-way interactions is most appropriate, both in case of the log-ratio as well as in case of the logit of the performance of RSR relative to the heritability. The model for the gain in accuracy of RR relative to RSR can explain approximately 12% of the variation in this measure on the basis of sample size and the other regressors. The model for the accuracy of RSR can explain about 61%.

A likely reason for the fact that we can explain far more variation in the R^2 of RSR than in the gain of RR relative to RSR is the following. In case both the R^2 of RR and RSR are to a large extent influenced by

Table 5.4: *Regressors used to explain the predictive accuracy of ridge regression and repeated simple regression in the simulation study.*

Regressor	**Captures**
$\log(N)$	Effect sample size
$\log(P)$	Effect number of SNPs
$\log(C)$	Effect of number of causal SNPs (C)
$\log(f_C)$	Effect of fraction of SNPs causal
$\log\left(h^2_{\text{SNP}}\right)$	Effect of SNP heritability

Table 5.5: *Fit of models explaining the predictive accuracy of ridge regression (RR) relative to repeated simple regression (RSR) and of RSR relative to SNP heritability across simulations and across simulation settings shown in Table 5.1. Measures of fit: Bayesian information criterion (BIC; reported in millions) and the proportion of variance explained* $\left(R^2_{model}\right)$*. Lowest BIC printed bold.*

Outcome	**Regressors (# regressors)**		**# obs.**	$\boldsymbol{R^2_{\text{model}}}$	**BIC/10^6**
Logarithm of the ratio of R^2 of RR and RSR					
$\log\left(\frac{R^2_{\text{RR}}}{R^2_{\text{RSR}}}\right)$	Intercept	(1)	1,254,168	0.0%	−3.998
$\log\left(\frac{R^2_{\text{RR}}}{R^2_{\text{RSR}}}\right)$	+ regressors Table 5.4	(+5)	1,254,168	4.7%	−4.058
$\log\left(\frac{R^2_{\text{RR}}}{R^2_{\text{RSR}}}\right)$	+ 2-way interactions	(+15)	1,254,168	8.4%	−4.107
$\log\left(\frac{R^2_{\text{RR}}}{R^2_{\text{RSR}}}\right)$	+ 3-way interactions	(+35)	1,254,168	12.4%	**−4.163**
Logit transformation of the ratio of R^2 of RSR and SNP heritability					
$\text{logit}\left(\frac{R^2_{\text{RSR}}}{h^2_{\text{SNP}}}\right)$	Intercept	(1)	1,239,721	0.0%	2.542
$\text{logit}\left(\frac{R^2_{\text{RSR}}}{h^2_{\text{SNP}}}\right)$	+ regressors Table 5.4	(+5)	1,239,721	48.6%	1.717
$\text{logit}\left(\frac{R^2_{\text{RSR}}}{h^2_{\text{SNP}}}\right)$	+ 2-way interactions	(+15)	1,239,721	56.3%	1.515
$\text{logit}\left(\frac{R^2_{\text{RSR}}}{h^2_{\text{SNP}}}\right)$	+ 3-way interactions	(+35)	1,239,721	60.9%	**1.379**

Table 5.6: *Predictions of the predictive accuracy of repeated simple regression (RSR) and of the gain in predictive accuracy of repeated simple regression (RR) compared to RSR in large-scale samples (e.g., $N \geq 100k$). Predictions are based on a model fitted to predictive accuracy results from the simulations, with sample size (N), the number of SNPs (P), the fraction of SNPs that is causal, and SNP heritability (h^2_{SNP}) as predictors. 95% confidence intervals (CI) are reported in parentheses, with the middle value indicating the point estimate. In the predictions 1% of the SNPs are assumed to be causal.*

N	P	h^2_{SNP}	95% CI R^2_{RSR}	95% CI $R^2_{\text{RR}}/R^2_{\text{RSR}}$
100k	500k	0.50	(0.139; 0.146; 0.153)	(1.062; 1.070; 1.079)
200k	500k	0.50	(0.324; 0.337; 0.349)	(1.094; 1.107; 1.121)
500k	500k	0.50	(0.473; 0.478; 0.482)	(1.142; 1.167; 1.193)
500k	100k	0.50	(0.486; 0.488; 0.490)	(1.218; 1.244; 1.270)
500k	200k	0.50	(0.482; 0.485; 0.488)	(1.193; 1.218; 1.244)
500k	500k	0.50	(0.473; 0.478; 0.482)	(1.142; 1.167; 1.193)
500k	500k	0.25	(0.205; 0.212; 0.218)	(1.110; 1.135; 1.160)
500k	500k	0.50	(0.473; 0.478; 0.482)	(1.142; 1.167; 1.193)
500k	500k	0.75	(0.733; 0.736; 0.739)	(1.191; 1.218; 1.245)

our factors in a similar way, taking the log-ratio basically eliminates these common effects. What then remains is a measure over which the factors have less predictive power than over the absolute R^2 measure.

Using the parameters estimates of the models we predict the log-ratio of R^2_{RR} and R^2_{RSR} as well as R^2_{RSR} for sample sizes between 100k and 500k individuals and the number of SNPs between 100k and 500k. For heritability and the fraction of causal SNPs we use the ranges considered in the initial simulations. The resulting predictions of the gain in accuracy are displayed in the heatmap in Figure 5.5.

In addition, point estimates of $R^2_{\text{RR}}/R^2_{\text{RSR}}$ and R^2_{RSR} are reported together with confidence intervals in Table 5.6. There are three groups of predictions. In the first group $P = 500\text{k}$, $h^2 = 50\%$, and N varies from 100k to 500k. In the second group $N = 500\text{k}$ and P varies from 100k to 500k. In the last group $P = N = 500\text{k}$ and h^2 ranges from 25 to 75%.

Results from Figure 5.5 and Table 5.6 indicate that in most cases RR is expected to yield a relative increases in R^2 between 10% and 20% for sample sizes ranging between 100k and 500k individuals. All increases in accuracy are greater than zero at a 5% significance level. Moreover, RSR attains values of R^2 ranging between 15% and 75%. As example, in case of 200k individuals and 500k SNPs, for a trait with

$h^2_{\text{SNP}} = 50\%$ the R^2 of RSR is expected to be 33.7% and the R^2 of RR 37.3%.

Regarding these findings, combining the R^2 attained by RSR with the relative increase by RR yields expected values of the R^2 of RR which in some cases surpass h^2. In practice this cannot be true. In case a trait has an h^2 of 50% it is not possible to consistently predict more than 50% of the phenotypic variation on the basis of SNP data. This seems to indicate that our estimates are somewhat optimistic. Nevertheless, for the ranges in which we actually simulated data (i.e., $N \leq 20\text{k}$ and $P \leq 500\text{k}$) RSR is able to attain a substantial R^2 when $N \approx P$ and RR is able to considerably increase the R^2. For instance, at $h^2 = 50\%$ and $N = P = 20\text{k}$, with 200 causal SNPs the median R^2 of RSR is 17%, and the median R^2 of RR is 20%. This constitutes a relative increase in R^2 of about 16%. As shown in Figure 5.4, this pattern seems to persist while $N \approx P$. Hence, at the very least, the expectation that RR improves the R^2 of RSR considerably for large samples (e.g., $N \approx P \approx 500\text{k}$) is not unreasonable.

5.10. CONCLUSIONS AND DISCUSSION

Ridge regression is a flexible technique that can be used to estimate the association between a set of P SNPs and an outcome observed for N individuals, even when $P \gg N$. When the ridge penalty is equal to the ratio of the noise variance and the variance of random SNP effects in an LMM, prediction using the weights from ridge regression is equivalent to the best linear unbiased prediction used in animal breeding, agricultural science, and more recently also human genetics.

Ridge regression can be perceived as method that partially accounts for linkage disequilibrium between markers. For a sufficiently low penalty the method fully accounts for linkage disequilibrium and is therefore equivalent to the OLS estimator of the multiple regression problem using all SNPs jointly. On the other hand, for a sufficiently high penalty, in terms of predictions ridge regression ignores linkage disequilibrium and is therefore equivalent to the approach of a simple regression per SNP, which is common in a GWAS.

Using standard results from, for instance, machine learning and animal breeding, prediction using ridge regression can be shown to constitute solving an equation with N unknown weights and applying

these weights to a measure of relatedness of individuals out-of-sample and in-sample. Formulating ridge regression this way makes it a computationally efficient technique, even for a large number of SNPs.

As with multiple regression and GWAS predictions, ridge regression can account for the presence of confounding variables, such as age, gender, and population structure. Moreover, such corrections can again be implemented at low computational costs.

When the shrinkage parameter is unknown ridge prediction can be formulated such that predictions for different values of this parameter can be generated in a single step, requiring the eigendecomposition of an $N \times N$ matrix only once. This expression allows the researcher to efficiently carry out procedures, such as cross-validation, to tune this parameter.

Finally, ridge regression prediction is amenable to a wide array of advanced techniques. First, using the kernel trick from machine learning, nonlinear effects such as dominance and epistasis can easily be incorporated in the prediction model. Moreover, in a Bayesian spirit, results from earlier studies can be used to give a prior weight to SNPs in the ridge regression prediction. Similarly, when prior information is not available, in-sample information can be used to discount SNPs differently, yielding a heteroskedastic ridge regression prediction.

Empirical findings so far seem to suggest that for current sample sizes the performance of plain vanilla ridge regression is very similar to that of the repeated simple regression approach used in a GWAS. This raises two questions. First, how do more advanced ridge regression approaches perform? Second, how will the plain version of ridge regression perform in upcoming large scale initiatives, such as biobanks?

Using a suite of simulations we consider the second question. We confirm the finding that for most current studies, with sample sizes usually below 10k individuals and more than 500k SNPs, ridge regression hardly outperforms the classical GWAS approach. For a sample of 10k observations, with 500k SNPs of which 5k causal, for a trait with a heritability of 50%, the median R^2 in 21 independently simulated datasets is 0.5% for repeated simple regression and 0.51% for ridge regression. This resonates with the finding that the main determinant of predictive accuracy of the polygenic score is the sample size of the training set (e.g., Dudbridge 2013, Warren et al. 2014). As long as $N \ll P$, there seems to be little advantage of advanced approaches,

such as ridge regression, over the classical GWAS approach (Warren et al., 2014).

However, by analyzing the difference in accuracy of the classical approach and ridge regression for different values of N, P, trait heritability, and the fraction of causal variants, we are able to extrapolate the performance of ridge regression for large scale initiatives. For a sample size of 200k individuals and 500k SNPs, we find that in a trait with 50% heritability and with 5k causal variants the polygenic score of a GWAS is expected to explain 34% of the phenotypic variation, whereas ridge regression is expected to capture about 37%. Thus, in this scenario ridge regression is expected to capture about 75% of the genetic variation, whereas the classical approach captures 67%.

However, these predictions are rather coarse. They depend highly on the model being fitted (e.g., by including interactions between the number of individuals, SNPs, heritability, etc.). This observation comes as no surprise; we extrapolate quite a bit outside the interior of the levels of the factors that were considered in the simulations (e.g., $N \leq 20$k). However, one thing that remains unchanged even under different specifications of the models that try to explain the accuracy of respective methods, is that ridge regression outperforms the repeated simple regression approach in all large scale samples considered.

A final note concerns the independence of the loci. In the present simulations at most 500k truly independent markers were used. As a result, all carry their own idiosyncratic bit of information about the genetic relationship of individuals in the data. As is shown, however, by Yang et al. (2010), in real data with linkage disequilibrium taking a random subset of 60% or more of the SNPs from a grand set of 295k SNPs yields heritability estimates of human height highly similar to estimates based on the full set; apparently adding more markers hardly changes the genetic relatedness estimates.

The findings of Yang et al. (2010) illustrate that there might be a limited number of SNPs that can make a meaningful contribution to the SNP-based measure of genetic relationship. After this 'effective number of SNPs' (Dudbridge and Gusnanto, 2008), new SNPs are primarily repeating the story that has been told by previous SNPs already. Therefore, even with many millions of SNPs (e.g., in imputed data), the resulting genetic relatedness estimates are highly similar to those obtained from a considerably smaller set of SNPs. Consequently,

if this 'effective number of SNPs' exists this implies that for large scale initiatives the performance of ridge regression relative to repeated simple regression might be similar to what we have observed in our simulations when $N \approx P$, even when in fact P is far greater still than N. Such a proposition would need to be tested either in empirical work or by means of simulations using actual genotype data in which linkage disequilibrium is present.

The use of GWAS data for the prediction of complex traits based on sample sizes far below 100k individuals yields genetic risk scores with little predictive accuracy, regardless of whether one applies the classical GWAS approach or ridge regression. However, as sample sizes approach the 'effective number of SNPs' we expect the polygenic risk score based on repeated simple regression to be able to explain a substantial proportion of the normal genetic variation. Moreover, under this scenario prediction using ridge regression is likely to outperform the classical GWAS predictions significantly. Bearing in mind that ridge regression is amenable to include non-additive genetic variance in the prediction model it is therefore not unlikely that ridge regression will make an even more substantial contribution to the accuracy of polygenic scores in traits where epistasis and dominance are expected to play an important role.

6

Multivariate Average-Information Constrained GREML

Work in progress by De Vlaming and Groenen

ABSTRACT

Genomic-relatedness-matrix restricted maximum-likelihood (GREML) estimation in a univariate linear mixed model (LMM) is often used to estimate SNP heritability, whereas GREML estimation in a bivariate LMM is frequently used for estimating the genetic correlation between traits. A natural extension of such bivariate GREML methods is a multivariate approach. However, instead of estimating multivariate LMMs, researchers often estimate pairwise bivariate LMMs for each combination of two phenotypes in a set of phenotypes. Moreover, even though a multivariate GREML approach is reported in the literature, little attention has yet been paid to (i) the statistical efficiency of joint estimation compared to pairwise bivariate estimation, (ii) the computational efficiency of multivariate GREML, and (iii) ensuring that the covariance matrices are always positive definite. Therefore, we present a multivariate average-information constrained GREML (MacGREML) estimation method. This method consists of an iterative procedure, based on a Newton-Raphson algorithm, to obtain unbiased estimates of the parameters of a multivariate SNP-based LMM for balanced data on P phenotypes observed for N individuals. We parametrize the LMM such that the $(NP) \times (NP)$ covariance matrices are positive (semi)-definite, irrespective of starting values and updates of the estimates throughout the iterations. We rewrite the log-likelihood, the gradient, and the average-information matrix in terms of the eigendecomposition of an $N \times N$ genomic-relatedness matrix and transformations of $P \times P$ matrices of parameters. In doing so, we are able to reduce the computational complexity of MacGREML estimation from the order $(NP)^3$ to an order of N^3. Therefore, the MacGREML estimation method we propose can – in theory – be applied to a large set of phenotypes, provided the data are balanced. Our parametrization is such that we can impose two basic factor restrictions: (i) such that all traits have a perfect genetic correlation and (ii) the traits have no genetic correlation. The significance of the additional fit of the saturated model we employ, compared to the two restricted versions of the model, can easily be tested using a likelihood ratio test. In terms of further research, analyses using simulations and real data are needed in order to assess the empirical merits of this method.

6.1. INTRODUCTION

Tools like genome-wide complex trait analysis (GCTA; Yang et al. 2011a) can be used to estimate the proportion of trait variation explained by SNPs (e.g., Yang et al. 2010). GCTA employs restricted maximum likelihood (REML) to estimate a linear mixed model (LMM) in which the effects of standardized SNPs are assumed to be independent draws from a normal distribution with mean zero and a homoskedastic variance. This approach is often referred to as univariate genomic-relatedness-matrix (GRM) REML or – in short – GREML estimation.

Similarly, based on a bivariate LMM, one can employ bivariate GREML estimation (Lee et al., 2012) to estimate the genetic covariance between two traits and – if the underlying samples of the traits overlap sufficiently – the environment covariance between traits. The bivariate GREML approach has been successful in identifying pleiotropy between traits (e.g., Wray et al. 2013a, Tropf et al. 2015) but also within traits across sexes and populations (Yang et al., 2015b, De Vlaming et al., 2017).

A natural extension of such bivariate GREML methods is a multivariate approach. Although REML has been used to estimate multivariate animal models (Meyer, 1985, 1991) and, in addition, multivariate structural models have also been widely used in the twin-study literature (e.g., using Mx software; Neale et al. 1994), estimation of multi-trait genetic and environment covariance matrices using SNP data have, so far, been used only sparingly. Interestingly, the work of Maier et al. (2015) does provide a multivariate GREML approach, but uses it primarily to improve accuracy of the best linear unbiased prediction (BLUP) of breeding values.

Little attention has yet been paid to the statistical efficiency of multivariate GREML in estimating variance and covariance components, compared to pairwise bivariate GREML. Instead of estimating multivariate LMMs, researchers often estimate pairwise bivariate LMMs for each combination of two phenotypes in a set of P phenotypes. Separate estimation of related quantities is often detrimental to statistical efficiency (e.g., Zellner and Huang 1962).

In addition, to the best of our knowledge, the parametrization used in the most recent multivariate GREML literature does not guarantee the resulting estimates of covariance matrices are real covariance

matrices (i.e., the unconstrained optimization typically used does not necessarily lead to positive (semi)-definite covariance matrices). This limitation implies that the current state-of-the-art implementation of this multivariate method may fail to converge and, therefore, may yield invalid estimates. The fact that parameter estimates in GREML estimation can leave the feasible parameter space in any iteration is currently dealt with in *ad hoc* manners (e.g., by searching for parameter estimates outside the parameter space in each iteration and – if found – replacing those estimates by values that lie just within the parameter space; e.g., Yang et al. 2011a).

By studying the parametrization of a saturated model, also referred to as a Cholesky model, seen frequently in the twin-study literature (e.g., Kaprio et al. 1982, Phillips and Fulker 1989) and incorporating such a parametrization in the multivariate GREML model, we obtain a model with an underlying parametrization such that the usual iterative optimization procedure (e.g., a Newton-Raphson algorithm) yields at the least positive semi-definite covariance matrices, and – if the parametrization is further amended – even positive definite covariance matrices.

The aforementioned pairwise bivariate approach – instead of a joint multivariate approach – does not address the fact that a large set of traits may have only a few sources of underlying genetic variation and/or environment variation (e.g., when the traits are highly correlated). Pairwise covariances merely indicate that trait A and B have a common genetic basis, and similarly traits B and C have a common genetic basis. However, the question whether traits A, B, and C jointly have a common basis goes unanswered. Parametrizing the model in terms of the elements of the proposed Cholesky decompositions of covariance matrices enables us to impose simple restrictions which force the model to be more parsimonious than the saturated one. That is, we can – for instance – impose a restriction where there is only one genetic factor affecting all traits. Such a factor restriction can be embedded in the classical framework of a likelihood-ratio test.

Regarding computational efficiency, in case one knows the variance components beforehand and one is only interested in obtaining BLUPs for the breeding values, it is possible to use a canonical transformation to strongly reduce the numerical complexity of the problem (Ducrocq and Chapuis, 1997). However, we are interested in estimating the

variance components themselves. Therefore, we cannot apply this transformation. Interestingly, much of our derivations are in a vein similar to a canonical transformation.

In the existing multivariate GREML literature, little attention is paid to numerical efficiency. Yet, under the assumption of balanced data (i.e., a fixed set of individuals in which all traits and relevant covariates are observed), careful inspection of the matrix algebra of the multivariate phenotypic covariance matrix reveals that expressions exist that decrease the numerical complexity from the order $(NP)^3$ to N^3, where N denotes the sample size and P the number of phenotypes in the multivariate analysis.

Our aim is to provide a theoretical framework for a multivariate GREML estimation method which is (i) parametrized such that parameter estimates always remain within a well-defined parameter space and (ii) such that computational requirements are minimized. Given this method and its parametrization, we show how it is related to a factor model, where – for instance – a small subset of genetic factors can drive all observed genetic covariances in a set of phenotypes.

In the following sections, we present the multivariate SNP-based LMM (Section 6.2). Subsequently, we discuss Cholesky decompositions for sets of parameters of the LMM, how to ensure that the parameter estimates are constrained to a well-defined parameter space, and discuss how this parametrization is related to a genetic factor model (Section 6.3). In Section 6.4, we derive an iterative multivariate average-information constrained GREML (MacGREML) estimation method based on a Newton-Raphson algorithm. In Section 6.5 we obtain computationally efficient expressions for the log-likelihood, the gradient, and the average-information (AI) matrix (Gilmour et al., 1995) used in the MacGREML estimation method. Finally, in Section 6.6 we propose a fixed-effect dummy-variable approach for dealing with slightly unbalanced data. We posit that the computational efficiency of our method may be preserved when dealing with slightly unbalanced data, by further inspection of the underlying matrix algebra. We summarize the main properties of our method and propose future applications in Section 6.7.

6.2. MULTIVARIATE SNP-BASED LINEAR MIXED MODELS

We first consider a univariate LMM for a set of traits, and introduce cross-phenotype covariation in SNP effects and environment effects. On this premise, we combine the univariate LMMs into a multivariate LMM. In addition, we briefly discuss how to formulate a multivariate LMM with multiple variance components (e.g., with one GRM based on common variants and one GRM based on rarer variants). Finally, we allude to dealing with missing data; this issue is discussed more thoroughly in Section 6.6.

6.2.1. *Univariate model*

Consider balanced data on P phenotypes. Let $\mathbf{y}_i$ denote the $N \times 1$ phenotype vector, for phenotype $i = 1,\ldots,P$, where N is the number of individuals for whom each phenotype of interest is observed. The LMM for phenotype i, with random SNP effects and fixed confounder effects, is defined as

$$\left.\begin{array}{rcl} \mathbf{y}_i & = & \mathbf{Z}\boldsymbol{\gamma}_i + \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \\ \boldsymbol{\beta}_i & = & \mathcal{N}\left(\mathbf{0},\, \sigma^2_{\boldsymbol{\beta}_i}\mathbf{I}_M\right) \\ \boldsymbol{\varepsilon}_i & = & \mathcal{N}\left(\mathbf{0},\, \sigma^2_{\mathbf{E}_i}\mathbf{I}_N\right) \end{array}\right\} \text{for } i = 1,\ldots,P, \tag{6.1}$$

where $\mathbf{X}$ is the $N \times M$ matrix of standardized genotypes with random effects $\boldsymbol{\beta}_i$, M is the number of SNPs, $\mathbf{Z}$ is the $N \times C$ matrix of confounders (e.g., age and sex) with fixed effects $\boldsymbol{\gamma}_i$, and C is the number of confounders. In this LMM independence between environment effects and SNP effects is assumed.

6.2.2. *Multivariate model*

We assume the existence of between-trait covariance in both environment effects as well as SNP effects. The former reflects the idea that a certain environment tends to affect many traits, rather than just one trait, whereas the latter covariance reflects pleiotropy: the effect of a certain SNP has on phenotype i is correlated with the effect is has on phenotype k, for $i \neq k$.

To be more specific, let β_{ij} denote the j-th element of $\boldsymbol{\beta}_i$ (i.e., the effect of SNP j on phenotype i) and let ε_{il} denote the environment noise for phenotype i, individual l. Now, let the cross-trait covariance be denoted as follows:

$$\begin{aligned}\mathrm{Cov}\left(\beta_{ij},\beta_{kj}\right) &= \sigma_{\boldsymbol{\beta}_{ik}} \text{ for } j = 1,\ldots,M \text{ and}\\ \mathrm{Cov}\left(\varepsilon_{il},\varepsilon_{kl}\right) &= \sigma_{\mathbf{E}_{ik}} \text{ for } l = 1,\ldots,N.\end{aligned}$$

Finally, assume there is no correlation between the effects of different SNPs and between the environment effects for different individuals (i.e., no close relatives in the data). That is,

$$\begin{aligned}\mathrm{Cov}\left(\beta_{ij},\beta_{km}\right) &= 0 \;\forall\; j \neq m \text{ and}\\ \mathrm{Cov}\left(\varepsilon_{il},\varepsilon_{kn}\right) &= 0 \;\forall\; l \neq n.\end{aligned}$$

With these assumptions, the univariate LMMs can be put into vector format, by defining $\sigma^2_{\boldsymbol{\beta}_i} = \sigma_{\boldsymbol{\beta}_{ii}}$, $\sigma^2_{\mathbf{E}_i} = \sigma_{\mathbf{E}_{ii}}$,

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \begin{pmatrix} \sigma^2_{\boldsymbol{\beta}_1} & \cdots & \sigma_{\boldsymbol{\beta}_{1P}} \\ \vdots & \ddots & \vdots \\ \sigma_{\boldsymbol{\beta}_{1P}} & \cdots & \sigma^2_{\boldsymbol{\beta}_P} \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathbf{E}} = \begin{pmatrix} \sigma^2_{\mathbf{E}_1} & \cdots & \sigma_{\mathbf{E}_{1P}} \\ \vdots & \ddots & \vdots \\ \sigma_{\mathbf{E}_{1P}} & \cdots & \sigma^2_{\mathbf{E}_P} \end{pmatrix},$$

and by defining the following stacked vectors

$$\begin{aligned}\mathbf{y} &= \left(\mathbf{y}_1^\top,\ldots,\mathbf{y}_P^\top\right)^\top,\\ \boldsymbol{\beta} &= \left(\boldsymbol{\beta}_1^\top,\ldots,\boldsymbol{\beta}_P^\top\right)^\top,\\ \boldsymbol{\varepsilon} &= \left(\boldsymbol{\varepsilon}_1^\top,\ldots,\boldsymbol{\varepsilon}_P^\top\right)^\top, \text{ and}\\ \boldsymbol{\gamma} &= \left(\boldsymbol{\gamma}_1^\top,\ldots,\boldsymbol{\gamma}_P^\top\right)^\top.\end{aligned}$$

From these definitions it follows that

$$\begin{aligned}\boldsymbol{\beta} &\sim \mathcal{N}\left(\mathbf{0},\, \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \otimes \mathbf{I}_M\right),\\ \boldsymbol{\varepsilon} &\sim \mathcal{N}\left(\mathbf{0},\, \boldsymbol{\Sigma}_{\mathbf{E}} \otimes \mathbf{I}_N\right),\\ \mathbf{y} &= \left(\mathbf{I}_P \otimes \mathbf{Z}\right)\boldsymbol{\gamma} + \left(\mathbf{I}_P \otimes \mathbf{X}\right)\boldsymbol{\beta} + \boldsymbol{\varepsilon},\end{aligned}$$

where '⊗' denotes the Kronecker product.

Consequently,

$$\begin{aligned}\mathbf{y} &\sim \mathcal{N}\left((\mathbf{I}_P \otimes \mathbf{Z})\boldsymbol{\gamma}, (\mathbf{I}_P \otimes \mathbf{X})\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}} \otimes \mathbf{I}_M\right)\left(\mathbf{I}_P \otimes \mathbf{X}^\top\right) + (\boldsymbol{\Sigma}_{\mathbf{E}} \otimes \mathbf{I}_N)\right) \\ &\sim \mathcal{N}\left((\mathbf{I}_P \otimes \mathbf{Z})\boldsymbol{\gamma}, \left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}} \otimes \mathbf{X}\mathbf{X}^\top\right) + (\boldsymbol{\Sigma}_{\mathbf{E}} \otimes \mathbf{I}_N)\right) \\ &\sim \mathcal{N}\left((\mathbf{I}_P \otimes \mathbf{Z})\boldsymbol{\gamma}, (\boldsymbol{\Sigma}_{\mathbf{A}} \otimes \mathbf{A}) + (\boldsymbol{\Sigma}_{\mathbf{E}} \otimes \mathbf{I}_N)\right),\end{aligned}$$

where $\mathbf{A} = M^{-1}\mathbf{X}\mathbf{X}^\top$ and $\boldsymbol{\Sigma}_{\mathbf{A}} = M\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$. The distribution of all phenotypes in $\mathbf{y}$ can be written compactly as

$$\begin{aligned}\mathbf{y} &\sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{V}\right), \\ \boldsymbol{\mu} &= (\mathbf{I}_P \otimes \mathbf{Z})\boldsymbol{\gamma}, \\ \mathbf{V} &= \boldsymbol{\Sigma}_{\mathbf{A}} \otimes \mathbf{A} + \boldsymbol{\Sigma}_{\mathbf{E}} \otimes \mathbf{I}_N,\end{aligned} \tag{6.2}$$

where

$$\boldsymbol{\Sigma}_{\mathbf{A}} = M\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \begin{pmatrix} \sigma^2_{\mathbf{A}_1} & \cdots & \sigma_{\mathbf{A}_{1P}} \\ \vdots & \ddots & \vdots \\ \sigma_{\mathbf{A}_{1P}} & \cdots & \sigma^2_{\mathbf{A}_P} \end{pmatrix}.$$

This model can also be formulated in terms of a contribution from (i) fixed effects of the control variables, (ii) genetic breeding values, and (iii) noise. If the matrices $\boldsymbol{\Sigma}_{\mathbf{A}}$ and $\boldsymbol{\Sigma}_{\mathbf{E}}$ are known and the aim is to obtain BLUPs for the breeding values and the fixed effects, a canonical transformation can be applied (Ducrocq and Chapuis, 1997). That is, one first constructs a $P \times P$ matrix $\mathbf{Q}$, such that $\mathbf{Q}\boldsymbol{\Sigma}_{\mathbf{A}}\mathbf{Q}^\top$ yields a diagonal matrix and $\mathbf{Q}\boldsymbol{\Sigma}_{\mathbf{E}}\mathbf{Q}^\top = \mathbf{I}_P$. Second, using this matrix, rather than modelling $\mathbf{y}$, one models $(\mathbf{Q} \otimes \mathbf{I}_N)\mathbf{y}$. This approach can greatly reduce the numerical complexity of a multivariate estimation of breeding values and fixed effects. However, we are interested in estimating the variance components. Therefore, we cannot apply this transformation. Nevertheless, much of our derivations are in a spirit akin to this canonical transformation.

6.2.3. *Multiple genetic variance components*

In addition to having a single variance component, there can also be multiple genetic variant components (e.g., a component for each chromosome or components based on different allele frequency bins). This extension can be implemented by changing the covariance matrix

$\mathbf{V}$ for the stacked phenotypes in $\mathbf{y}$. That is, by taking

$$\mathbf{V} = \sum_{k=0}^{K} (\boldsymbol{\Sigma}_k \otimes \mathbf{A}_k),$$

where $\mathbf{A}_k$ are the respective GRMs, with associated variance components in $\boldsymbol{\Sigma}_k$, for $k = 1,\ldots,K$. In addition, index $k = 0$ corresponds to environment effects; that is, $\mathbf{A}_0 = \mathbf{I}_N$, and $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_{\mathbf{E}}$. More specifically,

$$\mathbf{A}_k = \frac{1}{M_k}\mathbf{X}_k\mathbf{X}_k^\top \text{ for } k = 1,\ldots,K,$$

where $\mathbf{X}_k$ is the set of standardized SNPs underlying the k-th component, and where M_k is the number of SNPs in that set.

6.2.4. *Unbalanced data*

A more general version of this multivariate LMM allows for (partially) non-overlapping samples between different phenotypes (i.e., unbalanced data). Let N denote the number of genotyped individuals, N_i the number of individuals for whom phenotype i is observed, and $\mathbf{y}_i$ the associated $N_i \times 1$ phenotype vector.

To allow for unbalanced data, consider a $N_i \times N$ selection matrix $\mathbf{S}_i$, such that $\mathbf{S}_i\mathbf{A}_k\mathbf{S}_i^\top$ yields the submatrix of GRM $\mathbf{A}_k$ that corresponds to the individuals for whom phenotype i is observed. This selection is achieved by defining element $\{j,l\}$ of $\mathbf{S}_i$ as

$$\{\mathbf{S}_i\}_{j,l} = \begin{cases} 1 & \text{if element } j \text{ in } \mathbf{y}_i \text{ corresponds to row } l \text{ in } \mathbf{A}_k \,\forall\, k \\ 0 & \text{otherwise.} \end{cases}$$

By defining a block-diagonal matrix $\mathbf{S}$ as

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{S}_P \end{pmatrix},$$

we have that

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{S}(\mathbf{I}_P \otimes \mathbf{Z})\boldsymbol{\gamma},\, \mathbf{S}\left[\sum_{k=0}^{K} (\boldsymbol{\Sigma}_k \otimes \mathbf{A}_k)\right]\mathbf{S}^\top\right).$$

We defer the intricacies of unbalanced data until Section 6.6. For

now, we focus on balanced data with one genetic variance component. Consequently, our working model is as defined in Equation 6.2.

6.3. SATURATED MODELS

Figure 6.1 shows a factor model with P genetic factors and P environment factors, also known as a Cholesky model. This model is such that one factor influences all phenotypes, one factor influences $P-1$ phenotypes, etc., and the last factor influences only one phenotype. This parametrization is such that it can yield any permissible covariance matrices $\mathbf{\Sigma_A}$ and $\mathbf{\Sigma_E}$ (i.e., positive (semi)-definite matrices). Hence, this parametrization allows for the most degrees of freedom in fitting $P \times P$ covariance matrices, making it least parsimonious. Hence, this model has been referred to in the literature as a 'saturated model' (e.g., Kaprio et al. 1982).

We can now show that this saturated model is equivalent to the multivariate LMM under the assumption that the P latent genetic factors $\mathbf{g}$ are linear combinations of SNP data, where the weights are IID draws from a standard normal distribution, and that the P latent environment factors $\mathbf{e}$ are IID draws from a standard normal distribution. That is, we impose the following assumptions for factors $f = 1,\ldots,P$ and $h = 1,\ldots,P$:

$$\begin{aligned}
&\mathbf{g}_f = \frac{1}{\sqrt{M}}\mathbf{X}\boldsymbol{\omega}_f, \\
&\boldsymbol{\omega}_f \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_M) \text{ where } \mathbb{E}\left[\boldsymbol{\omega}_f \boldsymbol{\omega}_h^\top\right] = \mathbf{0} \text{ for } f \neq h, \text{ and} \\
&\mathbf{e}_f \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N) \text{ where } \mathbb{E}\left[\mathbf{e}_f \mathbf{e}_h^\top\right] = \mathbf{0} \text{ for } f \neq h \text{ and } \mathbb{E}\left[\boldsymbol{\omega}_f \mathbf{e}_h^\top\right] = \mathbf{0} \;\forall\; f,h.
\end{aligned}$$

As before, defining $\mathbf{A} = M^{-1}\mathbf{X}\mathbf{X}^\top$, under these assumption we have that

$$\mathbf{g}_f \sim \mathcal{N}(\mathbf{0}, \mathbf{A}) \text{ where } \mathbb{E}\left[\mathbf{g}_f \mathbf{g}_h^\top\right] = \mathbf{0} \text{ for } f \neq h \text{ and } \mathbb{E}\left[\mathbf{g}_f \mathbf{e}_h^\top\right] = \mathbf{0} \;\forall\; f,h.$$

Assuming that the expectation of phenotype $\mathbf{y}_i$ is affected by covariates $\mathbf{Z}$ with fixed effects $\boldsymbol{\gamma}_i$, then phenotype $\mathbf{y}_i$ can be written as

$$\mathbf{y}_i = \mathbf{Z}\boldsymbol{\gamma}_i + \sum_{f=1}^{i} \alpha_{fi}\mathbf{g}_f + \sum_{f=1}^{i} \eta_{fi}\mathbf{e}_f.$$

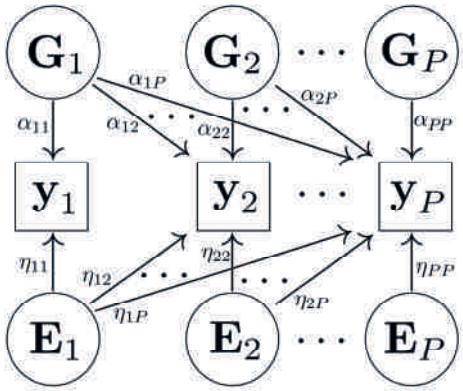


Figure 6.1: *Graphical representation of a saturated multivariate structural model underlying P phenotypes. Squares denote observed variables and circles latent variables.*

This expression for $\mathbf{y}_i$ implies that

$$\begin{aligned}\mathbf{y}_i &\sim \mathcal{N}\left(\mathbf{Z}\boldsymbol{\gamma}_i,\ \sum_{f=1}^{i}\alpha_{fi}^2\mathrm{Var}\left(\mathbf{g}_f\right)+\sum_{f=1}^{i}\eta_{fi}^2\mathrm{Var}\left(\mathbf{e}_f\right)\right)\\ &\sim \mathcal{N}\left(\mathbf{Z}\boldsymbol{\gamma}_i,\ \mathbf{A}\sum_{f=1}^{i}\alpha_{fi}^2+\mathbf{I}_N\sum_{f=1}^{i}\eta_{fi}^2\right).\end{aligned}$$

More generally, the covariance matrix of $\mathbf{y}_i$ and $\mathbf{y}_j$ is given by

$$\mathrm{Cov}\left(\mathbf{y}_i,\mathbf{y}_j\right)=\mathbf{A}\sum_{f=1}^{\min i,j}\alpha_{fi}\alpha_{fj}+\mathbf{I}_N\sum_{f=1}^{\min i,j}\eta_{fi}\eta_{fj}.$$

Consequently, the joint distribution of $\mathbf{y}=\left(\mathbf{y}_1^\top,\ldots,\mathbf{y}_P^\top\right)^\top$ is given by

$$\mathbf{y}\sim\mathcal{N}\left((\mathbf{I}_P\otimes\mathbf{Z})\boldsymbol{\gamma},\ \boldsymbol{\Sigma}_\mathbf{A}\otimes\mathbf{A}+\boldsymbol{\Sigma}_\mathbf{E}\otimes\mathbf{I}_N\right), \tag{6.3}$$

where $\boldsymbol{\Sigma}_\mathbf{A}=\boldsymbol{\Gamma}_\mathbf{A}^\top\boldsymbol{\Gamma}_\mathbf{A}$ and $\boldsymbol{\Sigma}_\mathbf{E}=\boldsymbol{\Gamma}_\mathbf{E}^\top\boldsymbol{\Gamma}_\mathbf{E}$, where

$$\begin{aligned}\boldsymbol{\Gamma}_\mathbf{A} &= \begin{pmatrix}\alpha_{11} & \ldots & \alpha_{1P}\\ & \ddots & \vdots\\ 0 & & \alpha_{PP}\end{pmatrix}\text{ and}\\ \boldsymbol{\Gamma}_\mathbf{E} &= \begin{pmatrix}\eta_{11} & \ldots & \eta_{1P}\\ & \ddots & \vdots\\ 0 & & \eta_{PP}\end{pmatrix}.\end{aligned} \tag{6.4}$$

Element $\{f,i\}$ of $\mathbf{\Gamma_A}$ corresponds to a path coefficient in Figure 6.1, from factor $\mathbf{g}_f$ to phenotype $\mathbf{y}_i$, and similarly, element $\{f,i\}$ of $\mathbf{\Gamma_E}$ to the path coefficient from factor $\mathbf{e}_f$ to phenotype $\mathbf{y}_i$. The preceding equations show that the saturated genetic factor model is equivalent to the multivariate LMM. Moreover, going back to the diagram of this model in Figure 6.1, removing pathways boils down to restricting corresponding elements in $\mathbf{\Gamma_A}$ and $\mathbf{\Gamma_E}$ to zero.

6.3.1. *A well-behaved parametrization*

A reasonable restriction on the variance matrix $\mathbf{V}$ in Equation 6.2, is that this matrix should be positive definite. For any two matrices, $\mathbf{B}$ and $\mathbf{C}$, a sum matrix $\mathbf{D}$, given by $\mathbf{D} = \mathbf{B} + \mathbf{C}$, is positive definite for sure, if $\mathbf{B}$ is at least positive semi-definite, and $\mathbf{C}$ is positive definite.

Since $\mathbf{V}$ is a sum of two Kronecker products, the requirement for $\mathbf{V}$ to be positive definite translates into the requirement that either $\mathbf{\Sigma_A} \otimes \mathbf{A}$ or $\mathbf{\Sigma_E} \otimes \mathbf{I}_N$ is positive definite and the other positive semi-definite. Note that the GRM, $\mathbf{A} = M^{-1}\mathbf{X}\mathbf{X}^\top$, is positive semi-definite by definition, and $\mathbf{I}_N$ is positive definite by definition.

A positive definite matrix has all eigenvalues greater than zero, whereas in a positive semi-definite matrix each eigenvalue is at least zero. Moreover, the eigenvalues of a Kronecker product are equal to the pairwise products of the eigenvalues of the two matrices used to construct the Kronecker product.

Consequently, the Kronecker product of two positive definite matrices is positive definite, and the Kronecker product of a positive definite and a positive semi-definite matrix is positive semi-definite. Therefore, by choosing $\mathbf{\Sigma_A}$ and $\mathbf{\Sigma_E}$ such that these are both positive definite, ensures that $\mathbf{V}$ is positive definite.

Since any positive semi-definite and definite matrix has a Cholesky decomposition, the genetic factor model approach provides a parametrization of the multivariate LMM, seen in Equations 6.3 and 6.4, such that $\mathbf{\Sigma_A}$ and $\mathbf{\Sigma_E}$ are both always positive semi-definite or definite. Two issues here are that (i) we need $\mathbf{\Sigma_A}$ and $\mathbf{\Sigma_E}$ to always be positive definite and never semi-definite, (ii) the mapping of a Cholesky decomposition $\mathbf{\Gamma}$ to a matrix of variance components $\mathbf{\Sigma}$ needs to be unique.

The unicity of the Cholesky decomposition can be imposed by restricting the diagonal elements to be nonnegative. In case the Cholesky decomposition $\mathbf{\Gamma}$ is not constrained, one can arbitrarily change the sign

of any row in $\mathbf{\Gamma}$, without changing the resulting matrix $\mathbf{\Gamma}^\top\mathbf{\Gamma}$. However, by restricting each diagonal element to be nonnegative, these changes in sign of a row can no longer be imposed. Hence, the nonnegative restriction on the diagonal of the Cholesky decomposition ensures unicity.

Moreover, a Cholesky decomposition $\mathbf{\Gamma}$ that has full rank can be shown to always yield a positive definite matrix $\mathbf{\Gamma}^\top\mathbf{\Gamma}$. The rank of a Cholesky decomposition is given by the number of non-zero diagonal elements. Given our restriction of having only nonnegative diagonal elements, the rank requirement implies that all diagonal elements should be positive.

Consequently, we reparametrize the genetic factor model slightly, such that each path from factor i to phenotype i is positive. We do so by replacing the diagonal elements of $\mathbf{\Gamma_A}$ and $\mathbf{\Gamma_E}$ by α^*_{ii} and η^*_{ii}, respectively, and by defining these diagonal elements as functions of underlying parameters $\alpha_{ii} \in \mathbb{R}$ and $\eta_{ii} \in \mathbb{R}$, where these functions are given by $\alpha^*_{ii} = \exp\{\alpha_{ii}\}$ and $\eta^*_{ii} = \exp\{\eta_{ii}\}$, for $i = 1,\ldots,P$. As a consequence of this reparametrization, our multivariate LMM is such that (i) any estimated variance component always lies within a parameter space that ensures the phenotypic covariance matrix is positive definite and (ii) the path coefficients of the underlying factor model, from factors to phenotypes are unique.

A less restrictive parametrization would enforce the positive restriction on the diagonal elements of $\mathbf{\Gamma_E}$ only, whereas for $\mathbf{\Gamma_A}$ nonnegative diagonal elements would suffice. This more lenient approach is particularly useful in case one wants to restrict some genetic path coefficients to be zero (e.g., when assessing how well less than P genetic factors would fit the model). The only restriction of this type that we consider, is a restriction where there is only one genetic factor underlying all P traits. This restriction can be formulated in terms of a Cholesky decomposition where only the first row, consisting of P parameters, is allowed to have entries different from zero. Such a restriction on $\mathbf{\Gamma_A}$ renders it to be rank deficient, causing $\left(\mathbf{\Gamma_A}^\top\mathbf{\Gamma_A}\right) \otimes \mathbf{A}$ to be positive semi-definite, which poses no problem so long as $\left(\mathbf{\Gamma_E}^\top\mathbf{\Gamma_E}\right) \otimes \mathbf{I}_N$ is positive definite. This positive definiteness is implemented by restricting all diagonal elements of $\mathbf{\Gamma_E}$ to be positive in the aforementioned fashion.

6.4. MULTIVARIATE GREML

The vector of stacked phenotypes is now distributed as follows:

$$\mathbf{y} \sim \mathcal{N}\left((\mathbf{I}_P \otimes \mathbf{Z})\boldsymbol{\gamma},\, \left(\left(\boldsymbol{\Gamma}_{\mathbf{A}}^{\top}\boldsymbol{\Gamma}_{\mathbf{A}}\right)\otimes\mathbf{A}\right) + \left(\left(\boldsymbol{\Gamma}_{\mathbf{E}}^{\top}\boldsymbol{\Gamma}_{\mathbf{E}}\right)\otimes\mathbf{I}_N\right)\right) \text{ where}$$

$$\boldsymbol{\Gamma}_{\mathbf{A}} = \begin{pmatrix} \exp\{\alpha_{11}\} & \alpha_{12} & \dots & \alpha_{1P} \\ & \exp\{\alpha_{22}\} & \ddots & \vdots \\ & & \ddots & \alpha_{(P-1),P} \\ 0 & & & \exp\{\alpha_{PP}\} \end{pmatrix} \text{ and}$$

$$\boldsymbol{\Gamma}_{\mathbf{E}} = \begin{pmatrix} \exp\{\eta_{11}\} & \eta_{12} & \dots & \eta_{1P} \\ & \exp\{\eta_{22}\} & \ddots & \vdots \\ & & \ddots & \eta_{(P-1),P} \\ 0 & & & \exp\{\eta_{PP}\} \end{pmatrix}.$$

In addition, the restriction of only one genetic factor (i.e., all cross-trait genetic correlations equal to $+1$ and/or -1) can be formulated as

$$\boldsymbol{\Gamma}_{\mathbf{A}} = \begin{pmatrix} \exp\{\alpha_{11}\} & \alpha_{12} & \dots & \alpha_{1P} \\ 0 & 0 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & 0 \end{pmatrix}$$

and the restriction of no cross-trait genetic correlation as

$$\boldsymbol{\Gamma}_{\mathbf{A}} = \begin{pmatrix} \exp\{\alpha_{11}\} & & 0 \\ & \ddots & \\ 0 & & \exp\{\alpha_{PP}\} \end{pmatrix}.$$

The essence of REML boils to premultiplying $\mathbf{y}$ by a matrix $\mathbf{K}$, which depends on the matrix of confounders $\mathbf{I}_P \otimes \mathbf{Z}$ and is such that $\mathbf{K}(\mathbf{I}_P \otimes \mathbf{Z}) = \mathbf{0}$, and applying maximum likelihood estimation to $\mathbf{y}^* = \mathbf{K}\mathbf{y}$. This matrix $\mathbf{K}$ is $(NP - r) \times (NP)$, where $r = \operatorname{rank}(\mathbf{I}_P \otimes \mathbf{Z}) = P \cdot \operatorname{rank}(\mathbf{Z})$. Provided $\mathbf{Z}$ is of full rank, $r = PC$, where C is the number of confounders.

The distribution of the transformed phenotype $\mathbf{y}^*$ is given by

$$\mathbf{y}^* \sim \mathcal{N}\left(\mathbf{0},\, \mathbf{K}\mathbf{V}\mathbf{K}^{\top}\right) \text{ where } \mathbf{V} = \left(\left(\boldsymbol{\Gamma}_{\mathbf{A}}^{\top}\boldsymbol{\Gamma}_{\mathbf{A}}\right)\otimes\mathbf{A}\right) + \left(\left(\boldsymbol{\Gamma}_{\mathbf{E}}^{\top}\boldsymbol{\Gamma}_{\mathbf{E}}\right)\otimes\mathbf{I}_N\right),$$

and the associated log-likelihood (up to a constant) by

$$l(\boldsymbol{\theta}_{\mathbf{A}}, \boldsymbol{\theta}_{\mathbf{E}}) = -\frac{1}{2}\log\left|\mathbf{K}\mathbf{V}_{\boldsymbol{\theta}_{\mathbf{A}},\boldsymbol{\theta}_{\mathbf{E}}}\mathbf{K}^{\top}\right| - \frac{1}{2}\mathbf{y}^{\top}\mathbf{K}^{\top}\left(\mathbf{K}\mathbf{V}_{\boldsymbol{\theta}_{\mathbf{A}},\boldsymbol{\theta}_{\mathbf{E}}}\mathbf{K}^{\top}\right)^{-1}\mathbf{K}\mathbf{y}, \quad (6.5)$$

where $\mathbf{V}_{\boldsymbol{\theta}_{\mathbf{A}},\boldsymbol{\theta}_{\mathbf{E}}}$ emphasizes the fact that $\mathbf{V}$ depends on the parameters $\boldsymbol{\theta}_{\mathbf{A}}$ and $\boldsymbol{\theta}_{\mathbf{E}}$, where

$$\boldsymbol{\theta}_{\mathbf{A}} = \left\{\left\{\alpha_{fi}\right\}_{i=f}^{P}\right\}_{f=1}^{P}, \text{ and } \boldsymbol{\theta}_{\mathbf{E}} = \left\{\left\{\eta_{fi}\right\}_{i=f}^{P}\right\}_{f=1}^{P}.$$

For the remainder of this section, we omit the subscripts for $\mathbf{V}$ indicating the dependence on the parameters of the model; rather we assume the reader is aware of this dependence, not only for $\mathbf{V}$, but also for its first and second derivatives with respect to the parameters of the model.

The gradient of the log-likelihood with respect to any of the model's parameters θ is now given by

$$\begin{aligned}\frac{\partial l(\boldsymbol{\theta}_{\mathbf{A}}, \boldsymbol{\theta}_{\mathbf{E}})}{\partial \theta} = &-\frac{1}{2}\operatorname{tr}\left(\mathbf{K}^{\top}\left(\mathbf{K}\mathbf{V}\mathbf{K}^{\top}\right)^{-1}\mathbf{K}\frac{\partial \mathbf{V}}{\partial \theta}\right)\ldots \\ &+\frac{1}{2}\mathbf{y}^{\top}\mathbf{K}^{\top}\left(\mathbf{K}\mathbf{V}\mathbf{K}^{\top}\right)^{-1}\mathbf{K}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{K}^{\top}\left(\mathbf{K}\mathbf{V}\mathbf{K}^{\top}\right)^{-1}\mathbf{K}\mathbf{y}.\end{aligned}$$

The second derivative with respect to any combination of two of the parameters, θ and ϕ is now given by

$$\begin{aligned}\frac{\partial^2 l(\boldsymbol{\theta}_{\mathbf{A}}, \boldsymbol{\theta}_{\mathbf{E}})}{\partial \theta \partial \phi} = &-\frac{1}{2}\operatorname{tr}\left(\mathbf{K}^{\top}\left(\mathbf{K}\mathbf{V}\mathbf{K}^{\top}\right)^{-1}\mathbf{K}\frac{\partial^2 \mathbf{V}}{\partial \theta \partial \phi}\right)\ldots \\ &+\frac{1}{2}\operatorname{tr}\left(\mathbf{K}^{\top}\left(\mathbf{K}\mathbf{V}\mathbf{K}^{\top}\right)^{-1}\mathbf{K}\frac{\partial \mathbf{V}}{\partial \phi}\mathbf{K}^{\top}\left(\mathbf{K}\mathbf{V}\mathbf{K}^{\top}\right)^{-1}\mathbf{K}\frac{\partial \mathbf{V}}{\partial \theta}\right)\ldots \\ &+\frac{1}{2}\mathbf{y}^{\top}\mathbf{K}^{\top}\left(\mathbf{K}\mathbf{V}\mathbf{K}^{\top}\right)^{-1}\mathbf{K}\frac{\partial^2 \mathbf{V}}{\partial \theta \partial \phi}\mathbf{K}^{\top}\left(\mathbf{K}\mathbf{V}\mathbf{K}^{\top}\right)^{-1}\mathbf{K}\mathbf{y}\ldots \\ &-\left[\mathbf{y}^{\top}\mathbf{K}^{\top}\left(\mathbf{K}\mathbf{V}\mathbf{K}^{\top}\right)^{-1}\mathbf{K}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{K}^{\top}\left(\mathbf{K}\mathbf{V}\mathbf{K}^{\top}\right)^{-1}\ldots\right. \\ &\qquad\left.\mathbf{K}\frac{\partial \mathbf{V}}{\partial \phi}\mathbf{K}^{\top}\left(\mathbf{K}\mathbf{V}\mathbf{K}^{\top}\right)^{-1}\mathbf{K}\mathbf{y}\right].\end{aligned}$$

Since these derivatives involve solving a set of $NP - r$ equations of the form $\left(\mathbf{K}\mathbf{V}\mathbf{K}^{\top}\right)\mathbf{a} = \mathbf{b}$, if we are able to invert $\mathbf{K}\mathbf{V}\mathbf{K}^{\top}$, we can easily apply a second-order method. Unfortunately, for large N and P, matrix $\mathbf{K}\mathbf{V}\mathbf{K}^{\top}$ tends to become prohibitively large for inversion.

Even though we can rewrite $\mathbf{V}$ such that computing its inverse is of

the order N^3 rather than $(NP)^3$, $\mathbf{K}$ is not a square matrix. Hence, the inverse of $\mathbf{KVK}^\top$ cannot be rewritten in terms of a transformation of $\mathbf{K}$ and the inverse of $\mathbf{V}$. However, by using the Casella-Searle identity (Casella and Searle, 1985, Searle et al., 1992), which states that

$$\begin{aligned}\mathbf{K}^\top\left(\mathbf{KVK}^\top\right)^{-1}\mathbf{K} &= \mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{R}\left(\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{R}\right)^{-1}\mathbf{R}^\top\mathbf{V}^{-1}\\ &= \mathbf{P},\end{aligned} \tag{6.6}$$

the situation can be simplified. In this identity, we have that $\mathbf{R}=\mathbf{I}_P\otimes\mathbf{Z}$, where $\mathbf{Z}$ are the confounders. Using the short-hand notation, $\mathbf{P}$, the first and second-order derivatives can be rewritten as

$$\begin{aligned}\frac{\partial l(\boldsymbol{\theta}_\mathbf{A},\boldsymbol{\theta}_\mathbf{E})}{\partial\theta} &= -\frac{1}{2}\mathrm{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta}\right)+\frac{1}{2}\mathbf{y}^\top\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{Py},\\ \frac{\partial^2 l(\boldsymbol{\theta}_\mathbf{A},\boldsymbol{\theta}_\mathbf{E})}{\partial\theta\partial\phi} &= -\frac{1}{2}\mathrm{tr}\left(\mathbf{P}\frac{\partial^2\mathbf{V}}{\partial\theta\partial\phi}\right)+\frac{1}{2}\mathrm{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\phi}\right)\ldots\\ &\quad+\frac{1}{2}\mathbf{y}^\top\mathbf{P}\frac{\partial^2\mathbf{V}}{\partial\theta\partial\phi}\mathbf{Py}-\mathbf{y}^\top\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\phi}\mathbf{Py}.\end{aligned}$$

According to Equation 6.6, matrix $\mathbf{P}$ can be expressed in terms of the inverse of $\mathbf{V}$ and the inverse of a $(PC)\times(PC)$ matrix. As we show later, the numerical complexity of inverting $\mathbf{V}$ is independent of P and manageable if N is not too large. Moreover, provided we have obtained $\mathbf{V}^{-1}$, the inversion of the $(PC)\times(PC)$ matrix $\left(\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{R}\right)$ is easy.

Given that we vectorize the parameters of the model in $\boldsymbol{\phi}$, and store the first and second derivatives of the log-likelihood function with respect to $\boldsymbol{\phi}$ in a gradient vector $\mathbf{g}$ and Hessian matrix $\mathscr{H}$, we can apply Newton's iterative method. This method is based on a current set of estimates, denoted by $\boldsymbol{\phi}^{(t)}$, and associated gradient $\mathbf{g}^{(t)}$ and Hessian $\mathscr{H}^{(t)}$. Updated parameter estimates are obtained as follows

$$\begin{aligned}\boldsymbol{\phi}^{(t+1)} &= \boldsymbol{\phi}^{(t)}-\left(\mathscr{H}^{(t)}\right)^{-1}\mathbf{g}^{(t)},\\ &= \boldsymbol{\phi}^{(t)}+\left(\mathscr{I}^{(t)}\right)^{-1}\mathbf{g}^{(t)},\end{aligned}$$

where $\mathscr{I}=-\mathscr{H}$, a matrix known as the information matrix.

The elements of the information matrix can be constructed in several ways. First, one can take each element to be the second derivative of the minus-log-likelihood with respect to the combination of two model parameters corresponding to that entry. The resulting informa-

tion matrix is known as the observed information matrix. Second, one can take expectation of each entry of the information matrix, yielding the expected or Fisher information matrix. Finally, one can average the observed and expected information matrices, yielding the AI matrix (Gilmour et al., 1995). The AI matrix is widely accepted in its use for REML estimation and is computationally less involved than the observed information matrix.

Using the fact that the expectation of a scalar is equal to the expectation of the trace of that scalar, we can rewrite the expectation of the second derivative of the minus log-likelihood to yield

$$\mathbb{E}\left[\frac{\partial^2[-l(\boldsymbol{\theta}_\mathbf{A},\boldsymbol{\theta}_\mathbf{E})]}{\partial\theta\partial\phi}\right] = \frac{1}{2}\mathrm{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\phi}\right).$$

Moreover, an element of the observed information matrix is given by

$$\begin{aligned}\frac{\partial^2[-l(\boldsymbol{\theta}_\mathbf{A},\boldsymbol{\theta}_\mathbf{E})]}{\partial\theta\partial\phi} = \frac{1}{2}\mathrm{tr}\left(\mathbf{P}\frac{\partial^2\mathbf{V}}{\partial\theta\partial\phi}\right) - \frac{1}{2}\mathrm{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\phi}\right)\ldots \\ - \frac{1}{2}\mathbf{y}^\top\mathbf{P}\frac{\partial^2\mathbf{V}}{\partial\theta\partial\phi}\mathbf{P}\mathbf{y} + \mathbf{y}^\top\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\phi}\mathbf{P}\mathbf{y}.\end{aligned}$$

Consequently, an element of the AI matrix is given by

$$\mathbf{AI}_{\{\theta,\phi\}} = \frac{1}{4}\mathrm{tr}\left(\mathbf{P}\frac{\partial^2\mathbf{V}}{\partial\theta\partial\phi}\right) - \frac{1}{4}\mathbf{y}^\top\mathbf{P}\frac{\partial^2\mathbf{V}}{\partial\theta\partial\phi}\mathbf{P}\mathbf{y} + \frac{1}{2}\mathbf{y}^\top\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\phi}\mathbf{P}\mathbf{y}.$$

6.4.1. *Simplified log-likelihood*

Before deriving efficient expressions for the gradient and AI matrix, we need a simplified expression for the log-likelihood; the log-likelihood will be used to assess convergence of our method. Based on the work of Casella and Searle (1985) and assuming – without loss of generality – that the rows of $\mathbf{K}$ are orthonormal, we can rewrite the log-determinant of $\mathbf{KVK}^\top$ as

$$\begin{aligned}\log\left|\mathbf{KVK}^\top\right| &= \log\left|\mathbf{KK}^\top\right| - \log\prod_{\lambda_i(\mathbf{P})\neq 0}\lambda_i(\mathbf{P}) \\ &= -\log\prod_{\lambda_i(\mathbf{P})\neq 0}\lambda_i(\mathbf{P}),\end{aligned}$$

where $\lambda_i(\mathbf{P}) \neq 0$ denotes the i-th non-zero eigenvalue of $\mathbf{P}$. Moreover, as shown by others (e.g., Harville 1977), this expression can be rewritten

as

$$\log \prod_{\lambda_i(\mathbf{P})\neq 0} \lambda_i(\mathbf{P}) = -\left(\log|\mathbf{V}| + \log\left|\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{R}\right|\right)$$

Hence, the simplified log-likelihood is given (up to a constant) by

$$l(\boldsymbol{\theta}_\mathbf{A}, \boldsymbol{\theta}_\mathbf{E}) = -\frac{1}{2}\left(\log|\mathbf{V}| + \log\left|\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{R}\right| + \mathbf{y}^\top\mathbf{P}\mathbf{y}\right), \tag{6.7}$$

Consequently, in Section 6.5 we will derive efficient expressions for $\log|\mathbf{V}|$, $\log\left|\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{R}\right|$, and $\mathbf{y}^\top\mathbf{P}\mathbf{y}$.

6.4.2. *Derivatives of the phenotypic covariance matrix*

In the remainder of this section consider the derivatives of the phenotypic covariance matrix. The first and second-order derivatives of $\mathbf{V}$ are given by

$$\frac{\partial\mathbf{V}}{\partial\alpha_{fi}} = \left(\left(\boldsymbol{\Gamma}_\mathbf{A}^\top\frac{\partial\boldsymbol{\Gamma}_\mathbf{A}}{\partial\alpha_{fi}}\right)^\top + \left(\boldsymbol{\Gamma}_\mathbf{A}^\top\frac{\partial\boldsymbol{\Gamma}_\mathbf{A}}{\partial\alpha_{fi}}\right)\right)\otimes\mathbf{A},$$

$$\begin{aligned}
\frac{\partial\mathbf{V}}{\partial\eta_{fi}} &= \left(\left(\boldsymbol{\Gamma}_\mathbf{E}^\top\frac{\partial\boldsymbol{\Gamma}_\mathbf{E}}{\partial\eta_{fi}}\right)^\top + \left(\boldsymbol{\Gamma}_\mathbf{E}^\top\frac{\partial\boldsymbol{\Gamma}_\mathbf{E}}{\partial\eta_{fi}}\right)\right)\otimes\mathbf{I}_N,\\
\frac{\partial^2\mathbf{V}}{\partial\alpha_{fi}\partial\alpha_{gj}} &= \left(\left(\boldsymbol{\Gamma}_\mathbf{A}^\top\frac{\partial^2\boldsymbol{\Gamma}_\mathbf{A}}{\partial\alpha_{fi}\partial\alpha_{gj}} + \frac{\partial\boldsymbol{\Gamma}_\mathbf{A}^\top}{\partial\alpha_{gj}}\frac{\partial\boldsymbol{\Gamma}_\mathbf{A}}{\partial\alpha_{fi}}\right)^\top\cdots\right.\\
&\qquad\left.+\left(\boldsymbol{\Gamma}_\mathbf{A}^\top\frac{\partial^2\boldsymbol{\Gamma}_\mathbf{A}}{\partial\alpha_{fi}\partial\alpha_{gj}} + \frac{\partial\boldsymbol{\Gamma}_\mathbf{A}^\top}{\partial\alpha_{gj}}\frac{\partial\boldsymbol{\Gamma}_\mathbf{A}}{\partial\alpha_{fi}}\right)\right)\otimes\mathbf{A},\\
\frac{\partial^2\mathbf{V}}{\partial\eta_{fi}\partial\eta_{gj}} &= \left(\left(\boldsymbol{\Gamma}_\mathbf{E}^\top\frac{\partial^2\boldsymbol{\Gamma}_\mathbf{E}}{\partial\eta_{fi}\partial\eta_{gj}} + \frac{\partial\boldsymbol{\Gamma}_\mathbf{E}^\top}{\partial\eta_{gj}}\frac{\partial\boldsymbol{\Gamma}_\mathbf{E}}{\partial\eta_{fi}}\right)^\top\cdots\right.\\
&\qquad\left.+\left(\boldsymbol{\Gamma}_\mathbf{E}^\top\frac{\partial^2\boldsymbol{\Gamma}_\mathbf{E}}{\partial\eta_{fi}\partial\eta_{gj}} + \frac{\partial\boldsymbol{\Gamma}_\mathbf{E}^\top}{\partial\eta_{gj}}\frac{\partial\boldsymbol{\Gamma}_\mathbf{E}}{\partial\eta_{fi}}\right)\right)\otimes\mathbf{I}_N,\\
\frac{\partial^2\mathbf{V}}{\partial\alpha_{fi}\partial\eta_{gj}} &= \frac{\partial^2\mathbf{V}}{\partial\eta_{fi}\partial\alpha_{gj}} = \mathbf{0}.
\end{aligned}$$

On the basis of these expression for the derivatives of $\mathbf{V}$, we need the derivatives of $\boldsymbol{\Gamma}_\mathbf{A}$ and $\boldsymbol{\Gamma}_\mathbf{E}$. Since the structure of $\boldsymbol{\Gamma}_\mathbf{A}$ and $\boldsymbol{\Gamma}_\mathbf{E}$ is identical, we only consider the derivatives of $\boldsymbol{\Gamma}_\mathbf{A}$. The derivatives of $\boldsymbol{\Gamma}_\mathbf{E}$ are then known by analogy.

Element $\{h,k\}$ of the first derivative of $\mathbf{\Gamma_A}$ is given by

$$\left\{\frac{\partial \mathbf{\Gamma_A}}{\partial \alpha_{fi}}\right\}_{hk} = \begin{cases} 1 & \text{if } f = h < i = k, \\ \exp\{\alpha_{ii}\} & \text{if } f = h = i = k, \\ 0 & \text{otherwise.} \end{cases} \tag{6.8}$$

Element $\{h,k\}$ of the second derivative of $\mathbf{\Gamma_A}$ is given by

$$\left\{\frac{\partial^2 \mathbf{\Gamma_A}}{\partial \alpha_{fi}\partial \alpha_{gj}}\right\}_{hk} = \begin{cases} \exp\{\alpha_{ii}\} & \text{if } f = i = g = j = h = k, \\ 0 & \text{otherwise.} \end{cases} \tag{6.9}$$

6.5. COMPUTATIONAL EFFICIENCY

In the previous section, we derived the gradient of the REML function and the AI matrix. In this section, we derive computationally efficient expressions for the log-likelihood, and for each element of the gradient and the AI matrix. Specifically, we need efficient expressions for

$$\log|\mathbf{V}|,\ \log\left|\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{R}\right|,\ \mathbf{y}^\top\mathbf{P}\mathbf{y},\ \mathbf{y}^\top\mathbf{P}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{P}\mathbf{y},$$

$$\mathbf{y}^\top\mathbf{P}\frac{\partial^2 \mathbf{V}}{\partial\theta\partial\phi}\mathbf{P}\mathbf{y},\ \mathbf{y}^\top\mathbf{P}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{P}\frac{\partial \mathbf{V}}{\partial \phi}\mathbf{P}\mathbf{y},\ \mathrm{tr}\left(\mathbf{P}\frac{\partial \mathbf{V}}{\partial \theta}\right),\ \text{and}\ \mathrm{tr}\left(\mathbf{P}\frac{\partial^2 \mathbf{V}}{\partial\theta\partial\phi}\right),$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{R}\left(\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{R}\right)^{-1}\mathbf{R}^\top\mathbf{V}^{-1}$ and $\mathbf{R} = \mathbf{I}_P \otimes \mathbf{Z}$. We now focus on finding efficient expressions for the constituents of $\mathbf{P}$.

6.5.1. *Inverse and log-determinant of* $\boldsymbol{V}$

In the multivariate LMM, the phenotypic variance matrix is given by

$$\mathbf{V} = \mathbf{\Sigma_A} \otimes \mathbf{A} + \mathbf{\Sigma_E} \otimes \mathbf{I}_N, \tag{6.10}$$

where $\mathbf{\Sigma_A}$ and $\mathbf{\Sigma_E}$ are both $P \times P$ matrices, and $\mathbf{A}$ is an $N \times N$ symmetric positive semi-definite matrix. Therefore, the eigendecomposition of $\mathbf{A}$ is given by

$$\mathbf{A} = \mathbf{Q}\mathbf{\Phi}^2\mathbf{Q}^\top,$$

such that $\mathbf{Q}\mathbf{Q}^\top = \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_N$.

Consequently, $\mathbf{V}$ can be rewritten as

$$\begin{aligned}\mathbf{V} &= \left(\boldsymbol{\Sigma}_{\mathbf{A}} \otimes \left(\mathbf{Q}\boldsymbol{\Phi}^2\mathbf{Q}^\top\right)\right) + \left(\boldsymbol{\Sigma}_{\mathbf{E}} \otimes \left(\mathbf{Q}\mathbf{Q}^\top\right)\right) \\ &= \left(\mathbf{I}_P \otimes \mathbf{Q}\right)\left(\boldsymbol{\Sigma}_{\mathbf{A}} \otimes \boldsymbol{\Phi}^2\right)\left(\mathbf{I}_P \otimes \mathbf{Q}^\top\right) + \left(\mathbf{I}_P \otimes \mathbf{Q}\right)\left(\boldsymbol{\Sigma}_{\mathbf{E}} \otimes \mathbf{I}_N\right)\left(\mathbf{I}_P \otimes \mathbf{Q}^\top\right) \\ &= \left(\mathbf{I}_P \otimes \mathbf{Q}\right)\left[\left(\boldsymbol{\Sigma}_{\mathbf{A}} \otimes \boldsymbol{\Phi}^2\right) + \left(\boldsymbol{\Sigma}_{\mathbf{E}} \otimes \mathbf{I}_N\right)\right]\left(\mathbf{I}_P \otimes \mathbf{Q}^\top\right).\end{aligned}$$

From this, we can deduce that the inverse of the covariance matrix is given by

$$\mathbf{V}^{-1} = \left(\mathbf{I}_P \otimes \mathbf{Q}\right)\left[\left(\boldsymbol{\Sigma}_{\mathbf{A}} \otimes \boldsymbol{\Phi}^2\right) + \left(\boldsymbol{\Sigma}_{\mathbf{E}} \otimes \mathbf{I}_N\right)\right]^{-1}\left(\mathbf{I}_P \otimes \mathbf{Q}^\top\right). \tag{6.11}$$

Under the parametrization for the environment factors discussed in the previous subsection, $\boldsymbol{\Sigma}_{\mathbf{E}}$ is symmetric and positive definite, and therefore invertible. Moreover, using its eigendecomposition, it easy to construct square-root matrix $\boldsymbol{\Sigma}_{\mathbf{E}}^{\frac{1}{2}}$, such that $\boldsymbol{\Sigma}_{\mathbf{E}}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{E}}^{\frac{1}{2}} = \boldsymbol{\Sigma}_{\mathbf{E}}$. This square root matrix allows us to take $\boldsymbol{\Sigma}_{\mathbf{E}} \otimes \mathbf{I}_N$ from the inverse on both sides of Equation 6.11, as follows

$$\mathbf{V}^{-1} = \left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}} \otimes \mathbf{Q}\right)\left[\left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{A}}\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right) \otimes \boldsymbol{\Phi}^2\right) + \mathbf{I}_{PN}\right]^{-1}\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}} \otimes \mathbf{Q}^\top\right).$$

The existence of $\left[\left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{A}}\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right) \otimes \boldsymbol{\Phi}^2\right) + \mathbf{I}_{PN}\right]^{-1}$ necessitates that none of its eigenvalues are zero. From the properties of the Kronecker product, it follows that the eigenvalues of $\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{A}}\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right) \otimes \boldsymbol{\Phi}^2$ are given by the pair-wise combinations of eigenvalues captured in the diagonal elements of $\boldsymbol{\Phi}^2$ and the eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{A}}\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}$.

By construction of $\mathbf{A}$, the eigenvalues in $\boldsymbol{\Phi}^2$ are non-negative. Moreover, by recognizing the symmetry of both $\boldsymbol{\Sigma}_{\mathbf{E}}$, $\boldsymbol{\Sigma}_{\mathbf{E}}^{-1}$, $\boldsymbol{\Sigma}_{\mathbf{E}}^{\frac{1}{2}}$, and $\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}$, we can rewrite $\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{A}}\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}$ as $\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right)^\top \boldsymbol{\Sigma}_{\mathbf{A}}\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}$, which shows $\boldsymbol{\Sigma}_{\mathbf{A}}$ is congruent to $\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{A}}\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}$.

By assuming that $\boldsymbol{\Sigma}_{\mathbf{A}}$ is parametrized such that it is at least positive semi-definite (as we have successfully done at the start of this section), and by using Sylvester's law stating that the number of negative, positive, and zero eigenvalues of two congruent matrices is equal, it follows that $\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{A}}\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}$ has no negative eigenvalues, since $\boldsymbol{\Sigma}_{\mathbf{A}}$ has none. Therefore, the eigenvalues of $\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{A}}\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right) \otimes \boldsymbol{\Phi}^2$ are

all non-negative. Moreover, the eigenvalues of any matrix $\mathbf{X}+\mathbf{I}$ are equal to the eigenvalues of $\mathbf{X}$ plus one. Therefore, the eigenvalues of $\left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{A}}\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right)\otimes\boldsymbol{\Phi}^2\right)+\mathbf{I}_{PN}$ are all greater than zero, making this matrix invertible. Hence, $\left[\left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{A}}\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right)\otimes\boldsymbol{\Phi}^2\right)+\mathbf{I}_{PN}\right]^{-1}$ exists.

Going back to the efficient computation of $\mathbf{V}^{-1}$, using the fact that $\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{A}}\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}$ is positive semi-definite and symmetric, we can take the eigendecomposition of this matrix, given by

$$\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{A}}\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}=\mathbf{U}\boldsymbol{\Theta}^2\mathbf{U}^\top.$$

Now the inverse of the variance matrix can be written as

$$\begin{aligned}\mathbf{V}^{-1}&=\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\otimes\mathbf{Q}\right)\left[\left(\mathbf{U}\boldsymbol{\Theta}^2\mathbf{U}^\top\otimes\boldsymbol{\Phi}^2\right)+\left(\left(\mathbf{U}\mathbf{U}^\top\right)\otimes\mathbf{I}_N\right)\right]^{-1}\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\otimes\mathbf{Q}^\top\right)\\&=\left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\mathbf{U}\right)\otimes\mathbf{Q}\right)\mathbf{D}^{-1}\left(\left(\mathbf{U}^\top\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right)\otimes\mathbf{Q}^\top\right),\end{aligned}$$

where $\mathbf{D}=\boldsymbol{\Theta}^2\otimes\boldsymbol{\Phi}^2+\mathbf{I}_{PN}$ denotes the diagonal matrix with transformations of combinations of eigenvalues, of which the inverse is also a diagonal matrix, such that the diagonal elements of the inverse are given by one over the corresponding element of $\mathbf{D}$.

This decomposition of the inverse of the variance matrix is written in terms of the eigendecomposition of $N\times N$ matrix $\mathbf{A}$. This eigendecomposition is independent of the current estimates of the parameters of the model. Therefore, the eigenvectors in $\mathbf{Q}$ and the eigenvalues in $\boldsymbol{\Phi}^2$ need only be computed once. Moreover, even for a considerable set of phenotypes (e.g., $P=100$), the computation of matrices such as $\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}$ and the eigendecomposition of $\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{A}}\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}$ are easy.

In addition, the inverse of diagonal matrix $\mathbf{D}$ can also be computed easily, since this inverse is a diagonal matrix with entries given by one over the diagonal entries of the original matrix $\boldsymbol{\Theta}^2\otimes\boldsymbol{\Phi}^2+\mathbf{I}_{PN}$. Finally, by further computational simplifications, we can completely avoid storing the covariance matrix of the stacked phenotypes and its inverse, in the memory at any point.

Regarding the log-determinant of $\mathbf{V}$, using several properties of the

determinant, we have that

$$\begin{aligned}
|\mathbf{V}| &= \left|\left(\boldsymbol{\Sigma}_{\mathbf{A}} \otimes \boldsymbol{\Phi}^2\right) + \left(\boldsymbol{\Sigma}_{\mathbf{E}} \otimes \mathbf{I}_N\right)\right| \\
&= \left|\left(\boldsymbol{\Sigma}_{\mathbf{A}} \otimes \boldsymbol{\Phi}^2\right) + \left(\boldsymbol{\Sigma}_{\mathbf{E}} \otimes \mathbf{I}_N\right)\right| \left|\boldsymbol{\Sigma}_{\mathbf{E}}^{-1}\right|^N |\boldsymbol{\Sigma}_{\mathbf{E}}|^N \\
&= \left|\left(\boldsymbol{\Sigma}_{\mathbf{A}} \otimes \boldsymbol{\Phi}^2\right) + \left(\boldsymbol{\Sigma}_{\mathbf{E}} \otimes \mathbf{I}_N\right)\right| \left|\boldsymbol{\Sigma}_{\mathbf{E}}^{-1} \otimes \mathbf{I}_N\right| |\boldsymbol{\Sigma}_{\mathbf{E}}|^N \\
&= \left|\left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}} \boldsymbol{\Sigma}_{\mathbf{A}} \boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right) \otimes \boldsymbol{\Phi}^2\right) + \mathbf{I}_{NP}\right| |\boldsymbol{\Sigma}_{\mathbf{E}}|^N .
\end{aligned}$$

Hence,

$$\log|\mathbf{V}| = \sum \log\left(\lambda_i\left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}} \boldsymbol{\Sigma}_{\mathbf{A}} \boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right) \otimes \boldsymbol{\Phi}^2\right) + 1\right) + N\log|\boldsymbol{\Sigma}_{\mathbf{E}}|,$$

where $\lambda_i\left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}} \boldsymbol{\Sigma}_{\mathbf{A}} \boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right) \otimes \boldsymbol{\Phi}^2\right)$ denotes the i-th eigenvalue of the Kronecker product between parentheses. Now using the properties of eigenvalues Kronecker products, we finally have that

$$\log|\mathbf{V}| = N\log|\boldsymbol{\Sigma}_{\mathbf{E}}| + \sum_{h=1}^{P}\sum_{j=1}^{N} \log\left(\theta_h^2 \phi_j^2 + 1\right), \tag{6.12}$$

where (overloading notation) θ_h^2 is the h-th eigenvalue of $\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}} \boldsymbol{\Sigma}_{\mathbf{A}} \boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}$ and ϕ_j^2 the j-th eigenvalue from the GRM. When θ and ϕ are preceded by a partial derivative symbol ('∂') we refer to derivatives with respect to parameters in the model. When θ^2 and ϕ^2 are followed by subscripts, we are refering to the diagonal elements of $\boldsymbol{\Theta}^2$ and $\boldsymbol{\Phi}^2$, that is, eigenvalues from the eigenvalue decompositions of $\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}} \boldsymbol{\Sigma}_{\mathbf{A}} \boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}$ and $\mathbf{A}$, respectively.

6.5.2. *Computing* $\log\left|\boldsymbol{R}^\top \boldsymbol{V}^{-1} \boldsymbol{R}\right|$ *and* $\left(\boldsymbol{R}^\top \boldsymbol{V}^{-1} \boldsymbol{R}\right)^{-1}$

We can write $\mathbf{R}^\top \mathbf{V}^{-1} \mathbf{R}$ as

$$\begin{aligned}
\mathbf{R}^\top \mathbf{V}^{-1} \mathbf{R} &= \left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}} \mathbf{U}\right) \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\left(\mathbf{U}^\top \boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right) \otimes \widetilde{\mathbf{Z}}\right) \\
&= \left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}} \mathbf{U}\right) \otimes \mathbf{I}_C\right) \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \ldots \\
&\quad \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \left(\left(\mathbf{U}^\top \boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right) \otimes \mathbf{I}_C\right)
\end{aligned}$$

where $\widetilde{\mathbf{Z}} = \mathbf{Q}^\top \mathbf{Z}$.

Now

$$\begin{aligned}\log\left|\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{R}\right| &= \log\left|\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\mathbf{U}\mathbf{U}^\top\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right)\otimes\mathbf{I}_C\right| + \sum_{h=1}^{P}\log\left|\widetilde{\mathbf{Z}}^\top\mathbf{D}_h^{-1}\widetilde{\mathbf{Z}}\right| \\ &= \log\left|\boldsymbol{\Sigma}_{\mathbf{E}}^{-1}\otimes\mathbf{I}_C\right| + \sum_{h=1}^{P}\log\left|\widetilde{\mathbf{Z}}^\top\mathbf{D}_h^{-1}\widetilde{\mathbf{Z}}\right| \\ &= -C\log|\boldsymbol{\Sigma}_{\mathbf{E}}| + \sum_{h=1}^{P}\log\left|\widetilde{\mathbf{Z}}^\top\mathbf{D}_h^{-1}\widetilde{\mathbf{Z}}\right|,\end{aligned}$$

where $\mathbf{D}_h^{-1}$ is an $N\times N$ diagonal matrix, with the j-th diagonal entry equal to $1/(\theta_h^2\phi_j^2+1)$. Now, the computationally efficient expression for the MacGREML log-likelihood (up to a constant) can be written as

$$l(\boldsymbol{\theta}_{\mathbf{A}},\boldsymbol{\theta}_{\mathbf{E}}) = -\frac{1}{2}\left((N-C)\log|\boldsymbol{\Sigma}_{\mathbf{E}}| + \sum_{h=1}^{P}\sum_{j=1}^{N}\log\left(\theta_h^2\phi_j^2+1\right)\ldots\right. \tag{6.13}$$

$$\left. + \sum_{h=1}^{P}\log\left|\widetilde{\mathbf{Z}}^\top\mathbf{D}_h^{-1}\widetilde{\mathbf{Z}}\right| + \mathbf{y}^\top\mathbf{P}\mathbf{y}\right), \tag{6.14}$$

Similar to the determinant, we can also derive an efficient expression for the inverse, which is given by

$$\left(\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{R}\right)^{-1} = \left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{\frac{1}{2}}\mathbf{U}\right)\otimes\mathbf{I}_C\right)\mathbf{B}\left(\left(\mathbf{U}^\top\boldsymbol{\Sigma}_{\mathbf{E}}^{\frac{1}{2}}\right)\otimes\mathbf{I}_C\right), \tag{6.15}$$

where $\mathbf{B}$ is a block-diagonal matrix defined as

$$\mathbf{B} = \begin{pmatrix} \left(\widetilde{\mathbf{Z}}^\top\mathbf{D}_1^{-1}\widetilde{\mathbf{Z}}\right)^{-1} & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \ldots & 0 & \left(\widetilde{\mathbf{Z}}^\top\mathbf{D}_P^{-1}\widetilde{\mathbf{Z}}\right)^{-1} \end{pmatrix}$$

Hence, computation of $\left(\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{R}\right)^{-1}$ requires inverting P matrices of size $C\times C$, putting these into sparse a block-diagonal matrix, and post- and premultiplying this sparse matrix by sparse Kronecker products. The resulting matrix is $(PC)\times(PC)$. Even for $P=50$ phenotypes and $C=20$ confounders, this merely implies storing a $1{,}000\times 1{,}000$ matrix in the memory each iteration.

6.5.3. *Computing tr$\left(\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial \theta}\right)$*

Regarding $\mathrm{tr}\left(\mathbf{P}\frac{\partial \mathbf{V}}{\partial \theta}\right)$, we first note that this can be rewritten as follows.

$$\mathrm{tr}\left(\mathbf{P}\frac{\partial \mathbf{V}}{\partial \theta}\right) = \mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}\right) - \mathrm{tr}\left(\left(\mathbf{R}^\top \mathbf{V}^{-1}\mathbf{R}\right)^{-1}\mathbf{R}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{V}^{-1}\mathbf{R}\right).$$

Before proceeding let us introduce additional notation for a derivative of $\mathbf{V}$ with respect to α_{fi} (i.e., a parameter pertaining to the genetic factors) and to η_{fi} (i.e., a parameter pertaining to the environment factors). Introducing the following notation

$$\frac{\partial \boldsymbol{\Sigma}_{\mathbf{A}}}{\partial \alpha_{fi}} = \left(\boldsymbol{\Gamma}_{\mathbf{A}}^\top \frac{\partial \boldsymbol{\Gamma}_{\mathbf{A}}}{\partial \alpha_{fi}}\right)^\top + \left(\boldsymbol{\Gamma}_{\mathbf{A}}^\top \frac{\partial \boldsymbol{\Gamma}_{\mathbf{A}}}{\partial \alpha_{fi}}\right) \text{ and} \tag{6.16}$$

$$\frac{\partial \boldsymbol{\Sigma}_{\mathbf{E}}}{\partial \eta_{fi}} = \left(\boldsymbol{\Gamma}_{\mathbf{E}}^\top \frac{\partial \boldsymbol{\Gamma}_{\mathbf{E}}}{\partial \eta_{fi}}\right)^\top + \left(\boldsymbol{\Gamma}_{\mathbf{E}}^\top \frac{\partial \boldsymbol{\Gamma}_{\mathbf{E}}}{\partial \eta_{fi}}\right), \tag{6.17}$$

we have that

$$\frac{\partial \mathbf{V}}{\partial \alpha_{fi}} = \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{A}}}{\partial \alpha_{fi}}\right) \otimes \mathbf{A} = \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{A}}}{\partial \alpha_{fi}}\right) \otimes \left(\mathbf{Q}\boldsymbol{\Phi}^2\mathbf{Q}^\top\right) \quad \text{and} \quad \frac{\partial \mathbf{V}}{\partial \eta_{fi}} = \left(\frac{\partial \boldsymbol{\Sigma}_{\mathbf{E}}}{\partial \eta_{fi}}\right) \otimes \mathbf{I}_N.$$

Now, for a genetic parameter α_{fi}, we have that

$$\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \alpha_{fi}}\right) = \mathrm{tr}\left(\left[\boldsymbol{\Theta}^2 \otimes \boldsymbol{\Phi}^2 + \mathbf{I}_{PN}\right]^{-1}\left(\mathbf{M} \otimes \boldsymbol{\Phi}^2\right)\right)$$

where $\mathbf{M} = \mathbf{U}^\top \boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}} \frac{\partial \boldsymbol{\Sigma}_{\mathbf{A}}}{\partial \alpha_{fi}} \boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}} \mathbf{U}$. This equation can be simplified to

$$\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \alpha_{fi}}\right) = \sum_{h=1}^{P}\sum_{j=1}^{N} \frac{\phi_j^2 m_{hh}}{\theta_h^2\phi_j^2 + 1}, \tag{6.18}$$

where m_{hh} is the h-th diagonal element of $\mathbf{M}$. Similarly, for η_{fi}

$$\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \eta_{fi}}\right) = \mathrm{tr}\left(\left[\boldsymbol{\Theta}^2 \otimes \boldsymbol{\Phi}^2 + \mathbf{I}_{PN}\right]^{-1}\left(\mathbf{L} \otimes \mathbf{I}_N\right)\right),$$

where $\mathbf{L} = \mathbf{U}^\top \boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}} \frac{\partial \boldsymbol{\Sigma}_{\mathbf{E}}}{\partial \eta_{fi}} \boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}} \mathbf{U}$. The equation can be simplified to

$$\mathrm{tr}\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \eta_{fi}}\right) = \sum_{h=1}^{P}\sum_{j=1}^{N} \frac{l_{hh}}{\theta_h^2\phi_j^2 + 1}, \tag{6.19}$$

where l_{hh} is the h-th diagonal element of $\mathbf{L}$.

Finally, the last term of $\mathrm{tr}\left(\mathbf{P}\frac{\partial \mathbf{V}}{\partial \theta}\right)$ is given by

$$\mathrm{tr}\left(\left(\mathbf{R}^\top \mathbf{V}^{-1}\mathbf{R}\right)^{-1}\mathbf{R}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{V}^{-1}\mathbf{R}\right).$$

The term $\mathbf{R}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{V}^{-1}\mathbf{R}$ can be simplified as follows:

$$\begin{aligned}\mathbf{R}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{V}^{-1}\mathbf{R} = &\left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\mathbf{U}\right)\otimes \widetilde{\mathbf{Z}}^\top\right)\left[\boldsymbol{\Theta}^2\otimes\boldsymbol{\Phi}^2+\mathbf{I}_{PN}\right]^{-1}\ldots\\ &\left(\left(\mathbf{U}^\top\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right)\otimes\mathbf{Q}^\top\right)\frac{\partial \mathbf{V}}{\partial \theta}\left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\mathbf{U}\right)\otimes\mathbf{Q}\right)\ldots\\ &\left[\boldsymbol{\Theta}^2\otimes\boldsymbol{\Phi}^2+\mathbf{I}_{PN}\right]^{-1}\left(\left(\mathbf{U}^\top\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right)\otimes\widetilde{\mathbf{Z}}\right).\end{aligned}$$

Considering this derivative with respect to a genetic parameter α_{fi} and environment parameter η_{fi}, we have that

$$\begin{aligned}\mathbf{R}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \alpha_{fi}}\mathbf{V}^{-1}\mathbf{R} &= \left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\mathbf{U}\right)\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{M}\otimes\boldsymbol{\Phi}^2\right)\mathbf{D}^{-1}\left(\left(\mathbf{U}^\top\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right)\otimes\widetilde{\mathbf{Z}}\right),\\ \mathbf{R}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \eta_{fi}}\mathbf{V}^{-1}\mathbf{R} &= \left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\mathbf{U}\right)\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{L}\otimes\mathbf{I}_N\right)\mathbf{D}^{-1}\left(\left(\mathbf{U}^\top\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right)\otimes\widetilde{\mathbf{Z}}\right),\end{aligned}$$

where $\mathbf{D}=\boldsymbol{\Theta}^2\otimes\boldsymbol{\Phi}^2+\mathbf{I}_{PN}$, for which the inverse is a diagonal matrix, with elements one over the diagonal elements of $\boldsymbol{\Theta}^2\otimes\boldsymbol{\Phi}^2+\mathbf{I}_{PN}$, and where $\mathbf{M}$ and $\mathbf{L}$ are as defined before.

Using the fact that for an $N\times M$ matrix $\mathbf{C}$ the following identity hold $\mathrm{tr}\left(\mathbf{C}\mathbf{C}^\top\right)=\mathrm{tr}\left(\mathbf{C}^\top\mathbf{C}\right)=\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\{\mathbf{C}\}_{i,j}\right)^2$ (where $\{\mathbf{C}\}_{ij}$ denotes the element of $\mathbf{C}$ in row i and column j) we can further reduce computational complexity for the remaining trace operations.

First we consider the remaining trace operator for a genetic parameter. That is,

$$\begin{aligned}&\mathrm{tr}\left(\left(\mathbf{R}^\top \mathbf{V}^{-1}\mathbf{R}\right)^{-1}\mathbf{R}^\top \mathbf{V}^{-1}\tfrac{\partial \mathbf{V}}{\partial \alpha_{fi}}\mathbf{V}^{-1}\mathbf{R}\right)=\ldots\\ &\quad\mathrm{tr}\left(\mathbf{B}^{\frac{1}{2}}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{M}^{\frac{1}{2}}\otimes\boldsymbol{\Phi}\right)\left(\mathbf{M}^{\frac{1}{2}}\otimes\boldsymbol{\Phi}\right)\mathbf{D}^{-1}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}\right)\mathbf{B}^{\frac{1}{2}}\right)\\ &=\sum_{k=1}^{(PC)}\sum_{l=1}^{(PN)}\left(\left\{\mathbf{B}^{\frac{1}{2}}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{M}^{\frac{1}{2}}\otimes\boldsymbol{\Phi}\right)\right\}_{k,l}\right)^2.\end{aligned}$$

In the last expression, we exploit the fact that the trace operator is effectively applied to a matrix times its transpose. The square-root

matrix $\mathbf{B}^{\frac{1}{2}}$ of block-diagonal matrix $\mathbf{B}$ defined earlier can simply be computed by taking the square-root matrix of each diagonal block in matrix $\mathbf{B}$. Although we omit the specificities here, given the sparse nature of $\mathbf{D}^{-1}\left(\mathbf{M}^{\frac{1}{2}} \otimes \mathbf{\Phi}\right)$ (i.e., NP^2 non-zero elements and $NP^2(N-1)$ elements equal to zero), the postmultiplication by this sparse matrix can also be performed efficiently.

For an environment parameter we obtain the following expression

$$\operatorname{tr}\left(\left(\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{R}\right)^{-1}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\eta_{fi}}\mathbf{V}^{-1}\mathbf{R}\right) = \ldots$$
$$\sum_{k=1}^{(PC)}\sum_{l=1}^{(PN)}\left(\left\{\mathbf{B}^{\frac{1}{2}}\left(\mathbf{I}_P \otimes \tilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{L}^{\frac{1}{2}} \otimes \mathbf{I}_N\right)\right\}_{k,l}\right)^2.$$

For this expression the same statements about sparsity and efficiency can be made as for the genetic parameters. Combined with Equations 6.18 and 6.19 we have a computationally efficient approach to compute $\operatorname{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta}\right)$.

6.5.4. *Computing tr* $\left(\boldsymbol{P}\frac{\partial^2\boldsymbol{V}}{\partial\theta\partial\phi}\right)$

Using the expressions for $\operatorname{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta}\right)$, it is relatively easy to find efficient expressions for $\operatorname{tr}\left(\mathbf{P}\frac{\partial^2\mathbf{V}}{\partial\theta\partial\phi}\right)$. First, let

$$\frac{\partial^2\mathbf{V}}{\partial\alpha_{fi}\partial\alpha_{gk}} = \left(\frac{\partial^2\mathbf{\Sigma_A}}{\partial\alpha_{fi}\partial\alpha_{gk}}\right) \otimes \left(\mathbf{Q}\mathbf{\Phi}^2\mathbf{Q}^\top\right) \text{ and } \frac{\partial^2\mathbf{V}}{\partial\eta_{fi}\partial\eta_{gk}} = \left(\frac{\partial^2\mathbf{\Sigma_E}}{\partial\eta_{fi}\partial\eta_{gk}}\right) \otimes \mathbf{I}_N.$$

Now, we can show that

$$\operatorname{tr}\left(\mathbf{P}\frac{\partial^2\mathbf{V}}{\partial\alpha_{fi}\partial\alpha_{gk}}\right) = \sum_{h=1}^{P}\sum_{j=1}^{N}\frac{\phi_j^2 m_{hh}^{(2)}}{\theta_h^2\phi_j^2+1}\cdots$$
$$-\sum_{k=1}^{(PC)}\sum_{l=1}^{(PN)}\left(\left\{\mathbf{B}^{\frac{1}{2}}\left(\mathbf{I}_P \otimes \tilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\left(\mathbf{M}^{(2)}\right)^{\frac{1}{2}} \otimes \mathbf{\Phi}\right)\right\}_{k,l}\right)^2,$$

where $m_{hh}^{(2)}$ is the h-th diagonal element of matrix

$$\mathbf{M}^{(2)} = \mathbf{U}^\top\mathbf{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\frac{\partial^2\mathbf{\Sigma_A}}{\partial\alpha_{fi}\partial\alpha_{gk}}\mathbf{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\mathbf{U}$$

and where $\left(\mathbf{M}^{(2)}\right)^{\frac{1}{2}}$ denotes the square root of $\mathbf{M}^{(2)}$.

Analogously, we can show that

$$\operatorname{tr}\left(\mathbf{P}\frac{\partial^2\mathbf{V}}{\partial\eta_{fi}\partial\eta_{gk}}\right) = \sum_{h=1}^{P}\sum_{j=1}^{N}\frac{l_{hh}^{(2)}}{\theta_h^2\phi_j^2+1}\cdots$$
$$-\sum_{k=1}^{(PC)}\sum_{l=1}^{(PN)}\left(\left\{\mathbf{B}^{\frac{1}{2}}\left(\mathbf{I}_P\otimes\tilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\left(\mathbf{L}^{(2)}\right)^{\frac{1}{2}}\otimes\mathbf{I}_N\right)\right\}_{k,l}\right)^2,$$

where $l_{hh}^{(2)}$ is defined as the h-th diagonal element of matrix

$$\mathbf{L}^{(2)} = \mathbf{U}^\top\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\frac{\partial^2\boldsymbol{\Sigma}_{\mathbf{E}}}{\partial\eta_{fi}\partial\eta_{gk}}\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\mathbf{U}$$

and where $\left(\mathbf{L}^{(2)}\right)^{\frac{1}{2}}$ denotes the square root of $\mathbf{L}^{(2)}$.

Finally, bearing in mind that $\frac{\partial^2\mathbf{V}}{\partial\alpha_{fi}\partial\eta_{gk}} = \frac{\partial^2\mathbf{V}}{\partial\eta_{fi}\partial\alpha_{gk}} = \mathbf{0}$, we have that

$$\operatorname{tr}\left(\mathbf{P}\frac{\partial^2\mathbf{V}}{\partial\alpha_{fi}\partial\eta_{gk}}\right) = 0 \text{ and } \operatorname{tr}\left(\mathbf{P}\frac{\partial^2\mathbf{V}}{\partial\eta_{fi}\partial\alpha_{gk}}\right) = 0.$$

6.5.5. *Computing* $\boldsymbol{y}^\top\boldsymbol{P}\boldsymbol{y}$

Let $\mathbf{C} = \left(\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{R}\right)^{-1}$. This notation allows us to rewrite $\mathbf{y}^\top\mathbf{P}\mathbf{y}$ as

$$\begin{aligned}\mathbf{y}^\top\mathbf{P}\mathbf{y} &= \mathbf{y}^\top\left(\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\right)\mathbf{y}\\ &= \mathbf{y}^\top\mathbf{V}^{-1}\mathbf{y}-\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{y}.\end{aligned}$$

In this equation and further equations, the expression $\mathbf{V}^{-1}\mathbf{y}$ recurs frequently. This expression can be simplified using the following identity for the Kronecker product. If matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{X}$ are conformable, in the sense that $\mathbf{AXB}$ exists, then

$$\left(\mathbf{B}^\top\otimes\mathbf{A}\right)\operatorname{vec}(\mathbf{X}) = \operatorname{vec}(\mathbf{AXB}) \;\Leftrightarrow\; \operatorname{vec}(\mathbf{X})^\top\left(\mathbf{B}\otimes\mathbf{A}^\top\right) = \operatorname{vec}(\mathbf{AXB})^\top,$$

where $\operatorname{vec}(\cdot)$ denotes the vectorization operator.

By recognizing that $\mathbf{y} = \operatorname{vec}(\mathbf{Y})$, where matrix $\mathbf{Y} = \left[\mathbf{y}_1\ldots\mathbf{y}_P\right]$, we can write $\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{y}$ as

$$\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{y} = \mathbf{w}^\top\mathbf{D}^{-1}\mathbf{w},$$

where $\mathbf{w} = \operatorname{vec}(\mathbf{W})$, where in turn, $\mathbf{W} = \mathbf{Q}^\top\mathbf{Y}\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\mathbf{U}$.

We can reshape the vector of diagonal elements of diagonal matrix $\mathbf{D}^{-1}$ into an $N \times P$ matrix $\mathbf{E}$, with element $\{j,h\}$ defined as

$$\{\mathbf{E}\}_{j,h} = \frac{1}{\phi_j^2 \theta_h^2 + 1}.$$

Now, we can rewrite $\mathbf{D}^{-1}\mathbf{w}$ as

$$\mathbf{D}^{-1}\mathbf{w} = \widetilde{\mathbf{w}} = \mathrm{vec}\left(\widetilde{\mathbf{W}}\right),$$

where $\widetilde{\mathbf{W}} = \mathbf{E} \circ \mathbf{W}$ and '$\circ$' denotes the element-wise (or Hadamard product). Hence,

$$\begin{aligned}
\mathbf{y}^\top \mathbf{V}^{-1}\mathbf{y} &= \mathbf{w}^\top \widetilde{\mathbf{w}} = \sum_{j=1}^{N}\sum_{h=1}^{P} \left(\{\mathbf{W}\}_{j,h}\right)^2 \{\mathbf{E}\}_{j,h} \text{ and} \\
\mathbf{R}^\top \mathbf{V}^{-1}\mathbf{y} &= \left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\mathbf{U}\right) \otimes \widetilde{\mathbf{Z}}^\top\right)\widetilde{\mathbf{w}} = \left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\mathbf{U}\right) \otimes \widetilde{\mathbf{Z}}^\top\right)\mathrm{vec}\left(\widetilde{\mathbf{W}}\right) \\
&= \mathrm{vec}\left(\mathbf{T}\mathbf{U}^\top \boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right),
\end{aligned}$$

where $\mathbf{T} = \widetilde{\mathbf{Z}}^\top \widetilde{\mathbf{W}}$. In the expression for $\mathbf{T}$, matrix $\widetilde{\mathbf{Z}}^\top$ is a relatively small $C \times N$ matrix which only needs to be computed in the first iteration. Now, substituting the numerical efficient expression for $\mathbf{C}$ and using the Kronecker identity, we have

$$\begin{aligned}
\mathbf{y}^\top \mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top \mathbf{V}^{-1}\mathbf{y} &= \mathrm{vec}\left(\mathbf{T}\mathbf{U}^\top \boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right)^\top \left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{\frac{1}{2}}\mathbf{U}\right) \otimes \mathbf{I}_C\right) \\
&\quad \mathbf{B}\left(\left(\mathbf{U}^\top \boldsymbol{\Sigma}_{\mathbf{E}}^{\frac{1}{2}}\right) \otimes \mathbf{I}_C\right)\mathrm{vec}\left(\mathbf{T}\mathbf{U}^\top \boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right) \\
&= \mathrm{vec}(\mathbf{T})^\top \mathbf{B}\,\mathrm{vec}(\mathbf{T})
\end{aligned}$$

Therefore,

$$\mathbf{y}^\top \mathbf{P}\mathbf{y} = \left(\sum_{j=1}^{N}\sum_{h=1}^{P} \left(\{\mathbf{W}\}_{j,h}\right)^2 \{\mathbf{E}\}_{j,h}\right) - \mathrm{vec}(\mathbf{T})^\top \mathbf{B}\,\mathrm{vec}(\mathbf{T})$$

The matrices in the above expression for $\mathbf{y}^\top \mathbf{P}\mathbf{y}$ are either $N \times P$, $C \times P$, and $(PC) \times (PC)$. The Kronecker product does not appear in this expression at all.

6.5.6. *Computing* $\boldsymbol{y}^\top \boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial \theta}\boldsymbol{P}\boldsymbol{y}$

Using the notation from the previous derivation, we have that

$$\begin{aligned}\mathbf{y}^\top\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{P}\mathbf{y} &= \mathbf{y}^\top\left(\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\right)\frac{\partial\mathbf{V}}{\partial\theta}\left(\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\right)\mathbf{y}\\ &= \mathbf{y}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{V}^{-1}\mathbf{y}-2\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{V}^{-1}\mathbf{y}\ldots\\ &\quad+\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{y}.\end{aligned}$$

Now

$$\begin{aligned}\mathbf{y}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\alpha_{fi}}\mathbf{V}^{-1}\mathbf{y} &= \sum_{j=1}^{N}\sum_{h=1}^{P}\left\{\widetilde{\mathbf{W}}\right\}_{j,h}\left\{\boldsymbol{\Phi}^2\widetilde{\mathbf{W}}\mathbf{M}\right\}_{j,h}\text{ and,}\\ \mathbf{y}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\eta_{fi}}\mathbf{V}^{-1}\mathbf{y} &= \sum_{j=1}^{N}\sum_{h=1}^{P}\left\{\widetilde{\mathbf{W}}\right\}_{j,h}\left\{\widetilde{\mathbf{W}}\mathbf{L}\right\}_{j,h}.\end{aligned}$$

Similarly, we obtain the following expressions for the other terms involved in $\mathbf{y}^\top\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{P}\mathbf{y}$,

$$\begin{aligned}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\alpha_{fi}}\mathbf{V}^{-1}\mathbf{y} &= \left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\mathbf{U}\right)\otimes\widetilde{\mathbf{Z}}^\top\right)\operatorname{vec}\left(\mathbf{E}\circ\left(\boldsymbol{\Phi}^2\widetilde{\mathbf{W}}\mathbf{M}\right)\right),\\ \mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\eta_{fi}}\mathbf{V}^{-1}\mathbf{y} &= \left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\mathbf{U}\right)\otimes\widetilde{\mathbf{Z}}^\top\right)\operatorname{vec}\left(\mathbf{E}\circ\left(\widetilde{\mathbf{W}}\mathbf{L}\right)\right),\\ \mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\alpha_{fi}}\mathbf{V}^{-1}\mathbf{R} &= \left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\mathbf{U}\right)\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{M}\otimes\boldsymbol{\Phi}^2\right)\mathbf{D}^{-1}\left(\left(\mathbf{U}^\top\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right)\otimes\widetilde{\mathbf{Z}}\right),\\ \mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\eta_{fi}}\mathbf{V}^{-1}\mathbf{R} &= \left(\left(\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\mathbf{U}\right)\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{L}\otimes\mathbf{I}_N\right)\mathbf{D}^{-1}\left(\left(\mathbf{U}^\top\boldsymbol{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\right)\otimes\widetilde{\mathbf{Z}}\right).\end{aligned}$$

Combining the various ingredients, again substituting $\mathbf{C}$, we now have

$$\begin{aligned}\mathbf{y}^\top\mathbf{P}\frac{\partial\mathbf{V}}{\partial\alpha_{fi}}\mathbf{P}\mathbf{y} &= \left(\sum_{j=1}^{N}\sum_{h=1}^{P}\left\{\widetilde{\mathbf{W}}\right\}_{j,h}\left\{\boldsymbol{\Phi}^2\widetilde{\mathbf{W}}\mathbf{M}\right\}_{j,h}\right)-2\operatorname{vec}(\mathbf{T})^\top\mathbf{B}\operatorname{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{M}}\right)\ldots\\ &\quad+\left[\operatorname{vec}(\mathbf{T})^\top\mathbf{B}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{M}\otimes\boldsymbol{\Phi}^2\right)\mathbf{D}^{-1}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}\right)\mathbf{B}\operatorname{vec}(\mathbf{T})\right]\text{ and}\\ \mathbf{y}^\top\mathbf{P}\frac{\partial\mathbf{V}}{\partial\eta_{fi}}\mathbf{P}\mathbf{y} &= \left(\sum_{j=1}^{N}\sum_{h=1}^{P}\left\{\widetilde{\mathbf{W}}\right\}_{j,h}\left\{\widetilde{\mathbf{W}}\mathbf{L}\right\}_{j,h}\right)-2\operatorname{vec}(\mathbf{T})^\top\mathbf{B}\operatorname{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{L}}\right)\ldots\\ &\quad+\left[\operatorname{vec}(\mathbf{T})^\top\mathbf{B}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{L}\otimes\mathbf{I}_N\right)\mathbf{D}^{-1}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}\right)\mathbf{B}\operatorname{vec}(\mathbf{T})\right],\end{aligned}$$

where $\widetilde{\mathbf{T}}_{\mathbf{M}}=\widetilde{\mathbf{Z}}^\top\left(\mathbf{E}\circ\left(\boldsymbol{\Phi}^2\widetilde{\mathbf{W}}\mathbf{M}\right)\right)$ and $\widetilde{\mathbf{T}}_{\mathbf{L}}=\widetilde{\mathbf{Z}}^\top\left(\mathbf{E}\circ\left(\widetilde{\mathbf{W}}\mathbf{L}\right)\right)$.

6.5.7. *Computing* $\boldsymbol{y}^\top \boldsymbol{P} \frac{\partial^2 \boldsymbol{V}}{\partial\theta\partial\phi} \boldsymbol{P}\boldsymbol{y}$

To obtain expressions for $\mathbf{y}^\top \mathbf{P} \frac{\partial^2 \mathbf{V}}{\partial\theta\partial\phi} \mathbf{P}\mathbf{y}$ we need to replace the matrices $\mathbf{M}$ and $\mathbf{L}$ in the expressions for $\mathbf{y}^\top \mathbf{P} \frac{\partial \mathbf{V}}{\partial\theta} \mathbf{P}\mathbf{y}$ by their second-order counterparts $\mathbf{M}^{(2)}$ and $\mathbf{L}^{(2)}$, yielding

$$\mathbf{y}^\top \mathbf{P} \frac{\partial^2 \mathbf{V}}{\partial\alpha_{fi}\partial\alpha_{gk}} \mathbf{P}\mathbf{y} = \left(\sum_{j=1}^{N} \sum_{h=1}^{P} \{\tilde{\mathbf{W}}\}_{j,h} \left\{ \boldsymbol{\Phi}^2 \tilde{\mathbf{W}} \mathbf{M}^{(2)} \right\}_{j,h} \right) - 2\mathrm{vec}(\mathbf{T})^\top \mathbf{B}\, \mathrm{vec}\left(\tilde{\mathbf{T}}_{\mathbf{M}^{(2)}} \right) \ldots$$
$$+ \left[\mathrm{vec}(\mathbf{T})^\top \mathbf{B} \left(\mathbf{I}_P \otimes \tilde{\mathbf{Z}}^\top \right) \mathbf{D}^{-1} \left(\mathbf{M}^{(2)} \otimes \boldsymbol{\Phi}^2 \right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \tilde{\mathbf{Z}} \right) \mathbf{B}\, \mathrm{vec}(\mathbf{T}) \right] \text{ and}$$

$$\mathbf{y}^\top \mathbf{P} \frac{\partial^2 \mathbf{V}}{\partial\eta_{fi}\partial\eta_{gk}} \mathbf{P}\mathbf{y} = \left(\sum_{j=1}^{N} \sum_{h=1}^{P} \{\tilde{\mathbf{W}}\}_{j,h} \left\{ \tilde{\mathbf{W}} \mathbf{L}^{(2)} \right\}_{j,h} \right) - 2\mathrm{vec}(\mathbf{T})^\top \mathbf{B}\, \mathrm{vec}\left(\tilde{\mathbf{T}}_{\mathbf{L}^{(2)}} \right) \ldots$$
$$+ \left[\mathrm{vec}(\mathbf{T})^\top \mathbf{B} \left(\mathbf{I}_P \otimes \tilde{\mathbf{Z}}^\top \right) \mathbf{D}^{-1} \left(\mathbf{L}^{(2)} \otimes \mathbf{I}_N \right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \tilde{\mathbf{Z}} \right) \mathbf{B}\, \mathrm{vec}(\mathbf{T}) \right],$$

where $\tilde{\mathbf{T}}_{\mathbf{M}^{(2)}} = \tilde{\mathbf{Z}}^\top \left(\mathbf{E} \circ \left(\boldsymbol{\Phi}^2 \tilde{\mathbf{W}} \mathbf{M}^{(2)} \right) \right)$ and $\tilde{\mathbf{T}}_{\mathbf{L}^{(2)}} = \tilde{\mathbf{Z}}^\top \left(\mathbf{E} \circ \left(\tilde{\mathbf{W}} \mathbf{L}^{(2)} \right) \right)$.

In the expressions for both $\mathbf{y}^\top \mathbf{P} \frac{\partial \mathbf{V}}{\partial\theta} \mathbf{P}\mathbf{y}$ and $\mathbf{y}^\top \mathbf{P} \frac{\partial^2 \mathbf{V}}{\partial\theta\partial\phi} \mathbf{P}\mathbf{y}$, one could formulate $\mathbf{B}\,\mathrm{vec}(\mathbf{T})$ as the vectorization of a $C \times P$ matrix. This formulation would allow us to reduce the computational complexity of $\left(\mathbf{I}_P \otimes \tilde{\mathbf{Z}} \right) \mathbf{B}\,\mathrm{vec}(\mathbf{T})$, again using the Kronecker identity. In fact, one could use the Kronecker identity in this fashion recursively, until all remaining Kronecker products in these and further equations have been eliminated. However, in order to keep notation tractable we do not pursue this possibility any further.

Finally, $\mathbf{y}^\top \mathbf{P} \frac{\partial^2 \mathbf{V}}{\partial\alpha_{fi}\partial\eta_{gk}} \mathbf{P}\mathbf{y} = 0$ and $\mathbf{y}^\top \mathbf{P} \frac{\partial^2 \mathbf{V}}{\partial\eta_{fi}\partial\alpha_{gk}} \mathbf{P}\mathbf{y} = 0$.

6.5.8. *Computing* $\boldsymbol{y}^\top \boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial \theta}\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial \phi}\boldsymbol{P}\boldsymbol{y}$

We can write $\mathbf{y}^\top \mathbf{P}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{P}\frac{\partial \mathbf{V}}{\partial \phi}\mathbf{P}\mathbf{y}$ as

$$
\begin{aligned}
\mathbf{y}^\top \mathbf{P}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{P}\frac{\partial \mathbf{V}}{\partial \phi}\mathbf{P}\mathbf{y} = &\dots \\
&\mathbf{y}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \phi}\mathbf{V}^{-1}\mathbf{y}\dots \\
&-\mathbf{y}^\top \mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \phi}\mathbf{V}^{-1}\mathbf{y}\dots \\
&-\mathbf{y}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \phi}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top \mathbf{V}^{-1}\mathbf{y}\dots \\
&-\mathbf{y}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \phi}\mathbf{V}^{-1}\mathbf{y}\dots \\
&+\mathbf{y}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \phi}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top \mathbf{V}^{-1}\mathbf{y}\dots \\
&+\mathbf{y}^\top \mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \phi}\mathbf{V}^{-1}\mathbf{y}\dots \\
&+\mathbf{y}^\top \mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \phi}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top \mathbf{V}^{-1}\mathbf{y}\dots \\
&-\mathbf{y}^\top \mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \phi}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top \mathbf{V}^{-1}\mathbf{y}.
\end{aligned}
$$

By introducing subscripts for $\mathbf{M}$ and $\mathbf{L}$, to indicate with respect to what parameter these matrices are, we can now find efficient expression for these terms. Using work from previous sections, we find that

$$
\begin{aligned}
\mathbf{y}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \alpha_{fi}}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \alpha_{gk}}\mathbf{V}^{-1}\mathbf{y} &= \operatorname{vec}\left(\mathbf{\Phi}^2\tilde{\mathbf{W}}\mathbf{M}_{fi}\right)^\top \mathbf{D}^{-1}\operatorname{vec}\left(\mathbf{\Phi}^2\tilde{\mathbf{W}}\mathbf{M}_{gk}\right), \\
\mathbf{y}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \eta_{fi}}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \eta_{gk}}\mathbf{V}^{-1}\mathbf{y} &= \operatorname{vec}\left(\tilde{\mathbf{W}}\mathbf{L}_{fi}\right)^\top \mathbf{D}^{-1}\operatorname{vec}\left(\tilde{\mathbf{W}}\mathbf{L}_{gk}\right), \\
\mathbf{y}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \alpha_{fi}}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \eta_{gk}}\mathbf{V}^{-1}\mathbf{y} &= \operatorname{vec}\left(\mathbf{\Phi}^2\tilde{\mathbf{W}}\mathbf{M}_{fi}\right)^\top \mathbf{D}^{-1}\operatorname{vec}\left(\tilde{\mathbf{W}}\mathbf{L}_{gk}\right), \text{ and} \\
\mathbf{y}^\top \mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \eta_{fi}}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \alpha_{gk}}\mathbf{V}^{-1}\mathbf{y} &= \operatorname{vec}\left(\tilde{\mathbf{W}}\mathbf{L}_{fi}\right)^\top \mathbf{D}^{-1}\operatorname{vec}\left(\mathbf{\Phi}^2\tilde{\mathbf{W}}\mathbf{M}_{gk}\right),
\end{aligned}
$$

where $\mathbf{M}_{fi} = \mathbf{U}^\top \mathbf{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\frac{\partial \mathbf{\Sigma}_{\mathbf{A}}}{\partial \alpha_{fi}}\mathbf{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\mathbf{U}$ and $\mathbf{L}_{fi} = \mathbf{U}^\top \mathbf{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\frac{\partial \mathbf{\Sigma}_{\mathbf{E}}}{\partial \eta_{fi}}\mathbf{\Sigma}_{\mathbf{E}}^{-\frac{1}{2}}\mathbf{U}$.

For the term $\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\phi}\mathbf{V}^{-1}\mathbf{y}$ we find, using preceding derivations, that

$$\begin{aligned}
\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\alpha_{fi}}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\alpha_{gk}}\mathbf{V}^{-1}\mathbf{y} &= \mathrm{vec}(\mathbf{T})^\top\mathbf{B}\ldots \\
&\quad \mathrm{vec}\left(\widetilde{\mathbf{Z}}^\top\left(\mathbf{E}\circ\left[\boldsymbol{\Phi}^2\left\{\mathbf{E}\circ\left(\boldsymbol{\Phi}^2\widetilde{\mathbf{W}}\mathbf{M}_{gk}\right)\right\}\mathbf{M}_{fi}\right]\right)\right), \\
\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\eta_{fi}}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\eta_{gk}}\mathbf{V}^{-1}\mathbf{y} &= \mathrm{vec}(\mathbf{T})^\top\mathbf{B}\ldots \\
&\quad \mathrm{vec}\left(\widetilde{\mathbf{Z}}^\top\left(\mathbf{E}\circ\left[\left\{\mathbf{E}\circ\left(\widetilde{\mathbf{W}}\mathbf{L}_{gk}\right)\right\}\mathbf{L}_{fi}\right]\right)\right), \\
\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\alpha_{fi}}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\eta_{gk}}\mathbf{V}^{-1}\mathbf{y} &= \mathrm{vec}(\mathbf{T})^\top\mathbf{B}\ldots \\
&\quad \mathrm{vec}\left(\widetilde{\mathbf{Z}}^\top\left(\mathbf{E}\circ\left[\boldsymbol{\Phi}^2\left\{\mathbf{E}\circ\left(\widetilde{\mathbf{W}}\mathbf{L}_{gk}\right)\right\}\mathbf{M}_{fi}\right]\right)\right), \text{ and} \\
\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\eta_{fi}}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\alpha_{gk}}\mathbf{V}^{-1}\mathbf{y} &= \mathrm{vec}(\mathbf{T})^\top\mathbf{B}\ldots \\
&\quad \mathrm{vec}\left(\widetilde{\mathbf{Z}}^\top\left(\mathbf{E}\circ\left[\left\{\mathbf{E}\circ\left(\boldsymbol{\Phi}^2\widetilde{\mathbf{W}}\mathbf{M}_{gk}\right)\right\}\mathbf{L}_{fi}\right]\right)\right).
\end{aligned}$$

$\mathbf{y}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\phi}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{y}$ can also be computed using the preceding expressions, since the latter expression is just the transpose of a scalar resulting from one of the four equations in the former set. One needs to pay special attention though, to the order of the derivatives of $\mathbf{V}$ with respect to θ and ϕ when transposing, making sure the intended expression is obtained.

Expressions for $\mathbf{y}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\phi}\mathbf{V}^{-1}\mathbf{y}$ are given by

$$\begin{aligned}
\mathbf{y}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\alpha_{fi}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\alpha_{gk}}\mathbf{V}^{-1}\mathbf{y} &= \mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{M}_{fi}}\right)^\top\mathbf{B}\mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{M}_{gk}}\right), \\
\mathbf{y}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\eta_{fi}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\eta_{gk}}\mathbf{V}^{-1}\mathbf{y} &= \mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{L}_{fi}}\right)^\top\mathbf{B}\mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{L}_{gk}}\right), \\
\mathbf{y}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\alpha_{fi}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\eta_{gk}}\mathbf{V}^{-1}\mathbf{y} &= \mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{M}_{fi}}\right)^\top\mathbf{B}\mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{L}_{gk}}\right), \text{ and} \\
\mathbf{y}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\eta_{fi}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\alpha_{gk}}\mathbf{V}^{-1}\mathbf{y} &= \mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{L}_{fi}}\right)^\top\mathbf{B}\mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{M}_{gk}}\right),
\end{aligned}$$

where $\widetilde{\mathbf{T}}_{\mathbf{M}_{fi}} = \widetilde{\mathbf{Z}}^\top\left\{\mathbf{E}\circ\left(\boldsymbol{\Phi}^2\widetilde{\mathbf{W}}\mathbf{M}_{fi}\right)\right\}$ and $\widetilde{\mathbf{T}}_{\mathbf{L}_{fi}} = \widetilde{\mathbf{Z}}^\top\left\{\mathbf{E}\circ\left(\widetilde{\mathbf{W}}\mathbf{L}_{fi}\right)\right\}$.

Expressions for $\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\phi}\mathbf{V}^{-1}\mathbf{y}$ are given by

$$\begin{aligned}
&\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\alpha_{fi}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\alpha_{gk}}\mathbf{V}^{-1}\mathbf{y} = \\
&\quad \mathrm{vec}(\mathbf{T})^\top\mathbf{B}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{M}_{fi}\otimes\boldsymbol{\Phi}^2\right)\mathbf{D}^{-1}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}\right)\mathbf{B}\mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{M}_{gk}}\right), \\
&\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\eta_{fi}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\eta_{gk}}\mathbf{V}^{-1}\mathbf{y} = \\
&\quad \mathrm{vec}(\mathbf{T})^\top\mathbf{B}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{L}_{fi}\otimes\mathbf{I}_N\right)\mathbf{D}^{-1}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}\right)\mathbf{B}\mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{L}_{gk}}\right), \\
&\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\alpha_{fi}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\eta_{gk}}\mathbf{V}^{-1}\mathbf{y} = \\
&\quad \mathrm{vec}(\mathbf{T})^\top\mathbf{B}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{M}_{fi}\otimes\boldsymbol{\Phi}^2\right)\mathbf{D}^{-1}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}\right)\mathbf{B}\mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{L}_{gk}}\right), \text{ and} \\
&\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\eta_{fi}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\alpha_{gk}}\mathbf{V}^{-1}\mathbf{y} = \\
&\quad \mathrm{vec}(\mathbf{T})^\top\mathbf{B}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{L}_{fi}\otimes\mathbf{I}_N\right)\mathbf{D}^{-1}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}\right)\mathbf{B}\mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{M}_{gk}}\right).
\end{aligned}$$

As before, note that $\mathbf{y}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\phi}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{y}$ can be computed using the preceding expressions, since the desired expressions are just the transpose of a scalar resulting from one of the four equations in the former set. Again, one needs to pay special attention to the order of the derivatives of $\mathbf{V}$ with respect to θ and ϕ when transposing, making sure the intended expression is obtained.

For $\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\phi}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{y}$ we have the following expressions

$$\begin{aligned}
&\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\alpha_{fi}}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\alpha_{gk}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{y} = \\
&\qquad \mathrm{vec}(\mathbf{T})^\top\mathbf{B}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{M}_{fi}\otimes\boldsymbol{\Phi}^2\right)\mathbf{D}^{-1}\ldots \\
&\qquad \left(\mathbf{M}_{gk}\otimes\boldsymbol{\Phi}^2\right)\mathbf{D}^{-1}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}\right)\mathbf{B}\mathrm{vec}(\mathbf{T}), \\
&\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\eta_{fi}}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\eta_{gk}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{y} = \\
&\qquad \mathrm{vec}(\mathbf{T})^\top\mathbf{B}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{L}_{fi}\otimes\mathbf{I}_N\right)\mathbf{D}^{-1}\ldots \\
&\qquad \left(\mathbf{L}_{gk}\otimes\mathbf{I}_N\right)\mathbf{D}^{-1}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}\right)\mathbf{B}\mathrm{vec}(\mathbf{T}),
\end{aligned}$$

$$\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\tfrac{\partial\mathbf{V}}{\partial\alpha_{fi}}\mathbf{V}^{-1}\tfrac{\partial\mathbf{V}}{\partial\eta_{gk}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{y} =$$
$$\mathrm{vec}(\mathbf{T})^\top\mathbf{B}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{M}_{fi}\otimes\boldsymbol{\Phi}^2\right)\mathbf{D}^{-1}\ldots$$
$$\left(\mathbf{L}_{gk}\otimes\mathbf{I}_N\right)\mathbf{D}^{-1}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}\right)\mathbf{B}\,\mathrm{vec}(\mathbf{T}),\text{ and}$$

$$\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\tfrac{\partial\mathbf{V}}{\partial\eta_{fi}}\mathbf{V}^{-1}\tfrac{\partial\mathbf{V}}{\partial\alpha_{gk}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{y} =$$
$$\mathrm{vec}(\mathbf{T})^\top\mathbf{B}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{L}_{fi}\otimes\mathbf{I}_N\right)\mathbf{D}^{-1}\ldots$$
$$\left(\mathbf{M}_{gk}\otimes\boldsymbol{\Phi}^2\right)\mathbf{D}^{-1}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}\right)\mathbf{B}\,\mathrm{vec}(\mathbf{T}).$$

Finally, for $\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\phi}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{y}$ we have that

$$\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\tfrac{\partial\mathbf{V}}{\partial\alpha_{fi}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\tfrac{\partial\mathbf{V}}{\partial\alpha_{gk}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{y} =$$
$$\mathrm{vec}(\mathbf{T})^\top\mathbf{B}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{M}_{fi}\otimes\boldsymbol{\Phi}^2\right)\mathbf{D}^{-1}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}\right)\mathbf{B}\ldots$$
$$\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{M}_{gk}\otimes\boldsymbol{\Phi}^2\right)\mathbf{D}^{-1}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}\right)\mathbf{B}\,\mathrm{vec}(\mathbf{T}),$$

$$\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\tfrac{\partial\mathbf{V}}{\partial\eta_{fi}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\tfrac{\partial\mathbf{V}}{\partial\eta_{gk}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{y} =$$
$$\mathrm{vec}(\mathbf{T})^\top\mathbf{B}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{L}_{fi}\otimes\mathbf{I}_N\right)\mathbf{D}^{-1}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}\right)\mathbf{B}\ldots$$
$$\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{L}_{gk}\otimes\mathbf{I}_N\right)\mathbf{D}^{-1}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}\right)\mathbf{B}\,\mathrm{vec}(\mathbf{T}),$$

$$\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\tfrac{\partial\mathbf{V}}{\partial\alpha_{fi}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\tfrac{\partial\mathbf{V}}{\partial\eta_{gk}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{y} =$$
$$\mathrm{vec}(\mathbf{T})^\top\mathbf{B}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{M}_{fi}\otimes\boldsymbol{\Phi}^2\right)\mathbf{D}^{-1}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}\right)\mathbf{B}\ldots$$
$$\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{L}_{gk}\otimes\mathbf{I}_N\right)\mathbf{D}^{-1}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}\right)\mathbf{B}\,\mathrm{vec}(\mathbf{T}),\text{ and}$$

$$\mathbf{y}^\top\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\tfrac{\partial\mathbf{V}}{\partial\eta_{fi}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\tfrac{\partial\mathbf{V}}{\partial\alpha_{gk}}\mathbf{V}^{-1}\mathbf{R}\mathbf{C}\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{y} =$$
$$\mathrm{vec}(\mathbf{T})^\top\mathbf{B}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{L}_{fi}\otimes\mathbf{I}_N\right)\mathbf{D}^{-1}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}\right)\mathbf{B}\ldots$$
$$\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}^\top\right)\mathbf{D}^{-1}\left(\mathbf{M}_{gk}\otimes\boldsymbol{\Phi}^2\right)\mathbf{D}^{-1}\left(\mathbf{I}_P\otimes\widetilde{\mathbf{Z}}\right)\mathbf{B}\,\mathrm{vec}(\mathbf{T}).$$

The eight expressions need to be combined into $\mathbf{y}^\top \mathbf{P} \frac{\partial \mathbf{V}}{\partial \theta} \mathbf{P} \frac{\partial \mathbf{V}}{\partial \phi} \mathbf{P}\mathbf{y}$. Substitutions yield the following expressions:

$$
\begin{aligned}
&\mathbf{y}^\top \mathbf{P} \tfrac{\partial \mathbf{V}}{\partial \alpha_{fi}} \mathbf{P} \tfrac{\partial \mathbf{V}}{\partial \alpha_{gk}} \mathbf{P}\mathbf{y} = \\
&\quad \mathrm{vec}\left(\boldsymbol{\Phi}^2 \widetilde{\mathbf{W}} \mathbf{M}_{fi}\right)^\top \mathbf{D}^{-1} \mathrm{vec}\left(\boldsymbol{\Phi}^2 \widetilde{\mathbf{W}} \mathbf{M}_{gk}\right) - \mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{M}_{fi}}\right)^\top \mathbf{B} \mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{M}_{gk}}\right) \ldots \\
&\quad -\mathrm{vec}(\mathbf{T})^\top \mathbf{B} \mathrm{vec}\left(\widetilde{\mathbf{Z}}^\top \left(\mathbf{E} \circ \left[\boldsymbol{\Phi}^2 \left\{\mathbf{E} \circ \left(\boldsymbol{\Phi}^2 \widetilde{\mathbf{W}} \mathbf{M}_{gk}\right)\right\} \mathbf{M}_{fi}\right]\right)\right) \ldots \\
&\quad -\mathrm{vec}(\mathbf{T})^\top \mathbf{B} \mathrm{vec}\left(\widetilde{\mathbf{Z}}^\top \left(\mathbf{E} \circ \left[\boldsymbol{\Phi}^2 \left\{\mathbf{E} \circ \left(\boldsymbol{\Phi}^2 \widetilde{\mathbf{W}} \mathbf{M}_{fi}\right)\right\} \mathbf{M}_{gk}\right]\right)\right) \ldots \\
&\quad +\mathrm{vec}(\mathbf{T})^\top \mathbf{B} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{M}_{fi} \otimes \boldsymbol{\Phi}^2\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{M}_{gk}}\right) \ldots \\
&\quad +\mathrm{vec}(\mathbf{T})^\top \mathbf{B} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{M}_{gk} \otimes \boldsymbol{\Phi}^2\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{M}_{fi}}\right) \ldots \\
&\quad +\mathrm{vec}(\mathbf{T})^\top \mathbf{B} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{M}_{fi} \otimes \boldsymbol{\Phi}^2\right) \mathbf{D}^{-1} \ldots \\
&\qquad \left(\mathbf{M}_{gk} \otimes \boldsymbol{\Phi}^2\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \mathrm{vec}(\mathbf{T}) \ldots \\
&\quad -\mathrm{vec}(\mathbf{T})^\top \mathbf{B} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{M}_{fi} \otimes \boldsymbol{\Phi}^2\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \ldots \\
&\qquad \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{M}_{gk} \otimes \boldsymbol{\Phi}^2\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \mathrm{vec}(\mathbf{T}),
\end{aligned}
$$

$$
\begin{aligned}
&\mathbf{y}^\top \mathbf{P} \tfrac{\partial \mathbf{V}}{\partial \eta_{fi}} \mathbf{P} \tfrac{\partial \mathbf{V}}{\partial \eta_{gk}} \mathbf{P}\mathbf{y} = \\
&\quad \mathrm{vec}\left(\widetilde{\mathbf{W}} \mathbf{L}_{fi}\right)^\top \mathbf{D}^{-1} \mathrm{vec}\left(\widetilde{\mathbf{W}} \mathbf{L}_{gk}\right) - \mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{L}_{fi}}\right)^\top \mathbf{B} \mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{L}_{gk}}\right) \ldots \\
&\quad -\mathrm{vec}(\mathbf{T})^\top \mathbf{B} \mathrm{vec}\left(\widetilde{\mathbf{Z}}^\top \left(\mathbf{E} \circ \left[\left\{\mathbf{E} \circ \left(\widetilde{\mathbf{W}} \mathbf{L}_{gk}\right)\right\} \mathbf{L}_{fi}\right]\right)\right) \ldots \\
&\quad -\mathrm{vec}(\mathbf{T})^\top \mathbf{B} \mathrm{vec}\left(\widetilde{\mathbf{Z}}^\top \left(\mathbf{E} \circ \left[\left\{\mathbf{E} \circ \left(\widetilde{\mathbf{W}} \mathbf{L}_{fi}\right)\right\} \mathbf{L}_{gk}\right]\right)\right) \ldots \\
&\quad +\mathrm{vec}(\mathbf{T})^\top \mathbf{B} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{L}_{fi} \otimes \mathbf{I}_N\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{L}_{gk}}\right) \ldots \\
&\quad +\mathrm{vec}(\mathbf{T})^\top \mathbf{B} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{L}_{gk} \otimes \mathbf{I}_N\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \mathrm{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{L}_{fi}}\right) \ldots \\
&\quad +\mathrm{vec}(\mathbf{T})^\top \mathbf{B} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{L}_{fi} \otimes \mathbf{I}_N\right) \mathbf{D}^{-1} \ldots \\
&\qquad \left(\mathbf{L}_{gk} \otimes \mathbf{I}_N\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \mathrm{vec}(\mathbf{T}) \ldots \\
&\quad -\mathrm{vec}(\mathbf{T})^\top \mathbf{B} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{L}_{fi} \otimes \mathbf{I}_N\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \ldots \\
&\qquad \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{L}_{gk} \otimes \mathbf{I}_N\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \mathrm{vec}(\mathbf{T}),
\end{aligned}
$$

$$\mathbf{y}^\top \mathbf{P} \tfrac{\partial \mathbf{V}}{\partial \alpha_{fi}} \mathbf{P} \tfrac{\partial \mathbf{V}}{\partial \eta_{gk}} \mathbf{P}\mathbf{y} =$$

$$\begin{aligned}
&\operatorname{vec}\left(\boldsymbol{\Phi}^2 \widetilde{\mathbf{W}} \mathbf{M}_{fi}\right)^\top \mathbf{D}^{-1} \operatorname{vec}\left(\widetilde{\mathbf{W}} \mathbf{L}_{gk}\right) - \operatorname{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{M}_{fi}}\right)^\top \mathbf{B} \operatorname{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{L}_{gk}}\right) \dots \\
&-\operatorname{vec}(\mathbf{T})^\top \mathbf{B} \operatorname{vec}\left(\widetilde{\mathbf{Z}}^\top \left(\mathbf{E} \circ \left[\boldsymbol{\Phi}^2 \left\{\mathbf{E} \circ \left(\widetilde{\mathbf{W}} \mathbf{L}_{gk}\right)\right\} \mathbf{M}_{fi}\right]\right)\right) \dots \\
&-\operatorname{vec}(\mathbf{T})^\top \mathbf{B} \operatorname{vec}\left(\widetilde{\mathbf{Z}}^\top \left(\mathbf{E} \circ \left[\left\{\mathbf{E} \circ \left(\boldsymbol{\Phi}^2 \widetilde{\mathbf{W}} \mathbf{M}_{fi}\right)\right\} \mathbf{L}_{gk}\right]\right)\right) \dots \\
&+\operatorname{vec}(\mathbf{T})^\top \mathbf{B} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{M}_{fi} \otimes \boldsymbol{\Phi}^2\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \operatorname{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{L}_{gk}}\right) \dots \\
&+\operatorname{vec}(\mathbf{T})^\top \mathbf{B} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{L}_{gk} \otimes \mathbf{I}_N\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \operatorname{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{M}_{fi}}\right) \dots \\
&+\operatorname{vec}(\mathbf{T})^\top \mathbf{B} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{M}_{fi} \otimes \boldsymbol{\Phi}^2\right) \mathbf{D}^{-1} \dots \\
&\quad \left(\mathbf{L}_{gk} \otimes \mathbf{I}_N\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \operatorname{vec}(\mathbf{T}) \dots \\
&-\operatorname{vec}(\mathbf{T})^\top \mathbf{B} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{M}_{fi} \otimes \boldsymbol{\Phi}^2\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \dots \\
&\quad \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{L}_{gk} \otimes \mathbf{I}_N\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \operatorname{vec}(\mathbf{T}), \text{ and}
\end{aligned}$$

$$\mathbf{y}^\top \mathbf{P} \tfrac{\partial \mathbf{V}}{\partial \eta_{fi}} \mathbf{P} \tfrac{\partial \mathbf{V}}{\partial \alpha_{gk}} \mathbf{P}\mathbf{y} =$$

$$\begin{aligned}
&\operatorname{vec}\left(\widetilde{\mathbf{W}} \mathbf{L}_{fi}\right)^\top \mathbf{D}^{-1} \operatorname{vec}\left(\boldsymbol{\Phi}^2 \widetilde{\mathbf{W}} \mathbf{M}_{gk}\right) - \operatorname{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{L}_{fi}}\right)^\top \mathbf{B} \operatorname{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{M}_{gk}}\right) \dots \\
&-\operatorname{vec}(\mathbf{T})^\top \mathbf{B} \operatorname{vec}\left(\widetilde{\mathbf{Z}}^\top \left(\mathbf{E} \circ \left[\left\{\mathbf{E} \circ \left(\boldsymbol{\Phi}^2 \widetilde{\mathbf{W}} \mathbf{M}_{gk}\right)\right\} \mathbf{L}_{fi}\right]\right)\right) \dots \\
&-\operatorname{vec}(\mathbf{T})^\top \mathbf{B} \operatorname{vec}\left(\widetilde{\mathbf{Z}}^\top \left(\mathbf{E} \circ \left[\boldsymbol{\Phi}^2 \left\{\mathbf{E} \circ \left(\widetilde{\mathbf{W}} \mathbf{L}_{fi}\right)\right\} \mathbf{M}_{gk}\right]\right)\right) \dots \\
&+\operatorname{vec}(\mathbf{T})^\top \mathbf{B} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{L}_{fi} \otimes \mathbf{I}_N\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \operatorname{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{M}_{gk}}\right) \dots \\
&+\operatorname{vec}(\mathbf{T})^\top \mathbf{B} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{M}_{gk} \otimes \boldsymbol{\Phi}^2\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \operatorname{vec}\left(\widetilde{\mathbf{T}}_{\mathbf{L}_{fi}}\right) \dots \\
&+\operatorname{vec}(\mathbf{T})^\top \mathbf{B} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{L}_{fi} \otimes \mathbf{I}_N\right) \mathbf{D}^{-1} \dots \\
&\quad \left(\mathbf{M}_{gk} \otimes \boldsymbol{\Phi}^2\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \operatorname{vec}(\mathbf{T}) \dots \\
&-\operatorname{vec}(\mathbf{T})^\top \mathbf{B} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{L}_{fi} \otimes \mathbf{I}_N\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \dots \\
&\quad \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}^\top\right) \mathbf{D}^{-1} \left(\mathbf{M}_{gk} \otimes \boldsymbol{\Phi}^2\right) \mathbf{D}^{-1} \left(\mathbf{I}_P \otimes \widetilde{\mathbf{Z}}\right) \mathbf{B} \operatorname{vec}(\mathbf{T}).
\end{aligned}$$

Summarizing, we now have efficient expressions for (a) $\log|\mathbf{V}|$, (b) $\log\left|\mathbf{R}^\top\mathbf{V}^{-1}\mathbf{R}\right|$, (c) $\mathbf{y}^\top\mathbf{P}\mathbf{y}$, (d) $\mathrm{tr}\left(\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta}\right)$, (e) $\mathrm{tr}\left(\mathbf{P}\frac{\partial^2\mathbf{V}}{\partial\theta\partial\phi}\right)$, (f) $\mathbf{y}^\top\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{P}\mathbf{y}$, (g) $\mathbf{y}^\top\mathbf{P}\frac{\partial^2\mathbf{V}}{\partial\theta\partial\phi}\mathbf{P}\mathbf{y}$, and (h) $\mathbf{y}^\top\mathbf{P}\frac{\partial\mathbf{V}}{\partial\theta}\mathbf{P}\frac{\partial\mathbf{V}}{\partial\phi}\mathbf{P}\mathbf{y}$, where (a–c) are needed for computing the MacGREML log-likelihood, (d) and (f) for computing the gradient of the log-likelihood, and (e), (g), and (h) for computing the AI matrix.

6.6. UNBALANCED DATA

In case data is unbalanced (i.e., not all phenotypes are measured in all respondents), the mathematical complexity of REML estimation increases. To illustrate this complexity, consider the full $(NP)\times 1$ phenotype vector $\mathbf{y}$, associated variance matrix $\mathbf{V}$, and matrix of covariates $\mathbf{Z}$. Missing data, in this framework, means that we select the subset of rows of $\mathbf{Z}$ and $\mathbf{y}$, as well as rows and columns from $\mathbf{V}$, for which both the phenotypic as well as data on the controls is available.

Letting M denote the total number of missing values across phenotypes and $\mathbf{S}$, an $(NP-M)\times(NP)$ design matrix, consisting of zeros and ones, with precisely a single one per row, and at most a single one per column, such that $\mathbf{S}\mathbf{S}^\top=\mathbf{I}_{NP-M}$. Let

$$\begin{aligned}\mathbf{y}^* &= \mathbf{S}\mathbf{y}\\ \mathbf{V}^* &= \mathbf{S}\mathbf{V}\mathbf{S}^\top\\ \mathbf{Z}^* &= \mathbf{S}(\mathbf{I}_P\otimes\mathbf{Z}) = \begin{pmatrix}\mathbf{P}_1^* & \mathbf{P}_0^*\end{pmatrix}\begin{pmatrix}\boldsymbol{\Theta}\\ \mathbf{0}\end{pmatrix}\mathbf{Q}^{*\top},\end{aligned}$$

where the last expression constitutes the singular value decomposition of $\mathbf{Z}^*$, such that $\mathbf{P}_0^{*\top}\mathbf{Z}^*=\mathbf{0}$. We define $\mathbf{K}^*=\mathbf{P}_0^{*\top}$. Letting $r=\mathrm{rank}(\mathbf{Z}^*)$ denote the rank of $\mathbf{Z}^*$, then $\mathrm{rank}\left(\mathbf{P}_0^*\right)=NP-M-r$.

Now the REML log-likelihood with missing data is given by

$$l(\boldsymbol{\theta}) = -\frac{1}{2}\log\left(\left|\mathbf{K}^*\mathbf{S}\mathbf{V}\mathbf{S}^\top\mathbf{K}^{*\top}\right|\right) - \frac{1}{2}\mathbf{y}^\top\mathbf{S}^\top\mathbf{K}^{*\top}\left(\mathbf{K}^*\mathbf{S}\mathbf{V}\mathbf{S}^\top\mathbf{K}^{*\top}\right)^{-1}\mathbf{K}^*\mathbf{S}\mathbf{y}.$$

In previous sections, we have been able to use the identity by Casella and Searle (1985) and Searle et al. (1992) to rewrite expressions such as $\mathbf{K}^\top\left(\mathbf{K}\mathbf{V}\mathbf{K}^\top\right)^{-1}\mathbf{K}$ in terms of a projection matrix, from which we derived computationally efficient expressions for the gradient and AI matrix. However, at a first glance, there exists no equivalent expression for

$\mathbf{S}^\top \mathbf{K}^{*\top} \left(\mathbf{K}^* \mathbf{S} \mathbf{V} \mathbf{S}^\top \mathbf{K}^{*\top}\right)^{-1} \mathbf{K}^* \mathbf{S}$, such that our derivations hold.

However, we posit that $\mathbf{K}^*\mathbf{S}$ in its entirety can be considered as a design matrix, such that it is orthogonal to $\mathbf{Z}$ (including the missing observations, coded to have an arbitrary value; e.g., zero) and a set of M dummies that code for the missing observations, which leads to an identity such that $\mathbf{S}^\top \mathbf{K}^{*\top} \left(\mathbf{K}^* \mathbf{S} \mathbf{V} \mathbf{S}^\top \mathbf{K}^{*\top}\right)^{-1} \mathbf{K}^* \mathbf{S}$ can also be expressed in terms of a projection matrix for which we have an efficient expression.

Formalizing, we have a $(NP) \times (PC)$ matrix of confounders $(\mathbf{I}_P \otimes \mathbf{Z})$, a $(NP-M) \times (NP)$ selection matrix $\mathbf{S}$, such that element i, j of matrix $\mathbf{S}$, denoted by S_{ij}, is either zero or one, and such that row-sum $\sum_{j=1}^{NP} S_{ij} = 1 \, \forall \, i$ and column-sum $\sum_{i=1}^{(NP-M)} S_{ij} \in \{0,1\} \, \forall \, j$, and we have an $(NP - M) \times (NP - M - r)$ matrix $\mathbf{P}_0^*$ with orthonormal columns (i.e., $\mathbf{P}_0^{*\top} \mathbf{P}_0^* = \mathbf{I}_{NP-M-r}$) that lie in the null-space of $(NP - M) \times (PC)$ matrix $\mathbf{Z}^* = \mathbf{S}(\mathbf{I}_P \otimes \mathbf{Z})$, such that $\mathbf{P}_0^{*\top} \mathbf{Z}^* = \mathbf{0}_{(NP-M-r)\times(PC)}$, and where $r = \text{rank}(\mathbf{Z}^*)$. Finally, $\mathbf{M}$ is an $M \times (NP)$ matrix, defined analogously to $\mathbf{S}$, in such a manner that it selects missing observations rather than non-missing observations as $\mathbf{S}$ does. We can show that $\mathbf{S}\mathbf{S}^\top = \mathbf{I}_{NP-M}$, $\mathbf{M}\mathbf{M}^\top = \mathbf{I}_M$, $\mathbf{S}\mathbf{M}^\top = \mathbf{0}_{(NP-M)\times M}$, $\mathbf{M}\mathbf{S}^\top = \mathbf{0}_{M\times(NP-M)}$, $\mathbf{S}^\top\mathbf{S} + \mathbf{M}^\top\mathbf{M} = \mathbf{I}_{NP}$.

Theorem 1. *$\widetilde{\boldsymbol{P}_0} = \boldsymbol{S}^\top \boldsymbol{P}_0^*$ lies in the null-space of $\boldsymbol{Z}_M = \left[(\boldsymbol{I}_P \otimes \boldsymbol{Z}), \, \boldsymbol{M}^\top\right]$ and has rank $\left(\widetilde{\boldsymbol{P}_0}\right) = \text{rank}\left(\boldsymbol{P}_0^*\right) \equiv NP - M - r$.*

Proof.

$$\begin{aligned}
\widetilde{\mathbf{P}_0}^\top \mathbf{Z}_M &= \mathbf{P}_0^{\top *} \mathbf{S} \left[(\mathbf{I}_P \otimes \mathbf{Z}), \, \mathbf{M}^\top\right] \\
&= \left[\mathbf{P}_0^{\top *} \mathbf{S}(\mathbf{I}_P \otimes \mathbf{Z}), \, \mathbf{P}_0^{\top *} \mathbf{S}\mathbf{M}^\top\right] \\
&= \left[\mathbf{P}_0^{\top *} \mathbf{Z}^*, \, \mathbf{P}_0^{\top *} \mathbf{0}_{(NP-M)\times M}\right] \\
&= \mathbf{0}_{(NP-M-r)\times(PC+M)}.
\end{aligned}$$

Hence, $\widetilde{\mathbf{P}_0}$ lies in the null-space of $\mathbf{Z}_M$.

$$\begin{aligned}
\text{rank}\left(\widetilde{\mathbf{P}_0}\right) &= \text{rank}\left(\widetilde{\mathbf{P}_0}^\top \widetilde{\mathbf{P}_0}\right) \\
&= \text{rank}\left(\mathbf{P}_0^{*\top} \mathbf{S}\mathbf{S}^\top \mathbf{P}_0^*\right) \\
&= \text{rank}\left(\mathbf{P}_0^{*\top} \mathbf{P}_0^*\right) \\
&= \text{rank}\left(\mathbf{P}_0^*\right) \equiv NP - M - r.
\end{aligned}$$

□

Theorem 2. *rank* $(\mathbf{Z}_M) = r + M$ *and, therefore,* $\widetilde{\boldsymbol{P}_0}$ *spans the null space of* $\mathbf{Z}_M$.

Proof. rank $(\mathbf{Z}_M)$ is the number of independent columns $\mathbf{Z}_M$. Hence, orthogonalizing columns of $\mathbf{Z}_M$ with respect to each other does not change the rank. Letting $\mathbf{I} - \mathbf{M}^\top \left(\mathbf{M}\mathbf{M}^\top\right)^{-1}\mathbf{M} = \mathbf{I} - \mathbf{M}^\top\mathbf{M} = \mathbf{S}^\top\mathbf{S}$ denote the orthogonal projection matrix removing the collinearity with columns of $\mathbf{M}^\top$, we have – based on the orthogonalization-argument – that,

$$\begin{aligned}
\text{rank}(\mathbf{Z}_M) &= \text{rank}\left(\left[(\mathbf{I}_P \otimes \mathbf{Z}),\ \mathbf{M}^\top\right]\right) \\
&= \text{rank}\left(\left[\mathbf{S}^\top\mathbf{S}(\mathbf{I}_P \otimes \mathbf{Z}),\ \mathbf{M}^\top\right]\right) \\
&= \text{rank}\left(\mathbf{S}^\top\mathbf{S}(\mathbf{I}_P \otimes \mathbf{Z})\right) + \text{rank}\left(\mathbf{M}^\top\right) \\
&= \text{rank}\left(\left(\mathbf{I}_P \otimes \mathbf{Z}^\top\right)\mathbf{S}^\top\mathbf{S}\mathbf{S}^\top\mathbf{S}(\mathbf{I}_P \otimes \mathbf{Z})\right) + \text{rank}\left(\mathbf{M}\mathbf{M}^\top\right) \\
&= \text{rank}\left(\left(\mathbf{I}_P \otimes \mathbf{Z}^\top\right)\mathbf{S}^\top\mathbf{S}(\mathbf{I}_P \otimes \mathbf{Z})\right) + \text{rank}(\mathbf{I}_M) \\
&= \text{rank}(\mathbf{S}(\mathbf{I}_P \otimes \mathbf{Z})) + M \\
&= \text{rank}\left(\mathbf{Z}^*\right) + M = r + M.
\end{aligned}$$

Given that $\mathbf{Z}_M$ is $(NP) \times (PC + M)$, where $(NP) \gg PC + M \geq r + M$, and has rank equal to $r + M$, its null space is spanned by $NP - M - r$ independent columns. □

The preceding two theorems show that $\widetilde{\mathbf{P}_0^\top} = \mathbf{K}^*\mathbf{S}$ can be regarded as a design matrix, for which the following identity holds

$$\mathbf{S}^\top\mathbf{K}^{*\top}\left(\mathbf{K}^*\mathbf{S}\mathbf{V}\mathbf{S}^\top\mathbf{K}^{*\top}\right)^{-1}\mathbf{K}^*\mathbf{S} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{Z}_M\left(\mathbf{Z}_M^\top\mathbf{V}^{-1}\mathbf{Z}_M\right)^{-1}\mathbf{Z}_M^\top\mathbf{V}^{-1},$$

where $\mathbf{Z}_M = \left[(\mathbf{I}_P \otimes \mathbf{Z}),\ \mathbf{M}^\top\right]$. Hence, MacGREML estimation for missing data can be performed by treating the data as balanced, and including dummy variables for missing values as fixed-effect covariates.

The downside of this approach is that the matrix of fixed-effects covariates can no longer be represented as a Kronecker product. Therefore, we cannot directly apply the computationally efficient expressions that we derived under the assumption of balanced data. However, we do not deem the task of finding more general computationally efficient expressions impossible. Hence, this open end may prove to be an interesting venue for further research.

Finally, in case the phenotypic data is strongly unbalanced (i.e., $M \gg PC$) the missing observations start to dominate matrices such as

$\left(\mathbf{Z}_M^\top \mathbf{V}^{-1} \mathbf{Z}_M\right)^{-1}$. A large number of missing values could therefore even render a more general MacGREML estimation method computationally infeasible.

6.7. SUMMARY

In this chapter, we presented a multivariate average-information constrained GREML (MacGREML) estimation method. This method consists of an iterative procedure, based on a Newton-Raphson algorithm, to obtain unbiased estimates of the parameters of a multivariate SNP-based LMM for balanced data on P phenotypes, observed for N individuals. The LMM has been parametrized such that the $(NP) \times (NP)$ phenotypic variance matrix is positive definite, irrespective of starting values and updates of the estimates throughout the iterations. Since a naïve approach would require the computation and inversion of full $(NP) \times (NP)$ matrices in each iteration, such an approach would become computationally infeasible as N and P grow large. Therefore, we have rewritten the log-likelihood, gradient, and average-information matrix in terms of the eigendecomposition of an $N \times N$ genomic-relatedness matrix $\mathbf{A}$ and transformations of $P \times P$ matrices of parameters. By combining various computationally efficient expressions, we are able to provide a MacGREML estimation method which, in terms of computational complexity, is of the order N^3. Moreover, the order does not depend on the number of phenotypes. Therefore, the MacGREML estimation method we propose can – in theory – be applied in a large set of phenotypes, provided the data are balanced and we consider only one GRM. In addition, we demonstrate that it might be possible to deal with a low degree of missing data without losing too much computational efficiency. Investigating whether this is truly possible within our framework may prove to be an interesting venue for further research. Finally, our parametrization is such that we can easily impose two basic factor restrictions. The first factor restriction is such that all traits have a perfect genetic correlation and the second factor restriction imposes that traits have no genetic correlation. The significance of the additional fit of the saturated model we employ, compared to the two restricted versions, can easily be tested using a likelihood-ratio test. Finally, analyses using simulations and real data are needed to assess the overall empirical merits of this method.

7

LD-Score-Regression Intercept in Individual-Level Data

Work in progress by De Vlaming and Visscher

ABSTRACT

Summary-statistics-based methods, such as LD-score regression, use GWAS results in order to disentangle the contribution of polygenic signal (i.e., SNP-based heritability) and population stratification to the inflation of the inferred χ^2-test statistics. Other methods for estimating SNP heritability, such as a Haseman-Elston regression and genomic-relatedness restricted-maximum-likelihood estimation, use individual-level data instead. Given the difference in data used by summary-statistics-based methods and individual-level-data methods, these two classes of methods are often considered to be profoundly different from one another. However, in recent work, the LD-score-regression estimate of SNP-based heritability has been shown to be equivalent to the SNP heritability inferred by a Haseman-Elston regression, provided population stratification is absent. This observation raises the question whether this equivalence can be extended to include the so-called LD-score-regression intercept, which measures the contribution of confounding stratification to the inflation in the GWAS χ^2-test statistics. We show that, in an admixed sample drawing from two discrete populations, the theoretical LD-score-regression intercept can be estimated using individual-level data (i.e., using the phenotype vector and leading principal component from the genomic-relatedness matrix). Using simulations we show that our estimator of the LD-score-regression intercept is approximately unbiased. Moreover, we posit the conjecture that under more complex forms of stratification (i.e., with $P > 2$ discrete populations) an equivalence principle holds for the LD-score-regression intercept and a transformation of the estimates of a regression using individual-level data.

7.1. INTRODUCTION

Population stratification can confound genome-wide association study (GWAS) summary statistics, as such stratification inflates the expected χ^2-test statistics from a GWAS (e.g., Bulik-Sullivan et al. 2015b). The 'LD-score-regression' method by Bulik-Sullivan et al. (2015b) incorporates a parameter, which they refer to as the intercept, that accounts for confounding stratification permeating GWAS summary statistics. By including linkage-disequilibrium (LD) scores as regressor, this method is able to disentangle the contribution of population admixture and the contribution of polygenic signal to the inferred χ^2-test statistics. Hence, LD-score regression is able to quantify the contribution of stratification to the inferred GWAS summary statistics.

Population stratification can also bias estimates of variance components (Browning and Browning, 2011) in a linear mixed model (LMM), as admixture affects the inferred genetic relatedness between individuals (e.g., Thornton et al. 2012, Conomos et al. 2016) and, thereby, relatedness matrices and their eigenvalues (Bryc et al., 2013). By including the leading principal components (PC) from the genomic-relatedness matrix (GRM; inferred e.g., using `GCTA`; Yang et al. 2011a, or `PLINK`; Purcell et al. 2007, Chang et al. 2015) as fixed-effect covariates in the restricted-maximum-likelihood (REML) estimation one can correct – at least partially – for the confounding effects of stratification on the inferred variance components (Browning and Browning, 2011).

Both LD-score regression and genomic-relatedness-matrix (GRM) REML or – in short – GREML estimation can be used to infer SNP-based heritability (h^2_{SNP}) and both methods account for LD (Bulik-Sullivan et al., 2015b, Yang et al., 2016). In fact, Bulik-Sullivan (2015) shows the equivalence of Haseman-Elston regression (Haseman and Elston 1972; which can be regarded as a simplified form of REML) and LD-score regression for estimating h^2_{SNP}, when population stratification is absent.

However, the analogy between summary-statistics-based methods, such as LD-score regression, and methods using individual-level data, such as GREML and Haseman-Elston regression, seems to break down when considering a parameter that reflects the amount of population stratification present in the data. Whereas LD-score regression has a parameter called the intercept – such that an affine transformation of

this parameter reflects the contribution of population stratification – an equivalent parameter is seemingly absent in methods such as GREML and Haseman-Elston regression. Hence, we raise the question whether we can use individual-level data (e.g., in GREML estimation) to directly estimate a parameter that is at least approximately equivalent to the LD-score-regression intercept and, therefore, also quantifies the amount of population stratification present in the data at hand.

Under the same type of population stratification considered by Bulik-Sullivan et al. (2015b) in their theoretical derivations where there are two discrete populations (described in Section 7.2), we derive an unconditional expected GRM (Section 7.3). Using the expected GRM, we are able to obtain an explicit eigendecomposition of this matrix (Section 7.4). We show that all eigenvalues – except the leading eigenvalue – are decreased by the same small amount due to stratification, whereas the leading eigenvalue is strongly increased by stratification. Analogous to LD-score regression, where the expected increase in χ^2-statistics due to stratification is proportional to sample size, the increase in the first eigenvalue of the GRM is also proportional to the sample size. Based on the eigendecomposition, we show that by including the leading PC as fixed-effect covariate in REML estimation, the corresponding eigenvalue will cease to affect the likelihood. Hence, we posit that under the type of stratification subsumed in LD-score regression, inclusion of the first PC of the GRM as fixed-effect covariate will fully account for that type of stratification, and, thereby, ensure REML estimates of h^2_{SNP} are not upwards biased as a result of the population structure.

In addition, in Section 7.5 we derive an explicit transformation of the estimated association between the first PC and the phenotype. This transformation provides an estimate of the LD-score-regression intercept. Using simulations (Section 7.6) we show that the theoretical LD-score-regression intercept and the proposed transformation of the estimated association of the first PC with the phenotype are equivalent.

As discussed in Section 7.7, further research is needed to assess whether this equivalence also persists (i) when comparing actual LD-score-regression estimates of the intercept to estimates from individual-level data and (ii) in empirical data where stratification is of a more complex nature than both our theory and the theory underlying LD-score regression consider. Specifically, further research should also

focus on cases where there are more than two discrete populations.

7.2. STRATIFICATION IN LD-SCORE REGRESSION

Bulik-Sullivan et al. (2015b) theoretically incorporate stratification by considering a GWAS sample consisting of individuals drawn from two independent populations, with allele frequencies differing across these populations due to drift. Specifically, using slightly modified notation, they assume

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{s} + \boldsymbol{\varepsilon}, \\ \boldsymbol{\varepsilon} &\sim \mathcal{N}\left(\mathbf{0}, \sigma_E^2 \mathbf{I}_{N_T}\right), \\ \boldsymbol{\beta} &\sim \mathcal{N}\left(\mathbf{0}, \tfrac{\sigma_A^2}{P}\mathbf{I}_P\right), \\ \{\mathbf{s}\}_i &= \begin{cases} \frac{\sigma_s}{2}, & i \in \mathscr{P}_1, \\ -\frac{\sigma_s}{2}, & i \in \mathscr{P}_2, \end{cases} \end{aligned} \tag{7.1}$$

where $\mathscr{P}_1$ and $\mathscr{P}_2$ denote the sets of individuals drawn from Populations 1 and 2, respectively. Moreover, they assume that $|\mathscr{P}_1| = |\mathscr{P}_2| = N$. That is, there are N individuals drawn from both populations, yielding $N_T = 2N$ observations in total. In this model, parameter a, found in Equation 2.14 of the Supplementary Note to Bulik-Sullivan et al. (2015b), is approximately equal to σ_s^2 (i.e., the squared difference between the two populations in the phenotypic mean). In the above expression, $\mathbf{X}$ denotes an $N_T \times P$ matrix of P standardized SNPs, with random effects in vector $\boldsymbol{\beta}$. In line with Yang et al. (2011a), Bulik-Sullivan et al. (2015b) impose the assumption that effects of standardized SNPs, $\boldsymbol{\beta}$, are independent draws from a normal homoskedastic distribution. In addition, Bulik-Sullivan et al. (2015b) assume that the effects in vector $\boldsymbol{\beta}$ are constant across the two populations. We follow these assumptions. Vector $\boldsymbol{\varepsilon}$ denotes the environmental effects. Finally, σ_A^2 denotes the additive genetic variance and σ_E^2 the environment variance.

7.2.1. *Genetic drift*

Let f_1 and f_2 denote the allele frequency for a given SNP in Populations 1 and 2, and $\overline{f} = (f_1 + f_2)/2$ the average allele frequency across the two populations. For now, assume that all distributions, expectations, and variances are conditional on f_1, f_2, and, thereby, on $\overline{f}$. The additively-

coded genotype for individual i (denoted by $g_i \in \{0,1,2\}$) then satisfies the following properties

$$g_i \mid i \in \mathscr{P}_1 \sim \text{Binom}(2, f_1) \text{ and}$$
$$g_i \mid i \in \mathscr{P}_2 \sim \text{Binom}(2, f_2).$$

Therefore,

$$\begin{aligned}
\mathbb{E}[g_i \mid i \in \mathscr{P}_1] &= 2f_1, \\
\mathbb{E}[g_i \mid i \in \mathscr{P}_2] &= 2f_2, \\
\text{Var}(g_i \mid i \in \mathscr{P}_1) &= 2f_1(1-f_1), \text{ and} \\
\text{Var}(g_i \mid i \in \mathscr{P}_2) &= 2f_2(1-f_2).
\end{aligned}$$

Moreover, as indicated, individuals are sampled from the two populations with equal chance. Hence, g_i is a draw from a mixture distribution with mean $\mathbb{E}[g_i] = 2\overline{f}$ and variance

$$\begin{aligned}
\text{Var}(g_i) &= \mathbb{E}[\text{Var}(g_i \mid i \in \mathscr{P})] + \text{Var}(\mathbb{E}[g_i \mid i \in \mathscr{P}]) \\
&= f_1(1-f_1) + f_2(1-f_2) + \frac{1}{2}\left(2f_1 - 2\overline{f}\right)^2 + \frac{1}{2}\left(2f_2 - 2\overline{f}\right)^2 \\
&= f_1(1-f_1) + f_2(1-f_2) + (f_1 - f_2)^2 \\
&= f_1 + f_2 - 2f_1f_2.
\end{aligned}$$

Bulik-Sullivan et al. (2015b) assume genotypes are standardized, such that the standardized genotype for individual i (denoted by x_i) satisfies the following properties

$$\mathbb{E}[x_i] = 0 \quad \text{and} \quad \text{Var}(x_i) = 1.$$

Therefore, they implicitly assume that

$$x_i = \frac{g_i - 2\overline{f}}{\sqrt{f_1 + f_2 - 2f_1f_2}},$$

where $2\overline{f}$ and $f_1 + f_2 - 2f_1f_2$ are the theoretical expectation and variance, respectively, of a random variable that is drawn from the aforementioned mixture distribution (i.e., 50% chance of a draw from a Binom$(2, f_1)$ distribution and 50% chance of a draw from Binom$(2, f_2)$).

Under this scenario, we have

$$\mathbb{E}[x_i \mid i \in \mathscr{P}_1] = -\mathbb{E}[x_i \mid i \in \mathscr{P}_2] = \frac{2(f_1 - \overline{f})}{\sqrt{f_1 + f_2 - 2f_1 f_2}}.$$

Now, treating f_1 and f_2 as random, and conditioning only on the pooled frequency $\overline{f}$ (also making the conditioning explicit from this point onwards), Bulik-Sullivan et al. (2015b) implicitly assume that

$$\frac{2(f_1 - \overline{f})}{\sqrt{f_1 + f_2 - 2f_1 f_2}} | \overline{f} \sim \mathscr{N}(0, F_{ST}),$$

where F_{ST} denotes Wright's F-statistic (Wright, 1949), which measures the amount of genetic drift across populations. The larger F_{ST}, the larger – on average – the allele frequency differences across populations will be.

Provided allele frequencies do not vary too much across the two populations, we note that $2\overline{f}(1-\overline{f}) \approx f_1 + f_2 - 2f_1 f_2$. For instance, when the difference in allele frequencies across the two populations of a given SNP is 10% and the pooled frequency is 50%, we have $2\overline{f}(1-\overline{f}) = 0.5$ and $f_1 + f_2 - 2f_1 f_2 = 0.505$. Thus, even under this substantial allele-frequency difference, the relative difference between the true variance, given by $f_1 + f_2 - 2f_1 f_2$, and the variance as inferred by $2\overline{f}(1-\overline{f})$ is only 1%.

Therefore, we make a slight amendment, by assuming

$$\frac{2(f_1 - \overline{f})}{\sqrt{2\overline{f}(1-\overline{f})}} | \overline{f} \sim \mathscr{N}(0, F_{ST}).$$

The reason for making this minor change is that it greatly simplifies our algebra in writing down the expected GRM. More precisely, the GRM (e.g., constructed using `GCTA`; Yang et al. 2011a) assumes each SNP is binomially distributed (i.e., in Hardy-Weinberg equilibrium; HWE). Therefore, each SNP is standardized according to its pooled – potentially admixed – allele frequency ($\overline{f}$) assuming HWE (i.e., in addition to correcting for the expected value of the raw genotype, given by $2\overline{f}$, the raw genotypes are also divided by $\sqrt{2\overline{f}(1-\overline{f})}$). By replacing $f_1 + f_2 - 2f_1 f_2$ by $2\overline{f}(1-\overline{f})$ in the preceding expression, when deriving the unconditional expectation of the GRM, the denominator in the

left-hand side of this expression and the standardizing coefficient of the SNP, when computing the GRM, cancel each other out.

Under the proposed change, the distribution of the difference in allele frequency between the first population and the pooled frequency can be written as

$$r \mid \overline{f} \sim \mathcal{N}\left(0, \frac{1}{2}\overline{f}(1-\overline{f})F_{ST}\right), \text{ where } r = f_1 - \overline{f}.$$

Hence, $\mathbb{E}\left[r \,\middle|\, \overline{f}\right] = 0$ and

$$\mathbb{E}\left[r^2 \,\middle|\, \overline{f}\right] = \text{Var}(r) = \frac{1}{2}\overline{f}(1-\overline{f})F_{ST}. \tag{7.2}$$

This expected squared difference between the population-specific and the pooled allele frequency is in line with the updated Nei estimator of F_{ST} reported by Bhatia et al. (2013), which is based on the work of Nei (1986). We would like to note here that the work of Bhatia et al. (2013) is also what Bulik-Sullivan et al. (2015b) point to when presenting the distribution of the standardized difference in allele frequency. Moreover, the preceding expression also aligns with an expression that Weir and Cockerham (1984) refer to as the "*most common explicit computational formula*" for F_{ST} (page 1361 of Weir and Cockerham 1984).

Interestingly, in this expression, Weir and Cockerham (1984) explicitly account for the loss of one degree of freedom across populations, when considering deviations from the pooled allele frequency. We should point out that in later work, the loss of this degree of freedom is ignored (Weir and Hill, 2002), which would correspond to

$$\mathbb{E}\left[r^2 \,\middle|\, \overline{f}\right] = \overline{f}(1-\overline{f})F_{ST},$$

which matches, for instance, with the approach adopted in the Supplementary Note to the work of Robinson et al. (2015). Rather than commenting on whether one should account for the lost degree of freedom or not, our focus should be to keep as close as possible to the approach adopted by Bulik-Sullivan et al. (2015b). Therefore, we assume that the variance of r is as given in Equation 7.2.

Although we have made the implicit distribution of $r \mid \overline{f}$ assumed by Bulik-Sullivan et al. (2015b) explicit, opposed to their work, without loss of comparability of our methods, we make no assumptions about

the precise distribution of $r \mid \overline{f}$. We only impose the condition that its expectation is zero and the variance as shown in Equation 7.2.

7.3. GENETIC RELATEDNESS UNDER STRATIFICATION

Without loss of generality, we can order the phenotype vector, $\mathbf{y}$, according to the populations from which the individuals are drawn (i.e., first the set of N individuals from Population 1 followed by the set of N individuals from Population 2). In addition, as indicated, we consider SNPs that are standardized according to cross-population allele frequencies. That is, for individual i and SNP k, the standardized genotype, from genotype matrix $\mathbf{X}$, is given by

$$\{\mathbf{X}\}_{ip} = x_{ip} = \frac{g_{ip} - 2\overline{f}_p}{\sqrt{2\overline{f}_p\left(1-2\overline{f}_p\right)}},$$

where $g_{ip} \in \{0,1,2\}$ denotes the additively-coded raw genotype.

Now we can rewrite the linear mixed model in Equation 7.1 in terms of variance components, the GRM, and differences in phenotypic mean across the two populations, as follows:

$$\mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\mu}, \sigma_A^2\mathbf{A} + \sigma_E^2\mathbf{I}_{N_T}\right) \text{ and}$$
$$\boldsymbol{\mu} = \frac{\sigma_s}{2}\begin{pmatrix} 1 \\ -1 \end{pmatrix} \otimes \boldsymbol{\iota}_N,$$

where $\mathbf{A} = P^{-1}\mathbf{X}\mathbf{X}^\top$ denotes the $N_T \times N_T$ GRM in the admixed sample, $\mathbf{I}_{N_T}$ the identity matrix of appropriate dimensions, and $\boldsymbol{\iota}_N$ as a column vector of ones, and where $N_T = 2N$ denotes the total sample size and N the number of individuals drawn from both populations.

7.3.1. *Conditonal genomic-relatedness matrix*

We now consider what the typical elements of a GRM look like under the stratification discussed in Section 7.2. In order to arrive at tractable expressions, we first derive the expected GRM based on a single SNP, conditional on the pooled and within-population allele frequencies (i.e.,

$\overline{f}, f_1$, and f_2), for a sample of four individuals (i.e., $N = 2$ individuals from each population). Subsequently, by applying the law of iterated expectations, we obtain the expected GRM independent of allele frequencies. Using this four-by-four expected unconditional GRM, we can generalize to an N_T-by-N_T GRM with N individuals from both populations.

As indicated, each SNP is standardized under the assumption of HWE. That is, the standardized genotype of individual i for a given SNP with pooled allele frequency $\overline{f}$, is given by

$$x_i = \frac{g_i - 2\overline{f}}{\sqrt{2\overline{f}(1-\overline{f})}}.$$

Under the aforementioned conditioning, a one-SNP GRM for four individuals, two from each population, can be written as

$$\mathbb{E}\left[\mathbf{A} \;\middle|\; \begin{array}{l} f_1, f_2, \overline{f}, \\ i, j \in \mathscr{P}_1, \\ k, l \in \mathscr{P}_2 \end{array}\right] = \mathbb{E}\left[\begin{pmatrix} x_i^2 & x_i x_j & x_i x_k & x_i x_l \\ x_i x_j & x_j^2 & x_j x_k & x_j x_l \\ x_i x_k & x_j x_k & x_k^2 & x_k x_l \\ x_i x_l & x_j x_l & x_k x_l & x_l^2 \end{pmatrix} \;\middle|\; \begin{array}{l} f_1, f_2, \overline{f}, \\ i, j \in \mathscr{P}_1, \\ k, l \in \mathscr{P}_2 \end{array}\right]$$

The expected standardized genotype value for $i \in \mathscr{P}_1$ is given by

$$\mathbb{E}\left[x_i \;\middle|\; i \in \mathscr{P}_1, f_1, f_2, \overline{f}\right] = \frac{f_1 - f_2}{\sqrt{2\overline{f}(1-\overline{f})}}.$$

The expected squared value for $i \in \mathscr{P}_1$ is given by

$$\begin{aligned} \mathbb{E}\left[x_i^2 \;\middle|\; i \in \mathscr{P}_1, f_1, f_2, \overline{f}\right] &= \frac{\mathbb{E}\left[g_i^2 - 4g_i\overline{f} + 4\overline{f}^2 \;\middle|\; i \in \mathscr{P}_1, f_1, f_2, \overline{f}\right]}{2\overline{f}(1-\overline{f})} \\ &= \frac{2f_1 - f_1^2 - 2f_1 f_2 + f_2^2}{2\overline{f}(1-\overline{f})}. \end{aligned}$$

Without loss of generality, we can interchange the indices of the populations. Hence, for $i \in \mathscr{P}_2$ we have

$$\mathbb{E}\left[x_i^2 \;\middle|\; i \in \mathscr{P}_2, f_1, f_2, \overline{f}\right] = \frac{2f_2 - f_2^2 - 2f_1 f_2 + f_1^2}{2\overline{f}(1-\overline{f})}.$$

For the expected genetic relatedness between individuals i and j, both

from Population 1, assuming the individuals within the respective subpopulations are no relatives, we have that

$$\mathbb{E}\left[x_i x_j \,\middle|\, \{i,j\} \in \mathscr{P}_1, f_1, f_2, \overline{f}\right] = \frac{f_1^2 + f_2^2 - 2f_1 f_2}{2\overline{f}(1-\overline{f})}.$$

Again, interchanging indices, we have that

$$\mathbb{E}\left[x_i x_j \,\middle|\, \{i,j\} \in \mathscr{P}_2, f_1, f_2, \overline{f}\right] = \frac{f_1^2 + f_2^2 - 2f_1 f_2}{2\overline{f}(1-\overline{f})}.$$

Hence, the stratification induces artificial genetic relatedness between individuals within a given subpopulation. Only in case (i) the individuals are unrelated and (ii) the frequencies across populations are equal, we have an expected relatedness equal to zero; in that scenario, the following holds for the numerator $f_1^2 + f_2^2 - 2f_1 f_2 = \overline{f}^2 + \overline{f}^2 - 2\overline{f}^2 = 0$.

Finally, we consider the expected genetic relatedness between two unrelated individuals from two different populations. We have that

$$\mathbb{E}\left[x_i x_j \,\middle|\, i \in \mathscr{P}_1, j \in \mathscr{P}_2, f_1, f_2, \overline{f}\right] = \frac{-f_1^2 - f_2^2 + 2f_1 f_2}{2\overline{f}(1-\overline{f})},$$

and conversely that

$$\mathbb{E}\left[x_i x_j \,\middle|\, i \in \mathscr{P}_2, j \in \mathscr{P}_1, f_1, f_2, \overline{f}\right] = \frac{-f_1^2 - f_2^2 + 2f_1 f_2}{2\overline{f}(1-\overline{f})}.$$

Defining $f_{12} = f_1 f_2$, our four-by-four GRM is given by

$$\mathbb{E}\left[\mathbf{A} \,\middle|\, f_1, f_2, \overline{f}\right] = \left(2\overline{f}\left(1-\overline{f}\right)\right)^{-1} \times \ldots$$

$$\begin{pmatrix}
2f_1 - f_1^2 - 2f_{12} + f_2^2 & f_1^2 + f_2^2 - 2f_{12} & -f_1^2 - f_2^2 + 2f_{12} & -f_1^2 - f_2^2 + 2f_{12} \\
f_1^2 + f_2^2 - 2f_{12} & 2f_1 - f_1^2 - 2f_{12} + f_2^2 & -f_1^2 - f_2^2 + 2f_{12} & -f_1^2 - f_2^2 + 2f_{12} \\
-f_1^2 - f_2^2 + 2f_{12} & -f_1^2 - f_2^2 + 2f_{12} & 2f_2 - f_2^2 - 2f_{12} + f_1^2 & f_1^2 + f_2^2 - 2f_{12} \\
-f_1^2 - f_2^2 + 2f_{12} & -f_1^2 - f_2^2 + 2f_{12} & f_1^2 + f_2^2 - 2f_{12} & 2f_2 - f_2^2 - 2f_{12} + f_1^2
\end{pmatrix}.$$

As before, $f_1 = \overline{f} + r$ and $f_2 = \overline{f} - r$, where $\mathbb{E}\left[r \,\middle|\, \overline{f}\right] = 0$, and, in line with

Equation 7.2, $\text{Var}\left(r \,\middle|\, \overline{f}\right) = \frac{1}{2}\overline{f}(1-\overline{f})F_{ST}$. Therefore,

$$\begin{aligned}
\mathbb{E}\left[f_1 \,\middle|\, \overline{f}\right] &= \mathbb{E}\left[f_2 \,\middle|\, \overline{f}\right] = \overline{f} \\
\mathbb{E}\left[f_1^2 \,\middle|\, \overline{f}\right] &= \mathbb{E}\left[f_2^2 \,\middle|\, \overline{f}\right] = \overline{f}^2 + \frac{1}{2}\overline{f}(1-\overline{f})F_{ST} \\
\mathbb{E}\left[f_{12} \,\middle|\, \overline{f}\right] &= \overline{f}^2 - \frac{1}{2}\overline{f}(1-\overline{f})F_{ST}.
\end{aligned}$$

Using these expressions, the expected GRM conditional on $\overline{f}$ can be rewritten as

$$\mathbb{E}\left[\mathbf{A} \,\middle|\, \overline{f}\right] = \begin{pmatrix} 1+\frac{1}{2}F_{ST} & F_{ST} & -F_{ST} & -F_{ST} \\ F_{ST} & 1+\frac{1}{2}F_{ST} & -F_{ST} & -F_{ST} \\ -F_{ST} & -F_{ST} & 1+\frac{1}{2}F_{ST} & F_{ST} \\ -F_{ST} & -F_{ST} & F_{ST} & 1+\frac{1}{2}F_{ST}. \end{pmatrix} \tag{7.3}$$

$$= \mathbb{E}[\mathbf{A}]. \tag{7.4}$$

This expression for the expectation of the GRM conditional on the pooled allele frequency is independent of the pooled allele frequency and the drift captured by r. The latter random variable (i.e., r) drops from the expression by the law of iterated expectations. After the iterated expectations, the only way in which r shapes the expected GRM is in terms of its variance (i.e., $\frac{1}{2}\overline{f}(1-\overline{f})F_{ST}$). The former random variable (i.e., $\overline{f}$) drops from the expression due to the SNP standardization prior to computing the GRM. Hence, the precise distributions of (i) the allele frequencies in the pooled sample and (ii) the genetic drift, are not relevant for this exercise.

7.4. PRINCIPAL COMPONENTS UNDER STRATIFICATION

Based on Equation 7.4, we can write the full N_T-by-N_T GRM in terms of the identity matrix and the Kronecker product of two real symmetric matrices. That is,

$$\mathbb{E}[\mathbf{A}] = \left(1-\frac{1}{2}F_{ST}\right)\mathbf{I}_{N_T} + \begin{pmatrix} F_{ST} & -F_{ST} \\ -F_{ST} & F_{ST} \end{pmatrix} \otimes \left(\boldsymbol{\iota}_N \boldsymbol{\iota}_N^\top\right) \tag{7.5}$$

Given the following matrices are real symmetric matrices, we can write down the following eigendecompositions:

$$\mathbf{P}\boldsymbol{\Phi}\mathbf{P}^\top = \begin{pmatrix} F_{ST} & -F_{ST} \\ -F_{ST} & F_{ST} \end{pmatrix} \text{ and}$$
$$\mathbf{Q}\boldsymbol{\Theta}\mathbf{Q}^\top = \boldsymbol{\iota}_N\boldsymbol{\iota}_N^\top, \text{ where}$$
$$\mathbf{P} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix},$$
$$\boldsymbol{\Phi} = \text{diag}(2F_{ST}, 0), \text{ and}$$
$$\boldsymbol{\Theta} = \text{diag}(N, 0, \ldots 0),$$

such that $\mathbf{P}\mathbf{P}^\top = \mathbf{P}^\top\mathbf{P} = \mathbf{I}_2$ and $\mathbf{Q}\mathbf{Q}^\top = \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}_N$. diag$(\cdot)$ denotes the operator that constructs a diagonal matrix with consecutive input arguments as elements on the diagonal.

Using properties of the Kronecker product and eigendecompositions of real symmetric matrices, the expected GRM can be written as

$$\mathbb{E}[\mathbf{A}] = \tag{7.6}$$
$$(\mathbf{P}\otimes\mathbf{Q})\,\text{diag}\left(1+F_{ST}\left(N_T-\tfrac{1}{2}\right), 1-\tfrac{1}{2}F_{ST}, \ldots, 1-\tfrac{1}{2}F_{ST}\right)\left(\mathbf{P}^\top\otimes\mathbf{Q}^\top\right).$$

Now $(\mathbf{P}\otimes\mathbf{Q})$ are the eigenvectors of the expected N_T-by-N_T GRM and diag$\left(1+F_{ST}\left(N_T-\frac{1}{2}\right), 1-\frac{1}{2}F_{ST}, \ldots, 1-\frac{1}{2}F_{ST}\right)$ its associated eigenvalues. Importantly, the first eigenvalue of the expected GRM is affected by the product of the total sample size and Wright's F-statistics. Even for fairly small values of F_{ST}, this quantity can grow large with increasing sample sizes. The remaining eigenvalues, however, are not affected by sample size; each remaining eigenvalue is merely decreased by $\frac{1}{2}F_{ST}$. As noted by Bulik-Sullivan et al. (2015b), values of F_{ST} are usually small; for populations on the same continent they report that typically $F_{ST} \approx 0.01$. Under this approximation, all remaining eigenvalues would be approximately equal to one.

7.4.1. *The leading principal component*

We will now study the log-likelihood function of GREML when including the first PC as fixed-effect covariate. We follow the notation by Yang et al. (2011a), where the log-likelihood ignoring the constant is given

by

$$l = -\frac{1}{2}\left(\log|\mathbf{V}| + \log\left|\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}\right| + \mathbf{y}^\top\mathbf{R}\mathbf{y}\right),$$

where $\mathbf{X}$ is the matrix of fixed-effects covariates, $\mathbf{V}$ is the phenotypic covariance matrix, and where

$$\mathbf{R} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}^\top\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{V}^{-1} = \mathbf{V}^{-\frac{1}{2}}\mathbf{M}\mathbf{V}^{-\frac{1}{2}},$$

where $\mathbf{M}$ is an idempotent matrix, projecting onto the null space of $\widetilde{\mathbf{X}} = \mathbf{V}^{-\frac{1}{2}}\mathbf{X}$, defined as

$$\mathbf{M} = \mathbf{I} - \widetilde{\mathbf{X}}\left(\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}}\right)^{-1}\widetilde{\mathbf{X}}^\top.$$

In our case $\mathbf{X}$ is merely a vector, defined as the first PC from the GRM. Hence, we switch to lower-case notation, and replace $\mathbf{x}$ by its theoretical expression, which is given by the Kronecker product of the first column of $\mathbf{P}$ and of $\mathbf{Q}$. That is,

$$\mathbf{x} = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix} \otimes \{\mathbf{Q}\}_{\bullet 1},$$

where $\{\mathbf{Q}\}_{\bullet 1}$ denotes the first column of $\mathbf{Q}$. Bearing in mind that matrix $\boldsymbol{\iota}_N\boldsymbol{\iota}_N^\top$ has rank one, its first eigenvalue and eigenvector are sufficient for reconstructing $\boldsymbol{\iota}_N\boldsymbol{\iota}_N^\top$. This observations implies that

$$\{\mathbf{Q}\}_{\bullet 1} = \frac{1}{\sqrt{N}}\boldsymbol{\iota}_N,$$

Therefore,

$$\mathbf{x} = \begin{pmatrix} \frac{1}{\sqrt{N_T}}\boldsymbol{\iota}_N \\ -\frac{1}{\sqrt{N_T}}\boldsymbol{\iota}_N \end{pmatrix}, \tag{7.7}$$

As before,

$$\mathbf{y} \sim \mathcal{N}\left(\frac{\sigma_s}{2}\begin{pmatrix} 1 \\ -1 \end{pmatrix} \otimes \boldsymbol{\iota}_N, \mathbf{V}\right),$$

where $\mathbf{V} = \sigma_A^2 \mathbf{A} + \sigma_E^2 \mathbf{I}_{2N}$, replacing $\mathbf{A}$ by its expectation under stratification, can be rewritten as

$$\mathbf{V} = (\mathbf{P} \otimes \mathbf{Q}) \mathbf{D} \left(\mathbf{P}^\top \otimes \mathbf{Q}^\top \right),$$

where

$$\mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_0, \ldots, \lambda_0),$$

$$\lambda_1 = \sigma_A^2 \left(1 + F_{ST} \left(N_T - \tfrac{1}{2}\right)\right) + \sigma_E^2, \quad \text{and} \quad \lambda_0 = \sigma_A^2 \left(1 - \tfrac{1}{2} F_{ST}\right) + \sigma_E^2.$$

Now,

$$\begin{aligned} \mathbf{V}^{-1} &= (\mathbf{P} \otimes \mathbf{Q}) \mathbf{D}^{-1} \left(\mathbf{P}^\top \otimes \mathbf{Q}^\top \right), \\ \mathbf{V}^{-\frac{1}{2}} &= (\mathbf{P} \otimes \mathbf{Q}) \mathbf{D}^{-\frac{1}{2}} \left(\mathbf{P}^\top \otimes \mathbf{Q}^\top \right), \end{aligned}$$

where the $\mathbf{D}^{-1}$ can be obtained by taking the element-wise reciprocal of the diagonal entries of $\mathbf{D}$, and, similarly, $\mathbf{D}^{-\frac{1}{2}}$ by taking the element-wise square-root of the diagonal elements $\mathbf{D}^{-1}$. Therefore, we have that

$$\begin{aligned} \mathbf{V}^{-\frac{1}{2}} \mathbf{x} &= (\mathbf{P} \otimes \mathbf{Q}) \left(\sqrt{\lambda_1^{-1}}, 0, \ldots, 0 \right)^\top, \\ \mathbf{x}^\top \mathbf{V}^{-1} \mathbf{x} &= \lambda_1^{-1}, \\ \log \left| \mathbf{x}^\top \mathbf{V}^{-1} \mathbf{x} \right| &= -\log(\lambda_1). \end{aligned}$$

Based on these expression, we can show that

$$\mathbf{M} = (\mathbf{P} \otimes \mathbf{Q}) \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & 1 \end{pmatrix} \left(\mathbf{P}^\top \otimes \mathbf{Q}^\top \right). \tag{7.8}$$

Moreover, we have that

$$\log |\mathbf{V}| = \sum_{i=N_T} \log(\{\mathbf{D}\}_{ii}) = \log(\lambda_1) + (N_T - 1) \log(\lambda_0). \tag{7.9}$$

Consequently, we can now write the log-likelihood as follows

$$l = -\frac{1}{2}\left((N_T - 1)\log(\lambda_0) + \mathbf{z}^\top \mathbf{M}\mathbf{z}\right), \text{ where} \tag{7.10}$$

$$\mathbf{z} = \mathbf{V}^{-\frac{1}{2}}\mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\mu}_\mathbf{z}, \mathbf{I}\right), \text{ where} \tag{7.11}$$

$$\boldsymbol{\mu}_\mathbf{z} = \frac{\sigma_s}{2}(\mathbf{P} \otimes \mathbf{Q}) \begin{pmatrix} \sqrt{N_T \lambda_1^{-1}} \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{7.12}$$

Exploiting the fact that $\mathbf{M}$ is idempotent, we have that

$$\mathbf{z}^\top \mathbf{M}\mathbf{z} = \mathbf{v}^\top \mathbf{v}, \text{ where} \tag{7.13}$$

$$\mathbf{v} = \mathbf{M}\mathbf{z} \sim \mathcal{N}\left(\boldsymbol{\mu}_\mathbf{v}, \mathbf{M}\right), \text{ where} \tag{7.14}$$

$$\boldsymbol{\mu}_\mathbf{v} = \mathbf{M}\boldsymbol{\mu}_\mathbf{z} = \frac{\sigma_s}{2}(\mathbf{P} \otimes \mathbf{Q}) \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & 1 \end{pmatrix} \begin{pmatrix} \sqrt{N_T \lambda_1^{-1}} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{7.15}$$

$$= \mathbf{0}. \tag{7.16}$$

The last equality shows that including the first PC as fixed-effect covariate in GREML estimation eliminates the population-dependent mean in the outcome variable. Moreover, the term $\mathbf{y}^\top \mathbf{R}\mathbf{y} \sim \chi^2(\text{tr}(\mathbf{M}))$, where $\text{tr}(\mathbf{M}) = N_T - 1$. By rewriting $\mathbf{M}$ in terms of individual PCs, rather than a Kronecker product, we can show that

$$\mathbf{M} = \mathbf{P}_{(1)}\mathbf{P}_{(1)}{}^\top, \tag{7.17}$$

where $\mathbf{P}_{(1)}$ denotes the matrix of all eigenvectors from the expected GRM except the first. Finally, in the last expression for the log-likelihood the first eigenvalue (i.e., $\lambda_1 = \sigma_A^2\left(1 + F_{ST}\left(N_T - \frac{1}{2}\right)\right) + \sigma_E^2$) has also been eliminated from the combined term $\log|\mathbf{V}| + \log\left|\mathbf{X}^\top \mathbf{V}^{-1}\mathbf{X}\right| = (N_T - 1)\log(\lambda_0)$.

Consequently, the GREML log-likelihood obtained by including the first PC from the GRM as fixed-effect covariate, is independent of the first eigenvalue and first eigenvector of the GRM. Since the effect of stratification on the first eigenvalue is of the order NF_{ST}, whilst

the effect on other eigenvalues is only of the order F_{ST}, it is obvious that this approach will remove the vast majority of any potential bias incurred due to stratification.

Hence, we posit the conjecture that GREML estimation including the first PC as fixed-effect covariate will be approximately unbiased, provided F_{ST} is small.

7.5. GRM-OLS-BASED INTERCEPT

Assuming we have a sufficiently large sample size for inference of σ^2_A and σ^2_E, such that sampling error hardly plays a role, and that these inferences are unbiased by inclusion of the first PC as fixed-effect covariate, we will now study the expected value of the fixed-effect estimate of the first PC and its relation with the LD-score-regression intercept.

The GLS (or REML fixed-effects) estimator is given by

$$\widehat{\beta}_{GLS} = \left(\mathbf{x}^\top \widehat{\mathbf{V}}^{-1}\mathbf{x}\right)^{-1}\mathbf{x}^\top \widehat{\mathbf{V}}^{-1}\mathbf{y} \tag{7.18}$$

$$= \left(\mathbf{x}^\top \widehat{\mathbf{V}}^{-1}\mathbf{x}\right)^{-1}\mathbf{x}^\top \widehat{\mathbf{V}}^{-1}\boldsymbol{\mu}_\mathbf{y} + \left(\mathbf{x}^\top \widehat{\mathbf{V}}^{-1}\mathbf{x}\right)^{-1}\mathbf{x}^\top \widehat{\mathbf{V}}^{-1}\boldsymbol{\varepsilon}, \text{ where} \tag{7.19}$$

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}), \tag{7.20}$$

where $\widehat{\mathbf{V}}$ denotes the estimate of the true covariance matrix $\mathbf{V}$, based on estimates $\widehat{\sigma}^2_A$ and $\widehat{\sigma}^2_E$ of the true variance components σ^2_A and σ^2_E of the model.

Now,

$$\mathbb{E}\left[\widehat{\beta}_{GLS}\right] = \left(\mathbf{x}^\top \widehat{\mathbf{V}}^{-1}\mathbf{x}\right)^{-1}\mathbf{x}^\top \widehat{\mathbf{V}}^{-1}\boldsymbol{\mu}_\mathbf{y} \tag{7.21}$$

Substituting previous expressions, we have that

$$\mathbb{E}\left[\widehat{\beta}_{GLS}\right] = \left(\widehat{\sigma}^2_A\left(1 + F_{ST}\left(N_T - \frac{1}{2}\right)\right) + \widehat{\sigma}^2_E\right)\mathbf{x}^\top \widehat{\mathbf{V}}^{-1}\boldsymbol{\mu}_\mathbf{y}, \text{ where} \tag{7.22}$$

$$\mathbf{x}^\top \widehat{\mathbf{V}}^{-1}\boldsymbol{\mu}_\mathbf{y} = \frac{\frac{\sigma_s}{2}\sqrt{N_T}}{\widehat{\sigma}^2_A\left(1 + F_{ST}\left(N_T - \frac{1}{2}\right)\right) + \widehat{\sigma}^2_E}. \tag{7.23}$$

Therefore,

$$\mathbb{E}\left[\widehat{\beta}_{GLS}\right] = \frac{\sigma_s}{2}\sqrt{N_T}, \tag{7.24}$$

where $N_T = 2N$ denotes the total sample size.

Moreover, the variance of this estimator is given

$$\mathrm{Var}\left(\widehat{\beta}_{GLS}\right) = \left(\mathbf{x}^\top \widehat{\mathbf{V}}^{-1}\mathbf{x}\right)^{-1} \mathbf{x}^\top \widehat{\mathbf{V}}^{-1}\mathbf{V}\widehat{\mathbf{V}}^{-1}\mathbf{x}^\top \left(\mathbf{x}^\top \widehat{\mathbf{V}}^{-1}\mathbf{x}\right)^{-1} \tag{7.25}$$

$$= \sigma_A^2 \left(1 + F_{ST}\left(N_T - \frac{1}{2}\right)\right) + \sigma_E^2. \tag{7.26}$$

Surprisingly, the ordinary-least-squares (OLS) estimate, which – owing to the unit length of the PCs – is given by

$$\widehat{\beta}_{OLS} = \mathbf{x}^\top \mathbf{y}, \tag{7.27}$$

has the same expectation and variance. Consequently, regressing the phenotype on the first PC using OLS is just as efficient as using GLS in this particular instance.

Using the fact that $\mathbb{E}\left[\widehat{\beta}_{OLS}\right] = \frac{\sigma_s}{2}\sqrt{N_T}$, we have that

$$1 + F_{ST}\mathbb{E}\left[\widehat{\beta}_{OLS}\right]^2 = 1 + \frac{\sigma_s^2}{4} F_{ST} N_T \tag{7.28}$$

$$\approx 1 + \frac{a}{4} F_{ST} N_T = \alpha_{LD}, \tag{7.29}$$

where $a \approx \sigma_s^2$ denotes the squared difference in phenotypic mean between the two subpopulations, and where α_{LD} denotes the theoretical LD-score-regression intercept, given a, F_{ST}, and N. This theoretical expression is based on Equation 2.7 in Section 2 of the Supplementary Note to Bulik-Sullivan et al. (2015b), after a minor correction in scale.

Exploiting the fact that the largest eigenvalue of the GRM (denoted by θ_1) is loosely expected to satisfy the following equality

$$\theta_1 = 1 + F_{ST}(N_T - \frac{1}{2}), \tag{7.30}$$

we have that

$$\widehat{F}_{ST} = \frac{\theta_1 - 1}{N_T - \frac{1}{2}}. \tag{7.31}$$

Therefore, our principal-components-based estimate of the LD-score-

regression intercept (denoted by $\widehat{\alpha}_{PC}$), is given by

$$\widehat{\alpha}_{PC} = 1 + \frac{\theta_1 - 1}{N_T - \frac{1}{2}} \widehat{\beta}^2_{OLS}, \tag{7.32}$$

where θ_1 denotes the first eigenvalue from the GRM, $\widehat{\beta}_{OLS}$ the OLS estimate of the regression of the phenotype on the first PC from the GRM, and N_T the total sample size.

Moreover, although beyond the scope of this chapter, under a more complex form of stratification where we have $P > 2$ discrete populations, we speculate that an individual-level data estimate of the intercept might approximately be given by the following expression

$$\widehat{\alpha}^{multi}_{PC} \approx 1 + \sum_{j=1}^{P-1} \frac{\theta_j - 1}{N_T} \widehat{\beta}^2_j,$$

where θ_j denotes the j-th largest eigenvalue of the GRM and $\widehat{\beta}_j$ the OLS estimate of the regression of the phenotype on the PC corresponding to eigenvalue θ_j.

7.6. SIMULATION STUDY

Although $\widehat{\beta}^2_{OLS}$ does not necessarily provide an unbiased estimator of $\mathbb{E}\left[\widehat{\beta}_{OLS}\right]^2$, an unbiased estimate of $\mathbb{E}\left[\widehat{\beta}_{OLS}\right]^2$ would typically hinge on knowing the true variance components. Since these are in general unknown, we will now use simulations in order to assess how well the squared OLS estimate for the 'effect' of the first PC predicts the theoretical intercept. That is, we assess whether the following relationship holds

$$\widehat{\alpha}_{PC} \overset{?}{\approx} 1 + \frac{a}{4} F_{ST} N_T = \alpha_{LD}. \tag{7.33}$$

Before we investigate the accuracy of this approximation, we would like to point out that

$$\mathrm{Var}\left(\widehat{\beta}_{OLS}\right) = \mathbb{E}\left[\widehat{\beta}^2_{OLS}\right] - \mathbb{E}\left[\widehat{\beta}_{OLS}\right]^2.$$

Hence, in our approximation, where the squared expectation of the OLS estimator is replaced by the squared OLS estimate, this squared

OLS estimate has the following expectation

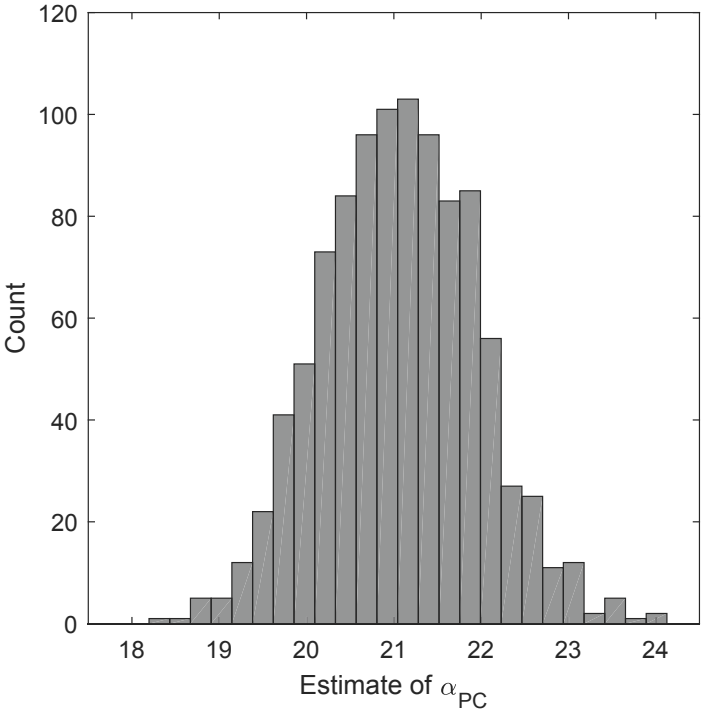
$$\begin{aligned}\mathbb{E}[\widehat{\beta}^2_{OLS}] &= \mathbb{E}[\widehat{\beta}_{OLS}]^2 + \mathrm{Var}(\widehat{\beta}_{OLS}) \\ &= \frac{\sigma_s^2}{4} N_T + \sigma_A^2 \left(1 + F_{ST}\left(N_T - \frac{1}{2}\right)\right) + \sigma_E^2 \\ &\overset{N \text{ large}}{\approx} \left(\delta^2 + \sigma_A^2 F_{ST}\right) N_T.\end{aligned}$$

Ideally, we want $\mathbb{E}[\widehat{\beta}^2_{OLS}] = \delta^2 N_T$, where $\delta^2 = \frac{\sigma_s^2}{4}$ denotes the squared deviation from the pooled mean in both populations. Bearing this in mind, when δ^2 is relatively large compared to the product $\sigma_A^2 F_{ST}$ our approximation of $\delta^2 N_T$ using the squared OLS estimator works well; since F_{ST} – as indicated – is typically small, only if $\sigma_A^2 \gg \delta^2$ does our approximation break down. However, we deem such a scenario unlikely. A minimum requirement for such a scenario would be much more within-population variation than between population variation in the phenotype. For instance, in case $\sigma_A^2 = \delta^2$ and $F_{ST} = 0.01$, the relatively deviation of $\left(\delta^2 + \sigma_A^2 F_{ST}\right) N_T$ from the desired quantity, $\delta^2 N_T$, is 1%.

7.6.1. *Description simulation*

We perform $R = 1\text{k}$ runs of simulations. In each simulation, we generate data for $P = 50\text{k}$ independent SNPs, all in Hardy-Weinberg equilibrium within populations. We have two populations, where we draw $N = 1\text{k}$ individuals from each population, yielding $N_T = 2N = 2\text{k}$ observations. We assume all SNPs are causal, and explain h^2/P of the phenotypic variation within each population. SNP heritability h^2 is set to 50%. The within-population phenotypic variance is set equal to one. Pooled allele frequencies are drawn from a $\mathcal{U}(0.1, 0.9)$ distribution, and for each pooled frequency $\overline{f}$ we draw an element $r \sim \mathcal{N}\left(0, \frac{1}{2}\overline{f}(1-\overline{f})F_{ST}\right)$, where F_{ST} denotes Wright's F-statistic. We then set the allele frequency in Population 1 equal to $\overline{f} + r$ and equal to $\overline{f} - r$ in Population 2. Any population-specific allele frequencies outside the interval $[2.5\%, 97.5\%]$ are constrained to the appropriate endpoint of this interval. We assume individuals within populations are unrelated. We set the phenotypic mean equal to $+1$ in Population 1 and to -1 in Population 2, yielding $a = 4$. We set $F_{ST} = 0.01$. Consequently, the theoretical intercept is given by $1 + 0.01 \cdot 2000 = 21$. In each simulation, we compute the estimate of the intercept obtained from the OLS regression of phenotype on

Figure 7.1: *Histogram of the GRM-OLS-based estimates (i.e., $\widehat{\alpha}_{PC}$ in Equation 7.32) of the theoretical LD-score-regression intercept (i.e., α_{LD} in Equation 7.29, where in our design $\alpha_{LD} = 21$) across 1,000 runs of simulations.*



the first PC from the GRM. That is, we compute $\widehat{\alpha}_{PC}$ as reported in Equation 7.32.

7.6.2. *Results*

Figure 7.1 shows the histogram of the intercept estimates from Equation 7.32. As can be seen these estimates seem centered around the theoretical value of 21. More precisely, the mean of the estimates across runs is 21.07, with a 95%-confidence interval given by [21.01, 21.12]. Thus, our GRM-OLS-based estimator $\widehat{\alpha}_{PC}$ of the theoretical intercept α_{LD} is very close to being unbiased. That is, $\mathbb{E}[\widehat{\alpha}_{PC}] \approx \alpha_{LD}$.

Moreover, when considering the GRM resulting from the last simulation and comparing this to the expected GRM in Equation 7.5, the

average of the diagonal elements 1.0047, which is approximately equal to the expected value of $1 + \frac{1}{2}F_{ST} = 1.005$. The average of the within-population off-diagonal elements is 0.0095, which lies close to the expected value of $F_{ST} = 0.010$. In addition, the average of the between-population elements is -0.0105, which is approximately equal to the expectation of $-F_{ST} = -0.010$. Hence, our expected GRM matches the realized GRM from simulations well.

When considering the eigenvalues of the last GRM and comparing these to the expected eigenvalues shown in eigendecomposition reported in Equation 7.6, the largest eigenvalue equals 21.005, which is reasonably close the expected $1 + F_{ST}\left(N_T - \frac{1}{2}\right) = 1 + 0.01 \cdot (2000 - 0.5) = 20.995$. Also, the average of the remaining eigenvalues is 0.9946, which is approximately equal to the predicted value $1 - \frac{1}{2}F_{ST} = 0.9950$. Finally, we find an R^2 of 99.79% when comparing the predicted first eigenvector in Equation 7.7 to the first eigenvector from the GRM resulting from the last simulation. Therefore, the eigendecomposition, resulting from a GRM based on simulated data, closely follows our theoretical predictions of what the eigendecomposition should look under the stratification at hand.

7.7. DISCUSSION

We have shown that in the data-generating process considered by Bulik-Sullivan et al. (2015b) – two discrete populations, admixed in one sample, with drift between the two populations being shaped by Wright's F-statistic (Wright, 1949) – there exists a closed-form expression that uses individual-level data (i.e., a phenotype vector and eigendecomposition of the GRM) which accurately estimates the theoretical LD-score-regression intercept, α_{LD}, derived by Bulik-Sullivan et al. (2015b).

In our simulation study, we find that our theoretical expressions for the expected GRM and its eigendecomposition closely matches the results from simulated data. More importantly, our estimator of the theoretical LD-score-regression intercept, $\widehat{\alpha}_{PC}$, is approximately unbiased. That is, $\mathbb{E}[\widehat{\alpha}_{PC}] \approx \alpha_{LD}$.

Therefore, these findings provide an encouraging perspective in terms of finding a quantity, based on individual-level data, which is equivalent to the LD-score-regression intercept. Nevertheless, we have

only compared our individual-level estimator $\widehat{\alpha}_{PC}$ to its theoretical value in simulations. In addition, we have not considered more complex types of stratification (e.g., where one has more than two admixed populations, and non-discrete populations).

Consequently, further research using simulated phenotypes and real genotypes, where two subpopulations can clearly be discerned, is needed. As real genotype data is in LD (opposed to our simulated data), such an exercise enables the application of LD-score regression to the summary statistics resulting from a GWAS on this simulated phenotype.

In addition, applications to real phenotype and genotype data are needed in order to establish whether the equivalence between the LD-score-regression intercept and our individual-level-data-based intercept persist under more complex types of stratification.

Should this equivalence break down under more complex population structures, we would like to point that our method – in principle – can be extended to $P > 2$ discrete populations, which would – in all likelihood – require an OLS regression of the phenotype on the first $P-1$ or P PCs from the GRM. Whether a transformation of the OLS estimates (e.g., a squared sum) can still be mapped back to a theoretical LD-score-regression intercept in this scenario, is a question that for now remains to be answered.

A

Appendices: Chapter 2

Based on De Vlaming et al. (2017)

A.1. DERIVATIONS POWER

In this section, we derive an expression for the power of a meta-analysis of GWAS results, under a design with many studies, with arbitrary sample sizes, SNP-based heritability, and cross-study genetic correlation (CGR).

First, the underlying assumptions are presented. Second, we write the GWAS Z statistics in terms of the true SNP effect and noise. Third, we incorporate cross-study genetic correlations by assuming a model with random SNP effects that are correlated imperfectly across studies. Using the Cholesky decomposition of the cross-study genetic correlation matrix, we write the correlated SNP effects in terms of a weighted sum of independent genetic factors. By means of this decomposition into independent factors, we derive the distribution of the Z statistic in a given study, as well as the distribution of the multi-study meta-analysis Z statistic. From the latter distribution we obtain a framework for performing multi-study power calculations.

It is important to note that models which incorporate random SNP effects have been widely used, for instance, to estimate variance components (Yang et al., 2011a) and genetic correlations across traits and samples (Lee et al., 2012), to control for cryptic relatedness and population structure in a GWAS (Yang et al., 2014), and to distill the constituents of genomic inflation (Yang et al., 2011b, Bulik-Sullivan et al., 2015b). Hence, the novelty in our work lies not in using random SNP-effect models to incorporate imperfect genetic correlations across studies. Instead the novelty lies in the subsequent step, *viz*., to use such models in order to perform power calculations under the presence of imperfect CGRs.

A.1.1. *Assumptions*

We derive an expression of statistical power for a quantitative trait in sample-size weighted meta-analysis (Willer et al., 2010). In order to arrive at a tractable expression of statistical power, we make the following assumptions.

1. When considering a given SNP in the GWAS, any phenotypic variance due to other SNPs gets absorbed by the normally, independent, and identically distributed residual term (which is what

happens when studying a sample of unrelated individuals, and which is in line with assumptions underlying most GWAS packages, except for family-based and mixed-linear-model-type GWAS software). This assumption keeps the algebra simple at the cost of a small loss in generality. In A.4 we show that violations of this assumption do not affect results.

2. The regressors (i.e., SNP data) in the meta-analysis studies are fixed (i.e., non-stochastic)—this assumption is equivalent to conditioning on the genotype data. This assumption also keeps the algebra simple at the cost of a small loss in generality. In A.4 we show that violations of this assumption do not affect results.

3. Each causal locus is shared across all studies. This assumption enables us to consider CGRs as a one-dimensional factor that is shaped solely by the cross-study correlation of the effects of trait-affecting haplotype blocks. In A.4 we show that violations of this assumption hardly affect results.

4. The genome can be divided into independent haplotype blocks, where for each block we have precisely one SNP that tags all the variation within this block. By means of this assumption, we can ignore linkage disequilibrium, thereby strongly reducing the complexity of our derivations. In addition, we assume that the effects of trait-affecting haplotype blocks are independent. The former assumption would imply that all trait-affecting variation in a haplotype block can be captured by the single tag SNP for that block. Although we make no claim that common SNPs perfectly tag all trait affecting variants, we do claim that a relatively small set of common SNPs can tag the heritability as estimated using common SNPs. Consequently, when using estimates of SNP heritability based on common SNPs, we deem this assumption and its implications to generate little bias in our theoretical predictions.

5. The effect sizes of SNPs are inversely related to SNP variance (i.e., rare variants have larger effects than common variants, such that the expected R^2 of each causal SNP, with respect to the phenotype, is equal regardless of allele frequency). This assumption makes it possible to compute statistical power without having

to specify the allele frequency and an *a priori* unknown effect size. Under this assumption, SNP heritability and the number of trait-affecting haplotype blocks replace a SNP-specific effect size and allele frequency. In A.4 we show that violations of this assumption hardly affect results.

A.1.2. *Single-SNP model*

Here, we write the GWAS Z statistic in a given study for a given SNP, as a function of the true effect and noise. This decomposition into true effect and noise helps to derive the distribution of the Z statistic.

For studies $j = 1, \ldots, C$ and SNPs $k = 1, \ldots, S$, let the model for a quantitative trait with a single SNP as predictor (Assumption 1) for the mean-centered phenotype $\mathbf{y}_j$ be given by

$$\mathbf{y}_j = \mathbf{x}_{jk}\beta_{jk} + \boldsymbol{\varepsilon}_j, \tag{A.1}$$

$$\boldsymbol{\varepsilon}_j \sim \mathcal{N}\left(\mathbf{0}, \sigma^2_{\boldsymbol{\varepsilon}_j}\mathbf{I}_{N_j}\right) \tag{A.2}$$

where $\mathbf{x}_{jk}$ denotes the mean-centered genotype vector of SNP k in study j, scaled such that $(\mathbf{x}_{jk}^\top \mathbf{x}_{jk})/N_j = 1$. In Equation A.1, β_{jk} is the effect of SNP k in study j. In Equation A.2, $\boldsymbol{\varepsilon}_j$ is the residual and $\mathbf{I}_{N_j}$ the $N_j \times N_j$ identity matrix, where N_j is the sample size of study j.

The GWAS estimate of β_{jk} for a quantitative trait is usually obtained by applying OLS. Hence, it can be written as

$$\widehat{\beta}_{jk} = \left(\frac{1}{N_j}\mathbf{x}_{jk}^\top \mathbf{x}_{jk}\right)^{-1} \frac{1}{N_j}\mathbf{x}_{jk}^\top \mathbf{y}_j \tag{A.3}$$

$$= \frac{1}{N_j}\mathbf{x}_{jk}^\top \mathbf{y}_j \tag{A.4}$$

$$= \frac{1}{N_j}\mathbf{x}_{jk}^\top \mathbf{x}_{jk}\beta_{jk} + \frac{1}{N_j}\mathbf{x}_{jk}^\top \boldsymbol{\varepsilon}_j \tag{A.5}$$

$$= \beta_{jk} + \frac{1}{N_j}\mathbf{x}_{jk}^\top \boldsymbol{\varepsilon}_j. \tag{A.6}$$

Using standard results from regression theory assuming fixed regressors (Assumption 2) and the aforementioned scaling of the genotype vector, the theoretical variance of the OLS-estimate of the SNP effect

is given by

$$\begin{aligned}\mathrm{Var}\left(\widehat{\beta}_{jk}\right) &= \sigma^2_{\boldsymbol{\varepsilon}_j}\left(\mathbf{x}_{jk}^\top \mathbf{x}_{jk}\right)^{-1} \\ &= \frac{\sigma^2_{\boldsymbol{\varepsilon}_j}}{N_j}.\end{aligned}$$

Therefore, the standard error of the OLS estimate is given by

$$\mathrm{s.d.}\left(\widehat{\beta}_{jk}\right) = \frac{\sigma_{\boldsymbol{\varepsilon}_j}}{\sqrt{N_j}}. \tag{A.7}$$

By taking the ratio of Equations A.6 and A.7 we obtain the Z statistic (instead of the commonly used and highly similar t-test statistics) for SNP k in study j. That is,

$$Z_{jk} = \frac{\widehat{\beta}_{jk}}{\mathrm{s.d.}\left(\widehat{\beta}_{jk}\right)} \tag{A.8}$$

$$= \frac{\sqrt{N_j}}{\sigma_{\boldsymbol{\varepsilon}_j}}\beta_{jk} + \frac{\mathbf{x}_{jk}^\top \boldsymbol{\varepsilon}_j}{\sigma_{\boldsymbol{\varepsilon}_j}\sqrt{N_j}}. \tag{A.9}$$

Let v_{jk} denote the last term in the right-hand side of Equation A.9. Under the aforementioned scaling of the regressor and the distribution of $\boldsymbol{\varepsilon}_j$, it follows from standard properties of the multivariate normal distribution that $v_{jk} \sim \mathcal{N}(0,1)$.

A.1.3. *Modelling cross-study genetic correlation*

We incorporate cross-study genetic correlations by considering a model with random SNP effects, correlated across studies. For ease of derivations, we assume that each causal SNP contributes across all studies (Assumption 3). In order to simplify further derivations, we use a Cholesky decomposition to write correlated SNP effects in terms of independent underlying factors. Using this independent-factor representation, we derive the distribution of a GWAS Z statistic, in terms of the study-specific noise and contributions of the underlying genetic factors.

Genetic correlation can be conceptualized as the correlation between SNP effects across different strata (e.g., across populations, studies, age groups, etc.). Taking studies as 'strata', a group of C

studies has $C \times C$ genetic correlation matrix, denoted by $\boldsymbol{P}_{\mathbf{G}}$.

When effects are normally distributed, a given correlation structure between effects is most straightforwardly obtained by constructing the Cholesky decomposition of the correlation matrix, and multiplying independent standard-normal random variables by this decomposition. An interpretation of this decomposition is that it provides a set of weights that transform a set of independent underlying genetic factors into correlated genetic effects.

First, we formalize how to transform independent standard-normal random variables into correlated normal random variables. Let $\boldsymbol{\Gamma}_{\mathbf{G}}$ be the lower-triangular Cholesky decomposition of the genetic correlation matrix, such that $\boldsymbol{\Gamma}_{\mathbf{G}}\boldsymbol{\Gamma}_{\mathbf{G}}{}^{\top} = \boldsymbol{P}_{\mathbf{G}}$, let $\mathscr{M}$ denote the set of M causal SNPs, let $\mathbf{E}$ be an $C \times M$ matrix of independent standard normal draws from different genetic factors (rows) for the different causal SNPs (columns), and let $\boldsymbol{\eta}_k$ be the column of $\mathbf{E}$ corresponding to causal SNP k. Then

$$\boldsymbol{\eta}_k = \begin{pmatrix} \eta_{1k} \\ \vdots \\ \eta_{Ck} \end{pmatrix} \sim \mathscr{N}(\mathbf{0}, \mathbf{I}_C),$$

where $\boldsymbol{\eta}_k$ is independent of $\boldsymbol{\eta}_l$ for $l \neq k$ (Assumption 4). Now, for SNP k in the set of causal SNPs, we can define the vector of effects across studies for the given SNP, such that it has correlation matrix $\boldsymbol{P}_{\mathbf{G}}$, as follows:

$$\boldsymbol{\beta}_k = \begin{pmatrix} \beta_{1k} \\ \vdots \\ \beta_{Ck} \end{pmatrix} = \text{diag}\left(\sigma_{\beta_1}, \ldots \sigma_{\beta_C}\right) \boldsymbol{\Gamma}_{\mathbf{G}} \boldsymbol{\eta}_k,$$

where diag() is a diagonal matrix with specified elements as diagonal entries, and

$$\sigma_{\beta_j} = \sqrt{\frac{h_j^2 \sigma_{\mathbf{y}_j}^2}{M}},$$

with h_j^2 (resp. $\sigma_{\mathbf{y}_j}^2$) denoting the SNP heritability (phenotypic variance) in study j. Under this design of study-specific SNP effects, we attain a CGR structure in line with $\boldsymbol{P}_{\mathbf{G}}$ and the desired study-specific SNP heritabilities.

Using this approach for constructing correlated SNP effects, we can write the effect of SNP k in study j (i.e., β_{jk}) as a linear combination of the independent underlying $\mathcal{N}(0,1)$ distributed random variables. That is,

$$\beta_{jk} = \sigma_{\beta_j} \sum_{i=1}^{j} \gamma_{ji} \eta_{ik}, \tag{A.10}$$

where γ_{ji} denotes element in row j column i of $\mathbf{\Gamma}$ and η_{ik} the i-th element of $\boldsymbol{\eta}_k$. Given our scaling of SNPs, the R^2 of each causal SNP in study j is given by $\sigma^2_{\beta_j}$, regardless of the allele frequency of the SNP of interest (Assumption 5).

We can now write the GWAS Z statistic for a given SNP in a given study, as a linear combination of independent random variables. For SNP k in the set of P non-causal SNPs, denoted by $\mathscr{P}$ (such that $\mathscr{M} \cap \mathscr{P} = \varnothing$), we have for all studies j that $\beta_{jk} = 0$. By substituting β in Equation A.9 according to Equation A.10 for causal SNPs and the preceding equality for non-causal SNPs, we obtain the following expression for the Z statistic of SNP k in study j:

$$Z_{jk} = \begin{cases} v_{jk} + \sqrt{N_j} \frac{\sigma_{\beta_j}}{\sigma_{\epsilon_j}} \sum_{i=1}^{j} \gamma_{ji} \eta_{ik} & \text{for} \quad k \in \mathscr{M}, \text{ and} \\ v_{jk} & \text{for} \quad k \in \mathscr{P}. \end{cases} \tag{A.11}$$

A.1.4. *Distribution meta-analysis Z statistic*

Here, we derive the distribution of the meta-analysis Z statistic and reduce the number of input parameters by appropriate substitutions. Finally, for intuition, we present the distribution of Z statistics from a meta-analysis of GWAS results from two studies.

For any SNP k in the set $\mathscr{S} = \mathscr{M} \cup \mathscr{P}$ consisting of S = M + P causal and non-causal SNPs, we use the sample-size-weighted meta-analysis Z statistic (Willer et al., 2010), defined as follows:

$$Z_k = \sum_{j=1}^{C} \frac{\sqrt{N_j}}{\sqrt{N_T}} Z_{jk}, \tag{A.12}$$

where $N_T = N_1 + \ldots + N_C$ denotes the total sample size. Plugging Equation A.11 for $k \in \mathscr{M}$ into Equation A.12, yields an expression for the meta-analysis Z statistic in terms of independent random variables.

That is,

$$Z_k = \begin{cases} \sum_{j=1}^{C} \frac{\sqrt{N_j}}{\sqrt{N_T}} v_{jk} + \sum_{j=1}^{C} \sum_{i=1}^{j} \frac{N_j}{\sqrt{N_T}} \frac{\sigma_{\beta_j}}{\sigma_{\varepsilon_j}} \gamma_{ji} \eta_{ik} & \text{for} \quad k \in \mathscr{M}, \text{ and} \\ \sum_{j=1}^{C} \frac{\sqrt{N_j}}{\sqrt{N_T}} v_{jk} & \text{for} \quad k \in \mathscr{P}. \end{cases} \tag{A.13}$$

As the v_{jk} terms in the preceding expression are independent standard-normal draws, it follows that

$$v_k = \sum_{j=1}^{C} \frac{\sqrt{N_j}}{\sqrt{N_T}} v_{jk} \sim \mathscr{N}(0,1).$$

In Equation A.13 we have a double sum over random variables. However, by changing the order of summation, this double sum can be rewritten as follows:

$$\sum_{j=1}^{C} \sum_{i=1}^{j} \frac{N_j}{\sqrt{N_T}} \frac{\sigma_{\beta_j}}{\sigma_{\varepsilon_j}} \gamma_{ji} \eta_{ik} = \sum_{i=1}^{C} \eta_{ik} \sum_{j=i}^{C} \frac{N_j}{\sqrt{N_T}} \frac{\sigma_{\beta_j}}{\sigma_{\varepsilon_j}} \gamma_{ji}.$$

Therefore, we can rewrite Equation A.13 as follows:

$$Z_k = \begin{cases} v_k + \sum_{i=1}^{C} \eta_{ik} \sum_{j=i}^{C} \frac{N_j}{\sqrt{N_T}} \frac{\sigma_{\beta_j}}{\sigma_{\varepsilon_j}} \gamma_{ji} & \text{for} \quad k \in \mathscr{M}, \text{ and} \\ v_k & \text{for} \quad k \in \mathscr{P}, \end{cases} \tag{A.14}$$

where the inner sum yields the weight for the random variable of interest.

Exploiting the fact that η_{ik} and v_k are independent standard-normal draws, the variance of the sum of terms is equal to the sum of the variance of the respective terms. Hence, we have that

$$Z_k \sim \begin{cases} \mathscr{N}(0, 1+d) & \text{for} \quad k \in \mathscr{M}, \text{ and} \\ \mathscr{N}(0,1) & \text{for} \quad k \in \mathscr{P}, \end{cases}$$

where

$$d = \sum_{i=1}^{C} \left(\sum_{j=i}^{C} \frac{N_j}{\sqrt{N_T}} \frac{\sigma_{\beta_j}}{\sigma_{\varepsilon_j}} \gamma_{ji} \right)^2 \tag{A.15}$$

$$= \frac{1}{N_T} \sum_{i=1}^{C} \left(\sum_{j=i}^{C} N_j \frac{\sigma_{\beta_j}}{\sigma_{\varepsilon_j}} \gamma_{ji} \right)^2 \tag{A.16}$$

The quantity d we refer to as the 'power parameter'. Since this parameter is a sum of squares, it is non-negative. The greater the power parameter is, the higher the statistical power of the meta-analysis of GWAS results is. Note that in case $\sigma_{\beta_j} = 0$ for all j (i.e., the trait is not heritable in any study), that $d = 0$, and hence the meta-analysis Z statistic reverts to a standard-normal test statistic, which matches the distribution under the null. However, as σ_{β_j} increases, d becomes larger, yielding a meta-analysis with higher statistical power.

Given SNP-based heritability, phenotypic variation, and the number of causal variants, we have that the effect size per causal SNP in a study is given by $\sigma^2_{\beta_j} = \frac{h^2_j \sigma^2_{\mathbf{y}_j}}{M}$, and the residual variance, absorbing the variance due to the omitted $M-1$ SNPs (Assumption 1), is given by $\sigma^2_{\boldsymbol{\varepsilon}_j} = \sigma^2_{\mathbf{y}_j} - \sigma^2_{\beta_j}$. Using these expressions, we can write the ratio of σ_{β_j} and $\sigma_{\boldsymbol{\varepsilon}_j}$, appearing in Equation A.16, as a function of only heritability and the number of causal SNPs. That is,

$$\frac{\sigma_{\beta_j}}{\sigma_{\boldsymbol{\varepsilon}_j}} = \sqrt{\frac{\frac{h^2_j \sigma^2_{\mathbf{y}_j}}{M}}{\sigma^2_{\mathbf{y}_j} - \frac{h^2_j \sigma^2_{\mathbf{y}_j}}{M}}} \tag{A.17}$$

$$= \sqrt{\frac{h^2_j}{M - h^2_j}}. \tag{A.18}$$

Plugging the last expression into Equation A.16 yields

$$d = \frac{1}{N_T} \sum_{i=1}^{C} \left(\sum_{j=i}^{C} N_j \sqrt{\frac{h^2_j}{M - h^2_j}} \gamma_{ji} \right)^2 \tag{A.19}$$

This expression for the power parameter shows that it is not affected by scaling due to phenotypic variance; the parameter is only affected by the cross-study genetic correlation matrix, the SNP-based heritability per study, and the sample size per study.

In case the number of studies is two, with sample size N in Study 1 and N in Study 2, SNP heritability h^2_{SNP}, and a genetic correlation $\rho_{\mathbf{G}}$ between the two studies, we have that the meta-analysis Z statistic, of

a trait-affecting SNP k, is normally distributed with mean zero and

$$\mathrm{Var}\left(Z_{k,C=2}\right) = 1 + \frac{h^2_{\mathrm{SNP}}}{M - h^2_{\mathrm{SNP}}} N\left(1 + \rho_{\mathbf{G}}\right).$$

Bearing in mind that the number of causal SNPs $M \gg 1$ under a highly polygenic model, while $h^2_{\mathrm{SNP}} \in [0,1]$, we have that under high polygenicity $M - h^2_{\mathrm{SNP}} \approx M$. Hence, the variance of Z_k can be approximated by

$$\mathrm{Var}\left(Z_{k,C=2,\text{high polygenicity}}\right) \approx 1 + \frac{h^2_{\mathrm{SNP}}}{M} N\left(1 + \rho_{\mathbf{G}}\right).$$

In the scenario where the cross-study genetic correlations equals one, we have that $\mathrm{Var}(Z_k) \approx 1 + \frac{h^2_{\mathrm{SNP}}}{M} N_T$ for a trait-affecting haplotype block and $\mathrm{Var}(Z_k) = 1$ for a non-causal haplotype block, where $N_T = 2N$. These expressions are equivalent to the expected value of the squared Z statistics from the linear regression analysis reported in Section 4.2 of the Supplementary Note to (Yang et al., 2014), as well as the first equation in (Bulik-Sullivan et al., 2015b) when assuming that confounding biases and linkage disequilibrium are absent.

A.1.5. *Adding genetically uncorrelated studies to the meta-analysis*

Here, we consider what happens to statistical power of a meta-analysis of GWAS results from several sets of studies, with genetic correlations between the studies within each set, but with no genetic correlation between the different sets. We first consider a scenario with one set consisting of $C-1$ studies and one other set consisting of only one study. We then generalize to a setting with multiple sets, each set containing at least one study. We show that the power parameter for a meta-analysis of several sets of studies with no genetic correlations between sets, can be written as a sample-size weighted sum of the power parameters within the respective sets.

In case one has $C-1$ studies with associated CGR matrix, the associated Cholesky decomposition denoted by $\mathbf{\Gamma}_{(C)}$, and an additional study indexed by C, which is genetically uncorrelated to the $C-1$ other studies, then the $C \times C$ Cholesky decomposition of the full CGR matrix

is given by

$$\mathbf{\Gamma_G} = \begin{pmatrix} \mathbf{\Gamma}_{(C)} & \mathbf{0} \\ \mathbf{0}^\top & 1 \end{pmatrix},$$

where $\mathbf{0}$ denotes a column vector of zeros.

Now, the quantity d in Equation A.19 can be decomposed as follows.

$$d = \frac{1}{N_T} \sum_{i=1}^{C-1} \left(\sum_{j=i}^{C-1} N_j \sqrt{\frac{h_j^2}{M - h_j^2}} \gamma_{ji} \right)^2 + \frac{1}{N_T} \left(N_C \sqrt{\frac{h_C^2}{M - h_C^2}} \right)^2 \tag{A.20}$$

$$= \frac{N_{(C)}}{N_T} \frac{1}{N_{(C)}} \sum_{i=1}^{C-1} \left(\sum_{j=i}^{C-1} N_j \sqrt{\frac{h_j^2}{M - h_j^2}} \gamma_{ji} \right)^2 + \frac{N_C}{N_T} \frac{1}{N_C} \left(N_C \sqrt{\frac{h_C^2}{M - h_C^2}} \right)^2 \tag{A.21}$$

$$= \frac{N_{(C)}}{N_T} d_{(C)} + \frac{N_C}{N_T} d_C, \tag{A.22}$$

where d_C denotes the power parameter in Equation A.19 had only study C (with sample-size N_C) been considered, and $d_{(C)}$ the power parameter in Equation A.19 had only the first $C-1$ studies (with total corresponding sample-size $N_{(C)}$) been considered. Hence, the power parameter in this scenario is the sample-size-weighted average of the power parameter of the first $C-1$ studies jointly and the power parameter of the last study.

Equation A.22 can be generalized, to reflect a situation where there are P disjoint sets of studies, denoted by $\mathscr{C}_1, \ldots, \mathscr{C}_P$, with genetic correlation within each set, but no genetic correlation between the sets. In this scenario, the power parameter d in Equation A.19 for a joint meta-analysis of all sets is given by

$$d_{\mathscr{C}_1 \cup \mathscr{C}_2 \cup \ldots \cup \mathscr{C}_P} = \sum_{p=1}^{P} \frac{N_{\mathscr{C}_p}}{N_T} d_{\mathscr{C}_p}, \tag{A.23}$$

where $N_{\mathscr{C}_p}$ denotes the total sample size in study-set $\mathscr{C}_p$ and $d_{\mathscr{C}_p}$ the power parameter in Equation A.19 for the meta-analysis of all studies in set $\mathscr{C}_p$, and N_T the total sample size when aggregating over all study sets. This equation states that power parameter for a meta-analysis of several sets of studies with CGR within each set, but no CGR between sets, is a weighted average of the power parameters in the underlying

sets.

Since the statistical power is a monotonically increasing function of the power parameter d, Equation A.23 leads to two corollaries under CGR equal to zero between sets of studies, namely that

$$\beta_{\mathscr{C}_1 \cup \mathscr{C}_2 \cup \ldots \cup \mathscr{C}_P} \leq \max\left\{\beta_{\mathscr{C}_p}\right\}_{p=1,\ldots,P} \text{ and} \tag{A.24}$$

$$\beta_{\mathscr{C}_1 \cup \mathscr{C}_2 \cup \ldots \cup \mathscr{C}_P} \geq \min\left\{\beta_{\mathscr{C}_p}\right\}_{p=1,\ldots,P}, \tag{A.25}$$

where $\beta_{\mathscr{A}}$ denotes the power in set of studies $\mathscr{A}$.

The implication of Equation A.23 is simple yet powerful; when several sets of studies with genetic correlation within each set, but no genetic correlation between sets, are considered for meta-analysis, one should not meta-analyze sets $\mathscr{C}_1, \ldots \mathscr{C}_P$ jointly, but rather meta-analyze only the set of studies which has the largest power parameter according to Equation A.19.

Only when $d_{\mathscr{C}_1 \cup \mathscr{C}_2 \cup \ldots \cup \mathscr{C}_P} > \max\left\{d_{\mathscr{C}_1}, \ldots, d_{\mathscr{C}_P}\right\}$, does the meta-analysis of all sets jointly have higher statistical power than a meta-analysis based on only one set of studies.

A.2. DERIVATIONS ACCURACY

This section extends the theoretical framework for meta-analytic power. Derivations are based on the same assumptions as in A.1. We consider the predictive accuracy of the polygenic score (PGS) including all S independent SNPs, with SNP-weights based on the meta-analysis results from the set of C study, in a hold-out sample indexed as 'study' $C + 1$. In this hold-out sample, we focus exclusively on the theoretical R^2 of the PGS; instead of considering multiple draws from the stochastic processes underlying the genotypes and treating these as fixed explanatory variables, we treat the phenotype, the PGS, and the underlying genotypes as random variables, and use probability theory to derive R^2. The hold-out sample is also allowed a study-specific SNP-based heritability, h^2_{C+1}, and genetic-correlations with the other C studies (thus extending both the CGR matrix and its Cholesky decomposition to $(C + 1)\times(C + 1)$ matrices).

First, we write the phenotype in hold-out sample as a function of noise and the independent genetic factors discussed in the preceding section. Second, we derive an expression for the PGS as a function

of the genetic factors. Third, using this representation we derive the theoretical covariance between the PGS and the phenotype. Fourth, using the theoretical variances and covariance, we obtain an expression for the theoretical R^2.

A.2.1. *Polygenic model*

Here, we derive an expression for the phenotype in the hold-out study as a function of independent genetic factors and an expression for the phenotypic variance.

Aggregating across causal SNP set $\mathscr{M}$ and the noise, the phenotype in study $C+1$ can be written as follows:

$$Y_{C+1} = \sum_{k \in \mathscr{M}} X_{C+1,k} \beta_{C+1,k} + \varepsilon_{C+1},$$

where, analogous to Equation A.10,

$$\beta_{C+1,k} = \sigma_{\beta_{C+1,k}} \sum_{i=1}^{C+1} \gamma_{C+1,i} \eta_{ik},$$

where η_{ik} now indicates the i-th element of the now $(C+1)$-dimensional vector of independent normal draws, $\boldsymbol{\eta}_k$, and where $\gamma_{C+1,i}$ describes an element of the Cholesky decomposition $\boldsymbol{\Gamma}_{\mathbf{G}}$ of the $(C+1)\times(C+1)$ cross-study genetic correlation matrix, incorporating the hold-out sample. Hence, the phenotype can be written as

$$Y_{C+1} = \varepsilon_{C+1} + \sum_{k \in \mathscr{M}} \left(X_{C+1,k} \sigma_{\beta_{C+1,k}} \sum_{i=1}^{C+1} \gamma_{C+1,i} \eta_{ik} \right).$$

Analogous to the scaling of SNPs in A.1 here, with genotypes treated as random variables, we assume

$$\mathbb{E}\left[X_{C+1,k}\right] = 0 \text{ and } \mathrm{Var}\left(X_{C+1,k}\right) = 1, \text{ for } k \in \mathscr{S}, \text{ and}$$
$$\mathrm{Cov}\left(X_{C+1,k}, X_{C+1,l}\right) = 0 \text{ for } k \neq l.$$

Consequently, the phenotypic variance in the hold-out sample is given by

$$\mathrm{Var}(Y_{C+1}) = M \sigma^2_{\beta_{C+1}} + \sigma^2_{\boldsymbol{\varepsilon}_{C+1}}. \tag{A.26}$$

A.2.2. *Polygenic score*

Here, we derive an expression for the PGS as a function of independent genetic factors, an expression for the PGS variance, and its covariance with the phenotype in the hold-out sample.

Since each SNP in each study in the meta-analysis has been scaled such that its dot product equals the sample size of that study, by analogy of the standard error of the SNP effect estimate in a single study, the standard-error of the meta-analytic effect estimate $\widehat{\beta}_{meta}$ for study $C + 1$ can be approximated by

$$\text{s.d.}\left(\widehat{\beta}_{meta}\right) \propto \frac{1}{\sqrt{N_T}} \propto 1,$$

where N_T denotes the total sample size of the meta-analysis.

Hence, the meta-analytic effect estimate is proportional to the meta-analysis Z statistic. Since any scalar multiple of the PGS will not affect its R^2 with respect to the phenotype, the Z statistics of the meta-analysis can be applied as SNP weights directly. Therefore, the PGS in the hold-out sample, including all SNPs, is given by

$$\widehat{Y}_{C+1} = \sum_{k \in \mathscr{S}} X_{C+1,k} Z_k. \tag{A.27}$$

Plugging the expression for Z_k from Equation A.14 into Equation A.27, and substitution of terms by means of the square root of Equation A.18, the PGS is given by

$$\widehat{Y}_{C+1} = \left(\sum_{k \in \mathscr{S}} X_{C+1,k} v_k \right) + \left(\sum_{k \in \mathscr{M}} X_{C+1,k} \sum_{i=1}^{C} \eta_{ik} \sum_{j=i}^{C} \frac{N_j}{\sqrt{N_T}} \sqrt{\frac{h_j^2}{M - h_j^2}} \gamma_{ji} \right).$$

Exploiting the fact that η_{ik}, v_k, and $X_{C+1,k}$ are all independent random variables, with mean zero and variance one, we find that the variance of the PGS is given by

$$\text{Var}\left(\widehat{Y}_{C+1}\right) = S + M \sum_{i=1}^{C} \left(\sum_{j=i}^{C} \frac{N_j}{\sqrt{N_T}} \sqrt{\frac{h_j^2}{M - h_j^2}} \gamma_{ji} \right)^2. \tag{A.28}$$

Again exploiting independence, zero mean, and unit variance of the respective terms, the covariance between the PGS and the phenotype

is given by

$$\begin{aligned}
\mathrm{Cov}\left(Y_{C+1}, \widehat{Y}_{C+1}\right) &= \mathbb{E}\left[Y_{C+1}\widehat{Y}_{C+1}\right] && \text{(A.29)}\\
&= \mathbb{E}\left[\left(\sum_{k\in\mathcal{M}} X_{C+1,k}\sigma_{\beta_{C+1,k}} \sum_{i=1}^{C+1} \gamma_{C+1,i}\eta_{ik}\right)\cdots\right. && \text{(A.30)}\\
&\qquad \left.\cdot\left(\sum_{k\in\mathcal{M}} X_{C+1,k} \sum_{i=1}^{C} \eta_{ik} \sum_{j=i}^{C} \frac{N_j}{\sqrt{N_T}} \sqrt{\frac{h_j^2}{M-h_j^2}}\gamma_{ji}\right)\right]\\
&= \mathbb{E}\left[\left(\sum_{k\in\mathcal{M}} X_{C+1,k}^2\sigma_{\beta_{C+1,k}} \left(\sum_{i=1}^{C} \gamma_{C+1,i}\eta_{ik}^2\cdots\right.\right.\right.\\
&\qquad \left.\left.\left.\cdot\left(\sum_{j=i}^{C} \frac{N_j}{\sqrt{N_T}} \sqrt{\frac{h_j^2}{M-h_j^2}}\gamma_{ji}\right)\right)\right)\right] && \text{(A.31)}\\
&= \sigma_{\beta_{C+1,k}} M\left(\sum_{i=1}^{C}\sum_{j=i}^{C} \frac{N_j}{\sqrt{N_T}} \sqrt{\frac{h_j^2}{M-h_j^2}}\gamma_{C+1,i}\gamma_{ji}\right). && \text{(A.32)}
\end{aligned}$$

A.2.3. *Theoretical R^2*

Here, we derive the theoretical R^2 between the PGS and the phenotype in a hold-out study. For intuition, we present the theoretical R^2 for a scenario with one study for discovery and one study as hold-out sample.

By combining Equations A.26, A.28, and A.32, the R^2, defined as the squared correlation of the outcome and the PGS in the hold-out sample, is now given by

$$\begin{aligned}
R^2\left(Y_{C+1}, \widehat{Y}_{C+1}\right) &= \frac{\left(\mathrm{Cov}\left(Y_{C+1}, \widehat{Y}_{C+1}\right)\right)^2}{\mathrm{Var}(Y_{C+1})\mathrm{Var}\left(\widehat{Y}_{C+1}\right)}\\
&= \frac{\sigma_{\beta_{C+1,k}}^2 M^2\left(\Sigma_{i=1}^{C}\Sigma_{j=i}^{C} \frac{N_j}{\sqrt{N_T}} \sqrt{\frac{h_j^2}{M-h_j^2}}\gamma_{C+1,i}\gamma_{ji}\right)^2}{\left(M\sigma_{\beta_{C+1}}^2 + \sigma_{\varepsilon_{C+1}}^2\right)\left(S + M\Sigma_{i=1}^{C}\left(\Sigma_{j=i}^{C} \frac{N_j}{\sqrt{N_T}} \sqrt{\frac{h_j^2}{M-h_j^2}}\gamma_{ji}\right)^2\right)}.
\end{aligned}$$

This expression can be simplified as follows:

$$R^2\left(Y_{C+1}, \widehat{Y}_{C+1}\right) = h_{C+1}^2 \frac{n}{\frac{S}{M} + d}, \qquad \text{(A.33)}$$

where d is the meta-analysis power parameter given in Equation A.19 and numerator n is given by

$$n = \frac{1}{N_T}\left(\sum_{i=1}^{C}\sum_{j=i}^{C} N_j \sqrt{\frac{h_j^2}{M-h_j^2}}\gamma_{C+1,i}\gamma_{ji}\right)^2,$$

where N is the total sample size in the meta-analysis.

The expression for R^2 in Equation A.33 is such that, in addition to the parameters needed for the power calculation, one only needs the genetic correlation between the hold-out sample and the meta-analysis samples and the heritability in the hold-out sample.

In case there is only one discovery study (i.e., $C = 1$) with sample size N, and with a genetic correlation $\rho_{\mathbf{G}}$ between the hold-out and discovery sample, we have that

$$R^2_{C=1} = h_2^2\rho_{\mathbf{G}}^2 \frac{\frac{Nh_1^2}{M-h_1^2}}{\frac{S}{M}+\frac{Nh_1^2}{M-h_1^2}}.$$

As in A.1, we have that under high polygenicity $M - h_1^2 \approx M$. Therefore, an easy approximation of R^2 in this scenario is given by

$$R^2_{C=1,\text{high polygenicity}} \approx h_2^2\rho_{\mathbf{G}}^2 \frac{h_1^2}{\frac{S}{N}+h_1^2}.$$

When $\rho_{\mathbf{G}}^2 = 1$, S=M, and $h_1^2 = h_2^2$, we obtain a known expression for PGS R^2 in terms of sample size, heritability, and the number of SNPs (Dudbridge, 2013). In case $\rho_{\mathbf{G}}^2 = 1$ and we consider the R^2 between the PGS and genetic value (i.e., the genetic component of the phenotype), both $\rho_{\mathbf{G}}^2$ and h_2^2 can be ignored, thereby making the last expression equivalent to the first equation in Daetwyler et al. (2008).

A.3. NOTE ON GENETIC CORRELATIONS

Consider, without loss of generality, a model for two phenotypes, Y_1 and Y_2. Similar to A.2, we treat phenotypes, genotypes, and SNP effects as random variables. In line with Assumption 5 in A.1, let each causal variant, for the phenotype of interest, have the same R^2 with respect

to that phenotype.

We can write the data-generating processes of the respective phenotypes as

$$Y_1 = \sum_{k \in \mathcal{M}_1} X_{k,1}\beta_{k,1} + \varepsilon_1 \text{ and}$$
$$Y_2 = \sum_{p \in \mathcal{M}_2} X_{p,2}\beta_{p,2} + \varepsilon_2,$$

where $\mathcal{M}_1$ (resp. $\mathcal{M}_2$) denotes the set of causal SNPs for Y_1 (Y_2) and where $\beta_{k,1}$ (resp. $\beta_{p,2}$) the effect of $X_{k,1}$ ($X_{p,2}$), that is, standardized SNP k (p), on phenotype 1 (2).

The genetic correlation at the genome-wide level can now be conceptualized as the correlation in the true genetic value for both phenotypes. That is

$$\begin{aligned}\rho_{\mathbf{G}} &= \text{Corr}\left(\sum_{k \in \mathcal{M}_1} X_{k,1}\beta_{k,1}, \sum_{p \in \mathcal{M}_2} X_{p,2}\beta_{p,2}\right) \\ &= \frac{\text{Cov}\left(\sum_{k \in \mathcal{M}_1} X_{k,1}\beta_{k,1}, \sum_{p \in \mathcal{M}_2} X_{p,2}\beta_{p,2}\right)}{\sqrt{\text{Var}\left(\sum_{k \in \mathcal{M}_1} X_{k,1}\beta_{k,1}\right)\text{Var}\left(\sum_{p \in \mathcal{M}_2} X_{p,2}\beta_{p,2}\right)}}\end{aligned}$$

Assuming independent haplotype blocks with independent effects (Assumption 4), where the effects have mean zero, this expression for the genetic correlation at the genome-wide level can be rewritten as

$$\begin{aligned}\rho_{\mathbf{G}} &= \frac{\sum_{k \in \{\mathcal{M}_1 \cap \mathcal{M}_2\}} \mathbb{E}\left[\beta_{k,1}\beta_{k,2}\right]}{\sqrt{|\mathcal{M}_1|\sigma^2_{\beta_1}|\mathcal{M}_2|\sigma^2_{\beta_2}}} \\ &= \frac{|\mathcal{M}_1 \cap \mathcal{M}_2|}{\sqrt{|\mathcal{M}_1||\mathcal{M}_2|}} \frac{\sigma_{\beta_{1,2}}}{\sqrt{\sigma^2_{\beta_1}\sigma^2_{\beta_2}}},\end{aligned}$$

where $|\mathscr{A}|$ denotes the number of elements in set $\mathscr{A}$.

Hence, the genetic correlation at the genome-wide level can be written as the product of overlap in causal loci between the two traits and the cross-trait correlation of the effects of these overlapping loci. That is,

$$\rho_{\mathbf{G}} = \frac{|\mathcal{M}_1 \cap \mathcal{M}_2|}{\sqrt{|\mathcal{M}_1||\mathcal{M}_2|}}\rho_{\beta}. \tag{A.34}$$

Equation A.34 is a generalization of the 'common-elements formula'

(Jensen, 1971), describing a correlation as a function of the number of overlapping elements and unique elements.

In particular, when $|\mathcal{M}_1| = |\mathcal{M}_2|$, we have that

$$\rho_{\mathbf{G}} = \frac{O}{O+D}\rho_\beta,$$

where O denotes the number of overlapping causal loci and D the number of idiosyncratic causal loci per trait.

We assume throughout the paper that all causal loci are shared across traits and studies (Assumption 3 in A.1). That is,

$$\frac{|\mathcal{M}_1 \cap \mathcal{M}_2|}{\sqrt{|\mathcal{M}_1||\mathcal{M}_2|}} = 1,$$

and that, consequently, the genetic correlation at the genome-wide level is equal to the correlation in the effects of overlapping causal SNPs. That is,

$$\rho_{\mathbf{G}} = \rho_\beta.$$

As we show in A.4, the theoretical predictions of GWAS power and predictive accuracy obtained under this assumption are quite accurate, even when an imperfect genetic correlation at the genome-wide level is shaped primarily by lack of overlap in causal loci, rather than a poor correlation in the effects of overlapping loci.

A.4. SIMULATION STUDIES

Using five simulation studies, we assess the accuracy of the MetaGAP calculator, which is based on the expressions for GWAS power and PGS R^2 derived in A.1 and A.2. Since the calculator is based on specific assumptions regarding the data-generating process, an important question is whether the calculator still provides accurate predictions of power and R^2 when the underlying assumptions are violated.

Hence, each simulation study has a different underlying data-generating process. The first study, Simulation 1, assumes that rare variants have larger effects than common variants to such an extent that each causal SNP, regardless of allele frequency, is expected to have the same R^2 with respect to the phenotype (Assumption 5 in A.1).

This simulation is entirely in line with the assumptions underlying the MetaGAP calculator. In the second study, Simulation 2, common variants have effects of the same magnitude as rare variants (leading a common causal variant to explain a larger proportion of the phenotypic variation that a rare causal variant). The third study, Simulation 3, also allows for differential R^2 between SNPs and, in addition, does not assume that SNP allele frequencies are uniformly distributed. Instead, the third study assumes that there are more variants in the lower minor allele frequency bins than in the higher minor allele frequency bins. In addition to the deviations from assumptions made in Simulations 2 and 3, Simulation 4 allows allele frequencies to be completely independent across studies. Finally, in Simulation 5, we go back to a data-generating process in line with the assumptions underlying the MetaGAP calculator, with one important difference; in Simulation 5, the genetic correlation as inferred at the genome-wide level is not only shaped by the correlation of SNP effects, but also by the degree of overlap of causal loci across studies. Thereby, Simulation 5 violates the assumption discussed in A.3, that the estimated CGR is shaped only by imperfect correlations of SNP effects across studies.

For each simulation study there are 100 independent runs. In each run data is simulated for $C = 3$ distinct samples for discovery as well as a fourth sample used as hold-out sample for prediction. The sample sizes of the respective studies are given by $N_1 = 20{,}000$, $N_2 = 15{,}000$, $N_3 = 10{,}000$, and $N_4 = 1{,}000$, where N_4 denotes the sample size of the hold-out sample. For Simulations 1–4, an 11×11 grid of equispaced values of $h^2_{\text{SNP}} \in [0,1]$ and $\rho_\beta \in [0,1]$ is considered. Similarly, for Simulation 5, an 11×11 grid of equispaced values of $s \in [0,1]$ and $\rho_\beta \in [0,1]$ is considered. Here, s denotes the fraction of causal SNPs that overlaps across studies and ρ_β the cross-study correlation of the effects of SNPs that are overlapping. In Simulations 1–4 we have that $s = 1$ and in Simulation 5 we have that $h^2_{\text{SNP}} = 0.5$. In all simulations there are $S = 100{,}000$ independent SNPs of which $M = 1{,}000$ have a causal influence. Moreover, when computing theoretical power and predictive accuracy, in line with A.3, we use $\rho_{\mathbf{G}} = s \cdot \rho_{\mathbf{G}}$ as value of the input parameter CGR. A detailed description of the data-generating process in each simulation study can be found in Table A.1.

For every run, data is simulated in accordance with the underlying data-generating process. Next, a GWAS is carried out in each of the

Table A.1: *Design of the simulations assessing the accuracy of the MetaGAP calculator. Settings identical to a preceding simulation design are denoted by 'idem', followed by the number of the first preceding simulation study with the same setting.*

Data-generating process*	**Notation**	**Simulation**				
		1	**2**	**3**	**4**	**5**
# studies for meta-analysis	$C =$	3	idem 1	idem 1	idem 1	idem 1
Index prediction sample	$C+1 =$	4	idem 1	idem 1	idem 1	idem 1
Sample size per study	$\{N_1, N_2, N_3, N_4\} =$	{20k; 15k; 10k; 1k}	idem 1	idem 1	idem 1	idem 1
Effective # SNPs	$\lvert\mathscr{S}\rvert = S =$	100k	idem 1	idem 1	idem 1	idem 1
Effective # causal SNPs, study j**	$\lvert\mathscr{M}_j\rvert = M =$	1k	idem 1	idem 1	idem 1	idem 1
Overlap causal SNPs, studies $\{j,h\}$***	$s_{jh} = \frac{\lvert\mathscr{M}_j \cap \mathscr{M}_h\rvert}{M} \in$	1	idem 1	idem 1	idem 1	$\{0, 0.1, \ldots, 1\}$
SNP-based heritability***	$h^2_{\text{SNP}} \in$	{0%, 10%, ..., 100%}	idem 1	idem 1	idem 1	50%
CGR of overlapping SNPs***	$\rho_\beta \in$	$\{0, 0.1, \ldots, 1\}$	idem 1	idem 1	idem 1	idem 1
Allele frequency SNP $k \in \mathscr{S}$, study j	$f_{jk} \sim$	$\mathscr{U}(0.05, 0.95)$	idem 1	$Beta(0.35,0.35)$****	idem 3	idem 1
Frequency SNP k, studies $j \neq h$	$=$ or $\perp\!\!\!\perp$*****	$=$	idem 1	idem 1	$\perp\!\!\!\perp$	idem 1
Genotype k, individual i, study j	$G_{jik} \sim$	$Binom(2, f_{jk})$	idem 1	idem 1	idem 1	idem 1
Effect SNP $k \notin \mathscr{M}, j$	$\beta_{jk} =$	0	idem 1	idem 1	idem 1	idem 1
Effect SNP $k \in \mathscr{M}, j$	$\beta_{jk} \sim$	$\mathscr{N}(0,1)$	idem 1	idem 1	idem 1	idem 1
Correlation SNP effect k, $j \neq h$	$\text{corr}\left(\beta_{jk}, \beta_{hk}\right) =$	$\rho_{\mathbf{G}}$	idem 1	idem 1	idem 1	idem 1
Residual i, j	$\varepsilon_{ji} \sim$	$\mathscr{N}\left(0, 1-h^2_{\text{SNP}}\right)$	idem 1	idem 1	idem 1	idem 1
Genetic value i, j	$GV_{ji} =$	$\sum_{k\in\mathscr{M}_j} \frac{G_{jik}-2f_{jk}}{\sqrt{2f_{jk}(1-f_{jk})}}\beta_{jk}$	$\sum_{k\in\mathscr{M}_j}(G_{jik}-2f_{jk})\beta_{jk}$	idem 2	idem 2	idem 1
GV coefficient j	$c_j =$	$\sqrt{h^2_{\text{SNP}}/M}$	$\sqrt{\frac{h^2_{\text{SNP}}}{\frac{1}{N_j}\sum_{i=1}^{N_j} GV_{ji}^2}}$	idem 2	idem 2	idem 1
Phenotype i, j	$Y_{ji} =$	$GV_{ji}c_j + \varepsilon_{ji}$	idem 1	idem 1	idem 1	idem 1
Number of runs	$R =$	100	idem 1	idem 1	idem 1	idem 1

* Correlations between all random quantities are zero unless otherwise specified

** Set of effective causal SNPs in study j, $\mathscr{M}_j \subset \mathscr{S}$, the set of effective SNPs

*** For each combination of values of h^2_{SNP}, fraction of overlapping causal SNPs, and CGR among overlapping SNPs, simulations are performed

**** The $Beta(0.35,0.35)$ distribution fits the empirical distribution of allele frequencies in the Health and Retirement Study data well

***** '$=$' denotes $f_{jk} = f_{hk}$ and '$\perp\!\!\!\perp$' denotes f_{jk} is independent of f_{hk}

three discovery samples. GWAS results are then meta-analyzed using sample-size weighting. The fraction of causal SNPs reaching genome-wide significance in the meta-analysis is the estimate of statistical power per SNP. The squared correlation between the meta-analysis-based PGS for the hold-out sample and the corresponding phenotype is the estimate of the PGS R^2.

Final estimates of power per causal SNPs and PGS R^2 are obtained by averaging the estimates across the runs. Figure A.1–A.2, show the resulting estimates of power per causal SNP in the meta-analysis and the R^2 of the PGS, for both Simulations 1–4 and Simulation 5. In addition, both figures report the power per causal SNP and R^2 predicted by the theoretical model, derived under the assumptions discussed in A.1. Inspection of Figure A.1 shows that there is no qualitative difference between the contour plots. Moreover, when computing the root-mean-square error (RMSE) between the theoretical predictions and the simulation-based estimates of power and R^2, even for the most extreme departures from our assumptions regarding allele frequencies and effects sizes (Simulations 3–4), the RMSE in power remains below 3% and the RMSE in R^2 of the PGS below 2%. Hence, the theoretical predictions of GWAS power and predictive accuracy – derived under assumptions of equal true R^2 of causal SNPs, with uniformly distributed allele frequencies that are equal across studies – are robust to violations of these assumptions.

Inspection of Figure A.2 shows that when CGRs are being shaped by a combination of poor overlap and poorly correlated effects of overlapping loci, there are some qualitative differences between predicted power and predictive accuracy compared to simulation-based estimates. However, the RMSE of theoretical power is only 1.2% with respect to the power estimated from simulations. Similarly, the RMSE of theoretical predictive accuracy is only 1.3%. Hence, the quantitative differences are small.

Simulation 5 shows that when low CGRs are induced by poor overlap of causal loci across studies instead of low correlations of the effects of overlapping loci, this leads to a slight downward bias in our theoretical predictions (i.e., making our theory conservative). Hence, we argue that if our calculator deems a study design well-powered, the analyses will be well-powered, potentially even more so than what our theory predicts (e.g., if some of the imperfect CGR can be attributed to causal

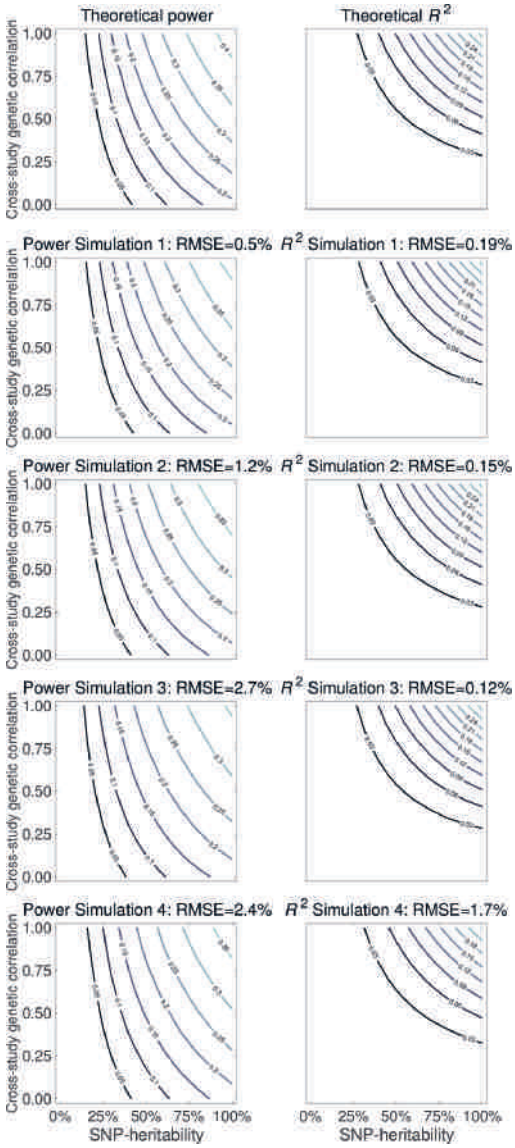


Figure A.1: *Contour plots of the statistical power per causal SNP (panels in first column) and out-of-sample polygenic-score R^2 (panels in second column), as predicted by theory (panels on first row) and as inferred by simulations (panels on subsequent rows), for various combinations of SNP heritability (x-axis within each plot) and cross-study genetic correlation (y-axis within each plot). Above each simulation-based plot, the root-mean-square error (RMSE) is reported for the difference between predictions from the theoretical model and the simulation-based estimates.*

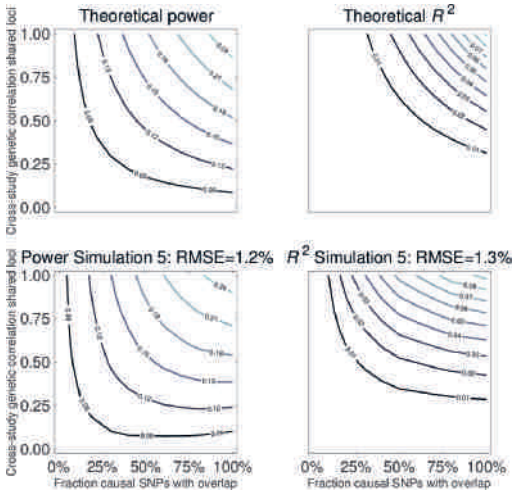


Figure A.2: *Contour plots of the statistical power per causal SNP (panels in first column) and out-of-sample polygenic-score R^2 (panels in second column), as predicted by theory (panels on first row) and as inferred by simulations (panels on second row), for various combinations of the fraction of causal loci that overlaps across studies (x-axis within each plot) and the cross-study correlation of the effects of overlapping loci (y-axis within each plot). Above each simulation-based plot, the root-mean-square error (RMSE) is reported for the difference between predictions from the theoretical model and the simulation-based estimates.*

loci that are not shared across studies).

A.5. DATA AND QUALITY CONTROL

A.5.1. *Genotype data*

In the bivariate and univariate genomic-relatedness-matrix restricted maximum likelihood (GREML) analyses we use genotype data from the Rotterdam Study (RS; Ergo waves 1–4 sample denoted by RS-I, Ergo Plus sample denoted by RS-II, and Ergo Jong sample denoted by RS-III), the Swedish Twin Registry (STR; TwinGene sample), and the Health and Retirement Study (HRS). For each study, details on the genotyping platform, quality control (QC) prior to imputation, the reference sample used for imputation, and imputation software, are listed in Table A.2.

To increase the overlap of SNPs across studies, we use genotypes imputed on the basis of the 1000 Genomes, Phase 1, Version 3 reference panel (McVean et al., 2012). We only consider the subset of HapMap3 SNPs available in the 1kG data. By using this subset we substantially reduce the computational burden of the analyses, while preserving overlap between the SNP-sets in the studies and still having a sufficiently dense set of both common and more rare SNPs (# SNPs after QC $\approx$ 1 million).

A.5.2. *Quality control*

Prior to QC, we extract HapMap3 SNPs (source: http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/hapmap3_r3/plink_format/, accessed: December 11, 2014) from the imputed genotype data of each study and convert the allele dosages to best-guess `PLINK` (Purcell et al., 2007, Chang et al., 2015) binary files by rounding dosages using `GCTA` (Yang et al., 2011a). Subsequently, we perform QC on the best-guess genotypes in two stages. In the first stage, we clean and harmonize the imputed genotype data at the study level. The cleaned and harmonized study genotypes are then merged into a pooled dataset. The second round of QC is aimed at cleaning the pooled dataset, on the basis of the samples for which the phenotype is available. Hence, the first QC stage is phenotype-independent, whereas the second stage depends on

Table A.2: *Genotyping and imputation of SNP data used in GREML analyses.*

Study	**Genotyping platform**	**SNP exclusions**			**Subject exclusions***	**Imputation****
		MAF <	**Call rate <**	**HWE *p*-value <**	**Call rate <**	**Software**
RS-I	Illumina 550K	0%	97.5%	10^{-7}	97.5%	MaCH/Minimac
RS-II	Illumina 550K	0%	97.5%	10^{-7}	97.5%	MaCH/Minimac
RS-III	Illumina 610K	0%	97.5%	10^{-7}	97.5%	MaCH/Minimac
STR	HumanOmniExpress 12v1A	1%	97.0%	10^{-7}	97.0%	MaCH/Minimac
HRS	Illumina Omni2.5	1%	98.0%	10^{-4}	98.0%	IMPUTE2

* Individuals are also excluded on the basis of sex mismatch, close relatives, duplicates and ancestry outliers (STR excepted), or autosomal heterozygosity outliers (HRS excepted).

** All samples have been imputed against the 1000Genomes, Phase 1, Version 3 haplotypes of all ancestries.

the phenotype of interest.

In the first QC stage (prior to merging), we filter out the following markers and individuals:

1. SNPs with imputation accuracy below 70%.
2. Non-autosomal SNPs.
3. SNPs with minor allele frequency below 1%.
4. SNPs with Hardy-Weinberg-Equilibrium p-value below 1%.
5. SNPs with missingness greater than 5%.
6. Individuals with missingness greater than 5%.
7. SNPs that are not present in all studies.
8. SNPs whose alleles cannot be aligned across studies.

Prior to the first QC stage, we apply the following two additional steps in HRS:

1. Switch alleles to address a strand-flip error due to incorrect annotation.
2. Drop individuals of non-European ancestry.

After the first round of QC, a set of roughly 1 million overlapping SNPs, available for about 30,000 individuals is left. Panel I in Table A.3 shows, for each study, the number of SNPs and individuals before and after the first round of QC.

The second QC stage, applied to the pooled data set, comprises the following steps:

1. Keep only individuals for whom the phenotype of interest and all corresponding control variables are available.
2. Drop SNPs with a minor allele frequency below 1%.
3. Drop SNPs with Hardy-Weinberg-Equilibrium p-value below 1%.
4. Drop SNPs with missingness greater than 5%.
5. Drop individuals with missingness greater than 5%.

6. Keep only one individual per pair of individuals with a genomic relatedness greater than 0.025.

Since the data in STR consists of twins and having highly related individuals can bias estimates of SNP-based heritability due to environment-sharing, we randomly select only one individual per twin pair after Step 1 in the second QC stage.

Table A.3: *Number of individuals and SNPs in data used for GREML analyses, before and after quality control (QC) at the study level (Panel I) and at the pooled level (Panel II).*

Panel I: study-level QC				
Study	***N***		**# SNPs**	
	pre-QC	**post-QC**	**pre-QC**	**post-QC**
RS-I	6,291	6,291	31,337,615	1,062,589
RS-II	2,157	2,157	31,337,615	1,062,589
RS-III	3,048	3,048	31,337,615	1,062,589
STR	9,617	9,617	31,326,389	1,062,589
HRS	12,454	8,652	21,632,048	1,062,589
Total		29,765		1,062,589
Panel II: pooled-level QC				
Phenotype	***N***		**# SNPs**	
	pre-QC	**post-QC**	**pre-QC**	**post-QC**
Height	29,765	20,458	1,062,589	1,052,572
BMI	29,765	20,449	1,062,589	1,052,600
EduYears	29,765	20,619	1,062,589	1,052,626
CurrCigt	29,765	20,686	1,062,589	1,052,524
CurrDrinkFreq	29,765	20,072	1,062,589	1,052,958
Self-rated health	29,765	19,184	1,062,589	1,053,190

Panel II in Table A.3 shows the sample size and the number of SNPs in the pooled dataset for the phenotypes discussed in the next subsection. We only consider phenotypes that attain a sample size of at least 18,000 individuals after all QC steps. For all phenotypes, the number of SNPs is slightly greater than one million.

A.5.3. *Phenotype data*

For HRS, we use the RAND HRS data, version N, to obtain the phenotypes of interest. These data consist of measurements from eleven

waves. RS-I consists of four data waves (Ergo 1–4). In both HRS and RS-I, data for some phenotypes are only available in a subset of the waves. RS-II, RS-III and STR do not have multiple measures over time for the phenotypes considered in this study. Table A.4 describes how the phenotypes are constructed in each of the five studies.

As Table A.4 shows, height, BMI, *EduYears*, and *CurrCigt* are measured quite consistently across waves. The self-rated health phenotype is also measured quite consistently, although in RS respondents are asked about health compared to members of the same age group, whereas a more absolute question is posed in STR and HRS. The drinking measure *CurrFreqDrink* is also measured somewhat heterogeneously; the threshold for what we treat as 'frequent drinking' is determined by how fine-grained the drinking frequency measure is in the respective studies.

A.6. GREML ESTIMATION

Height, BMI, *EduYears*, and self-rated health are treated as quantitative traits. *CurrCigt* and *CurrDrinkFreq* are treated as binary outcomes. In each study, (after aggregating across waves, if applicable) we regress quantitative phenotypes on age, squared age, sex, and an intercept. The residuals from the regression are standardized to have a sample-mean equal to zero and variance equal to one. For both binary and quantitative traits, the aforementioned covariates are also included in the GREML estimation. In addition, in bivariate GREML and pooled GREML estimation (i.e., considering multiple studies jointly), the intercept is replaced by indicator variables for the respective studies, capturing study-specific fixed effects. Finally, 20 principal components from the phenotype-specific genomic-relatedness matrix are added to the set of control variables in the GREML estimation, in order to correct for population stratification (Price et al., 2006).

A.7. GREML RESULTS

Details per phenotype on sample size, univariate estimates of SNP heritability, and bivariate estimates of genetic correlation, stratified

Table A.4: *Study-level measures for constructing the phenotypes used in GREML analyses.*

Phenotype	**Survey instrument in**				
	RS-I	**RS-II**	**RS-III**	**STR**	**HRS**
EduYears	Constructed in line with Rietveld et al. (2013b) in all studies.				
Height	Median height across waves 1-4.	Height	Height	Height	Median height across waves 1-11
BMI	Median BMI across waves 1-4.	BMI	BMI	BMI	Median BMI across waves 1-11
Currently smoking cigarettes (*CurrCigt*)	1 if stated to be a current smoker of cigarettes in the latest available measurement across waves 1-4.	1 if stated to be a current cigarette smoker.	Same as RS-II.	1 if stated to be a current cigarette smoker.	1 if responded positively to "currently smokes cigarettes?" in the latest available measurement across waves 1-11.
Currently drinking frequently (*CurrDrinkFreq*)	1 if indicated to "drink one or more alcoholic beverages per week" in the latest available measurement across waves 1-4.	1 if indicated to "drink one or more alcoholic beverages per week".	1 if indicated to "have drunk at least two alcoholic beverages a month during the the past year."	1 if indicated to "have drunk at least two alcoholic beverages in the past month".	1 if indicated to "drink alcohol once per week or more" in the latest available measurement across waves 3-11.
Self-rated health	Only available in wave 1: "How is your general health compared to members of your age group?" Response categories reverse-coded such that 0=worse, 1=same, and 2=better.	Same as RS-I.	*n.a.*	Rate their general health. Response categories re-coded such that 0=bad, 1=not so good, 2=average, 3=good, 4=excellent.	Mode of the 4-point self-reported health measure in HRS across waves 1-11. Responses reverse-coded such that 0=poor, 1=fair, 2=good, 3=very good, and 4=excellent.

across studies, and cross-study averages, are provided in Table A.5. Results stratified across sexes are listed in Table A.6.

A.8. LARGE-SCALE GWAS EFFORTS

Table A.7 shows the meta-analysis packages, and the assumptions underlying those packages, used in large-scale GWAS efforts for the traits considered in our attenuation study, reported in Table 2.2. Similarly, Table A.8 shows details and notes on the results from large-scale GWAS efforts that are used as input in the aforementioned attenuation study.

Table A.5: *GREML estimates of SNP heritability and genetic correlation across studies.*

Phenotype	*N*				Univariate estimates SNP heritability[1]				Bivariate estimates genetic correlation[1,2]			
	RS	STR	HRS	Total	RS	STR	HRS	Mean[3]	RS–STR	RS–HRS	STR–HRS	Mean[4]
Height	6,780	5,342	8,336	20,458	48.9% (4.9%) ***	50.8% (6.0%) ***	37.9% (4.1%) ***	44.9%	0.976 (0.102) ***	0.954 (0.095) ***	0.967 (0.106) ***	0.965
BMI	6,775	5,341	8,333	20,449	28.9% (4.9%) ***	16.4% (6.1%) ***	19.6% (4.1%) ***	21.9%	1.000 (0.269) ***	0.914 (0.172) ***	0.847 (0.246) ***	0.917
EduYears	6,735	5,543	8,341	20,619	17.5% (4.8%) ***	20.6% (5.8%) ***	17.3% (4.0%) ***	18.2%	0.690 (0.233) ***	0.659 (0.224) ***†	1.000 (0.263) ***	0.783
CurrCigt	6,803	5,579	8,304	20,686	17.8% (10.1%) **	18.7% (13.8%) *	20.4% (11.2%) **	19.1%	1.000 (0.643) ***	0.611 (0.448) *	1.000 (0.607) ***	0.858
CurrDrinkFreq	6,172	5,564	8,336	20,072	13.5% (8.7%) *	14.1% (9.5%) *	5.3% (6.3%)	10.3%	1.000 (0.666) ***	0.298 (0.670)	-0.056 (0.647)	0.381
Self-rated health	5,264	5,577	8,343	19,184	13.5% (6.2%) **	9.4% (5.7%) **	21.3% (4.0%) ***	15.7%	0.626 (0.439) **	0.363 (0.223) **††	0.447 (0.278) **	0.468

[1] Standard errors between parentheses.
[2] Significance of deviations from one only tested for genetic correlations.
[3] Sample-size weighted averages of univariate estimates across studies.
[4] Sample-size weighted averages of bivariate estimates across pairs of studies.

* > 0 at 10% sign. † < 1 at 10% sign.
** > 0 at 5% sign. †† < 1 at 5% sign.
*** > 0 at 1% sign. ††† < 1 at 1% sign.

Table A.6: *GREML estimates of SNP heritability and genetic correlation across sexes.*

Phenotype	*N*			Univariate estimates SNP heritability[1]			Bivariate estimates genetic correlation[1,2]
	Females	Males	Total	Females	Males	Mean[3]	Females–Males
Height	11,553	8,905	20,458	43.2% (3.0%) ***	45.1% (3.8%) ***	44.0%	0.981 (0.067) ***
BMI	11,542	8,907	20,449	22.1% (2.9%) ***	23.8% (3.8%) ***	22.8%	0.794 (0.122) *** †
EduYears	11,653	8,966	20,619	18.1% (2.9%) ***	18.9% (3.7%) ***	18.4%	0.832 (0.162) ***
CurrCigt	11,706	8,980	20,686	22.3% (7.1%) ***	26.7% (9.1%) ***	24.2%	0.543 (0.257) *** †
CurrDrinkFreq	11,312	8,760	20,072	14.1% (4.6%) ***	0.9% (6.0%)	8.3%	1.000 (2.068) *
Self-rated health	10,866	8,318	19,184	8.6% (3.1%) ***	10.8% (4.0%) ***	9.5%	1.000 (0.349) ***

[1] Standard errors between parentheses.
[2] Significance of deviations from one only tested for genetic correlations.
[3] Sample-size weighted averages of univariate estimates in females and males.

* > 0 at 10% sign. † < 1 at 10% sign.
** > 0 at 5% sign. †† < 1 at 5% sign.
*** > 0 at 1% sign. ††† < 1 at 1% sign.

Table A.7: *Meta-analysis methods used in large-scale GWAS efforts to date for traits considered in the GREML analyses. Traits are reported in order of appearance in Table 2.2.*

Phenotype	**Large-scale GWAS**	**Meta-analysis**			
		Software	**Weighting***	**Effects**	**Accounts for heterogeneity**
Height	Wood et al. (2014)	METAL	IV	Fixed	No
	Lango Allen et al. (2010)	METAL	IV	Fixed	No
	Weedon et al. (2008)**	*n.a.*	IV	Fixed	No
BMI	Locke et al. (2015)	METAL	IV	Fixed	No
	Speliotes et al. (2010)	METAL	IV and *N*	Fixed	No
	Willer et al. (2008)	METAL	*N*	Fixed	No
EduYears	Okbay et al. (2016b)	METAL	*N*	Fixed	No
	Okbay et al. (2016b)	METAL	*N*	Fixed	No
	Rietveld et al. (2013b)	METAL	*N*	Fixed	No
Self-rated health	Harris et al. (2016)***	*n.a.*	*n.a.*	*n.a.*	*n.a.*

* IV = inverse-variance weighting. *N* = sample-size weighting.
** No commonly-used meta-analysis tool is applied.
*** No meta-analysis is used since this concerns data from a single study.

Table A.8: *Notes on the design and outcomes of large-scale GWAS meta-analysis efforts to date for traits considered in the GREML analyses. Traits are reported in order of appearance in Table 2.2.*

Study	***N***			***C****			**Number of hits**			**PGS R^2**		
	Section	**Page**	**Note**	**Section**	**Page**	**Note**	**Section**	**Page**	**Note**	**Section**	**Page**	**Note**
Wood et al. (2014)	Abstract	1173		Results	1173		Abstract	1173		Fig 2d	1175	Approximation least stringent p-val threshold
Lango Allen et al. (2010)	Abstract	832		SI	2	Stages 1–2	Abstract	832		Fig 1a	833	Approximation least stringent p-val threshold
Weedon et al. (2008)	Abstract	575		Results	576		Table 1	577		Results	580	PGS based on 20 SNPs
Locke et al. (2015)	Abstract	197		Results	197		Abstract	197		Ext. Data Fig 3c		Approximation
Speliotes et al. (2010)	Abstract	937		Results	937		Results	937		Results	941	Least stringent p-val threshold
Willer et al. (2008)	Results	26	Stage 1	Results	26	Stage 1	Results	26	Stage 1	*n.a.*		
Okbay et al. (2016b)	SI	16		SI	12,16		SI	16		*n.a.*		
Okbay et al. (2016b)	Abstract	539		SI	12		Abstract	539		SI	108	
Rietveld et al. (2013b)	Abstract	1467		Results	1467		Results	1467	EduYears	Results	1469	Approximation average
Harris et al. (2016)	Abstract	2		Abstract	2		Abstract	2		*n.a.*		

*C denotes the number of studies in the meta-analysis; C is slightly subjective (e.g., RS I, II, and III can be considered as one study or as three).

Samenvatting

In de statistische genetica is men onder andere geïnteresseerd in het doorgronden van de genetische architectuur van fenotypes (waarneembare individuele eigenschappen) die beïnvloed worden door vele genetische varianten, zoals enkel-nucleotide polymorfismen (SNP's). Zo wil men in dit vakgebied onder andere weten wat de gezamenlijke bijdrage van alle SNP's aan fenotypische variatie is alsmede welke specifieke SNP's een robuuste en repliceerbare associatie met het fenotype hebben.

Hoewel de steekproefgroottes in genoombrede associatiestudies (GWAS's) de afgelopen jaren flink zijn toegenomen, is het aantal SNP's nog altijd een of meer ordes van grootte groter dan de steekproefomvang in een doorsnee-GWAS. Het menselijk genoom bestaat immers uit miljoenen SNP's, terwijl alleen de grootste GWAS-onderzoeken in de buurt van een steekproefgrootte van één miljoen observaties komen. Daarom is het – anno 2017 – nog altijd niet mogelijk middels standaardmethodes, zoals multipele regressie, de associatie tussen een genoombrede set SNP's en een fenotype gezamenlijk te schatten (dat wil zeggen: rekening houdend met de correlatie tussen regressoren). Mede om deze reden is het lineaire gemengde model (LMM) een belangrijk instrument in de statistische genetica. Doordat een LMM een redelijke *a-priori*-verdeling oplegt aan de effecten van SNP's, kan men in dit model de effecten van SNP's gezamenlijk schatten, zonder in de problemen te geraken door het feit dat men meer regressoren dan observaties heeft. Daarnaast kan een LMM worden ingezet om de gezamenlijke bijdrage van SNP's aan fenotypische variantie te schatten.

Dit proefschrift bestudeert een aantal eigenschappen van LMM's

nader. In het bijzonder worden onder andere de relatie tussen LMM's en ridge-regressie enerzijds en LMM's en LD-score-regressie anderzijds onder de loep genomen. Ook wordt een LMM in dit proefschrift ingezet om een online-tool, genaamd MetaGAP, te ontwikkelen waarmee men het onderscheidend vermogen en de voorspellende waarde van een GWAS kan berekenen wanneer er sprake is van heterogeniteit tussen subgroepen in de steekproef van een GWAS. Deze tool helpt daarmee *a priori* vast te stellen of een voorgenomen GWAS van een heterogeen fenotype kansrijk is.

Berekeningen van onderscheidend vermogen met behulp van MetaGAP laten zien dat we met de huidige steekproefgroottes, zelfs voor heterogene fenotypes, een goede kans van slagen hebben robuust geassocieerde SNP's te vinden. Dit werk onderschrijft daarmee de bevinding uit eerder onderzoek naar dergelijke heterogene uitkomsten in de sociale wetenschappen, zoals opleidingsniveau, dat een grootschalige GWAS in staat is geassocieerde SNP's te identificeren. Deze voorspelling wordt ondersteund door een GWAS die onderdeel uitmaakt van dit proefschrift. In deze GWAS worden, op basis van een steekproef van ruim 300.000 mensen, twaalf onafhankelijke SNP's gevonden die robuust geassocieerd zijn met twee maten voor reproductieve keuzes.

Deze bevindingen ondersteunen de these dat huidig GWAS-onderzoek in staat is de genetische architectuur van uitkomsten in de sociale wetenschappen deels te onthullen. Daarnaast ondersteunt dit werk de stelling dat, met verdere toenames in steekproefgrootte, GWAS-onderzoek in de nabije toekomst afdoende voorspellende waarde zal vergaren voor praktisch gebruik van GWAS-resultaten in de sociale wetenschappen als geheel.

Summary

One of the goals of statistical genetics is to elucidate the genetic architecture of phenotypes (i.e., observable individual characteristics) that are affected by many genetic variants, such as single-nucleotide polymorphisms (SNPs). More specifically, this field aims to assess the overall contribution of SNPs to phenotypic variation and to identify specific SNPs that are robustly associated with a given phenotype.

Although sample sizes in genome-wide association studies (GWASs) have increased strongly over the past decade, the number of SNPs is typically still several orders of magnitude larger than GWAS sample sizes; the human genome consists of millions of SNPs whilst the largest GWAS discovery samples are only beginning to approach a sample size of one million. For now, it remains infeasible to jointly infer the association between SNPs and a given phenotype (i.e., accounting for the correlation between regressors) using standard methods, such as multiple regression. Owing to this hurdle, the linear mixed model (LMM) has become a popular tool in statistical genetics. By placing a reasonable prior on SNP effects, an LMM can be used to jointly estimate the effect of each SNP and the overall contribution of SNPs to phenotypic variance.

This dissertation investigates several aspects of LMMs. More specifically, the relations between LMMs and methods such as ridge regression and LD-score regression are considered. In addition, an LMM is used to develop an online tool, called MetaGAP, which can be used to quantify statistical power and predictive accuracy of a GWAS in case there is heterogeneity (e.g., in phenotypic measurement and in the genetic architecture) between different subsamples used in the GWAS. Consequently, this tool helps to assess *a priori* whether an intended

GWAS of a heterogeneous phenotype is likely to yield meaningful outcomes.

Calculations of statistical power using MetaGAP show that current sample sizes yield good odds of finding associated SNPs, even for considerably heterogeneous traits. Therefore, this work supports the earlier empirical finding that large-scale GWAS efforts for heterogeneous traits in the social sciences, such as educational attainment, are able to identify robustly associated SNPs. This prediction is bolstered by a GWAS included in this dissertation, in which twelve independent SNPs are found that are robustly associated with reproductive choices in a sample of over 300,000 individuals.

These findings support two propositions that will have profound ramifications for the social sciences. First, current GWAS sample sizes already enable researchers to uncover parts of the genetic architecture of social-scientific traits. Second, results from GWAS efforts will attain sufficient predictive accuracy in the near future for useful applications in the social sciences as a whole.

References

Abraham, G., Kowalczyk, A., Zobel, J., and Inouye, M. Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genet Epidemiol*, 37:184–195, 2013.

Aizerman, A., Braverman, E. M., and Rozoner, L. I. Theoretical foundations of the potential function method in pattern recognition learning. *Autom Remote Control*, 25:821–837, 1964.

Albertsen, H. M., Chettier, R., Farrington, P., and Ward, K. Genome-wide association study link novel loci to endometriosis. *PLOS ONE*, 8:e58257, 2013.

Alford, J. R., Funk, C. L., and Hibbing, J. R. Are political orientations genetically transmitted? *Am Polit Sci Rev*, 99:153–167, 2005.

Altshuler, D. M., Gibbs, R. A., and the International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467:52–58, 2010.

Aronszajn, N. Theory of reproducing kernels. *Trans Am Math Soc*, 68: 337–404, 1950.

Avison, M. and Furnham, A. Personality and voluntary childlessness. *J Popul Res*, 32:45–67, 2015.

Balbo, N., Billari, F. C., and Mills, M. C. Fertility in advanced societies: a review of research. *Eur J Popul*, 29:1–38, 2013.

Barban, N., Jansen, R., De Vlaming, R., Vaez, A., Mandemakers, J. J., Tropf, F. C., Shen, X., et al. Genome-wide analysis identifies 12 loci

influencing human reproductive behavior. *Nat Genet*, 48:1462–1472, 2016.

Becker, G. S. Investment in human capital: A theoretical analysis. *J Polit Econ*, 70:9–49, 1962.

Benjamin, D. J., Cesarini, D., Chabris, C. F., Glaeser, E. L., Laibson, D. I., Guðnason, V., Harris, T. B., et al. The promises and pitfalls of genoeconomics. *Annu Rev Econom*, 4:627–662, 2012a.

Benjamin, D. J., Cesarini, D., Van der Loos, M. J. H. M., Dawes, C. T., Koellinger, P. D., Magnusson, P. K. E., Chabris, C. F., et al. The genetic architecture of economic and political preferences. *Proc Natl Acad Sci USA*, 109:8026–8031, 2012b.

Benner, A., Zucknick, M., Hielscher, T., Ittrich, C., and Mansmann, U. High-dimensional Cox models: the choice of penalty as part of the model building process. *Biom J*, 52:50–69, 2010.

Bentley, D. R., Foster, M. W., and the International HapMap Consortium. The international HapMap project. *Nature*, 426:789–796, 2003.

Berrington, A. Perpetual postponers? Women's, men's and couple's fertility intentions and subsequent fertility behaviour. *Popul Trends*, 117:9–19, 2004.

Bhatia, G., Patterson, N., Sankararaman, S., and Price, A. L. Estimating and interpreting F_{st}: the impact of rare variants. *Genome Res*, 23:1514–1521, 2013.

Bhattacharjee, S., Rajaraman, P., Jacobs, K. B., Wheeler, W. A., Melin, B. S., Hartge, P., Yeager, M., et al. A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am J Hum Genet*, 90:821–835, 2012.

Boivin, J., Bunting, L., Collins, J. A., and Nygren, K. G. International estimates of infertility prevalence and treatment-seeking: potential need and demand for infertility medical care. *Hum Reprod*, 22: 1506–1512, 2007.

Bøvelstad, H. M., Nygård, S., Størvold, H. L., Aldrin, M., Borgan, Ø., Frigessi, A., and Lingjærde, O. C. Predicting survival from microarray data—a comparative study. *Bioinformatics*, 23:2080–2087, 2007.

Brown, B. C., the Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, Ye, C. J., Price, A. L., and Zaitlen, N. Transethnic genetic-correlation estimates from summary statistics. *Am J Hum Genet*, 99:76–88, 2016.

Browning, S. R. and Browning, B. L. Population structure can inflate SNP-based heritability estimates. *Am J Hum Genet*, 89:191–193, 2011.

Bryc, K., Bryc, W., and Silverstein, J. W. Separation of the largest eigenvalues in eigenanalysis of genotype data from discrete subpopulations. *Theor Popul Biol*, 89:34–43, 2013.

Bulik-Sullivan, B. K. Relationship between LD score and Haseman-Elston regression. *bioRxiv*, 018283, 2015.

Bulik-Sullivan, B. K., Finucane, H. K., Anttila, V., Gusev, A., Day, F. R., Loh, P. R., Duncan, L., et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet*, 47:1236–1241, 2015a.

Bulik-Sullivan, B. K., Loh, P. R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet*, 47:291–295, 2015b.

Byars, S. G., Ewbank, D., Govindaraju, D. R., and Stearns, S. C. Natural selection in a contemporary human population. *Proc Natl Acad Sci USA*, 107:1787–1792, 2010.

Casella, G. and Searle, S. R. On a matrix identity useful in variance component estimation. Biometrics Unit Technical Reports; Number BU-875-M, 1985.

Cesarini, D., Dawes, C. T., Johannesson, M., Lichtenstein, P., and Wallace, B. Genetic variation in preferences for giving and risk taking. *Q J Econ*, 124:809–842, 2009.

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., and Lee, J. J. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4:1–16, 2015.

Collins, F. S., Lander, E. S., and the International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431:931–945, 2004.

Conomos, M. P., Reiner, A. P., Weir, B. S., and Thornton, T. A. Model-free estimation of recent genetic relatedness. *Am J Hum Genet*, 98: 127–148, 2016.

Crossa, J., De Los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., Makumbi, D., et al. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*, 186:713–724, 2010.

Daetwyler, H. D., Villanueva, B., and Woolliams, J. A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLOS ONE*, 3:e3395, 2008.

Day, F. R., Elks, C. E., Murray, A., Ong, K. K., and Perry, J. R. B. Puberty timing associated with diabetes, cardiovascular disease and also diverse health outcomes in men and women: the UK Biobank study. *Sci Rep*, 5:11208, 2015a.

Day, F. R., Hinds, D. A., Tung, J. Y., Stolk, L., Styrkarsdottir, U., Saxena, R., Bjonnes, A., et al. Causal mechanisms and balancing selection inferred from genetic associations with polycystic ovary syndrome. *Nat Commun*, 6:8464, 2015b.

Day, F. R., Ruth, K. S., Thompson, D. J., Lunetta, K. L., Pervjakova, N., Chasman, D. I., Stolk, L., et al. Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat Genet*, 47: 1294–1303, 2015c.

Day, F. R., Helgason, H., Chasman, D. I., Rose, L. M., Loh, P. R., Scott, R. A., Helgason, A., et al. Physical and neurobehavioral determinants of reproductive onset and success. *Nat Genet*, 48:617–623, 2016.

De Los Campos, G., Gianola, D., and Rosa, G. J. M. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *J Anim Sci*, 87:1883–1887, 2009.

De Vlaming, R. and Groenen, P. J. F. The current and future use of ridge regression for prediction in quantitative genetics. *BioMed Res Int*, 2015:143712, 2015.

De Vlaming, R., Okbay, A., Rietveld, C. A., Johannesson, M., Magnusson, P. K. E., Uitterlinden, A. G., Van Rooij, F. J. A., et al. Meta-GWAS accuracy and power (MetaGAP) calculator shows that hiding heritability is partially due to imperfect genetic correlations across studies. *PLOS Genet*, 13:e1006495, 2017.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B Stat Methodol*, 39:1–38, 1977.

Devlin, B., Roeder, K., and Wasserman, L. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol*, 60: 155–166, 2001.

Dhawan, V., Brookes, Z. L. S., and Kaufman, S. Long-term effects of repeated pregnancies (multiparity) on blood pressure regulation. *Cardiovasc Res*, 64:179–186, 2004.

Donnelly, P. Progress and challenges in genome-wide association studies in humans. *Nature*, 456:728–731, 2008.

Ducrocq, V. and Chapuis, H. Generalizing the use of the canonical transformation for the solution of multivariate mixed model equations. *Genet Sel Evol*, 29:205–224, 1997.

Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLOS Genet*, 9:e1003348, 2013.

Dudbridge, F. and Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol*, 32:227–234, 2008.

Eeles, R. A., Kote-Jarai, Z., Al Olama, A. A., Giles, G. G., Guy, M., Severi, G., Muir, K., et al. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat Genet*, 41:1116–1121, 2009.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *Ann Stat*, 32:407–499, 2004.

Ehret, G. B., Munroe, P. B., Rice, K. M., Bochud, M., Johnson, A. D., Chasman, D. I., Smith, A. V., et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478:103–109, 2011.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet*, 11:446–450, 2010.

Elks, C. E., Perry, J. R. B., Sulem, P., Chasman, D. I., Franceschini, N., He, C., Lunetta, K. L., et al. Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. *Nat Genet*, 42:1077–1085, 2010.

Endelman, J. B. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome*, 4:250–255, 2011.

Erken, H., Donselaar, P., and Thurik, A. R. Total factor productivity and the role of entrepreneurship. *J Technol Transf*, pages 1–29, 2016.

Evangelou, E. and Ioannidis, J. P. A. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet*, 14:379–389, 2013.

Evangelou, E., Fellay, J., Colombo, S., Martinez-Picado, J., Obel, N., Goldstein, D. B., Telenti, A., and Ioannidis, J. P. A. Impact of phenotype definition on genome-wide association signals: empirical evaluation in human immunodeficiency virus type 1 infection. *Am J Epidemiol*, 173:1336–1342, 2011.

Evans, D. M., Visscher, P. M., and Wray, N. R. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum Mol Genet*, 18: 3525–3531, 2009.

Finucane, H. K., Bulik-Sullivan, B. K., Gusev, A., Trynka, G., Reshef, Y., Loh, P. R., Anttila, V., et al. Partitioning heritability by functional

annotation using genome-wide association summary statistics. *Nat Genet*, 47:1228–1235, 2015.

Fowler, J. H. and Dawes, C. T. In defense of genopolitics. *Am Polit Sci Rev*, 107:362–374, 2013.

Fowler, J. H., Baker, L. A., and Dawes, C. T. Genetic variation in political participation. *Am Polit Sci Rev*, 102:233–248, 2008.

Frank, I. E. and Friedman, J. H. A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–135, 1993.

Gianola, D. and Van Kaam, J. B. C. H. M. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*, 178:2289–2303, 2008.

Gilmour, A. R., Thompson, R., and Cullis, B. R. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, 51:1440–1450, 1995.

González-Recio, O., Gianola, D., Long, N., Weigel, K. A., Rosa, G. J. M., and Avendaño, S. Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. *Genetics*, 178:2305–2313, 2008.

Gusev, A., Lee, S. H., Trynka, G., Finucane, H. K., Vilhjálmsson, B. J., Xu, H., Zang, C., et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am J Hum Genet*, 95:535–552, 2014.

Hakim, C. A new approach to explaining fertility patterns: preference theory. *Popul Dev Rev*, 29:349–374, 2003.

Han, B. and Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet*, 88:586–598, 2011.

Harris, S. E., Hagenaars, S. P., Davies, G., Hill, W. D., Liewald, D. C. M., Ritchie, S. J., Marioni, R. E., et al. Molecular genetic contributions to self-rated health. *Int J Epidemiol*, advance access:dyw219, 2016.

Harville, D. A. Discussion on a section on interpolation and estimation. In David, H. A. and David, H. T., editors, *Statistics, an appraisal:*

proceedings of a Conference Marking the 50th Anniversary of the Statistical Laboratory, pages 281–286. Iowa State University, Ames, Iowa, USA, 1983.

Harville, D. A. Maximum likelihood approaches to variance component estimation and to related problems. *J Am Stat Assoc*, 72:320–338, 1977.

Haseman, J. K. and Elston, R. C. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet*, 2:3–19, 1972.

Hastie, T. and Tibshirani, R. Efficient quadratic regularization for expression arrays. *Biostatistics*, 5:329–340, 2004.

Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning*. Springer, New York, New York, USA, 2nd edition, 2009.

He, C., Kraft, P., Chen, C., Buring, J. E., Paré, G., Hankinson, S. E., Chanock, S. J., et al. Genome-wide association studies identify loci associated with age at menarche and age at natural menopause. *Nat Genet*, 41:724–728, 2009.

Henderson, C. R. Estimation of genetic parameters (abstract). *Ann Math Stat*, 21:309–310, 1950.

Henderson, C. R. Estimation of variance and covariance components. *Biometrics*, 9:226–252, 1953.

Henderson, C. R. Selection index and expected genetic advance. In Hanson, W. D. and Robinson, H. F., editors, *Statistical genetics and plant breeding*, Publication 982, pages 141–163. National Academy of Sciences–National Research Council, Washington, D.C., USA, 1963.

Henderson, C. R. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31:423–447, 1975.

Henderson, C. R. Best linear unbiased prediction of nonadditive genetic merits. *J Anim Sci*, 60:111–117, 1985.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA*, 106:9362–9367, 2009.

Hoerl, A. E. and Kennard, R. W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

Hofheinz, N. and Frisch, M. Heteroscedastic ridge regression approaches for genome-wide prediction with a focus on computational efficiency and accurate effect estimation. *G3 (Bethesda)*, 4:539–546, 2014.

Hofheinz, N., Borchardt, D., Weissleder, K., and Frisch, M. Genome-based prediction of test cross performance in two subsequent breeding cycles. *Theor Appl Genet*, 125:1639–1645, 2012.

Jeffries, S. and Konnert, C. Regret and psychological well-being among voluntarily and involuntarily childless women and mothers. *Int J Aging Hum Dev*, 54:89–106, 2002.

Jensen, A. R. Note on why genetic correlations are not squared. *Psychol Bull*, 75:223–224, 1971.

Jolly, M., Sebire, N., Harris, J., Robinson, S., and Regan, L. The risks associated with pregnancy in women aged 35 years or older. *Hum Reprod*, 15:2433–2437, 2000.

Kaprio, J., Hammar, N., Koskenvuo, M., Floderus-Myrhed, B., Langinvainio, H., and Sarna, S. Cigarette smoking and alcohol use in Finland and Sweden: a cross-national twin study. *Int J Epidemiol*, 11:378–386, 1982.

Kimeldorf, G. S. and Wahba, G. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann Math Stat*, 41:495–502, 1970.

Kirk, K. M., Blomberg, S. P., Duffy, D. L., Heath, A. C., Owens, I. P. F., and Martin, N. G. Natural selection and quantitative genetics of life-history traits in Western women: a twin study. *Evolution*, 55: 423–435, 2001.

Koropeckyj-Cox, T. and Call, V. R. A. Characteristics of older childless persons and parents. Cross-national comparisons. *J Fam Issues*, 28: 1362–1414, 2007.

Lander, E. S. and Schork, N. J. Genetic dissection of complex traits. *Science*, 265:2037–2048, 1994.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., Willer, C. J., et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467:832–838, 2010.

Lebrec, J. J., Stijnen, T., and Van Houwelingen, H. C. Dealing with heterogeneity between cohorts in genomewide SNP association studies. *Stat Appl Genet Mol Biol*, 9:8, 2010.

Lee, J. J. and Chow, C. C. Conditions for the validity of SNP-based heritability estimation. *Hum Genet*, 133:1011–1022, 2014.

Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M., and Wray, N. R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, 28:2540–2542, 2012.

Lee, S., Teslovich, T. M., Boehnke, M., and Lin, X. General framework for meta-analysis of rare variants in sequencing association studies. *Am J Hum Genet*, 93:42–53, 2013.

Leridon, H. A new estimate of permanent sterility by age: sterility defined as the inability to conceive. *Popul Stud (Camb)*, 62:15–24, 2008.

Li, M. X., Yeung, J. M. Y., Cherny, S. S., and Sham, P. C. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum Genet*, 131:747–756, 2012.

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518:197–206, 2015.

Maher, B. Personal genomes: the case of the missing heritability. *Nature*, 456:18–21, 2008.

Maier, R., Moser, G., Chen, G. B., Ripke, S., the Cross-Disorder Working Group of the Psychiatric Genomics Consortium, Coryell, W., Potash, J. B., et al. Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am J Hum Genet*, 96:283–294, 2015.

Malo, N., Libiger, O., and Schork, N. J. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. *Am J Hum Genet*, 82:375–385, 2008.

Manolio, T. A., Brooks, L. D., and Collins, F. S. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest*, 118: 1590–1605, 2008.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., et al. Finding the missing heritability of complex diseases. *Nature*, 461:747–753, 2009.

Marchini, J. and Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet*, 11:499–511, 2010.

Martin, N. G., Eaves, L. J., Heath, A. C., Jardine, R., Feingold, L. M., and Eysenck, H. J. Transmission of social attitudes. *Proc Natl Acad Sci USA*, 83:4364–4368, 1986.

Mascarenhas, M. N., Flaxman, S. R., Boerma, T., Vanderpoel, S., and Stevens, G. A. National, regional, and global trends in infertility prevalence since 1990: a systematic analysis of 277 health surveys. *PLOS Med*, 9:e1001356, 2012.

McVean, G. A., Altshuler, D. M., and the 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56–65, 2012.

McVean, G., Daly, M. J., and the International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:851–861, 2007.

Mehta, D., Tropf, F. C., Gratten, J., Bakshi, A., Zhu, Z., Bacanu, S. A., Hemani, G., et al. Evidence for genetic overlap between schizophrenia and age at first birth in women. *JAMA Psychiatry*, 73:497–505, 2016.

Meigs, J. B., Shrader, P., Sullivan, L. M., McAteer, J. B., Fox, C. S., Dupuis, J., Manning, A. K., et al. Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N Engl J Med*, 359:2208–2219, 2008.

Menken, J. Age and fertility: how late can you wait? *Demography*, 22: 469–483, 1985.

Messerlian, C., Maclagan, L., and Basso, O. Infertility and the risk of adverse pregnancy outcomes: a systematic review and meta-analysis. *Hum Reprod*, 28:125–137, 2013.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157:1819–1829, 2001.

Meyer, K. Maximum likelihood estimation of variance components for a multivariate mixed model with equal design matrices. *Biometrics*, 41:153–165, 1985.

Meyer, K. Estimating variances and covariances for multivariate animal models by restricted maximum likelihood. *Genet Sel Evol*, 23:67–83, 1991.

Mills, M. C. and Tropf, F. C. The biodemography of fertility: a review and future research frontiers. In Hank, K. and Kreyenfeld, M., editors, *Social Demography – Forschung an der Schnittstelle von Soziologie und Demografie*, pages 397–424. Springer Fachmedien, Wiesbaden, Hessen, Germany, 2016.

Mills, M. C., Rindfuss, R. R., McDonald, P., and Te Velde, E., on behalf of the ESHRE Reproduction and Society Task Force. Why do people postpone parenthood? Reasons and social policy incentives. *Hum Reprod Update*, 17:848–860, 2011.

Mincer, J. A. Schooling and earnings. In Mincer, J. A., editor, *Schooling, experience, and earnings*, pages 41–63. National Bureau of Economic Research, Cambridge, Massachusetts, USA, 1974.

Montgomery, G. W., Zondervan, K. T., and Nyholt, D. R. The future for genetic studies in reproduction. *Mol Hum Reprod*, 20:1–14, 2014.

Morota, G. and Gianola, D. Kernel-based whole-genome prediction of complex traits: a review. *Front Genet*, 5:363, 2014.

Morota, G., Koyama, M., Rosa, G. J. M., Weigel, K. A., and Gianola, D. Predicting complex traits using a diffusion kernel on genetic markers with an application to dairy cattle and wheat data. *Genet Sel Evol*, 45:17, 2013.

Morris, A. P. Transethnic meta-analysis of genomewide association studies. *Genet Epidemiol*, 35:809–822, 2011.

Nalls, M. A., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., Saad, M., DeStefano, A. L., et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat Genet*, 46:989–993, 2014.

Neale, M. C., Walters, E. E., Eaves, L. J., Maes, H. H., and Kendler, K. S. Multivariate genetic analysis of twin-family data on fears: Mx models. *Behav Genet*, 24:119–139, 1994.

Nei, M. Definition and estimation of fixation indices. *Evolution*, 40: 643–645, 1986.

OECD. *Doing Better for Families*. OECD Publishing, Paris, France, 2011.

Ogutu, J. O., Schulz-Streeck, T., and Piepho, H. P. Genomic selection using regularized linear regression models: ridge regression, LASSO, elastic net and their extensions. *BMC Proc*, 6:Suppl 2 S10, 2012.

Okbay, A., Baselmans, B. M. L., De Neve, J. E., Turley, P., Nivard, M. G., Fontana, M. A., Meddens, S. F. W., et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat Genet*, 48:624–633, 2016a.

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., Turley, P., et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*, 533:539–542, 2016b.

Ollier, W., Sprosen, T., and Peakman, T. UK Biobank: from concept to reality. *Pharmacogenomics*, 6:639–646, 2005.

Painter, J. N., Anderson, C. A., Nyholt, D. R., Macgregor, S., Lin, J., Lee, S. H., Lambert, A., et al. Genome-wide association study identifies a locus at 7p15.2 associated with endometriosis. *Nat Genet*, 43:51–54, 2011.

Patterson, H. D. and Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554, 1971.

Perry, J. R. B., Corre, T., Esko, T., Chasman, D. I., Fischer, K., Franceschini, N., He, C., et al. A genome-wide association study of early menopause and the combined impact of identified variants. *Hum Mol Genet*, 22:1465–1472, 2013.

Perry, J. R. B., Day, F. R., Elks, C. E., Sulem, P., Thompson, D. J., Ferreira, T., He, C., et al. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature*, 514:92–97, 2014.

Pharoah, P. D. P., Antoniou, A., Bobrow, M., Zimmern, R. L., Easton, D. F., and Ponder, B. A. J. Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet*, 31:33–36, 2002.

Phillips, K. and Fulker, D. W. Quantitative genetic analysis of longitudinal trends in adoption designs with application to IQ in the Colorado Adoption Project. *Behav Genet*, 19:621–658, 1989.

Piepho, H. P. Ridge regression and extensions for genomewide selection in maize. *Crop Sci*, 49:1165–1176, 2009.

Polderman, T. J. C., Benyamin, B., De Leeuw, C. A., Sullivan, P. F., Van Bochoven, A., Visscher, P. M., and Posthuma, D. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat Genet*, 47:702–709, 2015.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38: 904–909, 2006.

Purcell, S. M., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., et al. `PLINK`: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81:559–575, 2007.

Purcell, S. M., Sklar, P., and the International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460:748–752, 2009.

Rahmioglu, N., Nyholt, D. R., Morris, A. P., Missmer, S. A., Montgomery, G. W., and Zondervan, K. T. Genetic variants underlying risk of

endometriosis: insights from meta-analysis of eight genome-wide association and replication datasets. *Hum Reprod Update*, 20:702–716, 2014.

Rietveld, C. A., Esko, T., Davies, G., Pers, T. H., Turley, P., Benyamin, B., Chabris, C. F., et al. Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proc Natl Acad Sci USA*, 111:13790–13794, 2014.

Rietveld, C. A., Cesarini, D., Benjamin, D. J., Koellinger, P. D., De Neve, J. E., Tiemeier, H., Johannesson, M., et al. Molecular genetics and subjective well-being. *Proc Natl Acad Sci USA*, 110:9692–9697, 2013a.

Rietveld, C. A., Medland, S. E., Derringer, J., Yang, J., Esko, T., Martin, N. W., Westra, H. J., et al. GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science*, 340:1467–1471, 2013b.

Ripke, S., Neale, B. M., and the Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511:421–427, 2014.

Robinson, G. E., Grozinger, C. M., and Whitfield, C. W. Sociogenomics: social life in molecular terms. *Nat Rev Genet*, 6:257–270, 2005.

Robinson, M. R., Hemani, G., Medina-Gomez, C., Mezzavilla, M., Esko, T., Shakhbazov, K., Powell, J. E., et al. Population genetic differentiation of height and body mass index across Europe. *Nat Genet*, 47: 1357–1362, 2015.

Sabourin, J., Nobel, A. B., and Valdar, W. Fine-mapping additive and dominant SNP effects using group-LASSO and fractional resample model averaging. *Genet Epidemiol*, 39:77–88, 2015.

Schaeffer, L. R. Strategy for applying genome-wide selection in dairy cattle. *J Anim Breed Genet*, 123:218–223, 2006.

Schwarz, G. Estimating the dimension of a model. *Ann Stat*, 6:461–464, 1978.

Searle, S. R., Casella, G., and McCulloch, C. E. *Variance Components*, chapter M.4, pages 451–452. John Wiley and Sons, Hoboken, New Jersey, USA, 1992.

Sham, P. C. and Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet*, 15:335–346, 2014.

Shen, X., Alam, M., Fikse, F., and Rönnegård, L. A novel generalized ridge regression method for quantitative genetics. *Genetics*, 193: 1255–1268, 2013.

Sherman, J. and Morrison, W. J. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Ann Math Stat*, 21:124–127, 1950.

Shi, J. and Lee, S. A novel random effect model for GWAS meta-analysis and its application to trans-ethnic meta-analysis. *Biometrics*, 72: 945–954, 2016.

Smoller, J. W., Kendler, K., Craddock, N., and the Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet*, 381:1371–1379, 2013.

Snieder, H., MacGregor, A. J., and Spector, T. D. Genes control the cessation of a woman's reproductive life: a twin study of hysterectomy and age at menopause. *J Clin Endocrinol Metab*, 83:1875–1880, 1998.

Speed, D., Hemani, G., Johnson, M. R., and Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet*, 91:1011–1021, 2012.

Speed, T. That BLUP is a good thing: the estimation of random effects: comment. *Stat Sci*, 6:42–44, 1991.

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Lango Allen, H., et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet*, 42:937–948, 2010.

Stearns, S. C., Byars, S. G., Govindaraju, D. R., and Ewbank, D. Measuring selection in contemporary human populations. *Nat Rev Genet*, 11:611–622, 2010.

Stolk, L., Zhai, G., Van Meurs, J. B. J., Verbiest, M. M. P. J., Visser, J. A., Estrada, K., Rivadeneira, F., et al. Loci at chromosomes 13, 19

and 20 influence age at natural menopause. *Nat Genet*, 41:645–647, 2009.

Stolk, L., Perry, J. R. B., Chasman, D. I., He, C., Mangino, M., Sulem, P., Barbalic, M., et al. Meta-analyses identify 13 loci associated with age at menopause and highlight DNA repair and immune pathways. *Nat Genet*, 44:260–268, 2012.

Sulem, P., Gudbjartsson, D. F., Rafnar, T., Holm, H., Olafsdottir, E. J., Olafsdottir, G. H., Jonsson, T., et al. Genome-wide association study identifies sequence variants on 6q21 associated with age at menarche. *Nat Genet*, 41:734–738, 2009.

Tanturri, M. L. and Mencarini, L. Childless or childfree? Paths to voluntary childlessness in Italy. *Popul Dev Rev*, 34:51–77, 2008.

Tarín, J. J., Brines, J., and Cano, A. Long-term effects of delayed parenthood. *Hum Reprod*, 13:2371–2376, 1998.

The Economist. Daily chart: scientific papers get more authors. *The Economist Group*, Nov 2016a.

The Economist. Why research papers have so many authors. *The Economist Group*, Nov 2016b.

Thornton, T. A., Tang, H., Hoffmann, T. J., Ochs-Balcom, H. M., Caan, B. J., and Risch, N. Estimating kinship in admixed populations. *Am J Hum Genet*, 91:122–138, 2012.

Tropf, F. C., Stulp, G., Barban, N., Visscher, P. M., Yang, J., Snieder, H., and Mills, M. C. Human fertility, molecular genetics, and natural selection in modern societies. *PLOS ONE*, 10:e0126821, 2015.

Tusell, L., Pérez-Rodríguez, P., Forni, S., and Gianola, D. Model averaging for genome-enabled prediction with reproducing kernel Hilbert spaces: a case study with pig litter size and wheat yield. *J Anim Breed Genet*, 131:105–115, 2014.

Usai, M. G., Goddard, M. E., and Hayes, B. J. LASSO with cross-validation for genomic selection. *Genet Res (Camb)*, 91:427–436, 2009.

Van Der Most, P. J., Vaez, A., Prins, B. P., Munoz, M. L., Snieder, H., Alizadeh, B. Z., and Nolte, I. M. QCGWAS: a flexible R package for automated quality control of genome-wide association results. *Bioinformatics*, 30:1185–1186, 2014.

Van Wieringen, W. N., Kun, D., Hampel, R., and Boulesteix, A. L. Survival prediction using gene expression data: a review and comparison. *Comput Stat Data Anal*, 53:1590–1603, 2009.

Vattikuti, S., Guo, J., and Chow, C. C. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLOS Genet*, 8:e1002637, 2012.

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am J Hum Genet*, 97: 576–592, 2015.

Visscher, P. M., Yang, J., and Goddard, M. E. A commentary on 'common SNPs explain a large proportion of the heritability for human height' by Yang et al. (2010). *Twin Res Hum Genet*, 13:517–524, 2010.

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. Five years of GWAS discovery. *Am J Hum Genet*, 90:7–24, 2012.

Visscher, P. M., Hemani, G., Vinkhuyzen, A. A. E., Chen, G. B., Lee, S. H., Wray, N. R., Goddard, M. E., and Yang, J. Statistical power to detect genetic (co) variance of complex traits using SNP data in unrelated samples. *PLOS Genet*, 10:e1004269, 2014.

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., Burtt, N. P., et al. The Metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLOS Genet*, 8:e1002793, 2012.

Warren, H., Casas, J. P., Hingorani, A., Dudbridge, F., and Whittaker, J. C. Genetic prediction of quantitative lipid traits: comparing shrinkage models to gene scores. *Genet Epidemiol*, 38:72–83, 2014.

Weedon, M. N., Lango Allen, H., Lindgren, C. M., Wallace, C., Evans, D. M., Mangino, M., Freathy, R. M., et al. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet*, 40: 575–583, 2008.

Weir, B. S. and Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution*, 38:1358–1370, 1984.

Weir, B. S. and Hill, W. G. Estimating F-statistics. *Annu Rev Genet*, 36: 721–750, 2002.

Wen, X. and Stephens, M. Bayesian methods for genetic association analysis with heterogeneous subgroups: from meta-analyses to gene-environment interactions. *Ann Appl Stat*, 8:176–203, 2014.

Whittaker, J. C., Thompson, R., and Denham, M. C. Marker-assisted selection using ridge regression. *Genet Res (Camb)*, 75:249–252, 2000.

Willer, C. J., Speliotes, E. K., Loos, R. J. F., Li, S., Lindgren, C. M., Heid, I. M., Berndt, S. I., et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet*, 41:25–34, 2008.

Willer, C. J., Li, Y., and Abecasis, G. R. `METAL`: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26: 2190–2191, 2010.

Winkler, T. W., Day, F. R., Croteau-Chonka, D. C., Wood, A. R., Locke, A. E., Mägi, R., Ferreira, T., et al. Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc*, 9:1192–1212, 2014.

Witte, J. S., Visscher, P. M., and Wray, N. R. The contribution of genetic variants to disease depends on the ruler. *Nat Rev Genet*, 15:765–776, 2014.

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet*, 46:1173–1186, 2014.

Woodbury, M. A. Inverting modified matrices. Memorandum Report 42, Statistical Research Group, Princeton University, Princeton, New Jersey, USA, 1950.

Wray, N. R. and Maier, R. Genetic basis of complex genetic disease: the contribution of disease heterogeneity to missing heritability. *Curr Epidemiol Rep*, 1:220–227, 2014.

Wray, N. R., Goddard, M. E., and Visscher, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res*, 17:1520–1528, 2007.

Wray, N. R., Lee, S. H., and Kendler, K. S. Impact of diagnostic misclassification on estimation of genetic correlations using genome-wide genotypes. *Eur J Hum Genet*, 20:668–674, 2012.

Wray, N. R., Lee, S. H., and the Cross-Disorder Group of the Psychiatric Genomics Consortium. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet*, 45:984–994, 2013a.

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., and Visscher, P. M. Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet*, 14:507–515, 2013b.

Wright, S. The genetical structure of populations. *Ann Hum Genet*, 15: 323–354, 1949.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*, 42:565–569, 2010.

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. `GCTA`: a tool for genome-wide complex trait analysis. *Am J Hum Genet*, 88:76–82, 2011a.

Yang, J., Weedon, M. N., Purcell, S. M., Lettre, G., Estrada, K., Willer, C. J., Smith, A. V., et al. Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet*, 19:807–812, 2011b.

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., and Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet*, 46:100–106, 2014.

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A. E., Lee, S. H., Robinson, M. R., et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet*, 47:1114–1120, 2015a.

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A. E., Nolte, I. M., Van Vliet-Ostaptchouk, J. V., et al. Genome-wide genetic homogeneity between sexes and populations for human height and body mass index. *Hum Mol Genet*, 24:7445–7449, 2015b.

Yang, J., Lee, S. H., Wray, N. R., Goddard, M. E., and Visscher, P. M. GCTA-GREML accounts for linkage disequilibrium when estimating genetic variance from genome-wide SNPs. *Proc Natl Acad Sci USA*, 113:E4579–E4580, 2016.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc Series B Stat Methodol*, 68:49–67, 2006.

Zellner, A. and Huang, D. S. Further properties of efficient estimators for seemingly unrelated regression equations. *Int Econ Rev*, 3:300–313, 1962.

Zietsch, B. P., Kuja-Halkola, R., Walum, H., and Verweij, K. J. H. Perfect genetic correlation between number of offspring and grand-offspring in an industrialized human population. *Proc Natl Acad Sci USA*, 111:1032–1036, 2014.

Zou, H. The adaptive LASSO and its oracle properties. *J Am Stat Assoc*, 101:1418–1429, 2006.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*, 67:301–320, 2005.

Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci USA*, 109:1193–1198, 2012.

About the Author

Ronald de Vlaming (1987) holds a M.Sc. degree *cum laude* in Econometrics and Management Science from the *Erasmus School of Economics* (ESE), *Erasmus University Rotterdam*. In 2013 Ronald started as a Ph.D. candidate under the supervision of professors A. Roy Thurik, Patrick J. F. Groenen, and Philipp D. Koellinger. He carried out his research within the *Department of Applied Economics* at the ESE as a member of the *Erasmus University Rotterdam Institute for Behavior and Biology* and of the *Social Science Genetic Association Consortium*. In 2016 Ronald visited the *Institute for Molecular Bioscience* at the *University of Queensland* for a period of three months. During his research visit he was supervised by professor Peter M. Visscher.

Ronald's research focusses on the application of statistical methods, matrix algebra, and numerical methods to estimation problems that arise in the study of the genetic architecture of complex traits. His work has been published in the following peer-reviewed journals: *PLOS Genetics*, *BioMed Research International*, *Nature Genetics*, and *Nature*. He has presented his work at meetings of the *Behavior Genetics Association*, the *Conference on Computational Statistics*, and the *European Human Genetics Conference*. Ronald will continue his career as postdoctoral researcher at the *VU University Amsterdam*.

Portfolio

Peer-Reviewed Publications

- **Meta-GWAS Accuracy and Power (MetaGAP) calculator shows that hiding heritability is partially due to imperfect genetic correlations across studies**. R. de Vlaming, A. Okbay, C.A. Rietveld, M. Johannesson, P.K.E. Magnusson, A.G. Uitterlinden, F.J.A. van Rooij, A. Hofman, P.J.F. Groenen, A.R. Thurik, and P.D. Koellinger, 2017, *PLOS Genetics*, **13**, e1006495.
- **Genome-wide analysis identifies 12 loci influencing human reproductive behavior**. N. Barban, R. Jansen, R. de Vlaming, A. Vaez, J.J. Mandemakers, F.C. Tropf, X. Shen, J.F. Wilson, D.I. Chasman, I.M. Nolte, V. Tragante, S.W. van der Laan, J.R.B. Perry, A. Kong, ... D.J. Benjamin, D. Cesarini, P.D. Koellinger, M. den Hoed, H. Snieder, M.C. Mills, 2016, *Nature Genetics*, **48**, 1462–1472.
- **Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses**. A. Okbay, B.M.L. Baselmans, J.E. de Neve, P. Turley, M.G. Nivard, M.A. Fontana, S.F.W. Meddens, R.K. Linner, C.A. Rietveld, J. Derringer, J. Gratten, J.J. Lee, J.Z. Liu, R. de Vlaming, T.S. Ahluwalia, ... R.F. Krueger, J.P. Beauchamp, P.D. Koellinger, D.J. Benjamin, M. Bartels, D. Cesarini, 2016, *Nature Genetics*, **48**, 624–633.
- **Genome-wide association study identifies 74 loci associated with educational attainment**. A. Okbay, J.P. Beauchamp, M.A. Fontana, J.J. Lee, T.H. Pers, C.A. Rietveld, P. Turley, G.B. Chen, V. Emilsson, S.F.W. Meddens, S. Oskarsson, J.K. Pickrell, K. Thom, P. Timshel, R. de Vlaming, A. Abdellaoui, ... P.M. Visscher, T. Esko, P.D. Koellinger, D. Cesarini, D.J. Benjamin, 2016, *Nature*, **533**, 539–542.
- **The current and future use of ridge regression for prediction in quantitative genetics**. R. de Vlaming and P.J.F. Groenen, 2015, *BioMed Research International*, **2015**, 143712.

Working Papers

- **Estimating the LD-score-regression intercept from individual-level data**. R. de Vlaming and P.M. Visscher.
- **Estimating the relation between allele frequency and effect size from GWAS summary statistics**. R. de Vlaming and P.M. Visscher.

- **Estimating the contribution of epistasis to variation in human height and body-mass index using GREML**. R. de Vlaming and C.A. Rietveld.
- **Multivariate average-information constrained GREML**. R. de Vlaming, E.A.W. Slob, P.J.F. Groenen, and C.A. Rietveld.
- **Large-scale genetic study of risk tolerance and risky behaviors identifies new loci and reveals shared genetic influences**. R. Karlsson Linnér, J.P. Beauchamp, et al.
- **Mega-analysis of 31,396 individuals from 6 countries uncovers strong gene-environment interaction for human fertility**. F. Tropf et al.

OTHER ACTIVITIES

- **Developed online MetaGAP calculator**, based on De Vlaming et al. (2017), available at www.devlaming.eu.
- **Advisory role** in the final stages before the publication of the manuscript titled *"Total factor productivity and the role of entrepreneurship"* by Erken et al. (2016).
- **Reviewing abstracts** for the conference of Research in Entrepreneurship and Small Business.
- **Represented PhD candidates** at *Erasmus School of Economics* during an assessment by an **international peer-review committee**.
- **Visited** the *University College London* for joint work on the **GWAS of subjective well-being**.
- Three-month **research visit** to prof. Peter M. Visscher's research group at the *Institute for Molecular Bioscience* at the *University of Queensland*.
- Setting up **ICT systems** for secure group access to **large-scale genetic data** from various studies.

REFEREED ARTICLES SUBMITTED TO

- *Behavior Genetics*
- *European Journal of Epidemiology*
- *Small Business Economics* (3×)
- *Management Science*

TEACHING

- Genoeconomics (2014/2015, *Tinbergen Institute*)
- Seminar Innovation and Entrepreneurship (2013/2014, *Erasmus School of Economics*)
- Seminar Innovation and Entrepreneurship (2012/2013, *Erasmus School of Economics*)
- Introduction to Entrepreneurship and Strategy Economics (2012/2013, *Erasmus School of Economics*)
- Supervised various bachelor and master's theses

PhD Courses and Certificates

- Complex Trait Genetics (*VU University*, Amsterdam)
- Statistical Genetic Methods for Human Complex (*Institute for Behavioral Genetics*)
- Erasmus Summer Programme (*Netherlands Institute for Health Sciences*, *Erasmus MC*)
- SNP Course IX (*Molecular Medicine*, *Erasmus MC*)
- Advanced Econometrics I, II, and III (*Tinbergen Institute*)
- Measure Theory and Stochastic Processes (*Tinbergen Institute*)
- Mathematics II and Principles of Programming in Econometrics (*Tinbergen Institute*)
- Stochastic Dynamic Optimization (*Erasmus Research Institute of Management*)
- Advances in the Economics of Entrepreneurship (*Erasmus Research Institute of Management*)
- Cambridge Certificate of Proficiency in English, Grade A (*Cambridge ESOL Examinations*)

Conferences, Workshops, and Meetings

- Workshop of the *Social Science Genetic Association Consortium* (Mallorca, Spain, 2015)
- Workshop of the *Social Science Genetic Association Consortium* (Bro, Sweden, 2014)
- Workshop of the *Social Science Genetic Association Consortium* (Ösmo, Sweden, 2013)
- Workshop of the *Social Science Genetic Association Consortium* (Reykjavik, Iceland, 2012; only attended, no presentation)
- Meeting of the *Behavior Genetics Association* (Brisbane, Queensland, Australia, 2016)
- Meeting of the *Behavior Genetics Association* (San Diego, California, USA, 2015)
- Meeting of the *Behavior Genetics Association* (Charlottesville, Virginia, USA, 2014)
- Meeting of the *Cohorts for Heart and Aging Research in Genomic Epidemiology* (Rotterdam, the Netherlands, 2013; only attended, no presentation))
- Meeting of the *Cohorts for Heart and Aging Research in Genomic Epidemiology* (Houston, Texas, USA, 2012; only attended, no presentation)
- Conference of the *European Society of Human Genetics* (Barcelona, Spain, 2016)
- Conference of the *European Research Consortium for Informatics and Mathematics* (London, UK, 2015)
- International Conference on *Computational Statistics* (Geneva, Switzerland, 2014)
- Meeting of the *Dutch-Flemish Classification Society* (Antwerp, Belgium, 2013)

EDUCATION

- **Master of Science in Econometrics and Management Science** (2010-2012, *Erasmus University Rotterdam*)
  - Specialization: Applied Econometric
  - Thesis: Forecasting advertisement traffic (grade: 9.0)
  - Grade-point average: 8.3 (*cum laude*)
- **Bachelor of Science in Econometrics and Operations Research** (2005-2010, *Erasmus University Rotterdam*)
  - Minor: Psychology, Major: Financial Econometrics
  - Thesis: Predicting short term horizon excess bond returns (grade: 8.0)
  - Grade-point average: 7.6

The ERIM PhD Series

The ERIM PhD Series contains PhD dissertations in the field of Research in Management defended at Erasmus University Rotterdam and supervised by senior researchers affiliated to the Erasmus Research Institute of Management (ERIM). All dissertations in the ERIM PhD Series are available in full text through the ERIM Electronic Series Portal: http://repub.eur.nl/pub. ERIM is the joint research institute of the Rotterdam School of Management (RSM) and the Erasmus School of Economics at the Erasmus University Rotterdam (EUR).

Dissertations in the last five years

Abbink, E.J., *Crew Management in Passenger Rail Transport*, Promotors: Prof. L.G. Kroon & Prof. A.P.M. Wagelmans, EPS-2014-325-LIS, http://repub.eur.nl/pub/76927

Acar, O.A., *Crowdsourcing for Innovation: Unpacking Motivational, Knowledge and Relational Mechanisms of Innovative Behavior in Crowdsourcing Platforms*, Promotor: Prof. J.C.M. van den Ende, EPS-2014-321-LIS, http://repub.eur.nl/pub/76076

Akin Ates, M., *Purchasing and Supply Management at the Purchase Category Level: Strategy, structure and performance*, Promotors: Prof. J.Y.F. Wynstra & Dr E.M. van Raaij, EPS-2014-300-LIS, http://repub.eur.nl/pub/50283

Akpinar, E., *Consumer Information Sharing*, Promotor: Prof. A. Smidts, EPS-2013-297-MKT, http://repub.eur.nl/pub/50140

Alexander, L., *People, Politics, and Innovation: A Process Perspective*, Promotors: Prof. H.G. Barkema & Prof. D.L. van Knippenberg, EPS-2014-331-S&E, http://repub.eur.nl/pub/77209

Alexiou, A., *Management of Emerging Technologies and the Learning Organization: Lessons from the Cloud and Serious Games Technology*, Promotors:

Prof. S.J. Magala, Prof. M.C. Schippers, & Dr I. Oshri, EPS-2016-404-ORG, http://repub.eur.nl/pub/93818

Almeida e Santos Nogueira, R.J. de, *Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty*, Promotors: Prof. U. Kaymak & Prof. J.M.C. Sousa, EPS-2014-310-LIS, http://repub.eur.nl/pub/51560

Bannouh, K., *Measuring and Forecasting Financial Market Volatility using High-frequency Data*, Promotor: Prof. D.J.C. van Dijk, EPS-2013-273-F&A, http://repub.eur.nl/pub/38240

Ben-Menahem, S.M., *Strategic Timing and Proactiveness of Organizations*, Promotors: Prof. H.W. Volberda & Prof. F.A.J. van den Bosch, EPS-2013-278-S&E, http://repub.eur.nl/pub/39128

Benschop, N., *Biases in Project Escalation: Names, frames & construal levels*, Promotors: Prof. K.I.M. Rhode, Prof. H.R. Commandeur, Prof. M. Keil & Dr A.L.P. Nuijten, EPS-2015-375-S&E, http://repub.eur.nl/pub/79408

Berg, W.E. van den, *Understanding Salesforce Behavior using Genetic Association Studies*, Promotor: Prof. W.J.M.I. Verbeke, EPS-2014-311-MKT, http://repub.eur.nl/pub/51440

Beusichem, H.C. van, *Firms and Financial Markets: Empirical Studies on the Informational Value of Dividends, Governance and Financial Reporting*, Promotors: Prof. A. de Jong & Dr G. Westerhuis, EPS-2016-378-F&A, http://repub.eur.nl/pub/93079

Bliek, R. de, *Empirical Studies on the Economic Impact of Trust*, Promotor: Prof. J. Veenman & Prof. Ph.H.B.F. Franses, EPS-2015-324-ORG, http://repub.eur.nl/pub/78159

Boons, M., *Working Together Alone in the Online Crowd: The Effects of Social Motivations and Individual Knowledge Backgrounds on the Participation and Performance of Members of Online Crowdsourcing Platforms*, Promotors: Prof. H.G. Barkema & Dr D.A. Stam, EPS-2014-306-S&E, http://repub.eur.nl/pub/50711

Brazys, J., *Aggregated Marcoeconomic News and Price Discovery*, Promotor: Prof. W.F.C. Verschoor, EPS-2015-351-F&A, http://repub.eur.nl/pub/78243

Byington, E., *Exploring Coworker Relationships: Antecedents and Dimensions of Interpersonal Fit, Coworker Satisfaction, and Relational Models*, Promotor: Prof. D.L. van Knippenberg, EPS-2013-292-ORG, http://repub.eur.nl/pub/41508

Cancurtaran, P., *Essays on Accelerated Product Development*, Promotors: Prof. F. Langerak & Prof. G.H. van Bruggen, EPS-2014-317-MKT,

http://repub.eur.nl/pub/76074

Caron, E.A.M., *Explanation of Exceptional Values in Multi-dimensional Business Databases*, Promotors: Prof. H.A.M. Daniels & Prof. G.W.J. Hendrikse, EPS-2013-296-LIS, http://repub.eur.nl/pub/50005

Carvalho, L. de, *Knowledge Locations in Cities: Emergence and Development Dynamics*, Promotor: Prof. L. Berg, EPS-2013-274-S&E, http://repub.eur.nl/pub/38449

Cranenburgh, K.C. van, *Money or Ethics: Multinational corporations and religious organisations operating in an era of corporate responsibility*, Prof. L.C.P.M. Meijs, Prof. R.J.M. van Tulder & Dr D. Arenas, EPS-2016-385-ORG, http://repub.eur.nl/pub/93104

Consiglio, I., *Others: Essays on Interpersonal and Consumer Behavior*, Promotor: Prof. S.M.J. van Osselaer, EPS-2016-366-MKT, http://repub.eur.nl/pub/79820

Cox, R.H.G.M., *To Own, To Finance, and To Insure - Residential Real Estate Revealed*, Promotor: Prof. D. Brounen, EPS-2013-290-F&A, http://repub.eur.nl/pub/40964

Darnihamedani, P., *Individual Characteristics, Contextual Factors and Entrepreneurial Behavior*, Promotors: Prof. A.R. Thurik & S.J.A. Hessels, EPS-2016-360-S&E, http://repub.eur.nl/pub/93280

Deng, W., *Social Capital and Diversification of Cooperatives*, Promotor: Prof. G.W.J. Hendrikse, EPS-2015-341-ORG, http://repub.eur.nl/pub/77449

Depecik, B.E., *Revitalizing brands and brand: Essays on Brand and Brand Portfolio Management Strategies*, Promotors: Prof. G.H. van Bruggen, Dr Y.M. van Everdingen, & Dr M.B. Ataman, EPS-2016-406-MKT, http://repub.eur.nl/pub/93507

Dollevoet, T.A.B., *Delay Management and Dispatching in Railways*, Promotor: Prof. A.P.M. Wagelmans, EPS-2013-272-LIS, http://repub.eur.nl/pub/38241

Duyvesteyn, J.G., *Empirical Studies on Sovereign Fixed Income Markets*, Promotors: Prof. P.Verwijmeren & Prof. M.P.E. Martens, EPS-2015-361-F&A, hdl.handle.net/1765/79033

Duursema, H., *Strategic Leadership: Moving Beyond the Leader-Follower Dyad*, Promotor: Prof. R.J.M. van Tulder, EPS-2013-279-ORG, http://repub.eur.nl/pub/39129

Elemes, A., *Studies on Determinants and Consequences of Financial Reporting Quality*, Promotor: Prof. E. Peek, EPS-2015-354-F&A, http://hdl.handle.net/1765/79037

Ellen, S. ter, *Measurement, Dynamics, and Implications of Heterogeneous Beliefs in Financial Markets*, Promotor: Prof. W.F.C. Verschoor, EPS-2015-343-F&A, http://repub.eur.nl/pub/78191

Erlemann, C., *Gender and Leadership Aspiration: The Impact of the Organizational Environment*, Promotor: Prof. D.L. van Knippenberg, EPS-2016-376-ORG, http://repub.eur.nl/pub/79409

Eskenazi, P.I., *The Accountable Animal*, Promotor: Prof. F.G.H. Hartmann, EPS-2015-355-F&A, http://repub.eur.nl/pub/78300

Evangelidis, I., *Preference Construction under Prominence*, Promotor: Prof. S.M.J. van Osselaer, EPS-2015-340-MKT, http://repub.eur.nl/pub/78202

Faber, N., *Structuring Warehouse Management*, Promotors: Prof. M.B.M. de Koster & Prof. A. Smidts, EPS-2015-336-LIS, http://repub.eur.nl/pub/78603

Fernald, K., *The Waves of Biotechnological Innovation in Medicine: Interfirm Cooperation Effects and a Venture Capital Perspective*, Promotors: Prof. E. Claassen, Prof. H.P.G. Pennings & Prof. H.R. Commandeur, EPS-2015-371-S&E, http://hdl.handle.net/1765/79120

Fisch, C.O., *Patents and trademarks: Motivations, antecedents, and value in industrialized and emerging markets*, Promotors: Prof. J.H. Block, Prof. H.P.G. Pennings & Prof. A.R. Thurik, EPS-2016-397-S&E, http://repub.eur.nl/pub/94036

Fliers, P.T., *Essays on Financing and Performance: The role of firms, banks and board*, Promotor: Prof. A. de Jong & Prof. P.G.J. Roosenboom, EPS-2016-388-F&A, http://repub.eur.nl/pub/93019

Fourne, S.P., *Managing Organizational Tensions: A Multi-Level Perspective on Exploration, Exploitation and Ambidexterity*, Promotors: Prof. J.J.P. Jansen & Prof. S.J. Magala, EPS-2014-318-S&E, http://repub.eur.nl/pub/76075

Gaast, J.P. van der, *Stochastic Models for Order Picking Systems*, Promotors: Prof. M.B.M de Koster & Prof. I.J.B.F. Adan, EPS-2016-398-LIS, http://repub.eur.nl/pub/93222

Glorie, K.M., *Clearing Barter Exchange Markets: Kidney Exchange and Beyond*, Promotors: Prof. A.P.M. Wagelmans & Prof. J.J. van de Klundert, EPS-2014-329-LIS, http://repub.eur.nl/pub/77183

Hekimoglu, M., *Spare Parts Management of Aging Capital Products*, Promotor: Prof. R. Dekker, EPS-2015-368-LIS, http://repub.eur.nl/pub/79092

Heyde Fernandes, D. von der, *The Functions and Dysfunctions of Reminders*, Promotor: Prof. S.M.J. van Osselaer, EPS-2013-295-MKT, http://repub.eur.nl/pub/41514

Hogenboom, A.C., *Sentiment Analysis of Text Guided by Semantics and Structure*, Promotors: Prof. U. Kaymak & Prof. F.M.G. de Jong, EPS-2015-369-LIS, http://repub.eur.nl/pub/79034

Hogenboom, F.P., *Automated Detection of Financial Events in News Text*, Promotors: Prof. U. Kaymak & Prof. F.M.G. de Jong, EPS-2014-326-LIS, http://repub.eur.nl/pub/77237

Hollen, R.M.A., *Exploratory Studies into Strategies to Enhance Innovation-Driven International Competitiveness in a Port Context: Toward Ambidextrous Ports*, Promotors: Prof. F.A.J. Van Den Bosch & Prof. H.W.Volberda, EPS-2015-372-S&E, http://repub.eur.nl/pub/78881

Hout, D.H. van, *Measuring Meaningful Differences: Sensory Testing Based Decision Making in an Industrial Context; Applications of Signal Detection Theory and Thurstonian Modelling*, Promotors: Prof. P.J.F. Groenen & Prof. G.B. Dijksterhuis, EPS-2014-304-MKT, http://repub.eur.nl/pub/50387

Houwelingen, G.G. van, *Something To Rely On*, Promotors: Prof. D. de Cremer & Prof. M.H. van Dijke, EPS-2014-335-ORG, http://repub.eur.nl/pub/77320

Hurk, E. van der, *Passengers, Information, and Disruptions*, Promotors: Prof. L.G. Kroon & Prof. P.H.M. Vervest, EPS-2015-345-LIS, http://repub.eur.nl/pub/78275

Iseger, P. den, *Fourier and Laplace Transform Inversion with Applications in Finance*, Promotor: Prof. R. Dekker, EPS-2014-322-LIS, http://repub.eur.nl/pub/76954

Jaarsveld, W.L. van, *Maintenance Centered Service Parts Inventory Control*, Promotor: Prof. R. Dekker, EPS-2013-288-LIS, http://repub.eur.nl/pub/39933

Khanagha, S., *Dynamic Capabilities for Managing Emerging Technologies*, Promotor: Prof. H.W. Volberda, EPS-2014-339-S&E, http://repub.eur.nl/pub/77319

Kil, J., *Acquisitions Through a Behavioral and Real Options Lens*, Promotor: Prof. H.T.J. Smit, EPS-2013-298-F&A, http://repub.eur.nl/pub/50142

Klooster, E. van't, *Travel to Learn: the Influence of Cultural Distance on Competence Development in Educational Travel*, Promotors: Prof. F.M. Go & Prof. P.J. van Baalen, EPS-2014-312-MKT, http://repub.eur.nl/pub/51462

Koendjbiharie, S.R., *The Information-Based View on Business Network Performance: Revealing the Performance of Interorganizational Networks*, Promotors: Prof. H.W.G.M. van Heck & Prof. P.H.M. Vervest, EPS-2014-315-LIS, http://repub.eur.nl/pub/51751

Koning, M., *The Financial Reporting Environment: The Role of the Media,*

Regulators and Auditors, Promotors: Prof. G.M.H. Mertens & Prof. P.G.J. Roosenboom, EPS-2014-330-F&A, http://repub.eur.nl/pub/77154

Konter, D.J., *Crossing Borders with HRM: An Inquiry of the Influence of Contextual Differences in the Adoption and Effectiveness of HRM*, Promotors: Prof. J. Paauwe & Dr L.H. Hoeksema, EPS-2014-305-ORG, http://repub.eur.nl/pub/50388

Korkmaz, E., *Bridging Models and Business: Understanding Heterogeneity in Hidden Drivers of Customer Purchase Behavior*, Promotors: Prof. S.L. van de Velde & Prof. D. Fok, EPS-2014-316-LIS, http://repub.eur.nl/pub/76008

Krämer, R., *A license to mine? Community organizing against multinational corporations*, Promotors: Prof. R.J.M. van Tulder & Prof. G.M. Whiteman, EPS-2016-383-ORG, http://repub.eur.nl/pub/94072

Kroezen, J.J., *The Renewal of Mature Industries: An Examination of the Revival of the Dutch Beer Brewing Industry*, Promotor: Prof. P.P.M.A.R. Heugens, EPS-2014-333-S&E, http://repub.eur.nl/pub/77042

Kysucky, V., *Access to Finance in a Cros-Country Context*, Promotor: Prof. L. Norden, EPS-2015-350-F&A, http://repub.eur.nl/pub/78225

Lee, C.I.S.G., *Big Data in Management Research: Exploring New Avenues*, Promotors: Prof. S.J. Magala & Dr W.A. Felps, EPS-2016-365-ORG, http://repub.eur.nl/pub/79818

Legault-Tremblay, P.O., *Corporate Governance During Market Transition: Heterogeneous responses to Institution Tensions in China*, Promotor: Prof. B. Krug, EPS-2015-362-ORG, http://repub.eur.nl/pub/78649

Lenoir, A.S., *Are You Talking to Me? Addressing Consumers in a Globalised World*, Promotors: Prof. S. Puntoni & Prof. S.M.J. van Osselaer, EPS-2015-363-MKT, http://repub.eur.nl/pub/79036

Leunissen, J.M., *All Apologies: On the Willingness of Perpetrators to Apologize*, Promotors: Prof. D. de Cremer & Dr M. van Dijke, EPS-2014-301-ORG, http://repub.eur.nl/pub/50318

Li, D., *Supply Chain Contracting for After-sales Service and Product Support*, Promotor: Prof. M.B.M. de Koster, EPS-2015-347-LIS, http://repub.eur.nl/pub/78526

Li, Z., *Irrationality: What, Why and How*, Promotors: Prof. H. Bleichrodt, Prof. P.P. Wakker, & Prof. K.I.M. Rohde, EPS-2014-338-MKT, http://repub.eur.nl/pub/77205

Liang, Q.X., *Governance, CEO Identity, and Quality Provision of Farmer Cooperatives*, Promotor: Prof. G.W.J. Hendrikse, EPS-2013-281-ORG,

http://repub.eur.nl/pub/39253

Liket, K., *Why 'Doing Good' is not Good Enough: Essays on Social Impact Measurement*, Promotors: Prof. H.R. Commandeur & Dr K.E.H. Maas, EPS-2014-307-STR, http://repub.eur.nl/pub/51130

Loos, M.J.H.M. van der, *Molecular Genetics and Hormones: New Frontiers in Entrepreneurship Research*, Promotors: Prof. A.R. Thurik, Prof. P.J.F. Groenen, & Prof. A. Hofman, EPS-2013-287-S&E, http://repub.eur.nl/pub/40081

Lu, Y., *Data-Driven Decision Making in Auction Markets*, Promotors: Prof. H.W.G.M. van Heck & Prof. W. Ketter, EPS-2014-314-LIS, http://repub.eur.nl/pub/51543

Ma, Y., *The Use of Advanced Transportation Monitoring Data for Official Statistics*, Promotors: Prof. L.G. Kroon & Dr J. van Dalen, EPS-2016-391-LIS, http://repub.eur.nl/pub/80174

Manders, B., *Implementation and Impact of ISO 9001*, Promotor: Prof. K. Blind, EPS-2014-337-LIS, http://repub.eur.nl/pub/77412

Mell, J.N., *Connecting Minds: On The Role of Metaknowledge in Knowledge Coordination*, Promotor: Prof. D.L. van Knippenberg, EPS-2015-359-ORG, http://hdl.handle.net/1765/78951

Meulen, D. van der, *The Distance Dilemma: the effect of flexible working practices on performance in the digital workplace*, Promotors: Prof. H.W.G.M. van Heck & Prof. P.J. van Baalen, EPS-2016-403-LIS, http://repub.eur.nl/pub/94033

Micheli, M.R., *Business Model Innovation: A Journey across Managers' Attention and Inter-Organizational Networks*, Promotor: Prof. J.J.P. Jansen, EPS-2015-344-S&E, http://repub.eur.nl/pub/78241

Milea, V., *News Analytics for Financial Decision Support*, Promotor: Prof. U. Kaymak, EPS-2013-275-LIS, http://repub.eur.nl/pub/38673

Moniz, A., *Textual Analysis of Intangible Information*, Promotors: Prof. C.B.M. van Riel, Prof. F.M.G de Jong & Dr G.A.J.M. Berens, EPS-2016-393-ORG, http://repub.eur.nl/pub/93001

Mulder, J., *Network design and robust scheduling in liner shipping*, Promotors: Prof. R. Dekker & Dr W.L. van Jaarsveld, EPS-2016-384-LIS, http://repub.eur.nl/pub/80258

Naumovska, I., *Socially Situated Financial Markets: A Neo-Behavioral Perspective on Firms, Investors and Practices*, Promotors: Prof. P.P.M.A.R. Heugens & Prof. A. de Jong, EPS-2014-319-S&E, http://repub.eur.nl/pub/76084

Neerijnen, P., *The Adaptive Organization: the socio-cognitive antecedents of ambidexterity and individual exploration*, Promotors: Prof. J.J.P. Jansen, P.P.M.A.R. Heugens & Dr T.J.M. Mom, EPS-2016-358-S&E, http://repub.eur.nl/pub/93274

Oord, J.A. van, *Essays on Momentum Strategies in Finance*, Promotor: Prof. H.K. van Dijk, EPS-2016-380-F&A, http://repub.eur.nl/pub/80036

Pennings, C.L.P., *Advancements in Demand Forecasting: Methods and Behavior*, Promotors: Prof. L.G. Kroon, Prof. H.W.G.M. van Heck & Dr J. van Dalen, EPS-2016-400-LIS, http://repub.eur.nl/pub/94039

Peters, M., *Machine Learning Algorithms for Smart Electricity Markets*, Promotor: Prof. W. Ketter, EPS-2014-332-LIS, http://repub.eur.nl/pub/77413

Porck, J., *No Team is an Island: An Integrative View of Strategic Consensus between Groups*, Promotors: Prof. P.J.F. Groenen & Prof. D.L. van Knippenberg, EPS-2013-299-ORG, http://repub.eur.nl/pub/50141

Pronker, E.S., *Innovation Paradox in Vaccine Target Selection*, Promotors: Prof. H.J.H.M. Claassen & Prof. H.R. Commandeur, EPS-2013-282-S&E, http://repub.eur.nl/pub/39654

Protzner, S., *Mind the gap between demand and supply: A behavioral perspective on demand forecasting*, Promotors: Prof. S.L. van de Velde & Dr L. Rook, EPS-2015-364-LIS, http://repub.eur.nl/pub/79355

Pruijssers, J.K., *An Organizational Perspective on Auditor Conduct*, Promotors: Prof. J. van Oosterhout & Prof. P.P.M.A.R. Heugens, EPS-2015-342-S&E, http://repub.eur.nl/pub/78192

Retel Helmrich, M.J., *Green Lot-Sizing*, Promotor: Prof. A.P.M. Wagelmans, EPS-2013-291-LIS, http://repub.eur.nl/pub/41330

Rietdijk, W.J.R., *The Use of Cognitive Factors for Explaining Entrepreneurship*, Promotors: Prof. A.R. Thurik & Prof. I.H.A. Franken, EPS-2015-356-S&E, http://repub.eur.nl/pub/79817

Rietveld, N., *Essays on the Intersection of Economics and Biology*, Promotors: Prof. A.R. Thurik, Prof. Ph.D. Koellinger, Prof. P.J.F. Groenen, & Prof. A. Hofman, EPS-2014-320-S&E, http://repub.eur.nl/pub/76907

Rösch, D., *Market Efficiency and Liquidity*, Promotor: Prof. M.A. van Dijk, EPS-2015-353-F&A, http://repub.eur.nl/pub/79121

Roza, L., *Employee Engagement in Corporate Social Responsibility: A collection of essays*, Promotor: L.C.P.M. Meijs, EPS-2016-396-ORG, http://repub.eur.nl/pub/93254

Rubbaniy, G., *Investment Behaviour of Institutional Investors*, Promotor: Prof. W.F.C. Verschoor, EPS-2013-284-F&A, http://repub.eur.nl/pub/40068

Schoonees, P., *Methods for Modelling Response Styles*, Promotor: Prof.dr P.J.F. Groenen, EPS-2015-348-MKT, http://repub.eur.nl/pub/79327

Schouten, M.E., *The Ups and Downs of Hierarchy: the causes and consequences of hierarchy struggles and positional loss*, Promotors; Prof. D.L. van Knippenberg & Dr L.L. Greer, EPS-2016-386-ORG, http://repub.eur.nl/pub/80059

Shahzad, K., *Credit Rating Agencies, Financial Regulations and the Capital Markets*, Promotor: Prof. G.M.H. Mertens, EPS-2013-283-F&A, http://repub.eur.nl/pub/39655

Smit, J., *Unlocking Business Model Innovation: A look through the keyhole at the inner workings of Business Model Innovation*, Promotor: H.G. Barkema, EPS-2016-399-S&E, http://repub.eur.nl/pub/93211

Sousa, M.J.C. de, *Servant Leadership to the Test: New Perspectives and Insights*, Promotors: Prof. D.L. van Knippenberg & Dr D. van Dierendonck, EPS-2014-313-ORG, http://repub.eur.nl/pub/51537

Spliet, R., *Vehicle Routing with Uncertain Demand*, Promotor: Prof. R. Dekker, EPS-2013-293-LIS, http://repub.eur.nl/pub/41513

Staadt, J.L., *Leading Public Housing Organisation in a Problematic Situation: A Critical Soft Systems Methodology Approach*, Promotor: Prof. S.J. Magala, EPS-2014-308-ORG, http://repub.eur.nl/pub/50712

Stallen, M., *Social Context Effects on Decision-Making: A Neurobiological Approach*, Promotor: Prof. A. Smidts, EPS-2013-285-MKT, http://repub.eur.nl/pub/39931

Szatmari, B., *We are (all) the champions: The effect of status in the implementation of innovations*, Promotors: Prof J.C.M & Dr D. Deichmann, EPS-2016-401-LIS, http://repub.eur.nl/pub/94633

Tarakci, M., *Behavioral Strategy: Strategic Consensus, Power and Networks*, Promotors: Prof. D.L. van Knippenberg & Prof. P.J.F. Groenen, EPS-2013-280-ORG, http://repub.eur.nl/pub/39130

Tuijl, E. van, *Upgrading across Organisational and Geographical Configurations*, Promotor: Prof. L. van den Berg, EPS-2015-349-S&E, http://repub.eur.nl/pub/78224

Tuncdogan, A., *Decision Making and Behavioral Strategy: The Role of Regulatory Focus in Corporate Innovation Processes*, Promotors: Prof. F.A.J. van den Bosch, Prof. H.W. Volberda, & Prof. T.J.M. Mom, EPS-2014-334-S&E, http://repub.eur.nl/pub/76978

Uijl, S. den, *The Emergence of De-facto Standards*, Promotor: Prof. K. Blind, EPS-2014-328-LIS, http://repub.eur.nl/pub/77382

Vagias, D., *Liquidity, Investors and International Capital Markets*, Promotor: Prof. M.A. van Dijk, EPS-2013-294-F&A, http://repub.eur.nl/pub/41511

Valogianni, K., *Sustainable Electric Vehicle Management using Coordinated Machine Learning*, Promotors: Prof. H.W.G.M. van Heck & Prof. W. Ketter, EPS-2016-387-LIS, http://repub.eur.nl/pub/93018

Veelenturf, L.P., *Disruption Management in Passenger Railways: Models for Timetable, Rolling Stock and Crew Rescheduling*, Promotor: Prof. L.G. Kroon, EPS-2014-327-LIS, http://repub.eur.nl/pub/77155

Venus, M., *Demystifying Visionary Leadership: In search of the essence of effective vision communication*, Promotor: Prof. D.L. van Knippenberg, EPS-2013-289-ORG, http://repub.eur.nl/pub/40079

Vermeer, W., *Propagation in Networks:The impact of information processing at the actor level on system-wide propagation dynamics*, Promotor: Prof. P.H.M.Vervest, EPS-2015-373-LIS, http://repub.eur.nl/pub/79325

Versluis, I., *Prevention of the Portion Size Effect*, Promotors: Prof. Ph.H.B.F. Franses & Dr E.K. Papies, EPS-2016-382-MKT, http://repub.eur.nl/pub/79880

Vishwanathan, P., *Governing for Stakeholders: How Organizations May Create or Destroy Value for their Stakeholders*, Promotors: Prof. J. van Oosterhout & Prof. L.C.P.M. Meijs, EPS-2016-377-ORG, http://repub.eur.nl/pub/93016

Visser, V.A., *Leader Affect and Leadership Effectiveness: How leader affective displays influence follower outcomes*, Promotor: Prof. D.L. van Knippenberg, EPS-2013-286-ORG, http://repub.eur.nl/pub/40076

Vries, J. de, *Behavioral Operations in Logistics*, Promotors: Prof. M.B.M de Koster & Prof. D.A. Stam, EPS-2015-374-LIS, http://repub.eur.nl/pub/79705

Wagenaar, J.C., *Practice Oriented Algorithmic Disruption Management in Passenger Railways*, Prof. L.G. Kroon & Prof. A.P.M. Wagelmans, EPS-2016-390-LIS, http://repub.eur.nl/pub/93177

Wang, P., *Innovations, status, and networks*, Promotors: Prof. J.J.P. Jansen & Dr V.J.A. van de Vrande, EPS-2016-381-S&E, http://repub.eur.nl/pub/93176

Wang, T., *Essays in Banking and Corporate Finance*, Promotors: Prof. L. Norden & Prof. P.G.J. Roosenboom, EPS-2015-352-F&A, http://repub.eur.nl/pub/78301

Wang, Y., *Corporate Reputation Management: Reaching Out to Financial Stakeholders*, Promotor: Prof. C.B.M. van Riel, EPS-2013-271-ORG,

http://repub.eur.nl/pub/38675

Weenen, T.C., *On the Origin and Development of the Medical Nutrition Industry*, Promotors: Prof. H.R. Commandeur & Prof. H.J.H.M. Claassen, EPS-2014-309-S&E, http://repub.eur.nl/pub/51134

Wolfswinkel, M., *Corporate Governance, Firm Risk and Shareholder Value*, Promotor: Prof. A. de Jong, EPS-2013-277-F&A, http://repub.eur.nl/pub/39127

Yang, S., *Information Aggregation Efficiency of Prediction Markets*, Promotor: Prof. H.W.G.M. van Heck, EPS-2014-323-LIS, http://repub.eur.nl/pub/77184

Ypsilantis, P., *The Design, Planning and Execution of Sustainable Intermodal Port-hinterland Transport Networks*, Promotors: Prof. R.A. Zuidwijk & Prof. L.G. Kroon, EPS-2016-395-LIS, http://repub.eur.nl/pub/94375

Yuferova, D. *Price Discovery, Liquidity Provision, and Low-Latency Trading*, Promotors: Prof. M.A. van Dijk & Dr D.G.J. Bongaerts, EPS-2016-379-F&A, http://repub.eur.nl/pub/93017

Zaerpour, N., *Efficient Management of Compact Storage Systems*, Promotor: Prof. M.B.M. de Koster, EPS-2013-276-LIS, http://repub.eur.nl/pub/38766

Zuber, F.B., *Looking at the Others: Studies on (un)ethical behavior and social relationships in organizations*, Promotor: Prof. S.P. Kaptein, EPS-2016-394-ORG, http://repub.eur.nl/pub/94388

One of the goals of statistical genetics is to elucidate the genetic architecture of phenotypes (i.e., observable individual characteristics) that are affected by many genetic variants (e.g., single-nucleotide polymorphisms; SNPs). A particular aim is to identify specific SNPs that are robustly associated with a given phenotype using a so-called genome-wide association study (GWAS).

Although GWAS sample sizes have increased in recent years, the number of SNPs still tends to vastly exceed sample sizes. Hence, multiple regression cannot be used to infer the association between SNPs and a phenotype jointly. Instead, the linear mixed model (LMM) has become a popular tool in statistical genetics. By placing a reasonable prior on SNP effects, LMMs can be used to jointly estimate SNP effects and to infer their contribution to phenotypic variance.

In this dissertation, I investigate several aspects of LMMs and related methods, such as ridge regression and LD-score regression. In addition, an LMM is used to develop an online tool, called MetaGAP, which quantifies the statistical power of a GWAS in case of heterogeneity in underlying subsamples. Using MetaGAP, I show that ongoing GWAS efforts are well-powered even for considerably heterogeneous phenotypes. This prediction is bolstered by a GWAS of reproductive choices, reported here, that finds twelve robustly associated SNPs.

I conclude that current GWAS sample sizes enable researchers to uncover parts of the genetic architecture of complex social-scientific outcomes and posit that GWAS efforts will soon attain sufficient predictive accuracy for useful applications throughout the social sciences.

ERiM

The Erasmus Research Institute of Management (ERIM) is the Research School (Onderzoekschool) in the field of management of the Erasmus University Rotterdam. The founding participants of ERIM are Rotterdam School of Management (RSM), and the Erasmus School of Economics (ESE). ERIM was founded in 1999 and is officially accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW). The research undertaken by ERIM is focused on the management of the firm in its environment, its intra- and interfirm relations, and its business processes in their interdependent connections.

The objective of ERIM is to carry out first rate research in management, and to offer an advanced doctoral programme in Research in Management. Within ERIM, over three hundred senior researchers and PhD candidates are active in the different research programmes. From a variety of academic backgrounds and expertises, the ERIM community is united in striving for excellence and working at the forefront of creating new business knowledge.

ERiM

ERIM PhD Series
Research in Management

Erasmus University Rotterdam (EUR)
Erasmus Research Institute of Management
Mandeville (T) Building
Burgemeester Oudlaan 50
3062 PA Rotterdam, The Netherlands

P.O. Box 1738
3000 DR Rotterdam, The Netherlands
T +31 10 408 1182
E info@erim.eur.nl
W www.erim.eur.nl