SIGNAL AND VARIANCE COMPONENT ESTIMATION IN HIGH-DIMENSIONAL LINEAR MODELS

BY RUIJUN MA

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Statistics and Biostatistics

Written under the direction of

Lee H. Dicker

And approved by

____

____

____

____

New Brunswick, New Jersey

May, 2018

ABSTRACT OF THE DISSERTATION

Signal and variance component estimation in high-dimensional linear models

by Ruijun Ma

Dissertation Director: Lee H. Dicker

Over the past several decades, dimensionalities of many datasets have grown exponentially as technology advances. Many approaches have been proposed to tackle high-dimensional problems, where dimensionality is much larger than sample size. This dissertation focuses on developing methodologies for signal and variance component estimations in three different areas, compressive sensing, genome-wide association studies and demand forecasting in e-commerce industry. In literatures, signal and variance component estimations are usually treated as independent tasks, and this work draws the connection between these estimation goals.

For the first problem in compressive sensing, we propose an algorithm that incorporates nonparametric empirical Bayes method with generalized approximate message passing (AMP). Generalized AMP is an effective algorithm for recovering signals from noisy linear measurements, assuming known a priori signal distributions. However, in practice, both the signal distribution and noise level are often unknown. We propose nonparametric maximum likelihood-AMP (NPML-AMP) for estimating an arbitrary signal distribution in this setting. In addition, we propose a simple noise variance estimator for use in conjunction with NPML-AMP.

For the second problem in genome-wide association studies, we focus on heritability estimation methods related to variance components estimation problems for linear mixed models (LMMs). Heritability is the proportion of phenotype variance explained by genetic variance, and standard approaches to LMM-based heritability estimation have some unresolved inconsistencies. We suggest that by adopting a slightly different statistical perspective, many of these inconsistencies can be seamlessly resolved. Moreover, with Mahalanobis kernel, we define a natural version of heritability, as a conditional variance under fixed-effects model.

The third problem is associated with predictions for online retailing demand forecasting and genetic risk prediction. In these big-data applications, regression-based linear dimension reduction technique performs well in minimizing out-of-sample error. We identify the asymptotic risk of such sharp estimate with model known to be misspecified. More importantly, we propose to estimate its asymptotic risk by variance component estimation discussed in the second problem. The risk evaluation technique can also be extended to model comparison between other methods with explicit asymptotic risk.

Acknowledgements

First and foremost, I owe my sincere gratitude to my thesis advisor Professor Lee Dicker. Without his insightful guidance and tremendous support, this thesis would not have been possible. Over the years working with him on high-dimensional methods, I am deeply influenced by his persistent desire to uncover patterns in data, devotion to work, and his unbounded thoughtfulness. I truly believe that his mentorship will benefit me in my future career and life. It is my privilege to have him as my advisor.

I would like to extend my gratitude to other members of the dissertation committee. I wish to thank Professor Steve Buyske and Professor Cun-Hui Zhang for their precious time and efforts on my manuscript. I am grateful to Professor Dean Foster for his mentorship when I interned at Amazon. I really appreciate their constructive comments on my research projects.

I also want to thank the Department of Statistics and Biostatistics of Rutgers University for keeping such a great research environment. Special thanks to the current and previous graduate directors Professor Harry Crane, Professor Tirthankar Dasgupta and Professor John Kolassa for their service. They are always patient and helpful when I go to them for advices.

Chapter 2 is an amended version of work published as Long Feng, Ruijun Ma and L. H. Dicker, "Nonparametric maximum likelihood approximate message passing," 2017 51st Annual Conference on Information Sciences and Systems (CISS), Baltimore, MD, 2017, pp. 1-6. (doi: 10.1109/CISS.2017.7926084). My thanks go to my friend and co-author Dr. Long Feng for his wisdom on this project, especially on deriving the approximate convexity and the interval of empirical nonparametric prior. AMP is my first research project and I appreciate the useful research techniques and experiences he shared with me.

In addition, a thank you to Xinyu Sun, Linglin He, Yilei Zhan and Liang Wang for our friendship and their help. My special thanks go to Yuting Chen for her company and encouragement.

Finally, I would like to express my endless appreciation to my father Enhui Ma, who had kept me interested in science at every stage of my youth. From him, I learned to be conscientious, persistent and responsible, as a way of life. I am deeply grateful for his lessons and unconditional support on my education. This manuscript is dedicated to the memory of him.

Dedication

In loving memory of my father Enhui.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Consider the linear model $y = \mathbf{x}^\top \boldsymbol{\beta} + \epsilon$ where $(y, \mathbf{x}^\top)^\top \in \mathbb{R}^{p+1}$ is the pair of observed response and covariates, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of effects, and ϵ is noise with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}(\epsilon) = \sigma_\epsilon^2$. Provided $(y_1, \mathbf{x}_1^\top)^\top, \ldots, (y_n, \mathbf{x}_n^\top)^\top$, we are generally interested in estimating $\boldsymbol{\beta}$, predicting unseen y_{n+1} provided $\mathbf{x}_{n+1}$, and estimating σ_ϵ^2 for the setting where $n << p$. While discussing methodologies in all these areas, this manuscript draws the crucial connection between signal estimation/prediction and variance component estimation.

1.1 Signal and variance component estimation

In this high-dimensional setting, the ordinary least squares (OLS) fails as the Gram matrix of the design matrix is not invertible. Then signals in big-data analysis are usually recovered under certain assumptions, such as structural smoothness and sparsity. A classical solution, ridge regression, adds a positive constant λ towards the diagonal entries of this Gram matrix (Tikhonov, 1963; Hoerl and Kennard, 1970) to gain the invertibility, where $\lambda > 0$ governs the strength of the regularization. From the Bayesian point of view, the ridge regression is equivalent to the posterior mean (and mode) of a model where $\boldsymbol{\beta}_j \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2/\lambda)$ for $j = 1, \ldots, p$. Similarly, the well-known LASSO (Tibshirani, 1996) solution is equivalent to the posterior mode provided that every entry in $\boldsymbol{\beta}$ follows a Laplace distribution. For both approaches, the optimal regularization parameter λ in practice is selected via cross-validation. A related random signal recovery method is discussed in Chapter 2.

Assuming $(y, \mathbf{x}^\top)^\top$is $p + 1$-dimensional random variable, then variance components, including noise variance and proportion of explained variance, are also parameters of interest. In genetics, understanding proportion of explained variance is crucial for animal breeding and investigating genetic merits. Moreover, in signal estimation and prediction problems, known knowledge of residual variance potentially improves estimation performance. Variance component estimation is also important for model evaluation purpose. For example, asymptotic risk of many high-dimensional methods are functions of noise variance and signal-to-noise ratio. Other popular model assessment criterions such as AIC and BIC (Akaike, 1974; Schwarz et al., 1978) also rely on plug-in estimates of residual variance. Chapter 3 mainly discusses estimating proportion of explained variation in the context of genetic data. The estimation method is applied in Chapter 4 for estimating quadratic out-of-sample error of several high-dimensional methods.

1.2 Dissertation in a nutshell

Chapter 2 introduces the least squares signal estimator for a model where $\boldsymbol{\beta}_j \sim F$ by incorporating nonparametric empirical Bayes method with approximate message passing (Feng et al., 2017). Generalized approximate message passing (GAMP) is an effective algorithm for recovering signals from noisy linear measurements, assuming known a priori signal distributions. However, in practice, both the signal distribution and noise level are often unknown. The EM-GM-AMP algorithm integrates GAMP with the EM algorithm to simultaneously estimate the signal distribution and noise variance while recovering the signal. EM-GM-AMP is built on the assumption that the signal is drawn from a sparse Gaussian mixture. In this paper, we propose nonparametric maximum likelihood-AMP (NPML-AMP) for estimating an arbitrary signal distribution in this setting. In addition to providing more flexibility (and performance improvements), we argue that the nonparametric approach actually simplifies implementation and improves stability by leveraging approximate convexity, which is not available in the sparse Gaussian mixture formulation of EM-GM-AMP. We also propose a simplified plug-in noise variance estimator for use in conjunction with NPML-AMP (or EM-GM-AMP). A comprehensive numerical study validates the performance of NPML-AMP algorithm in reaching nearly minimum mean squared error (MMSE) under various signal distributions, noise levels, and undersampling ratios.

Previous examples of regularized least squares commonly assume that effect-sizes are iid with mean 0. However, the random-effects assumption in many cases is questionable, and a misspecified assumption could potentially cause biased variance component estimation. Chapter 3 addresses this issue and discusses heritability estimation in genome-wide association studies (GWAS) data for linear models. Heritability is the proportion of phenotype variance explained by genetic variance (Falconer, 1960). Under linear models, it is defined as

$$h^2 := \frac{\mathrm{Var}(\mathbf{x}^\top \boldsymbol{\beta})}{\mathrm{Var}(y)}.$$

Existing linear model-based heritability methods generally require specification of a genetic relationship matrix (GRM), which measures genetic similarity between individuals. In literature, the GRM is frequently a sample correlation matrix constructed from the design matrix, which corresponds to a Euclidean distance kernel, or a random-effects assumption. However, standard approaches to heritability estimation have some unresolved inconsistencies caused by non-randomness in effect-sizes. In Chapter 3, we argue that the fixed-effects and random-effects heritabilities are equivalent if one adopts a Mahalanobis distance-based GRM. Moreover, with the Mahalanobis GRM, it's straightforward to define a natural version of heritability, interpreting the heritability coefficient as a conditional variance under the corresponding fixed-effects model. This builds a link between narrow-sense (or additive) heritability and broad-sense heritability, which is a more model-free measure of overall heritability defined in terms of the conditional variance of a phenotype given the genotype and other specified information.

In high-dimensional prediction problems, portion of explained variation of any out-of-sample prediction method is bounded by h^2. In many applications, reaching this upper bound (or even a fraction of it) is notoriously challenging. A different approach to this problem is via linear dimension reduction, where we regress y on finite dimensional projection of $\mathbf{x}$. Portion of explained variation for linear dimension reduction estimation is also bounded by the heritability of linearly transformed inputs. Although the upper bound of dimension reduction method is smaller than h^2, with advantages in dimensionality, estimation with transformed inputs easily achieves its theoretical limit in terms of out-of-sample error. In Chapter 4, we derive asymptotic quadratic risk for regression-based linear dimension reduction methods. Similar to many high-dimensional prediction methods, the explicit asymptotic risk of projected least squares is in terms of variance components. Hence, we propose to evaluate out-of-sample error of these methods by variance component estimation proposed in Chapter 3, and further compare risks of various models by heritability estimation and Wald test.

Chapter 2

Nonparametric Maximum Likelihood Approximate Message Passing[1]

2.1 Introduction

We consider reconstructing an N-dimensional signal $\boldsymbol{x} = (x_1, \ldots, x_N)^\top \in \mathbb{R}^N$ from $M < N$ linear measurements with noise

$$\boldsymbol{y} = \mathbf{A}\boldsymbol{x} + \boldsymbol{\varepsilon},$$

where $\mathbf{A} = (A_{ij}) \in \mathbb{R}^{M \times N}$ is a known transform matrix. The noise vector $\boldsymbol{\varepsilon} \in \mathbb{R}^N$ is assumed to be iid Gaussian $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \psi I_M)$, with ψ being the common noise variance.

We consider this high-dimensional problem in the setting where the values $x_1, \ldots, x_N$ are independently generated from a common probability distribution F. To estimate the signal $\boldsymbol{x}$, two natural estimators are the *maximum a posteriori* (MAP) and *minimum mean squared error* (MMSE) estimators. The MAP and MMSE estimators correspond to the posterior mode and mean of $\boldsymbol{x}$, respectively. Important special cases of MAP and MMSE estimators include LASSO (Tibshirani, 1996) and ridge regression (Tikhonov, 1963; Hoerl and Kennard, 1970). LASSO is the MAP estimator when the signal distribution F is Laplace; ridge regression is both the MAP and MMSE estimator when F is Gaussian (the MAP and MMSE estimators coincide for Gaussian F). Both LASSO and ridge regression can be formulated as convex optimization problems, and computation of these estimators is relatively tractable in many large-scale problems. On the other hand, for many other signal distributions F, the corresponding MAP and MMSE calculation problems are computationally prohibitive.

[1] ©2017 IEEE. reprinted, with permission, from Long Feng, Ruijun Ma and L. H. Dicker, "Nonparametric maximum likelihood approximate message passing," 2017 51st Annual Conference on Information Sciences and Systems (CISS), Baltimore, MD, March 2017, pp. 1-6.

To bypass these computational issues, numerous algorithms have been proposed. This chapter focuses on *approximate message passing* (AMP) algorithms for efficiently computing MMSE estimators for arbitrary unknown F. AMP algorithms have received significant attention over the past several years (e.g. Donoho et al., 2009, 2010a,b; Rangan, 2011; Montanari, 2012; Vila and Schniter, 2013; Donoho et al., 2011; Bayati et al., 2012) and form a class of loopy belief propagation algorithms for recovering $\boldsymbol{x}$. In 2009, Donoho et al. (2009) proposed an AMP algorithm to speed-up LASSO computations and proved statistical consistency under Gaussian $\mathbf{A}$ (Donoho et al., 2011). Building on this work and that of others (e.g. Guo and Wang, 2006, 2007; Rangan, 2010), Rangan (2011) proposed a generalized AMP (GAMP) algorithm to find the MMSE estimator for an arbitrary *known* signal distribution F. GAMP is a flexible, effective algorithm; however, the condition that F is known can be restrictive. To relax this assumption, Vila and Schniter (2013) proposed an EM-GM-AMP algorithm, which integrated GAMP with the EM algorithm to estimate the distribution F and the noise variance ψ, along with $\boldsymbol{x}$. Their work estimates F among the class of sparse Gaussian mixtures (i.e. F is a finite mixture of Gaussians and a point mass at 0).

This chapter builds on work of Vila and Schniter (2013). We propose nonparametric maximum likelihood-AMP (NPML-AMP) algorithm that can estimate arbitrary signal distributions F and simultaneously estimate the signal $\boldsymbol{x}$ via an approximate NPML-MMSE estimator. Our algorithm NPML-AMP allows for accurate signal recovery, even when F cannot be well approximated by a sparse Gaussian mixture. Addtionally, we argue that NPML-AMP may have computational advantages because the associated NPML optimization problem is *approximately convex*, in the scaling limit where $M/N \to c \in (0, \infty)$. On the other hand, a similar convexity argument does not seem possible for EM-GM-AMP, because of inherent non-convexity in the sparse Gaussian mixture likelihood. Finally, we propose a simplified method for estimating the noise variance ψ, which is used throughout as a part of NPML-AMP, but could also be used to improve the performance of EM-GM-AMP in settings with low signal to noise ratio.

The rest of the chapter is organized as follows. In Section 2.2, we describe the NPML-MMSE estimator. In Section 2.3, we discuss approximate convexity. Noise variance estimation is discussed in Section 2.4. The NPML-AMP algorithm is described in Section 2.5. Section 2.6 contains numerical results, and a concluding discussion may be found in Section 2.7.

Notation: Throughout the chapter, we let $\mathbf{A}_{i\cdot}$ and $\mathbf{A}_{\cdot j}$ be the i^{th} row and j^{th} column of $\mathbf{A}$, respectively. Denote the ℓ_2 norm of $\boldsymbol{x}$ by $\|\boldsymbol{x}\|_2 = (x_1^2 + \cdots + x_N^2)^{1/2}$ and the ℓ_0 norm of $\boldsymbol{x}$ by $\|\boldsymbol{x}\|_0 = |\{j;\ x_j \neq 0\}|$. The function ϕ is the probability density of the $\mathcal{N}(0,1)$ distribution. In general, for a probability distribution F, let f denote its corresponding density function. Additionally, if $\Omega \subseteq \mathbb{R}$, then $\mathcal{F}_\Omega$ denotes the collection of all probability distributions on Ω. Finally, for sequences of real-valued random variables $\{R_M\}$, $\{A_M\}$, we write $R_M = \mathcal{O}_p(A_M)$ (and say that "R_M is bounded in probability by A_M") if for every $\epsilon > 0$ there is a constant $C > 0$ such that $\limsup_M \mathbb{P}(|R_M/A_M| > C) < \epsilon$.

2.2 The NPML-MMSE Estimator for $\boldsymbol{x}$

In this section, assume that the noise variance ψ is known. Under the assumption that $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \psi I_M)$ is Gaussian, the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$ takes the form

$$p(\boldsymbol{y}|\boldsymbol{x}) \propto \psi^{-N/2} \exp\left(-\frac{\|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|_2^2}{2\psi}\right).$$

Recall the assumption on the signal

$$x_1, \ldots, x_N \overset{\text{iid}}{\sim} F.$$

Then the joint distribution of $(\boldsymbol{y}, \boldsymbol{x})$, the marginal distribution of $\boldsymbol{y}$, and the conditional distribution of $\boldsymbol{x} \mid \boldsymbol{y}$ are, respectively,

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{x}; F) &\propto \psi^{-N/2} \exp\left(-\frac{\|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|_2^2}{2\psi}\right) \prod_{j=1}^N f(x_j), \\ p(\boldsymbol{y}; F) &\propto \int_{\boldsymbol{x} \in \Omega^N} \psi^{-N/2} \exp\left(-\frac{\|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|_2^2}{2\psi}\right) \prod_{j=1}^N dF(x_j), \\ p(\boldsymbol{x} \mid \boldsymbol{y}; F) &= \frac{p(\boldsymbol{x}, \boldsymbol{y}; F)}{p(\boldsymbol{y}; F)}. \end{aligned}$$

Let $\Omega \subseteq \mathbb{R}$ and assume that $F \in \mathcal{F}_\Omega$. The MMSE estimator for $\boldsymbol{x}$ is given by

$$\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}(F) = \int_{\Omega^N} \boldsymbol{x}\ p(\boldsymbol{x} \mid \boldsymbol{y}; F)\ d\boldsymbol{x}. \tag{2.1}$$

If F is unknown, then $\hat{\boldsymbol{x}}$ cannot be implemented. However, a reasonable strategy is to replace F in (2.1) with an estimate $\hat{F}$. Still, there are two challenges in implementing $\hat{\boldsymbol{x}}(\hat{F})$: finding a good estimate $\hat{F}$ and evaluating the multiple integral in (2.1).

We propose to estimate F by maximum likelihood. That is, let $\hat{F} = \hat{F}_\Omega$, where

$$\hat{F}_\Omega = \underset{F \in \mathcal{F}_\Omega}{\operatorname{argmin}} \; -\frac{1}{N} \log p(\boldsymbol{y}; F). \tag{2.2}$$

The estimator (2.2) is the basis of this chapter. At this point, it still may not be evident that this approach is tractable: whenever Ω is infinite, the optimization problem (2.2) is infinite dimensional; moreover, evaluating the marginal likelihood $p(\boldsymbol{y}; F)$ is challenging because of the multiple integral. However, progress is enabled by first noting that the existing algorithms EM-GM-AMP and GAMP are essentially designed to handle the multiple integrals in (2.1)–(2.2). Second, the (infinite dimensional) nonparametric maximum likelihood problem (2.2) has been studied for independent data going back to the 1950s (Kiefer and Wolfowitz, 1956; Robbins, 1950) and is known to be well-behaved in many settings. For instance, if we were in an independent data setting where $N = M$ and the likelihood factored as $p(\boldsymbol{y} \mid \boldsymbol{x}) = \prod_i p(y_i \mid x_i)$ (e.g. if $\mathbf{A}$ was the identity matrix), then

$$\hat{F}_\Omega = \underset{F \in \mathcal{F}_\Omega}{\operatorname{argmin}} \; -\frac{1}{N} \sum_{i=1}^{N} \log \int p(y_i \mid x_i) \; dF(x_i) \tag{2.3}$$

and the objective function is convex (as a function of F — indeed, the integral is linear in F and $-\log$ is a convex function). The convex problem (2.3) is known to have a solution $\hat{F}_\Omega$ supported on at most N points (regardless of the size of Ω), with $\hat{F}_\Omega \to F$ as $N \to \infty$ under relatively weak conditions (Kiefer and Wolfowitz, 1956; Lindsay, 1995). While more detailed theoretical results for (2.3) are challenging and much remains unknown, recent work has focused on computationally feasible approximations to (2.3) that leverage convexity (Koenker and Mizera, 2014; Dicker and Zhao, 2016) — this is the jumping-off point for this chapter.

Following recent methods for the independent data NPML problem (Jiang and Zhang, 2010; Koenker and Mizera, 2014; Dicker and Zhao, 2016), our strategy for approximating (2.2) is to replace Ω with Λ, where $\Lambda \subseteq \Omega$ is a pre-specified finite subset of Ω. This reduces (2.2) to a finite dimensional optimization problem. Furthermore, in the analagous convex problem (2.3), replacing Ω with Λ preserves convexity and convexity of (2.3) is the key to the approximate convexity results in Section 2.3.

2.2.1 Relationship to EM-GM-AMP

In practice, our approach to reconstructing $\boldsymbol{x}$ is very similar to EM-GM-AMP. Let $\mathcal{F}_{\mathrm{GM}}$ denote the class of all sparse Gaussian mixtures on $\mathbb{R}$ with at most $L \in \mathbb{N}$ Gaussian components. EM-GM-AMP computes the estimator $\hat{\boldsymbol{x}}(\hat{F}_{\mathrm{GM}})$, where $\hat{F}_{\mathrm{GM}}$ solves (2.2), except that the minimization is over $\mathcal{F}_{\mathrm{GM}}$ instead of $\mathcal{F}_{\Omega}$. For EM-GM-AMP the user must specify L; in the literature it is typically taken to be 3 or 4. Perhaps the most significant difference between $\hat{F}_{\mathrm{GM}}$ and the NPML estimator $\hat{F}_{\Omega}$ is that the set $\mathcal{F}_{\mathrm{GM}}$ is not convex, unlike $\mathcal{F}_{\Omega}$ or $\mathcal{F}_{\Lambda}$. Consequently, the approximate convexity results for (2.2) do not appear to hold for EM-GM-AMP. We believe this leads to increased instability in $\hat{F}_{\mathrm{GM}}$ for large L, which in turn may limit the class of signal distributions that can be accurately reconstructed by EM-GM-AMP (specifically, EM-GM-AMP may be most effective for distributions that can be well-approximated by sparse Gaussian mixtures with only a few components). On the other hand, approximate convexity appears to enhance the stability of $\hat{F}_{\Lambda}$ as the size of $\Lambda \subseteq \Omega$ increases.

2.2.2 Choosing Λ

By increasing the size of $\Lambda \subseteq \Omega$, it is reasonable to expect that $\hat{F}_{\Lambda}$ becomes a more accurate approximation to $\hat{F}_{\Omega}$. We take the point of view that $\hat{F}_{\Lambda}$ inherits its properties from $\hat{F}_{\Omega}$ and, in practice, computational limitations appear to be the main issue in choosing the size of Λ (in all of the numerical experiments we take $|\Lambda| = 100$).

To gain some additional intuition on choosing Λ, define

$$\hat{x} = \frac{\sum_{j=1}^{N} \mathbf{A}_{\cdot j}^{\top} \boldsymbol{y}}{\sum_{1 \leq j_1, j_2 \leq N} \mathbf{A}_{\cdot j_1}^{\top} \mathbf{A}_{\cdot j_2}};$$

$$\tilde{x}^2 = \frac{\|\boldsymbol{y}\|_2^2 - \hat{x}^2 \sum_{j_1 \neq j_2} \mathbf{A}_{\cdot j_1}^{\top} \mathbf{A}_{\cdot j_2} - M\psi}{\sum_{j=1}^{N} \mathbf{A}_{\cdot j}^{\top} \mathbf{A}_{\cdot j}}.$$

Then we have the moment identities

$$\mathbb{E}(x_j) \quad = \quad \mathbb{E}(\hat{x} \mid \mathbf{A}),$$

$$\begin{aligned}\mathbb{E}(x_j^2) &= \mathbb{E}(\tilde{x}^2 \mid \mathbf{A}),\\ \mathrm{Var}(x_j^2) &= \mathbb{E}(\tilde{x}^2 \mid \mathbf{A}) - \mathbb{E}(\hat{x} \mid \mathbf{A})^2.\end{aligned}$$

The moment identities are verified in the appendix of this chapter. This suggests that for some $k \geq 0$, we should typically have

$$\hat{x} - k\sqrt{\tilde{x}^2 - \hat{x}^2} \leq x_j \leq \hat{x} + k\sqrt{\tilde{x}^2 - \hat{x}^2}.$$

Hence, it is reasonable to take

$$\Lambda \subseteq \Omega \cap \left[\hat{x} - k\sqrt{\tilde{x}^2 - \hat{x}^2}, \hat{x} + k\sqrt{\tilde{x}^2 - \hat{x}^2}\right]. \tag{2.4}$$

In the experiments in this chapter we have taken Λ to be a regular grid with 100 points, satisfying (2.4) with $k = 5$.

2.3 Approximate convexity

In Section 2.2, we discussed (2.3) — the independent data analogue of the NPML problem (2.2) — and noted that it is a convex optimization problem. In this section, we show that the original NPML problem (2.2) is approximately convex in the scaling limit where $M/N \to c \in (0, \infty)$, which lends additional structure to the problem. Specifically, the next theorem shows that the objective function in (2.2) can be approximated by a convex function.

Theorem 1. *Let $\Omega \in \mathbb{R}$ be a bounded set, so that $\Omega \subseteq [x_{\min}, x_{\max}]$ for some $-\infty < x_{\min} < x_{\max} < \infty$. Let $x^* = \max(|x_{\min}|, |x_{\max}|)$ and for $F \in \mathcal{F}_\Omega$ define*

$$\begin{aligned}\ell(F) &= -\frac{1}{N}\log p(\boldsymbol{y}; F),\\ \ell_{\mathrm{conv}}(F) &= -\frac{1}{N}\sum_{j=1}^{N}\log\int \frac{1}{\psi^{N/2}}\exp\left\{\frac{1}{\psi}(\mathbf{A}_{\cdot j}^\top \boldsymbol{y})x_j - \right.\\ &\qquad \left.\frac{1}{2\psi}\|\mathbf{A}_{\cdot j}\|_2^2 x_j^2\right\} dF(x_j) + \frac{\|\boldsymbol{y}\|_2^2}{2N\psi}.\end{aligned} \tag{2.5}$$

Assume A_{ij} are iid random variables with $\mathbb{E}(A_{ij}) = 0$ and $\mathbb{E}(A_{ij}^4) \leq C$ for some constant $C > 0$. If $M \to \infty$ and $M/N \to c \in (0, \infty)$, then,

$$\sup_{F \in \mathcal{F}_\Omega} |\ell(F) - \ell_{\mathrm{conv}}(F)| = \mathcal{O}_p\left(\frac{\{x^*\}^2}{\psi\sqrt{M}}\right). \tag{2.6}$$

A proof of Theorem 1 is contained in the appendix at the end of this chapter. Observe that $\ell_{\text{conv}}(F)$ is convex in F. Moreover, $\ell_{\text{conv}}(F)$ is the objective function for an independent data NPML problem, where the available data are

$$z_j = \mathbf{A}_{\cdot j}^\top \boldsymbol{y}/\|\mathbf{A}_{\cdot j}\|_2^2, \quad j = 1, \ldots, N,$$

and $z_j \mid x_j \sim \mathcal{N}(x_j, \psi/\|\mathbf{A}_{\cdot j}\|_2^2)$. In other words, the NPML objective function $\ell_{\text{conv}}(F)$ is obtained by replacing the data $(\boldsymbol{y}, \mathbf{A})$ with $z_1, \ldots, z_N$ and ignoring the correlation between z_i, z_j.

While Theorem 1 suggests that the objective function in (2.2) can be well-approximated by an independent data NPML objective function, it does not imply that $\hat{F}_\Omega$ can be found by optimizing $\ell_{\text{conv}}(F)$; indeed, preliminary numerical work suggests that optimizing $\ell_{\text{conv}}(F)$ directly leads to a significant loss in estimation accuracy.

2.4 Gaussian MLE for noise variance estimation

When the noise variance ψ is unknown, the EM-GM-AMP algorithm takes an additional EM step within each interation to provide updated estimates of ψ. This approach is also feasible for NPML-AMP. However, through numerical experimentation we have found that these estimates for ψ can be unstable, and that inaccurate estimates of ψ can lead to degraded performance in terms of signal recovery.

As a simple alternative to the approach described above, where estimating ψ is interwoven with the algorithm for estimating F and computing $\hat{\boldsymbol{x}}$, we propose to estimate ψ using the well-known Gaussian variance components MLE up front and then take this as a plug-in estimate for ψ throughout the AMP algorithm without any more updates for estimating ψ.

Specifically, for $\boldsymbol{\theta} = (\psi, \tau) \in \mathbb{R}^2$ with $\psi, \tau \geq 0$, define

$$l(\boldsymbol{\theta}) = \frac{1}{2} \log\det \left(\tau \mathbf{A}\mathbf{A}^\top + \psi I_M\right) + \frac{1}{2}\boldsymbol{y}^\top \left(\tau \mathbf{A}\mathbf{A}^\top + \psi I_M\right)^{-1} \boldsymbol{y}$$

and

$$\hat{\boldsymbol{\theta}} = (\hat{\psi}, \hat{\tau}) = \underset{\psi, \tau \geq 0}{\text{argmin}} \; l(\boldsymbol{\theta}). \tag{2.7}$$

Then $\hat{\boldsymbol{\theta}}$ is the MLE for $\boldsymbol{\theta}$ under the model where $x_j \sim \mathcal{N}(0, \tau)$ and we use $\hat{\psi}$ to estimate ψ in our AMP implementations. While $\hat{\psi}$ is derived under the assumption that the signal $\boldsymbol{x}$ is Gaussian, $\hat{\psi}$ is known to perform well for non-Gaussian $\boldsymbol{x}$ as long as the entries of $\mathbf{A}$ are centered (Jiang et al., 1996; Dicker and Erdogdu, 2016a) .

In numerical experiments, we have found that the MLE $\hat{\psi}$ is often substantially more accurate than other estimators for ψ — including the original EM-GM-AMP estimator for ψ — especially in settings where $N\mathbb{E}(x_j^2)/(M\psi)$ is small (i.e. small signal to noise ratio; results from these experiments ares not reported here due to space constraints). For the NPML-AMP algorithm, this approach to estimating ψ can also be interpreted as de-coupling the non-convex part of the problem from the (approximately) convex part: (2.7) is a non-convex problem, while (2.2) is approximately convex (by Theorem 1). Moreover, (2.7) is a relatively simple non-convex problem, which can be reduced to a univariate optimization problem by standard methods (Dicker and Erdogdu, 2016a; Jiang et al., 1996).

2.5 Implementation of NPML-AMP

Similar to EM-GM-AMP, the NPML-AMP algorithm alternates between GAMP and EM steps. The GAMP steps are exactly as described in (Rangan, 2011; Vila and Schniter, 2013). The EM steps of NPML-AMP for estimating F are described in Algorithm 1.

We proceed here under the assumption that the noise variance ψ is known, with the understanding that if it is unknown, then it should be estimated as in Section 2.4. Additionally, we assume that $\Lambda = \{\theta_1, \ldots, \theta_L\} \subseteq \Omega$ has been pre-determined (for instance, as described in Section 2.2.2). In Algorithm 1, $\boldsymbol{\omega}(t) = (\omega_1(t), \ldots, \omega_L(t))^\top \in \mathbb{R}^L$ denotes the probabilities corresponding to Λ, so that at step t of the algorithm, $x_j = \theta_l$ with probability $\omega_l(t)$; $\hat{\boldsymbol{x}}(t)$ is the value of the estimate $\hat{\boldsymbol{x}}$ at step t. Finally, $\hat{\mathbf{r}}(t)$ and $\boldsymbol{\mu}^r(t)$ are parameters generated by the GAMP algorithm described in Table I of (Vila and Schniter, 2013).

Algorithm 1 returns the estimated signal $\hat{\boldsymbol{x}}(t)$ and the weights $\boldsymbol{\omega}(t)$. The probability distribution $\sum_{l=1}^{L} \omega_l(t)\delta_{\theta_l}$ should be viewed as an approximation to $\hat{F}_\Lambda$, which in turn is

Algorithm 1 NPML-AMP: EM steps for estimating F.

```
Input $\Lambda$, $T_{\max}$, $\epsilon_{\text{break}}$.
Initialize $\boldsymbol{\omega}(0) = (1/L, \ldots, 1/L)$, $\hat{\boldsymbol{x}}(0) = \sum_{l=1}^{L} \omega_l(0)\theta_l$.
for $t = 1$ to $T_{\max}$ do
    Following Table I of Vila and Schniter (2013), use GAMP with inputs $\Lambda$, $\boldsymbol{\omega}(t-1)$,
    and $\psi$ to generate $\hat{\boldsymbol{x}}(t)$, $\hat{\boldsymbol{r}}(t)$ and $\boldsymbol{\mu}^r(t)$.
    if $\|\hat{\boldsymbol{x}}(t) - \hat{\boldsymbol{x}}(t-1)\|_2^2 / \|\hat{\boldsymbol{x}}(t-1)\|_2^2 < \epsilon_{\text{break}}$ then
        Break.
    end if
    for $l = 1$ to $L$ and $j = 1$ to $N$ do
        $q_{j,l}(t) \leftarrow \frac{\omega_l(t-1)\mu_j^r(t)^{-1/2}\phi\{\theta_l - \hat{r}_j(t)\}}{\sum_{l=1}^{L}\omega_l(t-1)\mu_j^r(t)^{-1/2}\phi\{\theta_l - \hat{r}_j(t)\}}$.
    end for
    for $l = 1$ to $L$ do
        $\omega_l(t) \leftarrow \frac{\sum_{j=1}^{N} q_{j,l}(t)}{\sum_{l'=1}^{L}\sum_{j=1}^{N} q_{j,l'}(t)}$.
    end for
end for
Output $\boldsymbol{\omega}(t)$, $\hat{\boldsymbol{x}}(t)$.
```

our proposed approximation to $\hat{F}_\Omega$.

2.6 Numerical results

In this section, we compare the performance of NPML-AMP, EM-GM-AMP and GAMP in several numerical experiments. In each setting, the GAMP estimates can be viewed as the optimal or oracle solution, since it makes use of the true signal distribution and noise variance. In the experiments, the performance of an estimator $\hat{\boldsymbol{x}}$ is assessed by

$$\text{NMSE[dB]} := 10\log_{10}\left(\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2 / \|\boldsymbol{x}\|_2^2\right).$$

Throughout, we set the tolerance rates for the algorithms to be $\epsilon_{\text{break}} = 10^{-6}$; the maximum number of iterations for each GAMP call is $T_{\max} = 50$ (see Table I from Vila and Schniter (2013)); and $T_{\max} = 200$ for the EM loops in EM-GM-AMP and NPML-AMP (in Algorithm 1 above). For NPML-AMP, the number of grid points in Λ

is fixed at 100; for EM-GM-AMP, the number of Gaussian components L was fixed at 3 for the sparse Gaussian mixture experiment describe in Section 2.6.1 (Figs. 2.1–2.2) and was 4 for the remaining settings. Finally, in all of the experiments in this section $\mathbf{A}$ has iid $\mathcal{N}(0, 1/M)$ entries. Our focus on iid Gaussian measurement matrices follows much of the earlier work on AMP algorithms (e.g. Donoho et al., 2009). More recent work has focused on extending AMP to other measurement ensembles (e.g. Rangan et al., 2014, 2017); we expect that similar extensions are possible for NPML-AMP and this is a topic for future research.

2.6.1 Signal recovery at various M/N ratio

In the first set of experiments, we fixed the signal to noise ratio

$$\mathrm{SNR[dB]} := 10 \log_{10}\{N\mathbb{E}(x_j^2)/(M\psi)\}$$

at 10 dB in all settings and fixed $N = 1000$. We varied M so that the undersampling ratio was $M/N \in \{0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5\}$ and considered three different signal distributions — a sparse Gaussian mixture, an exponential signal, and a discrete uniform signal. For each undersampling ratio and signal distribution, we generated $I = 100$ independent datasets.

For each dataset, we computed several estimators for $\boldsymbol{x}$. Specifically, we computed the GAMP estimator; the EM-GM-AMP estimator with unknown variance (following the algorithm described in (Vila and Schniter, 2013) exactly); the EM-GM-AMP estimator with *known* noise variance ψ, denoted EM-GM-AMP-KV; the NPML-AMP with unknown variance (using the method in Section 2.4 to estimate ψ); and the NPML-AMP estimator with known variance, denoted NPML-AMP-KV. For reference, we also computed the LASSO estimator for one of the signal distributions (Figs. 2.1–2.2; for each of the other signal distributions, all of the AMP methods dramatically out-perform LASSO and the LASSO NMSE[dB] would be off the scale in the plots). Finally, we recorded the estimated noise variance for the methods where the noise variance was unknown.

For LASSO, we used 10-fold cross validation to choose the regularization parameter and estimated the variance parameter ψ by

$$\hat{\psi}_{\mathrm{L}} = \|\boldsymbol{y} - \mathbf{A}\hat{\boldsymbol{x}}_{\mathrm{L}}\|_2^2/(M - \hat{N}_{\mathrm{L}}),$$

where $\hat{\boldsymbol{x}}_{\mathrm{L}}$ is the LASSO estimated signal and $\hat{N}_{\mathrm{L}} = \|\hat{\boldsymbol{x}}_{\mathrm{L}}\|_0$ is the estimated sparsity.

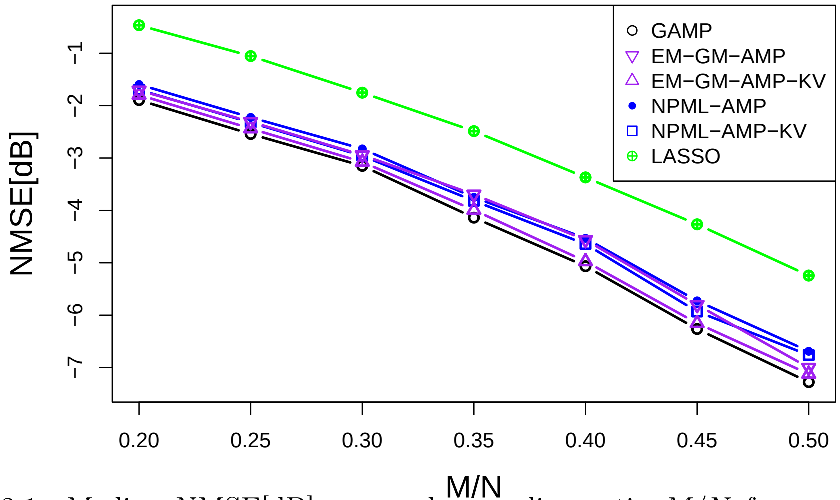

Figure 2.1: Median NMSE[dB] vs. undersampling ratio M/N for sparse Gaussian mixture signal.

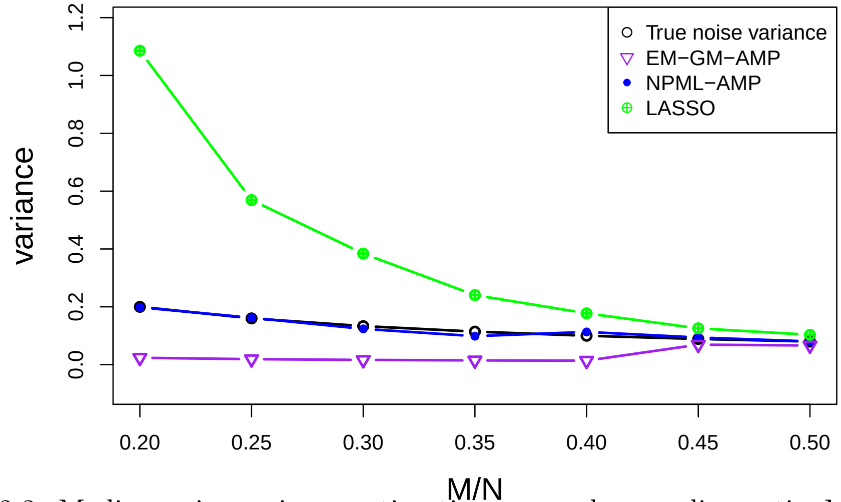

Figure 2.2: Median noise variance estimation vs. undersampling ratio M/N ratio for sparse Gaussian mixture signal.

Figures 2.1–2.2 depict the median NMSE[dB] and median values of the noise variance estimates $\hat{\psi}$, respectively, computed over 100 independent datasets, when the signal distribution was a sparse Gaussian mixture satisfying $\mathbb{P}(x_j = 0) = 0.8$ and $x_j \sim \mathcal{N}(1, 1)$ with probability 0.2. From Fig. 2.1, it appears that EM-GM-AMP slightly out-performs NPML-AMP, and both out-perform LASSO. This setting is favorable to EM-GM-AMP, since EM-GM-AMP is designed to estimate sparse Gaussian mixtures. In terms of NMSE[dB], there is little difference between the AMP methods when the noise variance is known or unknown. On the other hand, in Fig. 2.2, it appears that NPML-AMP gives the most accurate estimates of ψ (for NPML-AMP, recall that $\hat{\psi}$ is just the Gaussian variance components MLE for ψ).

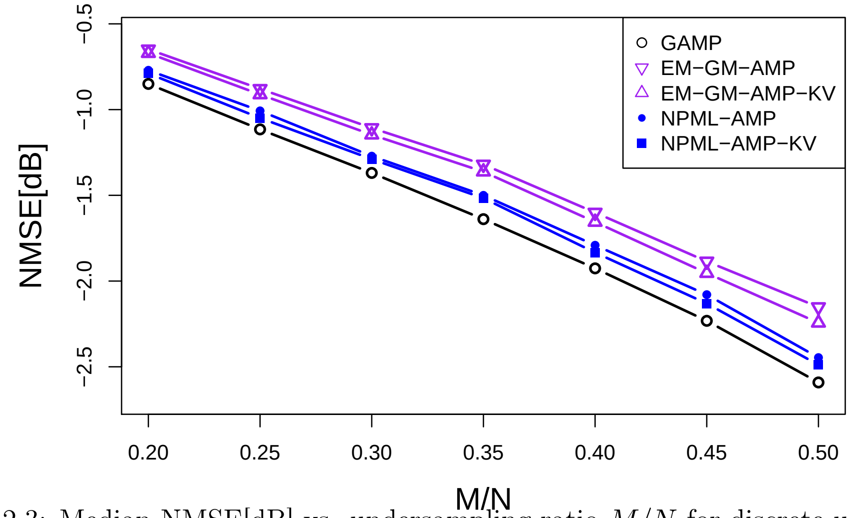

Figure 2.3: Median NMSE[dB] vs. undersampling ratio M/N for discrete uniform signal.

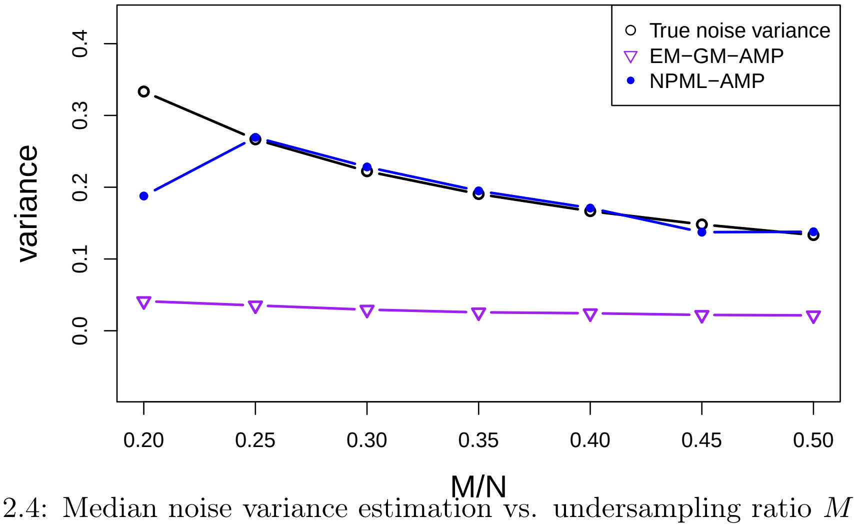

Figure 2.4: Median noise variance estimation vs. undersampling ratio M/N ratio for discrete uniform signal.

In Figs. 2.3–2.4, $x_j \in \{-1, 0, 1\}$ follows a discrete uniform distribution, and in Figs. 2.5–2.6, $x_j \sim$ exponential(1). In these settings, NPML-AMP outperforms EM-GM-AMP in terms of NMSE[dB]. As in Figs. 2.1–2.2, the NMSE[dB] results are similar for the known and unknown variance methods, while the NPML-AMP variance estimation

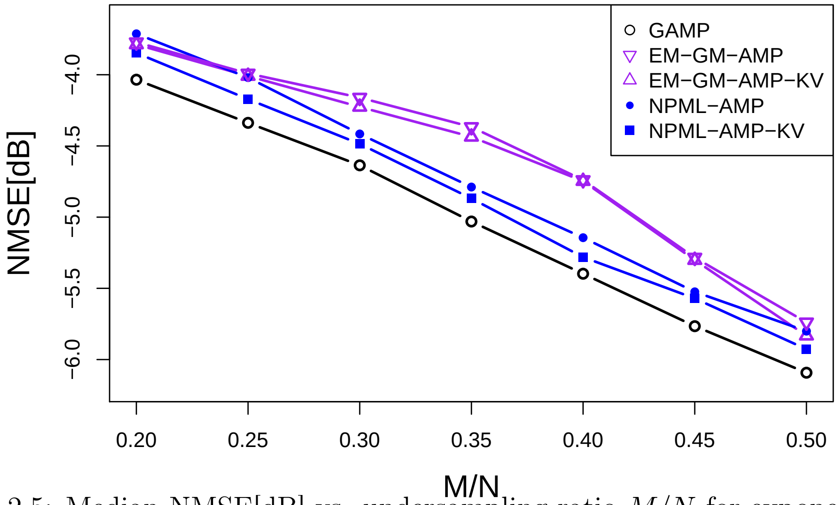

Figure 2.5: Median NMSE[dB] vs. undersampling ratio M/N for exponential signal.

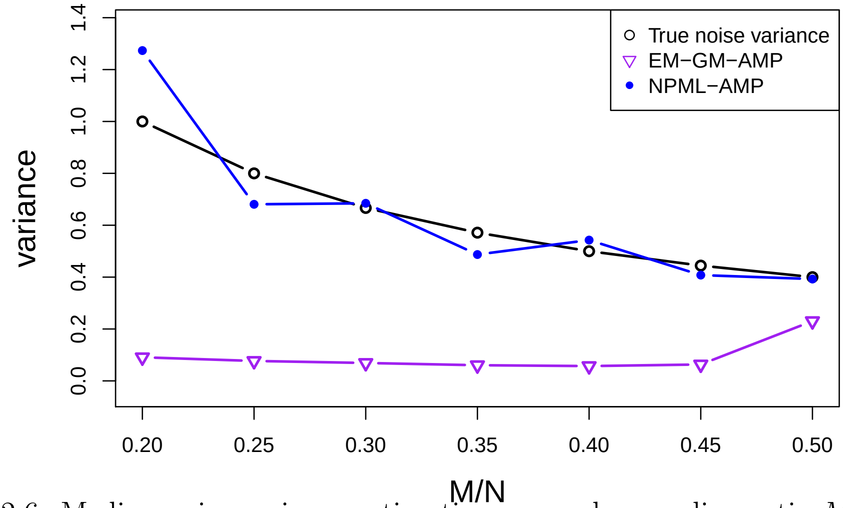

Figure 2.6: Median noise variance estimation vs. undersampling ratio M/N ratio for exponential signal.

appears to be significantly more accurate than EM-GM-AMP.

2.6.2 Signal Recovery at various SNR

We also conducted experiments where we varied the signal to noise ratio, while fixing the undersampling ratio $M/N = 0.4$ (with $N = 1000$ and $M = 400$ throughout). Here, we report results for the discrete uniform signal distribution with $x_j \in \{-1, 0, 1\}$ and SNR[dB] across SNR[dB] $\in \{0, 5, 10, 15, 20, 25\}$. As in the previous section, we generated $I = 100$ independent datasets for each setting and computed several estimators for each dataset. Summary statistics for NMSE[dB] and variance estimation are reported in Figs.

2.7–2.8.

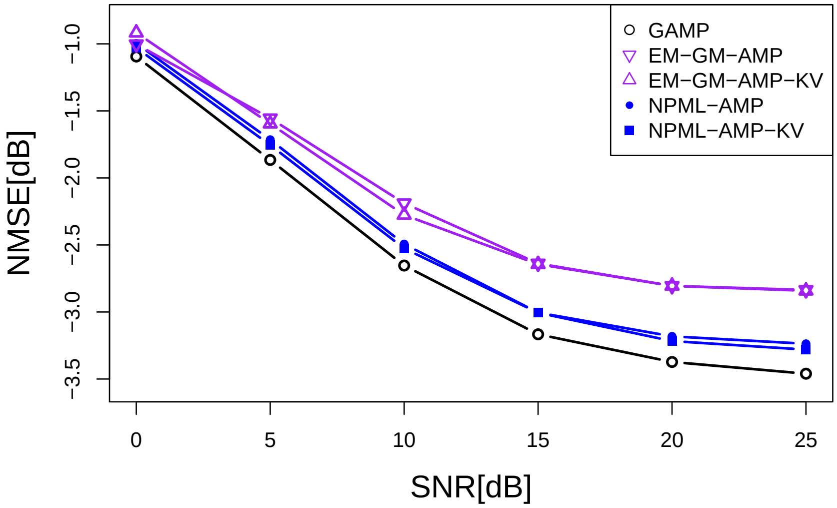

Figure 2.7: Median NMSE[dB] vs. SNR[dB] for discrete uniform signal.

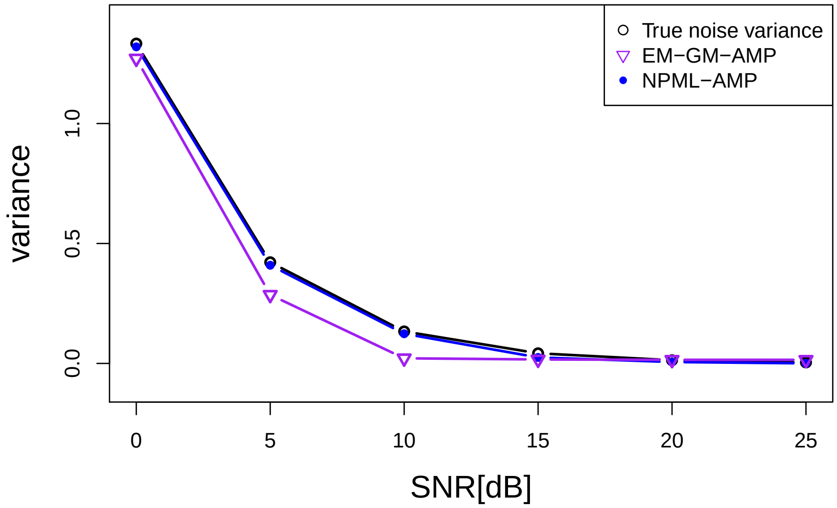

Figure 2.8: Median noise variance estimation vs. SNR for discrete uniform signal

In Fig. 2.7, NPML-AMP has smaller median NMSE[dB] than EM-GM-AMP across the entire range of signal to noise ratios. Fig. 2.8 indicates that for noise variance estimation, NPML-AMP may outperform EM-GM-AMP in this setting.

2.7 Conclusion

We have proposed a signal recovery algorithm for generic signal distributions, building upon GAMP and EM-GM-AMP. Our numerical results confirm that NPML-AMP provides nearly MMSE solutions and closes the gap between EM-GM-AMP and the oracle GAMP with known signal distributions. Although the experiments in this chapter focus on settings where $\mathbf{A}$ has iid Gaussian entries, suitable modifications of NPML-AMP should continue to perform well other measurement matrices. In fact, recently vector AMP (VAMP) algorithm (Rangan et al., 2017) is established for right-rotationally invariant measurement matrices. We expect similar performance results hold for a broader class of measurement matrices modifying NPML-AMP accordingly. Other interesting areas for future research include algorithmic refinements for NPML-AMP; deriving theoretical results on the convergence of NPML-AMP by leveraging approximate convexity; and deriving statistical properties of NPML-AMP and the NPML estimator $\hat{F}_\Omega$.

2.8 Appendix

2.8.1 Proof of Theorem 1

Theorem 1. *Let $\Omega \in \mathbb{R}$ be a bounded set, so that $\Omega \subseteq [x_{\min}, x_{\max}]$ for some $-\infty < x_{\min} < x_{\max} < \infty$. Let $x^* = \max(|x_{\min}|, |x_{\max}|)$ and for $F \in \mathcal{F}_\Omega$ define*

$$\begin{aligned} \ell(F) &= -\frac{1}{N}\log p(\boldsymbol{y}; F), \\ \ell_{\text{conv}}(F) &= -\frac{1}{N}\sum_{j=1}^{N}\log\int\frac{1}{\psi^{N/2}}\exp\left\{\frac{1}{\psi}(\mathbf{A}_{\cdot j}^\top\boldsymbol{y})x_j - \right. \\ &\qquad \left. \frac{1}{2\psi}\|\mathbf{A}_{\cdot j}\|_2^2 x_j^2\right\}\, dF(x_j) + \frac{\|\boldsymbol{y}\|_2^2}{2N\psi}. \end{aligned}$$

Assume A_{ij} are iid random variables with $\mathbb{E}(A_{ij}) = 0$ and $\mathbb{E}(A_{ij}^4) \le C$ for some constant $C > 0$. If $M \to \infty$ and $M/N \to c \in (0, \infty)$, then,

$$\sup_{F\in\mathcal{F}_\Omega} |\ell(F) - \ell_{\text{conv}}(F)| = \mathcal{O}_p\left(\frac{\{x^*\}^2}{\psi\sqrt{M}}\right).$$

Proof. First we consider the term $\|\sum_{j=1}^{N}\mathbf{A}_{\cdot j}\boldsymbol{x}_j\|^2$. Since $\mathbf{A}_{ij} \sim N\left(0, \frac{1}{M}\right)$,

$$\begin{aligned} \mathbb{E}\mathbf{A}_{\cdot j_1}^\top\mathbf{A}_{\cdot j_2} &= \mathbb{E}\sum_{i=1}^{M}\mathbf{A}_{i,j_1}\mathbf{A}_{i,j_2} = 0, \\ \mathbb{E}(\mathbf{A}_{\cdot j_1}^\top\mathbf{A}_{\cdot j_2})^2 &= \mathbb{E}\sum_{i=1}^{M}(\mathbf{A}_{i,j_1}\mathbf{A}_{i,j_2})^2 = \frac{1}{M}. \end{aligned}$$

Moreover, for $j_1 \neq j_1'$,

$$\mathrm{Cov}(\mathbf{A}_{\cdot j_1}^{\top}\mathbf{A}_{\cdot j_2}, \mathbf{A}_{\cdot j_1'}^{\top}\mathbf{A}_{\cdot j_2}) = \mathbb{E}\left(\sum_{i=1}^{M}\mathbf{A}_{i,j_1}\mathbf{A}_{i,j_2}\right)\left(\sum_{i=1}^{M}\mathbf{A}_{i,j_1'}\mathbf{A}_{i,j_2}\right) = 0.$$

It follows that

$$\mathrm{Var}\Big(\sum_{1\leq j_1<j_2\leq N}\mathbf{A}_{\cdot j_1}^{\top}\mathbf{A}_{\cdot j_2}x_{j_1}x_{j_2}\Big) \quad = \quad \frac{1}{M}\sum_{1\leq j_1<j_2\leq N}x_{j_1}^2x_{j_2}^2. \tag{2.8}$$

Then

$$\left|\sum_{1\leq j_1<j_2\leq N}\mathbf{A}_{\cdot j_1}^{\top}\mathbf{A}_{\cdot j_2}x_{j_1}x_{j_2}\right| \leq \frac{C_1}{\sqrt{M}}\big(\sum_{1\leq j_1<j_2\leq N}x_{j_1}^2x_{j_2}^2\big)^{1/2} \tag{2.9}$$

for certain constant C_1. Also, by the law of large numbers, $\|\mathbf{A}_{\cdot j}\|_2^2 \to 1$, or

$$(1-\epsilon)\sum_{j=1}^{N}x_j^2 \leq \sum_{j=1}^{N}\|\mathbf{A}_{\cdot j}\|_2^2x_j^2 \leq (1+\epsilon)\sum_{j=1}^{N}x_j^2$$

for any $\epsilon > 0$. Thus

$$\frac{|\sum_{1\leq j_1<j_2\leq N}\mathbf{A}_{\cdot j_1}^{\top}\mathbf{A}_{\cdot j_2}x_{j_1}x_{j_2}|}{\sum_{j=1}^{N}\|\mathbf{A}_{\cdot j}\|_2^2x_j^2} \leq \frac{C_1}{(1-\epsilon)\sqrt{M}}\frac{\big(\sum_{1\leq j_1<j_2\leq N}x_{j_1}^2x_{j_2}^2\big)^{1/2}}{\sum_{j=1}^{N}x_j^2} \leq \frac{C_1}{(1-\epsilon)\sqrt{M}}.$$

Let $C_2 = C_1/(1-\epsilon)$ and

$$\sum_{j=1}^{N}\|\mathbf{A}_{\cdot j}\|_2^2\boldsymbol{x}_j^2\left(1-\frac{2C_2}{\sqrt{M}}\right) \leq \|\sum_{j=1}^{N}A_{\cdot j}\boldsymbol{x}_j\|_2^2 \leq \sum_{j=1}^{N}\|\mathbf{A}_{\cdot j}\|_2^2\boldsymbol{x}_j^2\left(1+\frac{2C_2}{\sqrt{M}}\right). \tag{2.10}$$

It then follows that for any $\boldsymbol{\omega}$,

$$\begin{aligned}
\ell(\boldsymbol{\omega}) \quad \leq& \quad -\frac{1}{N}\log\sum_{\boldsymbol{x}\in\Lambda^N}\frac{1}{\psi^N}\exp\Big\{\frac{1}{2\psi}\Big(2\sum_{j=1}^{N}x_j\mathbf{A}_{\cdot j}^{\top}\boldsymbol{y}-\sum_{j=1}^{N}\|\mathbf{A}_{\cdot j}\|_2^2\boldsymbol{x}_j^2\big(1+\frac{2C_2}{\sqrt{M}}\big)\Big\}\\
&\quad \prod_{j=1}^{N}\Big(\sum_{l=1}^{L}\omega_l1(x_j=\theta_l)\Big)\\
=& \quad -\frac{1}{N}\sum_{j=1}^{N}\log\sum_{l=1}^{L}\exp\Big\{\frac{1}{\psi}(\mathbf{A}_{\cdot j}^{\top}\boldsymbol{y})\theta_l-\frac{1}{2\psi}\|\mathbf{A}_{\cdot j}\|_2^2\theta_l^2\big(1+\frac{2C_2}{\sqrt{M}}\big)\Big\}w_l\\
\leq& \quad -\frac{1}{N}\sum_{j=1}^{N}\log\sum_{l=1}^{L}\exp\Big\{\frac{1}{\psi}(\mathbf{A}_{\cdot j}^{\top}\boldsymbol{y})\theta_l-\frac{1}{2\psi}\|\mathbf{A}_{\cdot j}\|_2^2\theta_l^2\big)\Big\}\exp\Big\{-\frac{\|\mathbf{A}_{\cdot j}\|_2^2\{\theta^*\}^2}{2\psi}\frac{2C_2}{\sqrt{M}}\Big\}w_l\\
\leq& \quad -\frac{1}{N}\sum_{j=1}^{N}\log\sum_{l=1}^{L}\exp\Big\{\frac{1}{\psi}(\mathbf{A}_{\cdot j}^{\top}\boldsymbol{y})\theta_l-\frac{1}{2\psi}\|\mathbf{A}_{\cdot j}\|_2^2\theta_l^2\big)\Big\}w_l+\frac{1}{N}\sum_{j=1}^{N}\frac{\|\mathbf{A}_{\cdot j}\|_2^2\{\theta^*\}^2}{\psi}\frac{C_2}{\sqrt{M}}\\
\leq& \quad -\frac{1}{N}\sum_{j=1}^{N}\log\sum_{l=1}^{L}\exp\Big\{\frac{1}{\psi}(\mathbf{A}_{\cdot j}^{\top}\boldsymbol{y})\theta_l-\frac{1}{2\psi}\|\mathbf{A}_{\cdot j}\|_2^2\theta_l^2\big)\Big\}w_l+\frac{(1+\epsilon)C_2\{\theta^*\}^2}{\psi\sqrt{M}}.
\end{aligned} \tag{2.11}$$

The last inequality holds due to $(1/N)\sum_{i=1}^{N}\|A_{\cdot j}\|_2^2 \le 1+\epsilon$. Let $C=(1+\epsilon_2)C_2$, we have

$$\ell(\boldsymbol{\omega})-h(\boldsymbol{\omega}) \le \frac{C\{\theta^*\}^2}{\psi\sqrt{M}}, \qquad \forall \boldsymbol{\omega}\in\Delta^{L-1}.$$

Similarly, $\ell(\boldsymbol{\omega})-h(\boldsymbol{\omega}) \ge -\frac{C\{\theta^*\}^2}{\psi\sqrt{M}}$ for any $\boldsymbol{\omega}\in\Delta^{L-1}$, then (2.6) holds.

Moreover, we notice that

$$\mathbb{E}\|A_{\cdot j}^{\top}y\|_2 = \frac{N+M+1}{N}\tau^2+\psi, \quad \tau^2=\frac{1}{M}\sum_{j=1}^{N}x_j^2.$$

It then follows that

$$\begin{aligned}
h(\boldsymbol{\omega}) &= -\frac{1}{N}\sum_{j=1}^{N}\log\sum_{l=1}^{L}\exp\Big\{\frac{1}{\psi}(\mathbf{A}_{\cdot j}^{\top}\boldsymbol{y})\theta_l-\frac{1}{2\psi}\|\mathbf{A}_{\cdot j}\|_2^2\theta_l^2)\Big\}w_l \\
&= -\frac{1}{N}\sum_{j=1}^{N}\log\sum_{l=1}^{L}\exp\Big\{-\frac{\|\mathbf{A}_{\cdot j}\|_2^2}{2\psi}\Big(\theta_l-\frac{\mathbf{A}_{\cdot j}^{\top}\boldsymbol{y}}{\|\mathbf{A}_{\cdot j}\|_2^2}\Big)^2+\frac{\big(\mathbf{A}_{\cdot j}^{\top}\boldsymbol{y}\big)^2}{\psi\|\mathbf{A}_{\cdot j}\|_2^2}\Big\}w_l \\
&\ge -\frac{1}{N}\sum_{j=1}^{N}\frac{\big(\mathbf{A}_{\cdot j}^{\top}\boldsymbol{y}\big)^2}{\psi\|\mathbf{A}_{\cdot j}\|_2^2} \\
&\to -\frac{(N+M+1)\tau^2}{N\psi}-1.
\end{aligned}$$

Moreover,

$$h(\boldsymbol{\omega}) \;\le\; \max(a,b),$$

where

$$\begin{aligned}
a &= -\frac{1}{N}\sum_{j=1}^{N}\frac{1}{\psi}(\mathbf{A}_{\cdot j}^{\top}\boldsymbol{y})\theta_*-\frac{1}{2\psi}\|\mathbf{A}_{\cdot i}\|_2^2\theta_*^2, \\
b &= -\frac{1}{N}\sum_{j=1}^{N}\frac{1}{\psi}(\mathbf{A}_{\cdot j}^{\top}\boldsymbol{y})\theta^*-\frac{1}{2\psi}\|\mathbf{A}_{\cdot j}\|_2^2\{\theta^*\}^2.
\end{aligned}$$

We note that

$$\begin{aligned}
h(\boldsymbol{\omega}) &= -\frac{1}{N}\sum_{j=1}^{N}\log\Big\{\mathbb{E}_{\boldsymbol{x}}\exp\Big(\frac{1}{\psi}(A_{\cdot j}^{T}\boldsymbol{y})\boldsymbol{x}-\frac{1}{2\psi}\|\mathbf{A}_{\cdot j}\|_2^2\boldsymbol{x}^2\Big)\Big\} \\
&< -\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_X\log\Big\{\exp\Big(\frac{1}{\psi}(A_{\cdot j}^{\top}\boldsymbol{y})\boldsymbol{x}-\frac{1}{2\psi}\|\mathbf{A}_{\cdot j}\|_2^2\boldsymbol{x}^2\Big)\Big\} \\
&= -\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_X\Big(\frac{1}{\psi}(A_{\cdot j}^{\top}\boldsymbol{y})\boldsymbol{x}-\frac{1}{2\psi}\|\mathbf{A}_{\cdot j}\|_2^2\boldsymbol{x}^2\Big) \\
&= -\Big(\frac{(N+M+1)\tau^2}{N\psi}+1\Big)\mathbb{E}\boldsymbol{x}+\frac{1}{2\psi}\mathbb{E}\boldsymbol{x}^2.
\end{aligned}$$

□

2.8.2 Moment identities

Let

$$\begin{aligned}\hat{x} &= \frac{\sum_{j=1}^{N} \mathbf{A}_{\cdot j}^{\top} \boldsymbol{y}}{\sum_{1\leq j_1,j_2\leq N} \mathbf{A}_{\cdot j_1}^{\top} \mathbf{A}_{\cdot j_2}};\\ \tilde{x}^2 &= \frac{\|\boldsymbol{y}\|_2^2 - \hat{x}^2 \sum_{j_1\neq j_2} \mathbf{A}_{\cdot j_1}^{\top} \mathbf{A}_{\cdot j_2} - M\psi}{\sum_{j=1}^{N} \mathbf{A}_{\cdot j}^{\top} \mathbf{A}_{\cdot j}}.\end{aligned}$$

we would like to show that

$$\begin{aligned}\mathbb{E}(x_j) &= \mathbb{E}(\hat{x} \mid \mathbf{A}),\\ \mathbb{E}(x_j^2) &= \mathbb{E}(\tilde{x}^2 \mid \mathbf{A}),\\ \mathrm{Var}(x_j^2) &= \mathbb{E}(\tilde{x}^2 \mid \mathbf{A}) - \mathbb{E}(\hat{x} \mid \mathbf{A})^2.\end{aligned}$$

Proof. The linear model can be written as

$$\boldsymbol{y} = \sum_{j=1}^{N} \mathbf{A}_{\cdot j} x_j + \boldsymbol{\varepsilon}.$$

Equivalently,

$$\begin{aligned}\sum_{j=1}^{N} \mathbb{E}_{X,\varepsilon} \mathbf{A}_{\cdot j}^{\top} \boldsymbol{y} &= \sum_{j=1}^{N} \mathbb{E}_{X,\varepsilon} \left[\mathbf{A}_{\cdot i}^{\top} \left(\sum_{j=1}^{N} \mathbf{A}_{\cdot j} x_j + \mathbf{A}_{\cdot j}^{\top} \boldsymbol{\varepsilon} \right) \right]\\ &= \sum_{j=1}^{N} \mathbb{E}_{X} \left[\mathbf{A}_{\cdot j}^{\top} \left(\sum_{j=1}^{N} \mathbf{A}_{\cdot j} x_j \right) \right]\\ &= (\mathbb{E}_X x_1) \sum_{j_1,j_2=1}^{N} \mathbf{A}_{\cdot j_1}^{\top} \mathbf{A}_{\cdot j_2}.\end{aligned}$$

Let

$$\bar{X} = \frac{\sum_{j=1}^{N} \mathbf{A}_{\cdot j}^{\top} \boldsymbol{y}}{\sum_{j_1,j_2=1}^{N} \mathbf{A}_{\cdot j_1}^{\top} \mathbf{A}_{\cdot j_2}},$$

This lead to

$$\mathbb{E}_X x_1 = \frac{\sum_{j=1}^{N} \mathbb{E}_{X,\varepsilon} \mathbf{A}_{\cdot j}^{\top} \boldsymbol{y}}{\sum_{j_1,j_2=1}^{N} \mathbf{A}_{\cdot j_1}^{\top} \mathbf{A}_{\cdot j_2}} \approx \bar{X}. \tag{2.12}$$

Moreover,

$$\mathbb{E}_{X,\varepsilon} \|\boldsymbol{y}\|_2^2 = \mathbb{E}_{X,\varepsilon} \left\| \sum_{j=1}^{N} \mathbf{A}_{\cdot i} x_j + \boldsymbol{\varepsilon} \right\|_2^2$$

$$= \quad (\mathbb{E}_X x_1^2) \sum_{j=1}^{N} \mathbf{A}_{\cdot j}^\top \mathbf{A}_{\cdot j} + (\mathbb{E}_X x_1)^2 \sum_{j_1 \neq j_2} \mathbf{A}_{\cdot j_1}^\top \mathbf{A}_{\cdot j_2} + M\psi.$$

This leads to

$$\mathbb{E}_X x_1^2 = \frac{\mathbb{E}_{X,\varepsilon} \|\boldsymbol{y}\|_2^2 - M\psi - (\mathbb{E}_X x_1)^2 \sum_{j_1 \neq j_2} \mathbf{A}_{\cdot j_1}^\top \mathbf{A}_{\cdot j_2}}{\sum_{j=1}^N \mathbf{A}_{\cdot j}^\top \mathbf{A}_{\cdot j}} \approx \widetilde{X^2},$$

where

$$\widetilde{X^2} = \frac{\|\boldsymbol{y}\|_2^2 - M\psi - \bar{X}^2 \sum_{j_1 \neq j_2} \mathbf{A}_{\cdot j_1}^\top \mathbf{A}_{\cdot j_2}}{\sum_{j=1}^N \mathbf{A}_{\cdot j}^\top \mathbf{A}_{\cdot j}}.$$

Furthermore,

$$\mathrm{Var}_X(X_1) = \mathbb{E}_X x_1^2 - (\mathbb{E}_X x_1)^2 \approx \widetilde{X^2} - \bar{X}^2. \tag{2.13}$$

Combine (2.12) and (2.13), we can approximate the range of X by

$$X \in \left(\bar{X} - k\sqrt{\widetilde{X^2} - \bar{X}^2}, \quad \bar{X} + k\sqrt{\widetilde{X^2} - \bar{X}^2}\right).$$

As $N, M \to \infty$,

$$\begin{aligned} \bar{X} \quad &\to \quad \mathbb{E}X; \\ \widetilde{X^2} - \bar{X}^2 \quad &\to \quad \mathrm{Var}(X). \end{aligned}$$

In the case where ψ is unknown, we let the term $M\psi$ in $\widetilde{X^2}$ be 0 to further relax the interval.

By Chebyshev inequality, we can assure that asymptotically,

$$P\left(X \in \left(\bar{X} - k\sqrt{\widetilde{X^2} - \bar{X}^2}, \quad \bar{X} + k\sqrt{\widetilde{X^2} - \bar{X}^2}\right)\right) \geq 1 - \frac{1}{k^2}.$$

□

2.8.3 Derivation of M-step

In this subsection, we derive the maximization step for the EM algorithm.

$$\begin{aligned} \boldsymbol{\omega}(t+1) \quad &= \quad \underset{\boldsymbol{\omega}>0:\sum_l \omega_l = 1}{\mathrm{argmax}} \quad \hat{\mathbb{E}}\{\ln p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\omega}(t))\} \\ &= \quad \underset{\boldsymbol{\omega}>0:\sum_l \omega_l = 1}{\mathrm{argmax}} \quad \hat{\mathbb{E}}\Bigg\{ \ln \left(\prod_{j=1}^{N} p_X(x_j | \boldsymbol{y}, \boldsymbol{\omega}(t)) \right) \Bigg\} \end{aligned}$$

$$= \underset{\boldsymbol{\omega}>0:\sum_l \omega_l=1}{\operatorname{argmax}} \sum_{j=1}^{N} \hat{\mathbb{E}}\{\ln p_X(x_j;\boldsymbol{\omega},\psi)|\boldsymbol{y};\boldsymbol{\omega}(t)\} \tag{2.14}$$

The Lagrange multiplier problem can be written as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\omega},\lambda) &= \sum_{j=1}^{N} \hat{\mathbb{E}}\{\ln p_X(x_j;\boldsymbol{\omega})|\boldsymbol{y};\boldsymbol{\omega}(t)\} - \lambda\big(\sum_{l=1}^{L} w_l - 1\big) \\ &= \sum_{j=1}^{N}\sum_{l=1}^{L} p_{X|\boldsymbol{Y}}(x_j=\theta_l|\boldsymbol{y};\boldsymbol{\omega}(t)) \ln p_X(x_j=\theta_l;\boldsymbol{\omega}) - \lambda\big(\sum_{l=1}^{L} w_l - 1\big) \qquad (2.15) \end{aligned}$$

We set $\frac{d\mathcal{L}}{d\omega_l}=0$ and derive that

$$\begin{aligned} \sum_{j=1}^{N} p_{X|\boldsymbol{Y}}(x_j=\theta_l|\boldsymbol{y};\boldsymbol{\omega}(t)) \frac{d}{d\omega_l} \ln p_X(x_j=\theta_l;\boldsymbol{\omega}) &= \lambda \\ \sum_{j=1}^{N} \frac{1}{\omega_l} \frac{\omega_l^t \mu_j^{-1/2} \phi(\theta_l - \hat{r}_j)}{\sum_{l=1}^{L} \omega_l^t \mu_j^{-1/2} \phi(\theta_l - \hat{r}_j)} &= \lambda \qquad (2.16) \end{aligned}$$

, where $\frac{\omega_l^t \mu_j^{-1/2} \phi(\theta_l - \hat{r}_j)}{\sum_{l=1}^{L} \omega_l^t \mu_j^{-1/2} \phi(\theta_l - \hat{r}_j)}$ is previously defined as $\mathcal{Q}_{j,l}^t$.

Multiplying both sides by ω_l for $l = 1,\ldots,L$, since $\sum_{l=1}^{L} \omega_l = 1$,

$$\lambda = \sum_{l=1}^{L}\sum_{j=1}^{N} \mathcal{Q}_{j,l}^t. \tag{2.17}$$

Plugging (2.17) back to (2.16), we have

$$\omega_l(t+1) = \frac{\sum_{j=1}^{N} \mathcal{Q}_{j,l}^t}{\sum_{l=1}^{L}\sum_{j=1}^{N} \mathcal{Q}_{j,l}^t}. \tag{2.18}$$

Chapter 3

Heritability estimation in genome-wide association studies: fixed-effects vs random-effects methods

3.1 Introduction

Heritability is the proportion of phenotype variance explained by genetic variance (Falconer, 1960; Lynch et al., 1998). There are many different definitions of heritability and different methods for estimating heritability from data (e.g. Yang et al., 2010, 2011a; Golan et al., 2014; Bulik-Sullivan et al., 2015). This chapter is focused on heritability estimation methods related to variance components estimation problems for linear mixed models (LMMs). LMM-based methods for heritability estimation have been used since the 1950s (Henderson, 1950); additionally, over the last ten years they have emerged as one of the most widely-used methods for estimating heritability with genome-wide association studies (GWAS) data (Hindorff et al., 2009; Yang et al., 2010; Kang et al., 2010; Zaitlen and Kraft, 2012). However, standard approaches to heritability estimation with LMMs have some unresolved inconsistencies related to important topics in genetics, including linkage disequilibrium (LD), the distribution of causal variants, and partitioning heritability (Zaitlen and Kraft, 2012; Speed et al., 2012; Gusev et al., 2013, 2014). This chapter contains new statistical results, which suggest that by adopting a slightly different statistical perspective, many of these inconsistencies with LMM-based heritability estimation can be seamlessly resolved.

LMM-based heritability methods typically require specification of a genetic relationship matrix (GRM), which measures genetic similarity between subjects in a study. The GRM may be based on familial or other information; in GWAS, the GRM is frequently a sample correlation matrix constructed from study participant's single nucleotide polymorphism (SNP) values, which corresponds to a Euclidean distance kernel (Yang et al., 2010; Zaitlen and Kraft, 2012). In this chapter, we argue that if one adopts a Mahalanobis distance-based genetic relationship matrix for LMM analysis, then many of the previously noted LMM inconsistencies related to LD and causal variants are immediately resolved. Moreover, with the Mahalanobis GRM, it's straightforward to define a natural version of partitioned heritability, which avoids some of the pitfalls that have been noted for other approaches (Speed et al., 2012; Gusev et al., 2013, 2014). While the Mahalanobis GRM resolves these consistency questions at the modeling level, it also heightens the importance of understanding and estimating the LD structure for the study population – indeed, the LD matrix is required for computing the Mahalanobis kernel.

Our arguments for the Mahalanobis kernel touch on several fundamental aspects of statistical modeling in modern genetics. Questions about fixed- and random-effects modeling have been raised repeatedly in research on heritability estimation (Gibson, 2012). Many of these questions can be summarized as follows: should genetic effects be modeled as fixed or random quantities? To resolve this question, we argue that for the Mahalanobis kernel, the fixed- and random-effects models are essentially equivalent. Furthermore, under the Mahalanobis kernel, we show that the LMM heritability coefficient can also be interpreted as a conditional variance – which we refer to as the *C-heritability* (C for "conditional") – under the corresponding fixed-effects model (the random-effects interpretation is more standard in the literature, to date). This builds a link between narrow-sense (or additive) heritability, which LMM-based methods have traditionally been designed to estimate, and broad-sense heritability, which is a more model-free measure of overall heritability defined in terms of the conditional variance of a phenotype given the genotype and other specified information.

3.2 LMMs for heritability estimation

3.2.1 Additive decomposition: From GRMs to LMMs

In this section, we describe a statistical model that forms the basis for many LMM heritability methods (Yang et al., 2010; Zaitlen and Kraft, 2012). Let $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ be a vector of centered, real-vaued outcomes, where y_i represents the phenotypic value of individual i in some population. Assume that

$$\mathbf{y} = \mathbf{g} + \boldsymbol{e} \tag{3.1}$$

can be decomposed as the sum of an additive genetic effect $\mathbf{g} = (g_1, \ldots, g_n)^\top \in \mathbb{R}^n$ and an uncorrelated noise vector $\boldsymbol{e} = (e_1, \ldots, e_n)^\top \in \mathbb{R}^n$, which may contain other non-additive genetic effects, environmental noise, and measurement error. Further assume that the data are centered, so that $\mathbb{E}(\mathbf{g}) = \mathbb{E}(\boldsymbol{e}) = 0$, and that $\mathrm{Cov}(\mathbf{g}) = \sigma_g^2 K$ and $\mathrm{Cov}(\boldsymbol{e}) = \sigma_e^2 I$, where $\sigma_g^2, \sigma_e^2 \geq 0$ are genetic and environmental real-valued variance components, respectively, and K is the $n \times n$ GRM.

The diagonal noise covariance matrix assumption comes from two extra steps in data collecting. On one hand, individuals $i = 1, \ldots, n$ are selected by removing ones whose pairwise relatedness are greater than certain threshold. The relatedness is usually measured by the kernel defined according to K, and maximum pairwise relatedness in the sample is corresponding to cousins 2-3 times removed (Yang et al., 2010; Speed et al., 2012). On the other hand, $\mathbf{y}$ is a projected response that is orthogonal to the space of demographic covariates such as sex, age and handedness (Yang et al., 2011a; Bonnet et al., 2015; Lee et al., 2016).

Then

$$\mathbf{y} \sim \mathcal{MV}(0, \sigma_g^2 K + \sigma_e^2 I) \tag{3.2}$$

and the (narrow-sense) heritability is defined to be

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}. \tag{3.3}$$

The GRM is typically normalized so that its diagonal entries all equal 1; in this case, the correlation matrix for $\mathbf{y}$ is $\mathrm{Corr}(\mathbf{y}) = h^2 K + (1 - h^2) I$ and the heritability parameter h^2

represents the extent of correlation between individuals in the population determined by genetic relatedness.

With GWAS data, genetic relatedness can be encoded by similarities between sequences of single nucleotide polymorphisms (up to hundreds of thousands or millions of SNPs). Let $\mathbf{z}_i = (z_{i1}, \ldots, z_{im})^\top$ be the vector of normalized SNPs for individual i, in the sense that

$$\mathbf{z}_{ij} = \frac{f_{ij} - 2p_j}{\sqrt{2p_j(1-p_j)}},$$

where $f_{ij} = 0, 1, 2$ respectively if the genotype of individual i at SNP j is aa, Aa or AA, and p_j is the minor allele frequency (MAF) of SNP j (Meuwissen and Goddard, 2001; Hayes et al., 2009; Zaitlen and Kraft, 2012). Then the ij-entry of the GRM $K = (K_{ij})$ is determined by some kernel function $K : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$, whereby $K_{ij} = K(\mathbf{z}_i, \mathbf{z}_j)$.

Traditionally, the GRM (or kinship matrix) indicates proportion of identical genetic regions that individual i and j inherited from common ancestors. This identity-by-descent (IBD) kernel is defined with respect to a pedigree, and knowledge of an explicit pedigree for the population in the study is usually infeasible. In the absense of pedigree information, GRM is frequently defined by the identity-by-state (IBS) GRM, where

$$K(\mathbf{z}_i, \mathbf{z}_j) = \frac{1}{m}\mathbf{z}_i^\top \mathbf{z}_j. \tag{3.4}$$

The IBS GRM definition corresponds to the normalized linear kernel, and measures average allelic correlations (Powell et al., 2010; Speed and Balding, 2015). Other kernel functions have been proposed for GWAS heritability estimation problems (e.g. the Gaussian kernel or higher-order polynomial kernels (Akdemir and Jannink, 2015)). However, to date, there appears to be limited evidence for preferring these kernels over linear kernels.

The linear kernel (3.4) corresponds to a linear random effects model — or, a LMM before projecting out fixed covariates — hence, the term LMM-based heritability estimation. In this corresponding random effects model, $\mathbf{g} = Z\mathbf{u}$ in (3.1), where $\mathbf{u} = (u_1, \ldots, u_m)^\top \in \mathbb{R}^m$ is a vector of independent random genetic effects with $u_i \sim \mathcal{MV}(0, \sigma_g^2/m)$ and $Z = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^\top$ is the $n \times m$ matrix of genotypes. Thus, under these assumptions, we can rewrite (3.1) as

$$\mathbf{y} = Z\mathbf{u} + \boldsymbol{e}. \tag{3.5}$$

In this model, the data from each subject is the (phenotype, genotype)-pair $(y_i, \mathbf{z}_i) \in \mathbb{R}^{m+1}$.

The main focus of this paper is the Mahalanobis kernel. Let Σ be the $m \times m$ positive semi-definite matrix representing the population-level covariance (linkage disequilibrium) matrix for the SNPs $\mathbf{z}_i$. The Mahalanobis kernel is defined by $K(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i^\top \Sigma^{-1} \mathbf{z}_j$. The Mahalanobis kernel has been widely used in other applications involving genetics, e.g. genetic association testing (Majumdar et al., 2015). However, it appears to have received little attention in the context of heritability estimation. The Mahalanobis kernel also corresponds to a linear model with correlated random effects with $\mathbf{u} \sim \mathcal{MV}(0, \tau_g^2/m\Sigma^{-1})$ in (3.5). We would like to show that whitening the design matrix is the key to unbiased heritability estimation. In the remainder of the paper, we will derive several attractive features of the Mahalanobis kernel for heritability estimation.

3.2.2 Estimating h^2

Moment methods and maximum likelihood are two widely used classes of methods for estimating h^2 under (3.2). Note that these methods are applicable for any GRM K. In this subsection, suppose the GRM is normalized with its diagonal entries all equal 1, suppose $\mathbf{y}$ is centered. Then

$$\mathrm{Cov}(\mathbf{y}) = \sigma_g^2 K + \sigma_e^2 I. \tag{3.6}$$

One of the classical moment estimators for h^2 comes from observing that σ_g^2 is the least squares coefficient for regressing $y_i y_j$ on K_{ij} for all $i < j$. This is because (3.6) implies that

$$\mathbb{E}(y_i y_j | K) = \sigma_g^2 K_{ij}, \quad i \neq j$$

The corresponding estimator for σ_g^2 is

$$\tilde{\sigma}_g^2 = \left(\widehat{\mathrm{Var}}(K_{ij})\right)^{-1} \widehat{\mathrm{Cov}}(y_i y_j, K_{ij}),$$

where

$$\begin{aligned} \widehat{\mathrm{Var}}(K_{ij}) &= \frac{2}{n(n-1)} \sum_{i<j} K_{ij}^2 \\ \widehat{\mathrm{Cov}}(y_i y_j, K_{ij}) &= \frac{2}{n(n-1)} \sum_{i<j} y_i y_j K_{ij}. \end{aligned}$$

Henderson used least squares in this way to estimate h^2 with

$$\tilde{h}^2 = \frac{\tilde{\sigma}_g^2}{\|\mathbf{y}\|_2^2/n},$$

and variants of this method are still used today (Haseman and Elston, 1972; Henderson, 1984; Golan et al., 2014); this approach is also referred to as Haseman-Elston regression.

In the standard maximum likelihood approach, one assumes that $\mathbf{y}$ is Gaussian, i.e.

$$\mathbf{y} \sim \mathcal{N}(0, \sigma_g^2 K + \sigma_e^2 I),$$

and estimates σ_g^2, σ_e^2 and, subsequently, h^2, by maximizing the Gaussian likelihood for this model. Specifically, let $\eta^2 = \sigma_g^2/\sigma_e^2$. The maximum likelihood estimator for (σ_e^2, η^2) is

$$(\hat{\sigma}_e^2, \hat{\eta}^2) = \underset{\sigma_e^2, \eta^2 > 0}{\operatorname{argmax}} \; l(\sigma_e^2, \eta^2),$$

where

$$\begin{aligned} l(\sigma_e^2, \eta^2) &= -\frac{1}{2}\log(\sigma_e^2) - \frac{1}{2n}\log\det(\eta^2/mK + I) \\ &\quad -\frac{1}{2n\sigma_e^2}\mathbf{y}^\top(\eta^2/mK + I)^{-1}\mathbf{y}. \end{aligned}$$

Hence, the MLE of h^2 is

$$\hat{h}^2 = \frac{\hat{\eta}^2}{\hat{\eta}^2 + 1}.$$

Established by Yang et al. (2010, 2011a), the MLE with linear kernel-based GRM has become the landmark approach in estimating h^2.

Both maximum likelihood and moment estimators for h^2 often have nice statistical properties (e.g. consistency). In some circumstances, maximum likelihood estimators may have advantages over moment estimators in terms of efficiency (reduced variance). On the other hand, moment estimators have been the subject of renewed interest recently because of potential advantages related to computation and data privacy (as many data only disclose summary GWAS statistics for the population (Finucane et al., 2015)).

3.3 Challenges for LMMs

LMM methods for estimating h^2 may give inconsistent or unexpected results when used in settings where the generative model for $\mathbf{y}$ differs from (3.2), i.e. under model misspecification. This has been noted repeatedly in the heritability literature (Zaitlen and Kraft, 2012), and is important because many of the leading generative models from genetics for linking phenotypes $\mathbf{y}$ and SNP values $\mathbf{z}$ differ substantially from (3.2) (Barrett et al., 2009; Stahl et al., 2010; Gibson, 2012). We outline two such examples in Sections 3.3.1 and 3.3.2 below.

3.3.1 Causal variants and linkage disequilibrium

Many genetics models hypothesize a collection of causal loci (or causal variants), which are fixed locations along the genome, where the specific nucleotide combination impacts the phenotype — other, non-causal loci are assumed to have no direct impact on the phenotype (Pritchard, 2001). In the context of the LMM (3.5), this is frequently encoded by taking $\mathcal{S} \subseteq [m]$ to be the collection of causal loci and assuming:

$$\begin{aligned} &u_j \sim F \text{ are independent for } j \in \mathcal{S}, \\ &u_j = 0 \text{ if } j \notin \mathcal{S}. \end{aligned} \tag{3.7}$$

The assumptions (3.7) violate the exchangeability assumptions on the coordinates of $\mathbf{u}$ that are required under (3.5).

It turns out that violating exchangeability in (3.7) alone is not necessarily problematic for LMM heritability estimation. However, it has been noted previously that under the linear GRM, estimates of h^2 can be systematically unreliable when non-exchangeable genetic effects are coupled with linkage disequilibrium. Specifically, estimators with linear GRM is biased when LD-level of causal variants is substantially different from average LD-level for variants in the study (Zaitlen and Kraft, 2012; Speed et al., 2012; Gusev et al., 2013; Yang et al., 2015).

If the SNP values $\mathbf{z}_i$ are modeled to be random, then LD is measured by $\text{Cov}(\mathbf{z}_i) = \Sigma$. In particular, if Σ is diagonal, then there is no LD; if z_{ij} and $z_{ij'}$ are highly correlated, then LD between SNPs j and j' is high. If $\mathcal{S}$ is associated with the LD structure, e.g. if $\mathcal{S}$

is concentrated among a group of SNPs with relatively low or high LD, then heritability estimates may be biased (Speed et al., 2012).

Suppose $\mathbf{y}$ is normalized (i.e. $\sigma_e^2 = 1 - \sigma_g^2$), conditioning on $\mathbf{u}$, heritability measures

$$\mathbf{u}^\top \Sigma \mathbf{u} = \sum_{i,j}^{m} u_i u_j \Sigma_{ij}.$$

The random-effects assumption suggests that $\mathbb{E}_u(\mathbf{u}^\top \Sigma \mathbf{u}) = \sigma_g^2 \mathrm{tr}(\Sigma)/m$; in other words, it assumes that $\mathbb{E}_u\left(u_i u_j \Sigma_{ij}\right) = 0$ for $i \neq j$. Omitting these terms is reasonable as long as either $\mathcal{S}$ is not associated with the LD structure or there is no LD. Otherwise, when $\mathcal{S}$ is within a group of SNPs with different levels of LD, $\mathbb{E}_u\left[u_i u_j \Sigma_{ij}\right] \neq 0$ when $i \neq j$. As a result, random-effects heritability can cause potential bias because they ignore the impact of up to $m(m-1)$ cross terms.

There are many approaches to account for LD issue in linear GRM-based heritability estimation. One simple treatment is pruning. To achieve a diagonal Σ, one of every pair of correlated SNPs are romoved from the analysis (Purcell et al., 2007; Stahl et al., 2012). Without information of $\mathcal{S}$, causal loci could potentially be removed during the pruning step, and causes bias in estimating h^2. Other treatments focus on transforming and re-weighting the design matrix. For example Gusev et al. (2013) built on work of Patterson et al. (2006), and proposed to transform the design matrix such that each genotype is regressed on all preceeding SNPs. Each genotype is then replaced by the regression residuals. The LDAK method suggests to assign different weights to SNPs. The optimal SNP weights are computed by considering local LD and distance with neighboring SNPs and solving a linear programming problem. After the scaling step, SNPs with high LD is downweighted, and the lost signal can be compensated by its neighboring SNPs (Speed et al., 2012).

All of the previous LD adjustment methods are based on the normalized linear kernel. Although they target on mitigating h^2 estimation bias caused by LD issue, there are combinations of $\mathbf{u}$ and Σ where LD residual and LDAK adjustments fail to estimate h^2 consistently (Yang et al., 2015). We show that the Mahalanobis estimator also effectively resolves the uneven LD issue in the empirical example below. Moreover, in Section 3.4.1, we will discuss why the Mahalanobis estimator works for any Σ and $\mathbf{u}$.

Consider the case where e_j is Gaussian and $u_j \sim \mathcal{N}(0, \sigma_g^2/|\mathcal{S}|)$. Let $Z_\mathcal{S}$ denote the $n \times |\mathcal{S}|$ matrix obtained by extracting the columns of Z corresponding to $\mathcal{S}$. Then

$$\mathbf{y} \sim \mathcal{MV}\left(0, \frac{\sigma_g^2}{|\mathcal{S}|} Z_\mathcal{S} Z_\mathcal{S}^\top + \sigma_e^2 I\right)$$

follows the model (3.2) with $K_{i,j} = \mathbf{z}_{i,\mathcal{S}}^\top \mathbf{z}_{j,\mathcal{S}}/|\mathcal{S}|$ and $\mathbf{z}_{i,\mathcal{S}} = (z_{ik})_{k\in\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$, so the heritability coefficient is h^2. On the other hand, in the absence of additional information about $\mathcal{S}$, the least squares and maximum likelihood estimators for h^2 are frequently fit according to the model (3.5), with the linear kernel (3.4). We carry out some simulations below. We also fit the Mahalanobis estimator, assuming (3.5) using the Mahalanobis kernel, which requires additional information about LD. We ran a simple simulation study under this setting, with:

(i) $n = 500$, $m = 1000$.

(ii) $\mathcal{S} = \{1, \ldots, m/2\}$.

(iii) $\sigma_g^2 = \sigma_e^2 = 0.5$.

(iv) $\mathbf{z}_1, \ldots, \mathbf{z}_n \sim \mathcal{N}(0, \Sigma)$, where

$$\Sigma = \begin{pmatrix} \text{AR}(0.3) & 0 \\ 0 & \text{AR}(0.7) \end{pmatrix}$$

and AR(ρ) is the $m/2 \times m/2$ matrix with ij-entry $\rho^{|i-j|}$.

In this model, $h^2 = 0.5$. We simulated 50 independent datasets specified according to this model, and for each dataset maximum likelihood estimator with the linear kernel and the Mahalanobis maximum likelihood estimator are computed. The least squares estimator is not considered in this experiment because it relys on the approximation $\mathbf{u}^\top \Sigma^2 \mathbf{u} \approx \|\mathbf{u}\|^2 \text{tr}(\Sigma^2)/m$, however, under this simulation setting the approximation does not hold. Summary statistics are reported in Table 3.1.

From Table 3.1, it's evident that the estimator based on the linear kernel is substantially biased, and the Mahalanobis estimator is not. We defer a more comprehensive numerical analysis to Section 3.5.1.

Table 3.1: Means and confidence intervals for estimates of h^2. Based on results from 50 independent datasets. h^2 is estimated by MLE with linear and Mahalanobis kernels.

h^2	Linear MLE		Mahalanobis MLE	
0.5	Mean:	0.454	Mean:	0.495
	95% CI:	(0.427,0.482)	95% CI:	(0.468, 0.522)

3.3.2 Partitioning heritability

Studies on partitioning heritability seek to identify the heritability $h^2_{\mathcal{S}}$, which is attributable to a subset of SNPs $\mathcal{S} \subseteq [m]$ (Gusev et al., 2014; Finucane et al., 2015). Usually the SNPs are partitioned by functional areas such as genomes, levels of MAF and functional annotations (Davis et al., 2013). Care must be taken when disentangling the effects of SNPs in $\mathcal{S}$ with SNPs that are in linkage disequilibrium with $\mathcal{S}$. In particular, if LD is ignored, then estimates of $h^2_{\mathcal{S}}$ can be badly biased.

In (Yang et al., 2011b; Kostem and Eskin, 2013; Gusev et al., 2014), $\mathbf{y}$ is assumed to follow a LMM with two variance components

$$\mathbf{y} = Z_{\mathcal{S}}\mathbf{u}_{\mathcal{S}} + Z_{\mathcal{S}^c}\mathbf{u}_{\mathcal{S}^c} + \boldsymbol{e}, \tag{3.8}$$

where

$$u_i \sim \begin{cases} \mathcal{MV}\left(0, \frac{\sigma^2_{\mathcal{S}}}{|\mathcal{S}|}\right), & \text{if } i \in \mathcal{S}, \\ \mathcal{MV}\left(0, \frac{\sigma^2_{\mathcal{S}^c}}{m-|\mathcal{S}|}\right), & \text{if } i \notin \mathcal{S}. \end{cases}$$

Under this model, the heritability due to $\mathcal{S}$ is defined as

$$h^2_{\mathcal{S}} = \frac{\sigma^2_{\mathcal{S}}}{\sigma^2_{\mathcal{S}} + \sigma^2_{\mathcal{S}^c} + \sigma^2_e}, \tag{3.9}$$

and it can be estimated using maximum likelihood (further assuming a Gaussian model for the variance components, and then jointly estimate $\sigma^2_{\mathcal{S}}$, $\sigma^2_{\mathcal{S}^c}$ and σ^2_e (Yang et al., 2011a,b; Davis et al., 2013; Gusev et al., 2014)). However, if the hypothesis of causal loci holds in the total heritability LMM, then in this two variance components LMM, effect-sizes should really follow

$$u_i \sim \begin{cases} \mathcal{MV}\left(0, \frac{\sigma^2_{\mathcal{S}}}{|\mathcal{A}_1|}\right), & \text{if } i \in \mathcal{A}_1 \subseteq \mathcal{S}, \\ \mathcal{MV}\left(0, \frac{\sigma^2_{\mathcal{S}^c}}{|\mathcal{A}_2|}\right), & \text{if } i \in \mathcal{A}_2 \subseteq S^c, \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathcal{A}_1$ and $\mathcal{A}_2$ are the sets of causal loci in $\mathcal{S}$ and $\mathcal{S}^c$, with $\mathcal{A}_1 \cap \mathcal{A}_2 = \emptyset$. If the causal loci are concentrated in a high/low LD region, then similar bias observed in total heritability estimation is expected for partitioned heritability estimation with the landmark estimator restricted maximum likelihood (REML) approach with linear GRM (Yang et al., 2011a). The simulation results regarding the bias is discussed in Section 3.5.2, after introducing the Mahalanobis estimator for partitioned heritability.

With existence of non-exchangeable genetic effects, the current definition of partitioned heritability (3.9) is invalid because for extreme Σ and $\mathbf{u}$, it is possible to have $h^2_{\mathcal{S}}(\Sigma, \mathbf{u}) > h^2$. This phenomenon is unreasonable because contribution of any subset in explained variation should not exceed the contribution of the universal set. Provided Σ and fixed $\mathbf{u}$, we define the heritability attributable to $\mathcal{S}$ be

$$h^2_{\mathcal{S}} = 1 - \frac{\operatorname{Var}(y \mid \mathbf{z}_{\mathcal{S}})}{\operatorname{Var}(y)} = \frac{\mathbf{u}^\top \Sigma \mathbf{u} - \mathbf{u}_{\mathcal{S}^c}^\top \Sigma_{\mathcal{S}^c|\mathcal{S}} \mathbf{u}_{\mathcal{S}^c}}{\mathbf{u}^\top \Sigma \mathbf{u} + \sigma_e^2}.$$

This definition is a natural consequence of several reasonable properties of partitioned heritability, and we defer the formal discussion of the definition to Section 3.4.2. However, it is worth noting that the modified partitioned heritability definition may be different from the estimand defined in (3.9) even for iid random effects in $\mathcal{S}$ and $\mathcal{S}^c$.

Due to difference in estimands, estimators designed for (3.9) are potentially biased in estimating $h^2_{\mathcal{S}}$ when LD exists. Let $\operatorname{Cov}(\mathbf{z}_i) = \Sigma$, for $\mathcal{S} \subseteq \{1, \ldots, m\}$, let $\Sigma_{\mathcal{S},\mathcal{S}^c}$ be the submatrix of Σ with rows and columns selected according to $\mathcal{S}, \mathcal{S}^c \subseteq \{1, \ldots, m\}$, respectively. Linear kernel-based heritability estimation measures $\sigma^2_{\mathcal{S}}$ (assuming $\operatorname{Var}(y) = 1$), however, it automatically ignores the interaction between SNPs in group $\mathcal{S}$ and SNPs in $\mathcal{S}^c$ through $\Sigma_{\mathcal{S},\mathcal{S}^c}$. Thus, standard linear kernel approach to $h^2_{\mathcal{S}}$ create large bias. The numerical results regarding the bias is shown in Section 3.5.2.

3.4 Fixed-effects models and C-heritability

3.4.1 Fixed-effects heritability

The previous section illustrates some LD-related pitfalls that may arise in LMM-based approaches to heritability estimation. In this section we propose an alternative approach:

we consider a fixed effects model with Gaussian data. Based on the fixed effects model, we propose new definitions of heritability and partitioned heritability, which lean more heavily on the concept of broad-sense heritability. We also show that these concepts coincide with the random-effects approach under a Mahalanobis GRM or kernel.

In this section, we assume that the linear model (3.5) holds with some fixed (non-random) $\mathbf{u}$. We additionally assume that

$$\mathbf{z}_1, \ldots, \mathbf{z}_n \sim \mathcal{N}(0, \Sigma) \text{ and } e_1, \ldots, e_n \sim \mathcal{N}(0, \sigma_e^2) \tag{3.10}$$

are independent. These normality assumptions are unrealistic in practice (the entries of $\mathbf{z}_i$ are typically discrete). However, instead of taking these assumptions literally, we rely on them for motivation for the methods proposed in this section. Many other high-dimensional variance component estimation with fixed effects model (e.g. Dicker, 2014; Janson et al., 2017) require the same multivariate Gaussian random-design (3.10), for its invariance property under orthogonal transformations. Work of Bai et al. (2007) in random matrix theory has shown that in the large limit where $n, m \to \infty$, the invariance property holds for a broader class of random matrices. We expect our estimator to be robust asymptotically for reasonable trinary random designs, similar to simulation results in (Janson et al., 2017). Theoretical results on relaxing the Gaussian random design assumptions for the Mahalanobis estimator (by building on results in (Dicker and Erdogdu, 2016b)) would be an interesting future research direction.

Let $(y, \mathbf{z})$ be a generic draw from the study population. We define the (fixed-effects) heritability to be

$$h^2 = 1 - \frac{\mathrm{Var}(y \mid \mathbf{z})}{\mathrm{Var}(y)} = \frac{\mathbf{u}^\top \Sigma \mathbf{u}}{\mathbf{u}^\top \Sigma \mathbf{u} + \sigma_e^2}. \tag{3.11}$$

The fixed-effects heritability (3.11) captures correlation between SNPs and LD through the quadratic form $\mathbf{u}^\top \Sigma \mathbf{u}$. We also have the following bound for fixed-effects heritability:

$$\frac{\lambda_m \|\mathbf{u}\|^2}{\lambda_m \|\mathbf{u}\|^2 + \sigma_e^2} \leq h^2 \leq \frac{\lambda_1 \|\mathbf{u}\|^2}{\lambda_1 \|\mathbf{u}\|^2 + \sigma_e^2},$$

where λ_1 and λ_m are the largest and smallest eigenvalues of Σ, respectively.

For the sake of simplicity we assume that $\mathrm{Var}(\mathbf{y}) = 1$. Consider the case where $\mathrm{Cov}(\mathbf{z}_i) = \Sigma$ where diagonal of Σ are all 1. Then (3.11) becomes $h^2_{fe} = \mathrm{Var}(\mathbf{z}^\top \mathbf{u} | \mathbf{u}) =$

$\mathbf{u}^\top \Sigma \mathbf{u}$. The normalized linear kernel is equivalent to a random-effects assumption such that effects in $\mathbf{u}$ are iid with $\mathrm{Var}(\mathbf{u}) = \sigma_g^2/mI$, then its corresponding random-effects heritability is $h_{lk}^2 = \mathrm{Var}(\mathbf{z}^\top \mathbf{u}|\mathbf{z}) = \sigma_g^2/m\mathbb{E}(\mathbf{z}^\top \mathbf{z}) = \sigma_g^2$. When $\Sigma = I$, $h_{fe}^2 = \|\mathbf{u}\|_2^2$. then fixed- and random-effects heritabilities are equivalent in the sense that $\mathbb{E}_u(h_{fe}^2) = h_{lk}^2$. However, when $\Sigma \neq I$, $h_{fe}^2 = \|\mathbf{u}\|_2^2 + 2\sum_{i<j} u_i u_j \Sigma_{ij}$, and the linear kernel-based h_{lk}^2 could not pick up the cross-terms.

With Mahalanobis kernel, it is equivalent to assume that effects in $\Sigma^{1/2}\mathbf{u}$ are iid centered at 0 and with $\mathrm{Var}(\Sigma^{1/2}\mathbf{u}) = \tau^2/m$, then the Mahalanobis-based random-effects heritability is

$$h_{mk}^2 = \mathrm{Var}(\mathbf{z}^\top \mathbf{u}|\mathbf{z}) = \mathrm{Var}(\mathbf{z}^\top \Sigma^{-1/2}\Sigma^{1/2}\mathbf{u}|\mathbf{z}) = \tau^2/m\mathbb{E}(\mathbf{z}^\top \Sigma^{-1}\mathbf{z}) = \tau^2.$$

Hence, for any $\Sigma > 0$, random-effects heritability is similar to fixed-effects heritability such that $\mathbb{E}_u(h_{fe}^2) = h_{mk}^2$.

3.4.2 Partitioning heritability

Broad-sense heritability and the fixed-effects linear model described in Section 3.4.1 also motivate a natural definition of partitioned heritability. For $\mathcal{S} \subseteq \{1, \ldots, m\}$, we define the heritability attributable to $\mathcal{S}$ to be

$$h_{\mathcal{S}}^2 = 1 - \frac{\mathrm{Var}(y \mid \mathbf{z}_{\mathcal{S}})}{\mathrm{Var}(y)} = \frac{\mathbf{u}^\top \Sigma \mathbf{u} - \mathbf{u}_{\mathcal{S}^c}^\top \Sigma_{\mathcal{S}^c|\mathcal{S}} \mathbf{u}_{\mathcal{S}^c}}{\mathbf{u}^\top \Sigma \mathbf{u} + \sigma_e^2}, \tag{3.12}$$

where $\Sigma_{\mathcal{S}^c|\mathcal{S}} = \Sigma_{\mathcal{S}^c,\mathcal{S}^c} - \Sigma_{\mathcal{S}^c,\mathcal{S}}\Sigma_{\mathcal{S},\mathcal{S}}^{-1}\Sigma_{\mathcal{S},\mathcal{S}^c}$ and $\Sigma_{\mathcal{S}_1,\mathcal{S}_2}$ is the submatrix of Σ with rows and columns selected according to $\mathcal{S}_1, \mathcal{S}_2 \subseteq \{1, \ldots, m\}$, respectively. This definition for parititoned heritability consistently accounts for correlation between LD and SNPs.

The definition (3.12) makes sense in the context of broad-sense heritability, and when the genetic features are Gaussian (or approximately Gaussian). As discussed in Section 3.4.1, the Gaussian assumption almost never holds in practice. In the following proposition, we argue that (3.12) is also a natural consequence of three reasonable properties we might expect of any quadratic form-based estimator for partitioned heritabilty for linear models.

Proposition 1. *Assume that the linear model* (3.5) *holds, that* $\mathbf{u} \in \mathbb{R}^m$ *is a fixed vector and that* $\mathbb{E}(\mathbf{z}) = 0$, $\mathrm{Var}(\mathbf{z}) = \Sigma$*. If the heritability attributable to* $\mathcal{S}$, $h^2_{\mathcal{S}} = h^2_{\mathcal{S}}(\mathbf{u}; \Sigma)$*, is a quadratic form in* $\mathbf{u}$ *satisfying the following properties*

(i) $0 \leq h^2_{\mathcal{S}}(\mathbf{u}; \Sigma) \leq h^2(\mathbf{u}; \Sigma)$ *for all* $\mathbf{u} \in \mathbb{R}^m$*, where* $h^2 = h^2(\mathbf{u}; \Sigma)$ *is the fixed-effects heritability* (3.11)*,*

(ii) $h^2_{\mathcal{S}}(\mathbf{u}; \Sigma) = h^2(\mathbf{u}; \Sigma)$ *if and only if* $\mathbf{u}_{\mathcal{S}^c} = 0$*, and*

(iii) $h^2_{\mathcal{S}}(u; \Sigma)$ *does not depend on* $\Sigma_{\mathcal{S}^c, \mathcal{S}^c}$*,*

then we must have

$$h^2_{\mathcal{S}} = \frac{\mathbf{u}^\top \Sigma \mathbf{u} - \mathbf{u}^\top_{\mathcal{S}^c} \Sigma_{\mathcal{S}^c|\mathcal{S}} \mathbf{u}_{\mathcal{S}^c}}{\mathbf{u}^\top \Sigma \mathbf{u} + \sigma^2_e}.$$

Proposition 1 is proved in Section 3.6. Condition (i) in Proposition 1 says that the heritability attributable to a subset of SNPs $\mathcal{S}$ must be smaller than the total heritability (i.e. the heritability attributable to all measured SNPs); condition (ii) means that the heritability attributable to $\mathcal{S}$ is equal to the total heritability if and only if all causal loci are contained in $\mathcal{S}$; condition (iii) means that the heritability attributable to $\mathcal{S}$ should not depend on LD amongst SNPs that are not $\mathcal{S}$ (though it certainly may depend on LP between SNPs in $\mathcal{S}$ and those not in $\mathcal{S}$). We defer discussion of how to estimate $h^2_{\mathcal{S}}$ until the following sub-section.

3.4.3 C-heritability with projections

In addition to focusing on the heritability attributable to a subset of SNPs $\mathcal{S}$ with partitioned heritability, we can extend the definition of heritability to variation explained by any linear projection $C^\top \mathbf{z}$, for $m \times k$ matrices C with rank k:

$$h^2_C = h^2_C(\mathbf{u}; \Sigma) = 1 - \frac{\mathrm{Var}(y \mid C^\top \mathbf{z})}{\mathrm{Var}(y)}.$$

Under the linear model with Gaussian data (3.10) and the additional assumption that $\mathrm{Var}(y) = 1$, we have

$$h^2_C = \mathbf{u}^\top \Sigma C (C^\top \Sigma C)^{-1} C^\top \Sigma \mathbf{u}. \tag{3.13}$$

The following lemma summarizes some useful facts about h^2_C.

Lemma 1. *Assume* (3.5) *and* (3.10) *and that* $\mathrm{Var}(y) = 1$. *Then*

$$h^2_C(\mathbf{u}; I) = \mathbf{u}^\top C(C^\top C)^{-1}C^\top \mathbf{u} \tag{3.14}$$

and

$$h^2_C(\mathbf{u}; \Sigma) = h^2_{\Sigma^{1/2}C}(\Sigma^{1/2}\mathbf{u}; I). \tag{3.15}$$

If, furthermore, $m = k$, *then*

$$h^2_C(\mathbf{u}; \Sigma) = \mathbf{u}^\top \Sigma \mathbf{u}. \tag{3.16}$$

The proof of Lemma 1 is trivial. The second identity in the lemma (3.15) helps to explain the connection between LD and heritability – it implies that heritability in a model with LD structure Σ is equivalent to heritability in a model where LD has been removed through a whitening transformation $\mathbf{z} \mapsto \Sigma^{-1/2}\mathbf{z}$. The third identity (3.16) implies that C-heritability is invariant under (full rank) change-of-basis for the genotype $\mathbf{z} \mapsto C^{-1}\mathbf{z}$.

The projected C-heritability (3.4.3) is equivalent to partitioned heritability from the previous section.

Lemma 2. *Assume* (3.5) *and* (3.10) *and that* $\mathrm{Var}(y) = 1$. *Let* $\mathcal{S} \subseteq \{1, \ldots, m\}$ *and let* $\Pi_{\mathcal{S}}$ *be the projection matrix onto coordinates indexed by* $\mathcal{S}$. *Then* $h^2_{\mathcal{S}} = h^2_{\Pi_{\mathcal{S}}}$.

The proof of Lemma 2 is trivial. However, the lemma is useful because it provides a direct method for estimating $h^2_{\mathcal{S}}$ when combined with the results in the next sub-section.

3.4.4 Estimating C-heritability

In this section, we assume fixed-effects linear model and suppose $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$. We would like to estimate $h^2_C(\mathbf{u}; \Sigma)$, for some full rank matrix $C \in \mathbb{R}^{m \times k}$.

Let U_C be a $m \times k$ matrix with orthonormal columns such that $\Sigma^{1/2}C(C^\top \Sigma C)^{-1}C^\top \Sigma^{1/2} = U_C U_C^\top$. Let $U_{C^\perp}$ be a corresponding $m \times (m - k)$ matrix with orthonormal columns satisyfing $U_C^\top U_{C^\perp} = 0$ and $I = U_C U_C^\top + U_{C^\perp} U_{C^\perp}^\top$. Then

$$\begin{aligned}
\mathbf{y} &= Z\mathbf{u} + \boldsymbol{e} \\
&= Z\Sigma^{-1/2} U_C U_C^\top \Sigma^{1/2}\mathbf{u} + Z\Sigma^{-1/2} U_{C^\perp} U_{C^\perp}^\top \Sigma^{1/2}\mathbf{u} + \boldsymbol{e} \\
&= W_C \mathbf{v}_C + W_{C^\perp}\mathbf{v}_{C^\perp} + \boldsymbol{e} \\
&= Z_C \mathbf{v}_C + \boldsymbol{e}_C, \qquad (3.17)
\end{aligned}$$

where

$$\begin{aligned}
&W_C = Z\Sigma^{-1/2} U_C = ZC(C^\top \Sigma C)^{-1/2}, \quad W_{C^\perp} = Z\Sigma^{-1/2} U_{C^\perp}, \\
&\mathbf{v}_C = U_C^\top \Sigma^{1/2}\mathbf{u}, \quad \mathbf{v}_{C^\perp} = U_{C^\perp}^\top \Sigma^{1/2}\mathbf{u}, \\
&\boldsymbol{e}_C = Z_{C^\perp}\mathbf{v}_{C^\perp} + \boldsymbol{e}.
\end{aligned}$$

Thus, we've transformed the original linear model with data $(\mathbf{y}, Z)$ into the linear model (3.17) with data $(\mathbf{y}, W_C)$, where

$$W_C \sim \mathcal{N}(0, I), \quad \boldsymbol{e}_C \sim \mathcal{N}\left(0, (\|\mathbf{v}_{C^\perp}\|^2 + \sigma_e^2) I\right). \qquad (3.18)$$

Moreover, h^2 for the model (3.17) is equivalent to the C-heritability $h^2(\mathbf{u}; \Sigma)$ for the original linear model. Thus, to estimate $h^2(\mathbf{u}; \Sigma)$, we simply esimate h^2 under (3.17).

Let $\sigma_{C^\perp}^2 = \|\mathbf{v}_{C^\perp}\|^2 + \sigma_e^2$ and $\tau_C^2 = \|\mathbf{v}_C\|^2$. Consequently, we can estimate τ_C^2 and σ_C^2 with Gaussian maximum likelihood, for a random-effects model where $\mathbf{v}_C \sim \mathcal{N}\{0, (\tau_C^2/k)I\}$ and independent of $\boldsymbol{e}_C$ and W_C (Dicker and Erdogdu, 2016b). Let $\eta_C^2 = \tau_C^2/\sigma_{C^\perp}^2$, then our proposed fixed-effects heritability estimator is

$$\hat{h}_C^2 = \frac{\hat{\eta}_C^2}{1 + \hat{\eta}_C^2},$$

where

$$\begin{aligned}
(\hat{\eta}_C^2, \hat{\sigma}_{C^\perp}^2) := \underset{\eta_C^2,\ \sigma_{C^\perp}^2}{\operatorname{argmax}} \qquad & -\frac{1}{2}\log(\sigma_{C^\perp}^2) - \frac{1}{2n}\log\det\left(\eta^2/kW_C W_C^\top + I\right) \\
& -\frac{1}{2n\sigma_{C^\perp}^2}\mathbf{y}^\top \left(\eta^2/kW_C W_C^\top + I\right)^{-1}\mathbf{y}.
\end{aligned}$$

Despite $\hat{h}_C^2$ is motivated by maximizing likelihood of multivariate Gaussian variable $\mathbf{y} \mid W_C$, it is consistent in estimating fixed $\mathbf{u} \in \mathbb{R}^m$ under the assumption that $k/n \to \rho \in (0, \infty)/\{1\}$.

Proposition 2. *Assume (3.17)-(3.18) hold, and suppose $(\sigma_C^2, \eta_C^2) \in \mathcal{K}$ for some compact set $\mathcal{K} \subseteq (0, \infty)$ and $\rho \in (0, \infty)/\{1\}$, then as $k/n \to \rho$,*

$$\hat{h}_C^2 = \frac{\hat{\eta}_C}{\hat{\eta}_C + 1} \to h_C^2$$

in probability. Moreover, define $\mathcal{I} = \frac{1}{k} W_C W_C^\top$ and $\mathcal{J} = \frac{\eta_C^2}{k} W_C W_C^\top + I$, then

$$\sqrt{n}(\hat{h}_C^2 - h_C^2) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \quad \frac{2\sigma_{C^\perp}^4}{(1+\eta_C^2)^4}\left(1 - \frac{\mathrm{tr}(\mathcal{I}\mathcal{J}^{-1})^2}{n\mathrm{tr}(\mathcal{I}^2\mathcal{J}^{-2})}\right)^{-1}\right)^{-1}.$$

Consistency of $\hat{h}_C^2$ is an immediate result of Theorem 1 from (Dicker and Erdogdu, 2016b) and Slutsky's theorem. In addition asymptotic normality of $\hat{h}_C^2$ can be easily derived by Theorem 2 of (Dicker and Erdogdu, 2016b) and the delta method. The asymptotic variance is derived in Section 3.6 .

This estimator performs well for the fixed effects model, however, under Mahalanobis kernel, it is also the standard Gaussian variance component maximum likelihood approach for random-effects model such that the vector of effects follows $\mathbf{v}_C \sim \mathcal{N}(0, \tau_C^2/kI)$. Work of Dicker and Erdogdu (2016a,b) has shown that the MLE approach is reliable not only for Gaussian iid effects, but also for correlated random effects and even fixed effects by a coupling argument and a concentration bound for the MLE. Hence, our proposed estimator is also very flexible in terms of model specification. As previously mentioned, the least squares approach is sometimes preferred due to data privacy. Building on (3.17), method of moments approach to C-heritability estimation is also possible. We refer the reader to (Dicker, 2014) for details.

3.5 Numerical experiments

In this section, we run comprehensive numerical experiments on total and partitioned heritability estimation. Performance of total heritability estimators are compared while varying LD-level and sparsity of causal variants. Performance of partitioned heritability estimators are compared when either LD exists between partitioned sets of SNPs or LD-level of causal variants is uneven.

3.5.1 Total heritability estimation

For total heritability estimation, we consider the following model

$$\mathbf{y} \sim \mathcal{MV}\left(0, \frac{\sigma_g^2}{|\mathcal{S}|} Z_{\mathcal{S}} Z_{\mathcal{S}}^\top + \sigma_e^2 I\right)$$

where $e_j \sim \mathcal{N}(0, \sigma_e^2)$ and $u_j \sim \mathcal{N}(0, \sigma_g^2/|\mathcal{S}|)$. In the absence of additional information about $\mathcal{S}$, estimators for h^2 are usually fit according to the linear model (3.5). The simulation setting is very similar to that in Section 3.3.1, such that:

(i) $n = 500$, $m = 1000$.

(ii) $\sigma_e^2 = 1 - \sigma_g^2$.

(iii) $\mathbf{z}_1, \ldots, \mathbf{z}_n \sim \mathcal{N}(0, \Sigma)$, where

$$\Sigma = \begin{pmatrix} \text{AR}(0.3) & 0 \\ 0 & \text{AR}(0.7) \end{pmatrix}$$

and AR(ρ) is the $m/2 \times m/2$ matrix with ij-entry $\rho^{|i-j|}$.

Let $\mathcal{R}_l = \{1, \ldots, m/2\}$ be the low LD region and $\mathcal{R}_h = \{m/2+1, \ldots, m\}$ be the high LD region. We vary location of causal variants and $\sigma_g^2 = 0.3, 0.5, 0.7$ for experiment $d = 1, \ldots, 50$, such that $u_j^{(d)} \sim \mathcal{N}(0, \sigma_g^2/|\mathcal{S}|)$ for $j \in \mathcal{S}$ with

(i) $\mathcal{S} = \mathcal{R}_h \cup \mathcal{R}_l$;

(ii) $\mathcal{S} = \mathcal{R}_h$;

(iii) $\mathcal{S} = \mathcal{R}_l$.

In each of these senarios, we simulated 50 independent datasets specified according to this model, and for each dataset we compute the Mahalanobis estimator, maximum likelihood estimator with linear kernel proposed by (Yang et al., 2010, 2011a). We also include LD-adjusted kinship (LDAK) approach proposed by (Speed et al., 2012), which is designed to improve the total heritability estimation performance of linear MLE when uneven LD structure exists. For LDAK, the Linear GRM is adjusted by re-weighing each predictor, and the modified REML method takes new inputs $\mathbf{y}$ and LD-adjusted X.

Distributional summary statistics are reported in Figure 3.1.

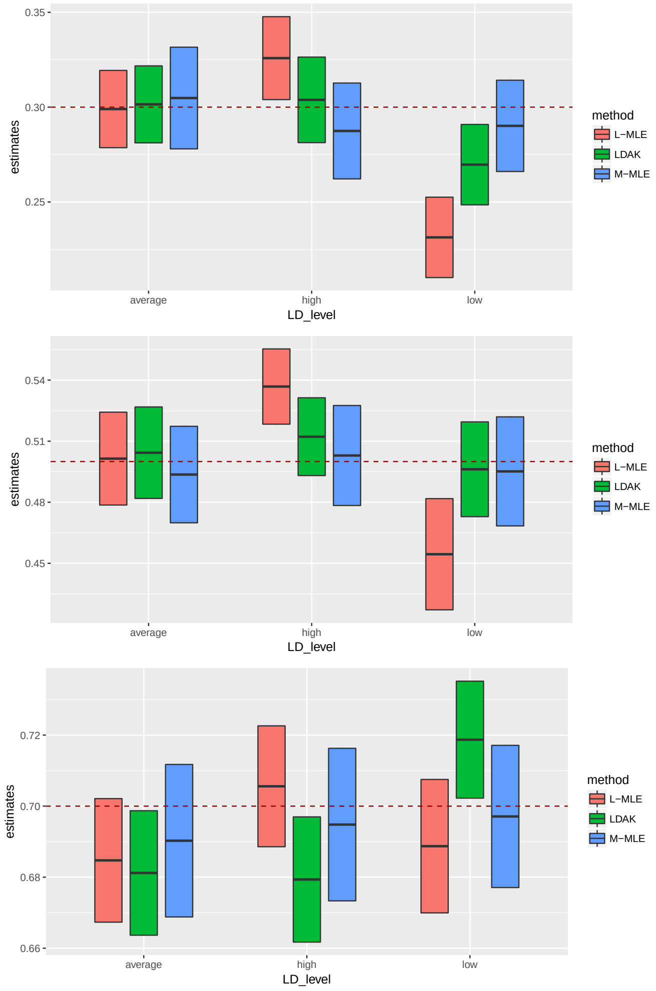

Figure 3.1: Confidence intervals of linear kernel-based maximum likelihood estimator (L-MLE), MLE with LD adjusted linear GRM (LDAK) and Mahalanobis kernel-based MLE (M-MLE) with causal variants from different LD-level regions. Underlying h^2 in panels from top to bottom are respectively $0.3, 0.5, 0.7$ and marked in red dashed line.

In Figure 3.1, maximum likelihood estimator with linear kernel is generally biased when causal effects are generated from high or low LD regions. The estimator does not show obvious bias in either direction when all SNPs are causal. LDAK attempts to adjust the biased linear MLE towards the underlying h^2. It shows some improvement in adjusting the biased linear MLE when uneven LD structure exists. However, when the bias of linear MLE is small (when $h^2 = 0.7$), the LDAK over-adjusts the linear MLE and becomes badly biased. With Mahalanobis GRM, the MLE is not biased when LD is coupled with causal loci. Since all three methods belong to the same maximum likelihood family, widths of their confidence intervals are very similar.

In the next experiment, we vary sparsity of effects. We let $n = 500$, $m = 1000$, $\sigma_g^2 = \sigma_e^2 = 0.5$ and $\mathbf{z}_1, \ldots, \mathbf{z}_n$ follow the same Gaussian distribution in the previous experiment setting. Given $|\mathcal{S}|$, we let $u_j \sim \mathcal{N}(0, \sigma_g^2/|\mathcal{S}|)$ for $j \in \mathcal{S}$ with elements in $\mathcal{S}$ uniformly sampled without replacement from region $\mathcal{A}$. For $|\mathcal{S}| = 25, 50, 100, 200$, we vary $\mathcal{A}$ as follows

(i) $\mathcal{A} = \mathcal{R}_h \cup \mathcal{R}_l$;

(ii) $\mathcal{A} = \mathcal{R}_h$;

(iii) $\mathcal{A} = \mathcal{R}_l$.

For each setting above, we simulate 50 independent datasets with common effect-sizes according to this model. For each dataset, we compute the same three estimators for the fixed-effects heritability.

In Figure 3.2, the underlying h^2 is a piece-wise constant function because of the fixed effects setting. The experiments show that the linear MLE is less stable when effect-sizes become more sparse (consistent with results in (Speed et al., 2012)). As number of causal variants decreases, it is more likely that the collection of causal variants is enriched in either high or low LD region. As a result, the top panel shows that there is obvious downward bias even when effects are randomly sampled from average LD region (indicates that the majority of causal effects are in low LD region). For all sparsity settings, linear kernel-based MLE is biased in the direction corresponding to the LD-level. Moreover, we observe that LDAK adjusts the bias of linear MLE towards the underlying h^2, however, it under-adjusts the estimate, and remains biased in panel 1 and 4. Regardless of sparsity, the Mahalanobis estimator is free of LD issues in estimating fixed-effects heritability.

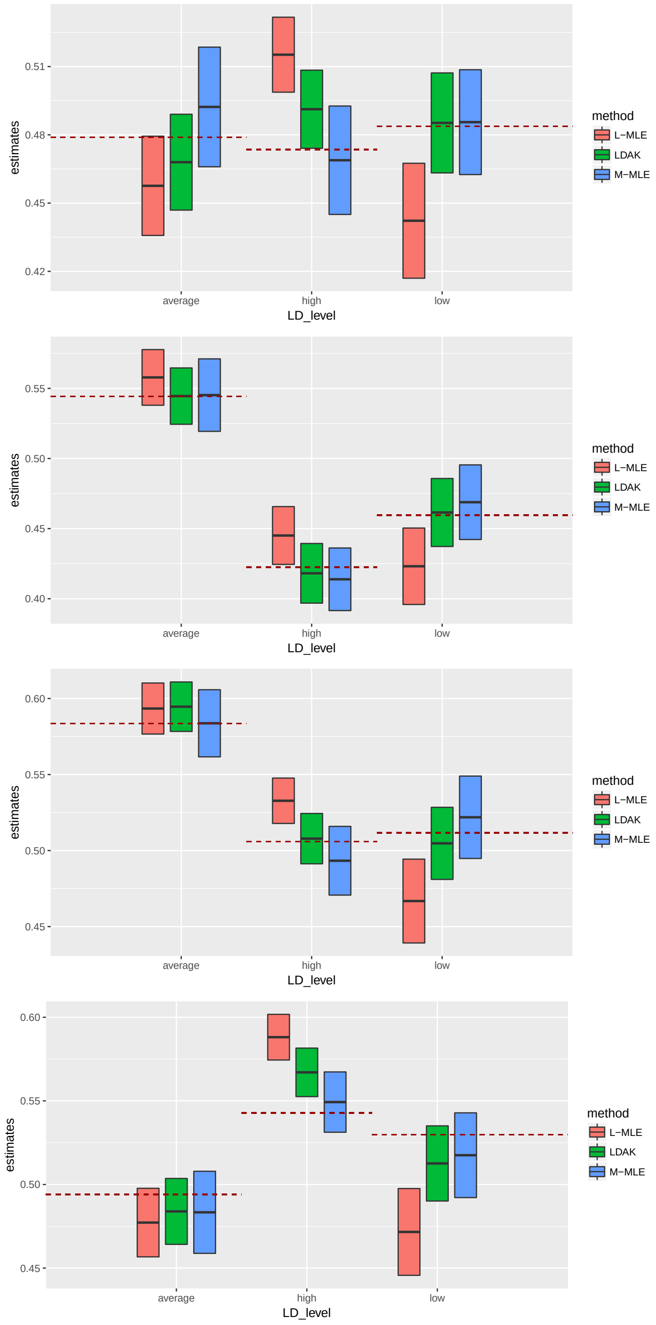

Figure 3.2: Confidence intervals of maximum likelihood estimator (L-MLE), MLE with LD adjusted linear GRM (LDAK) and Mahalanobis kernel-based MLE (M-MLE) with causal variants from different LD-level regions. Underlying h^2 is marked in red dashed line. Effect sparsities from top panel to bottom are $25, 50, 100, 200$, respectively.

3.5.2 Partitioned heritability estimation

Consider the two variance components linear model

$$\mathbf{y} = Z_{\mathcal{S}}\mathbf{u}_{\mathcal{S}} + Z_{\mathcal{S}^c}\mathbf{u}_{\mathcal{S}^c} + \boldsymbol{e}, \tag{3.19}$$

where

$$u_i \sim \begin{cases} \mathcal{N}\left(0, \frac{\sigma^2_{\mathcal{S}}}{|\mathcal{A}_1|}\right), & \text{if } i \in \mathcal{A}_1 \subseteq \mathcal{S}, \\ \mathcal{N}\left(0, \frac{\sigma^2_{\mathcal{S}^c}}{|\mathcal{A}_2|}\right), & \text{if } i \in \mathcal{A}_2 \subseteq S^c, \\ 0, & \text{otherwise}, \end{cases}$$

provided $e_j \sim \mathcal{N}(0, \sigma_e^2)$ and $\sigma_e^2 = 0.5$. We would like to estimate h^2 associated with $\mathcal{S}$.

The first experiment considers the setting where all variants are causal and LD exists between two partitions of SNPs. Specifically, we let

(i) $n = 500$, $m = 1000$.

(ii) $\mathcal{S} = \{i \in [m]; 1 \equiv i(\text{mod}4)\}$.

(iii) $\mathcal{A}_1 = \mathcal{S}$ and $\mathcal{A}_2 = \mathcal{S}^c$.

(iv) $\mathbf{z}_1, \ldots, \mathbf{z}_n \sim \mathcal{N}(0, \Sigma)$ where

$$\Sigma = \begin{pmatrix} \text{AR}(0.3) & 0 \\ 0 & \text{AR}(0.7) \end{pmatrix}$$

and AR(ρ) is the $m/2 \times m/2$ matrix with ij-entry $\rho^{|i-j|}$.

We vary size of $\sigma^2_{\mathcal{S}} = 0.1, 0.3, 0.5$. For each setting, we simulated 50 independent datasets specified according to this model, and for each dataset we compute the Mahalanobis estimator and restricted maximum likelihood with linear kernel (Gilmour et al., 1995; Yang et al., 2011a). REML finds the maximum likelihood estimator for two variance components linear model. Distributional summary statistics are reported in Figure 3.3.

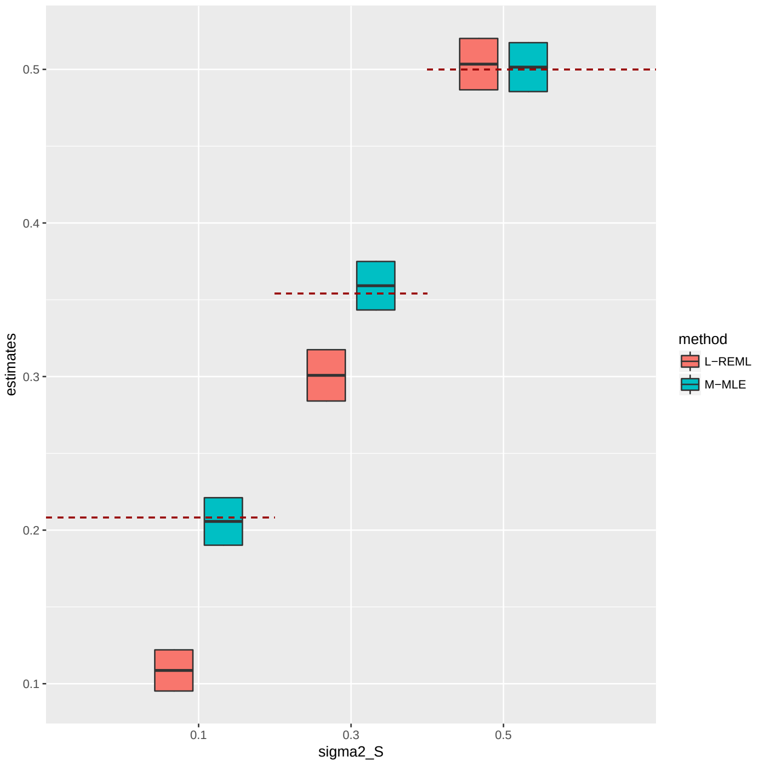

Figure 3.3: Confidence intervals of linear REML (L-REML) and Mahalanobis kernel-based MLE (M-MLE) with different signal strength in $\mathcal{S}$ ($\sigma^2_{\mathcal{S}} = 0.1, 0.3, 0.5$) when partitions of SNPs are in high LD. Underlying $h^2_{\mathcal{S}}$ is marked in red dashed line.

Since elements in $\mathcal{S}$ is chosen such that they are spreaded out, the LD between two partitions of SNPs is very significant. It is shown in Figure 3.3 that the linear kernel-based MLE is consistent in estimating $\sigma^2_{\mathcal{S}}$, however, it has downward bias in estimating $h^2_{\mathcal{S}}$ when $\sigma^2_{\mathcal{S}} < 0.5$; the bias is due to difference in estimand, and estimand of linear REML $\sigma^2_{\mathcal{S}}$ is less than $h^2_{\mathcal{S}}$ when $\sigma^2_{\mathcal{S}^c} \neq 0$. When $\sigma^2_{\mathcal{S}} = 0.5$, $h^2_{\mathcal{S}} = \sigma^2_{\mathcal{S}}$ because all causal loci are contained in $\mathcal{S}$. As $\sigma^2_{\mathcal{S}}$ decreases, the downward bias of linear REML becomes larger due to larger difference between $\sigma^2_{\mathcal{S}}$ and $h^2_{\mathcal{S}}$. The Mahalanobis estimator is consistent in estimating $h^2_{\mathcal{S}}$ as it considered the LD between two partitions. Due to similar nature, the confidence intervals of two methods are roughly the same.

In the next experiment, causal loci are selected to be from uneven LD regions. We let

(i) $n = 500$, $m = 1000$.

(ii) $\mathcal{S} = \{m/4+1, \dots, 3m/4\}$.

(iii) $\mathcal{A}_1 = \{m/4+1, m/2\}$.

(iv) $\mathbf{z}_1, \dots, \mathbf{z}_n \sim \mathcal{N}(0, \Sigma)$ where

$$\Sigma = \begin{pmatrix} \mathrm{AR}(0.3) & 0 & 0 & 0 \\ 0 & \mathrm{AR}(0.3) & 0 & 0 \\ 0 & 0 & \mathrm{AR}(0.7) & 0 \\ 0 & 0 & 0 & \mathrm{AR}(0.7) \end{pmatrix}$$

and $\mathrm{AR}(\rho)$ is the $m/4 \times m/4$ matrix with ij-entry $\rho^{|i-j|}$.

Denote $\mathcal{R}_l = \{1, \dots, m/4\}$ and $\mathcal{R}_h = \{3m/4+1, \dots, m\}$ be the set of indices in low and high LD regions, respectively. We vary locations of $\mathcal{A}_2$, such that

(i) $\mathcal{A}_2 = \mathcal{R}_l \cup \mathcal{R}_h$;

(ii) $\mathcal{A}_2 = \mathcal{R}_h$;

(iii) $\mathcal{A}_2 = \mathcal{R}_l$.

For each setting, we simulated 200 random-effects vectors and independent datasets, and for each dataset we compute the Mahalanobis estimator and REML with linear kernel. Distributional summary statistics are reported in Figure 3.4.

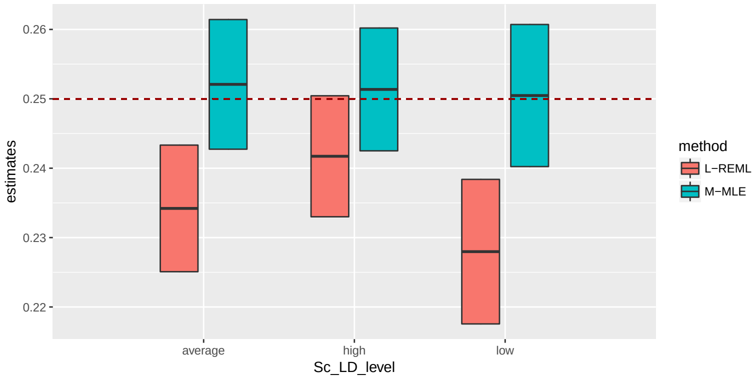

Figure 3.4: Confidence intervals of linear REML (L-REML) and Mahalanobis kernel-based MLE (M-MLE) with causal loci of partitions from uneven LD regions. Underlying $h^2_{\mathcal{S}}$ is marked in red dashed line. Causal loci $\mathcal{S}$ is from low-LD region while LD-level of causal loci in $\mathcal{S}^c$ varies.

In all three cases, the Mahalanobis estimator is consistent in estimating

$$h^2_{\mathcal{S}} = \frac{\sigma^2_{\mathcal{S}}}{\sigma^2_{\mathcal{S}} + \sigma^2_{\mathcal{S}^c} + \sigma^2_e} = 0.25,$$

while the linear REML shows downward bias (as expected). The objective of increasing the number of experiments is to shorten the confidence intervals in three cases and show the subtle difference of REML with linear GRM when LD-level of causal loci in $\mathcal{S}^c$ varies.

In the first setting, causal loci in $\mathcal{S}$ are from low LD region, while causal loci in $\mathcal{S}^c$ are from average LD region. When causal loci in $\mathcal{S}^c$ are instead from high-LD area, estimates of $h^2_{\mathcal{S}}$ increases surprisingly. Notice that linear REML's estimated $\hat{h}^2_{\mathcal{S}}$ is derived by plugging in estimated $\hat{\sigma}^2_{\mathcal{S}}, \hat{\sigma}^2_{\mathcal{S}^c}, \hat{\sigma}^2_e$. A possible explanation is that in the first setting, the overall LD-level of all causal variants is below average, and σ^2_e is over-estimated. As LD-level in $\mathcal{S}^c$ increases, the overall LD-level of all causal variants becomes even, therefore, upward bias in $\hat{\sigma}^2_e$ is reduced. Although $\hat{\sigma}^2_{\mathcal{S}^c}$ becomes upward biased because of uneven LD in $\mathcal{S}^c$, due to magnitude difference between σ^2_e and $\sigma^2_{\mathcal{S}^c}$, the decrease in $\hat{\sigma}^2_e$ outweigh the increase in $\hat{\sigma}_{\mathcal{S}^c}$. Thus, the overall $\hat{h}^2_{\mathcal{S}}$ shows less downward bias. When causal loci is from low LD region, similar reasoning explains why $\hat{h}^2_{\mathcal{S}}$ performs worse than the first setting.

3.6 Appendix

3.6.1 Proof of Proposition 1

Proposition 1. *Assume that the linear model* (3.5) *holds, that* $\mathbf{u} \in \mathbb{R}^m$ *is a fixed vector and that* $\mathbb{E}(\mathbf{z}) = 0$, $\mathrm{Var}(\mathbf{z}) = \Sigma$. *If the heritability attributable to* $\mathcal{S}$, $h^2_{\mathcal{S}} = h^2_{\mathcal{S}}(\mathbf{u}; \Sigma)$, *is a quadratic form in* $\mathbf{u}$ *satisfying the following properties*

(i) $0 \le h^2_{\mathcal{S}}(\mathbf{u}; \Sigma) \le h^2(\mathbf{u}; \Sigma)$ *for all* $\mathbf{u} \in \mathbb{R}^m$, *where* $h^2 = h^2(\mathbf{u}; \Sigma)$ *is the fixed-effects heritability* (3.11),

(ii) $h^2_{\mathcal{S}}(\mathbf{u}; \Sigma) = h^2(\mathbf{u}; \Sigma)$ *if and only if* $\mathbf{u}_{\mathcal{S}^c} = 0$, *and*

(iii) $h^2_{\mathcal{S}}(u; \Sigma)$ *does not depend on* $\Sigma_{\mathcal{S}^c, \mathcal{S}^c}$,

then we must have

$$h^2_{\mathcal{S}} = \frac{\mathbf{u}^\top \Sigma \mathbf{u} - \mathbf{u}^\top_{\mathcal{S}^c} \Sigma_{\mathcal{S}^c|\mathcal{S}} \mathbf{u}_{\mathcal{S}^c}}{\mathbf{u}^\top \Sigma \mathbf{u} + \sigma_e^2}.$$

Proof. For the sake of simplicity, assume $\mathrm{Var}(y) = 1$. Without loss of generality, assume that $\mathcal{S} = \{1, \ldots, |\mathcal{S}|\}$, let $\mathbf{u} = (\mathbf{u}^\top_{\mathcal{S}}, \mathbf{u}^\top_{\mathcal{S}^c})^\top$ and $\Sigma = \begin{pmatrix} \Sigma_{\mathcal{S}} & \Sigma_{\mathcal{S},\mathcal{S}^c} \\ \Sigma^\top_{\mathcal{S},\mathcal{S}^c} & \Sigma_{\mathcal{S}^c}. \end{pmatrix}$ Then the quadratic form based heritability $h^2_{\mathcal{S}}(\mathbf{u}, \Sigma) = \mathbf{u}^\top \Gamma \mathbf{u}$ where $\Gamma = \begin{pmatrix} \Gamma_{\mathcal{S}} & \Gamma_{\mathcal{S},\mathcal{S}^c} \\ \Gamma^\top_{\mathcal{S},\mathcal{S}^c} & \Gamma_{\mathcal{S}^c} \end{pmatrix}$ is a $p \times p$. Moreover, due to property (i), $0 \leq \Gamma \leq \Sigma$.

If $\mathbf{u}^c_{\mathcal{S}} = 0$, $\forall\ \mathbf{u}_{\mathcal{S}} \in \mathcal{R}^{|\mathcal{S}|}$,

$$\begin{aligned} h^2 &= \mathbf{u}^\top_{\mathcal{S}} \Sigma_{\mathcal{S}} \mathbf{u}_{\mathcal{S}}, \\ h^2_{\mathcal{S}} &= \mathbf{u}^\top_{\mathcal{S}} \Gamma_{\mathcal{S}} \mathbf{u}_{\mathcal{S}}. \end{aligned}$$

By property (ii), this implies $\Gamma_{\mathcal{S}} = \Sigma_{\mathcal{S}}$. If $\mathbf{u}_{\mathcal{S}} = 0$ and $\mathbf{u}_{\mathcal{S}^c} \neq 0$,

$$\begin{aligned} h^2 &= \mathbf{u}^\top_{\mathcal{S}^c} \Sigma_{\mathcal{S}^c} \mathbf{u}_{\mathcal{S}^c}, \\ h^2_{\mathcal{S}} &= \mathbf{u}^\top_{\mathcal{S}^c} \Gamma_{\mathcal{S}^c} \mathbf{u}_{\mathcal{S}^c}. \end{aligned}$$

Then by Property (i) and (ii), $\Gamma_{\mathcal{S}^c} < \Sigma_{\mathcal{S}^c}$. Next, Property (i) suggests that

$$\Sigma - \Gamma = \begin{pmatrix} 0 & \Sigma_{\mathcal{S},\mathcal{S}^c} - \Gamma_{\mathcal{S},\mathcal{S}^c} \\ (\Sigma_{\mathcal{S},\mathcal{S}^c} - \Gamma_{\mathcal{S},\mathcal{S}^c})^\top & \Sigma_{\mathcal{S}^c} - \Gamma_{\mathcal{S}^c} \end{pmatrix} \geq 0$$

Since $\Sigma_{\mathcal{S}^c} - \Gamma_{\mathcal{S}^c} > 0$, $\Sigma - \Gamma \geq 0$ is equivalent to

$$\begin{aligned} 0 - (\Sigma_{\mathcal{S},\mathcal{S}^c} - \Gamma_{\mathcal{S},\mathcal{S}^c})(\Sigma_{\mathcal{S}^c} - \Gamma_{\mathcal{S}^c})^{-1}(\Sigma_{\mathcal{S},\mathcal{S}^c} - \Gamma_{\mathcal{S},\mathcal{S}^c})^\top &\geq 0 \\ (\Sigma_{\mathcal{S},\mathcal{S}^c} - \Gamma_{\mathcal{S},\mathcal{S}^c})(\Sigma_{\mathcal{S}^c} - \Gamma_{\mathcal{S}^c})^{-1}(\Sigma_{\mathcal{S},\mathcal{S}^c} - \Gamma_{\mathcal{S},\mathcal{S}^c})^\top &\leq 0 \end{aligned}$$

However, $(\Sigma_{\mathcal{S},\mathcal{S}^c} - \Gamma_{\mathcal{S},\mathcal{S}^c})(\Sigma_{\mathcal{S}^c} - \Gamma_{\mathcal{S}^c})^{-1}(\Sigma_{\mathcal{S},\mathcal{S}^c} - \Gamma_{\mathcal{S},\mathcal{S}^c})^\top \geq 0$ because $\Sigma - \Gamma \geq 0$ and $\Sigma_{\mathcal{S}^c} - \Gamma_{\mathcal{S}^c} > 0$. Therefore,

$$(\Sigma_{\mathcal{S},\mathcal{S}^c} - \Gamma_{\mathcal{S},\mathcal{S}^c})(\Sigma_{\mathcal{S}^c} - \Gamma_{\mathcal{S}^c})^{-1}(\Sigma_{\mathcal{S},\mathcal{S}^c} - \Gamma_{\mathcal{S},\mathcal{S}^c})^\top = 0$$

Thus, $\Gamma_{\mathcal{S},\mathcal{S}^c} = \Sigma_{\mathcal{S},\mathcal{S}^c}$. Moreover, $\Gamma \geq 0$ implies

$$\Gamma_{\mathcal{S}^c} - \Sigma^\top_{\mathcal{S},\mathcal{S}^c} \Sigma^{-1}_{\mathcal{S}} \Sigma_{\mathcal{S},\mathcal{S}^c} \geq 0 \tag{3.20}$$

We then let

$$\Gamma_{\mathcal{S}^c} = \Sigma_{\mathcal{S},\mathcal{S}^c}^\top \Sigma_{\mathcal{S}}^{-1} \Sigma_{\mathcal{S},\mathcal{S}^c} + M, \quad M \geq 0$$

Finally, we would like to prove $M = 0$ by contradiction. Suppose that there exist some $\mathbf{u}_{\mathcal{S}^c} = \boldsymbol{\beta}$ and Σ such that $\boldsymbol{\beta}^\top M \tilde{\boldsymbol{\beta}} > 0$. Let $\mathbf{u} = (0, \ldots, 0,\ \boldsymbol{\beta})^\top$, then

$$h_{\mathcal{S}}^2 = \boldsymbol{\beta}^\top \Sigma_{\mathcal{S},\mathcal{S}^c}^\top \Sigma_{\mathcal{S}}^{-1} \Sigma_{\mathcal{S},\mathcal{S}^c} \boldsymbol{\beta} + \boldsymbol{\beta}^\top M \boldsymbol{\beta}$$

Now let

$$\tilde{\Sigma} = \begin{pmatrix} \Sigma_{\mathcal{S}} & \Sigma_{\mathcal{S},\mathcal{S}^c} \\ \Sigma_{\mathcal{S},\mathcal{S}^c}^\top & \Sigma_{\mathcal{S},\mathcal{S}^c}^\top \Sigma_{\mathcal{S}}^{-1} \Sigma_{\mathcal{S},\mathcal{S}^c} + \frac{1}{2} M + \frac{\boldsymbol{\beta}^\top M \boldsymbol{\beta}}{4\|\boldsymbol{\beta}\|_2^2} I \end{pmatrix} > 0$$

By property (iii), $h_{\mathcal{S}}^2(\mathbf{u}, \tilde{\Sigma}) = h_{\mathcal{S}}^2(\mathbf{u}, \Sigma)$. However,

$$\begin{aligned} h^2(\mathbf{u}, \tilde{\Sigma}) &= \boldsymbol{\beta}^\top \Sigma_{\mathcal{S},\mathcal{S}^c}^\top \Sigma_{\mathcal{S}}^{-1} \Sigma_{\mathcal{S},\mathcal{S}^c} \boldsymbol{\beta} + \frac{3}{4} \boldsymbol{\beta}^\top M \boldsymbol{\beta} \\ &= h_{\mathcal{S}}^2(\mathbf{u}, \tilde{\Sigma}) - \frac{1}{4} \boldsymbol{\beta}^\top M \boldsymbol{\beta} \end{aligned}$$

This contradicts with Property (i). Therefore, $M = 0$ and $\Gamma_{\mathcal{S}^c} = \Sigma_{\mathcal{S},\mathcal{S}^c}^\top \Sigma_{\mathcal{S}}^{-1} \Sigma_{\mathcal{S},\mathcal{S}^c}$. □

3.6.2 Proof of Proposition 2

Proposition 2. *Assume (3.17)-(3.18) hold, and suppose* $(\sigma_C^2, \eta_C^2) \in \mathcal{K}$ *for some compact set* $\mathcal{K} \subseteq (0, \infty)$ *and* $\rho \in (0, \infty)/\{1\}$, *then as* $k/n \to \rho$,

$$\hat{h}_C^2 = \frac{\hat{\eta}_C}{\hat{\eta}_C + 1} \to h_C^2$$

in probability. Moreover, define $\mathcal{I} = \frac{1}{k} W_C W_C^\top$ *and* $\mathcal{J} = \frac{\eta_C^2}{k} W_C W_C^\top + I$, *then*

$$\sqrt{n}(\hat{h}_C^2 - h^2) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \quad \frac{2\sigma_{C^\perp}^4}{(1 + \eta_C^2)^4} \left(1 - \frac{\operatorname{tr}(\mathcal{I}\mathcal{J}^{-1})^2}{n \operatorname{tr}(\mathcal{I}^2 \mathcal{J}^{-2})}\right)^{-1}\right)^{-1}.$$

Proof. By Theorem 2 of (Dicker and Erdogdu, 2016b),

$$\sqrt{n}(\hat{\eta}_C^2 - \eta_C^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \psi).$$

where $\psi = (\iota_2 - \iota_3^2/\iota_4)^{-1}$ and

$$\iota_\alpha = \frac{1}{2n\sigma_{C^\perp}^{2(4-\alpha)}} \operatorname{tr}\left\{ \left(\frac{1}{k} W_C W_C^\top\right)^{\alpha-2} \left(\frac{\eta_C^2}{k} W_C W_C^\top + I\right)^{2-\alpha} \right\}.$$

Let $\mathcal{I} = \frac{1}{k} W_C W_C^\top$ and $\mathcal{J} = \frac{\eta_C^2}{k} W_C W_C^\top + I$, It follows that

$$\psi = 2\sigma_{C^\perp}^4 \left(1 - \frac{\mathrm{tr}(\mathcal{I}\mathcal{J}^{-1})^2}{n\mathrm{tr}(\mathcal{I}^2\mathcal{J}^{-2})}\right)^{-1}.$$

By the Delta method,

$$\sqrt{n}(\hat{h}_C^2 - h^2) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\psi}{(1+\eta_C^2)^4}\right).$$

□

Chapter 4

Projected least squares and risk estimation for out-of-sample prediction

4.1 Introduction

Model complexity in high-dimensional data analysis makes regularization essential for estimating the parameters. On one hand, outcomes in these big data problems can be predicted via the shrinkage approach with additional structural smoothness or sparsity assumptions, for example, as discussed in Chapter 2. On the other hand, for non-random signals, a simple but effective approach is via dimension reduction. Moreover, the optimal approach is usually chosen by minimizing the risk, and heritability estimation discussed in Chapter 3 provides consistent risk estimation for fixed effects models. This chapter contains new model evaluation approach that identifies and assesses simple dimension reduction techniques for high-dimensional data analysis, with models that are known to be misspecified.

4.1.1 The model

Consider the linear model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ is centered and real-valued response vector, $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \in \mathbb{R}^{n\times p}$ is centered full rank design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ is parameter vector of length p, and $\boldsymbol{\epsilon}$ is noise vector of length n, such that $\mathbb{E}(\boldsymbol{\epsilon}) = 0$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma_\epsilon^2 I$. We are interested in the out-of-sample prediction problem under the setting where $(y_i, \mathbf{x}_i) \overset{\text{iid}}{\sim} F$ and $p/n \to \rho \in (0, \infty)$ when $n, p \to \infty$. Under quadratic loss, portion of explained variation for out-of-sample prediction method has an upper bound of

$\mathrm{Var}(\mathbf{x}^\top \boldsymbol{\beta})/\mathrm{Var}(y)$. In the context of genome-wide association studies (GWAS), the upper bound is interpreted as the narrow-sense heritability in Chapter 3. However, in practice, reaching this upper bound (or even a fraction of it) is notoriously challenging for this high-dimensional prediction problem as most of effect-sizes are small and insignificant, but not sparse (Consortium et al., 2009; Wray et al., 2013).

4.1.2 Motivation

In many high-dimensional data problems, out-of-sample prediction are optimized by complicated feature engineering methods (Zhang et al., 2016). These methods target at pushing portion of explained variation towards the upper bound — total heritability. However, in some applications, effects can be naturally partitioned into finitely many clusters. Instead of estimating each effect, we could simply estimate the (weighted) average effect within each cluster. Such strategy is equivalent to linearly transform the input vector. The average effects within clusters can be estimated easily when sample size is sufficiently large, and this linear dimension reduction method is easy to interpret and implement; however, since the linear dimension reduction approach ignores within-cluster variations, the model becomes misspecified. Comparing to portion of explained variation of correct model methods, that of misspecified dimension reduction has a smaller upper bound — C-heritability (defined as (3.13)). Although the upper bound C-heritability is smaller than the total heritability, it is much easier to achieve because misspecified dimension reduction deals with finitely many parameters.

Intuitively, when effects behave nicely within clusters (i.e., same direction and/or small variation), the misspecified error is small and misspecified dimension reduction should perform well. This approach has shown comparatively good performance in several areas such as demand forecasting, genetic risk prediction and asset return forecasting (Mark and Sul, 2011). In online retailing, the weekly demand of every product sold online is required for inventory planning. Training a common univariate time series model for all products has shown higher prediction accuracy under various metrics than other high-dimensional time series methods. Moreover, genetic risk prediction is an important topic in genetics for animal breeding and human disease prevention/intervention (Henderson, 1984; Consortium et al., 2009). The (multi-)polygenic score approach to quantitative trait prediction has drawn much attention recently (e.g. Krapohl et al., 2017; Selzam et al., 2017). The polygenic score (also known as genetic risk score) summarizes all genetic variants associated with a given risk factor by weighted sum over all genotypes (weights are derived by repeated simple regression from a large repetition sample). The prediction model is then fitted against multiple polygenic scores. Multi-polygenic score approach has shown to be effective in predicting educational attainment (Krapohl et al., 2017). Both pooling/stacking data in demand forecasting and multi-polygenic score approach are applications of misspecified dimension reduction method, and the goal of this chapter is to investigate the conditions for this approach to perform well, comparing to alternative methods.

4.1.3 Related work

The linear model (4.1) is sometimes derived by combining (stacking) many individual linear models despite heteroscedasticity and correlated error, especially for high-dimensional time-series analysis such as online-retailing data. For these datasets, generalized least squares is usually performed to transform the data to achieve uncorrelated and homoscedastic error, provided that covariance matrix of error is known up to a scalar factor (Fahrmeir et al., 2007). Comparing to the classical ordinary least squares, generalized least squares has advantage in terms of reduced variance.

However, generalized least squares does not handle problems of high-dimensionality and collinearity among predictors. If the majority of features are assumed to be zero, then the parameters can possibly be estimated consistently with some feature selection method. For non-sparse signal, its dimensionality is usually reduced by feature extraction. After reducing the model complexity, novel parameters could be easily estimated the by least squares method.

Principal component regression (PCR) is a classical feature extraction approach that handles high-dimensionality issue by regressing the response on top principal components of predictors. When PCR was first introduced by Hotelling (1957), the PCR was treated as a change-of-basis variation of the original least squares problem, where all principal components (PCs) remain in the least squares model regardless their corresponding magnitudes of variances. However, for dimension reduction purpose, low-variance PCs are usually dropped from the model assuming that the loss of prediction power is very little. Many real data examples noted in (Jolliffe, 1982) contradicts with this assumption, by showing the significance of low-variance PCs. Since PCR drops low-variance PCs without considering the responses, the approach could potentially result in large misspecification error.

Partial Least squares method is another linear dimension reduction approach. It is developed by (Wold, 1975; Wold et al., 1984) primarily for models with high dimensionality and multicollinearity issues in the field of chemometrics. The method has been widely applied in genomic data for regression and classification purposes (Boulesteix and Strimmer, 2006). In contrast to PCR, partial least squares creates a sequence of orthogonal input directions by iteratively maximizing variance between response and predictors (Hastie et al., 2009). The dimensionality of the problem can be reduced by dropping trailing input directions in the sequence, and the majority of prediction power retains in the misspecified model.

Both PCR and partial least squares project the data to some lower dimensional subspace. The response is then predicted by least squares. We refer *projected least squares* to be the class of least squares estimators that takes a linearly transformed covariates. The projection-based dimension reduction methods are based on distances calculated under the linear kernel. With different kernels, many linear dimension reduction methods can be extended to non-linear manifold methods (Nasrabadi, 2007).

4.1.4 Contribution

Previously mentioned linear dimension reduction approaches discuss specific implementations of finding the projection directions for design matrix. However, to the best of our knowledge, risk estimation for fitting these misspecified model is not yet available for methods with derived input directions. This chapter derives asymptotic out-of-sample error for projected least squares approach rather than focuses on deriving the transformed directions. Similar to many high-dimensional prediction methods (Bayati and Montanari, 2012; Dicker, 2016), the explicit asymptotic risk of projected least squares is in terms of variance components.

Recall that in Chapter 3, we have discussed an estimation method for evaluating proportion of explained variance in a model with random-design assumption. Hence, we propose to evaluate out-of-sample error by variance component estimation, and furthermore compare risks of various models by heritability estimation. The evaluation tool can be directly applied to risk estimation and model evaluation in multi-polygenic score method for genetic risk prediction.

Rest of the chapter is organized as follows: Section 4.2 discusses out-of-sample error of ordinary least squares (OLS) estimator. Section 4.3 presents asymptotic risk of projected least squares. Risk estimation for various high-dimensional methods with explicit asymptotic risk is contained in Section 4.4. Section 4.5 and 4.6 discuss numerical analysis regarding genetic risk prediction and demand forecasting, respectively.

Notations. For $\mathbf{v} \in \mathbb{R}^n$ and $M \in R^{n\times n}$, denote the l_2 norm $\|\mathbf{v}\|_2 := \sqrt{\mathbf{v}^\top \mathbf{v}}$, denote the matrix norm $\|\mathbf{v}\|_M := \sqrt{\mathbf{v}^\top M\mathbf{v}}$, and denote the spectral norm $\|M\| := \max_{\|\mathbf{x}\|_2=1} \|M\mathbf{x}\|_2^2$.

4.2 Out-of-sample prediction and OLS

4.2.1 Loss function and random-design assumption

For in-sample prediction only, we usually pursue an estimator such that it minimizes the Euclidean distance with $\mathbf{y}$, i.e. minimizing

$$l_{in}(\boldsymbol{\delta}; X, \mathbf{y}) = \|X\boldsymbol{\delta} - \mathbf{y}\|_2^2. \tag{4.2}$$

When $\rho < 1$, the OLS estimator of (4.1), by definition, is the minimizer of (4.2), where

$$\hat{\boldsymbol{\beta}}_{ols} = (X^\top X)^{-1} X^\top \mathbf{y}.$$

Although $\hat{\boldsymbol{\beta}}_{ols}$ achieves the minimum in-sample error (4.2), supervised learning is primarily used to predict unseen data. For out-of-sample prediction, random-design assumption is usually required. We assume that for $i = 1, \ldots, n$, $\mathbf{z}_i = (y_i, \mathbf{x}_i^\top)^\top$ is the i^{th} sample drawn from the population with a distribution F, where $\mathbb{E}(\mathbf{x}_i) = 0$ and $\mathrm{Var}(\mathbf{x}_i) = \Sigma$. Therefore, we would like to find an estimator that minimizes the expected l_2 loss for unseen data $\mathbf{z}_{n+1}$. Then, the out-of-sample error of an estimator $\hat{\boldsymbol{\delta}}$ conditioning on the training set $(\mathbf{y}, X)$ is

$$l(\hat{\boldsymbol{\delta}}; \mathbf{y}, X, \Sigma) = \|\hat{\boldsymbol{\delta}} - \boldsymbol{\beta}\|_\Sigma^2. \tag{4.3}$$

Throughout the chapter, our goal is to minimize (4.3).

In this chapter, we have strong distributional assumption on F, such that

$$\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma) \quad \text{and} \quad \epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2) \tag{4.4}$$

are independent, where Σ is a positive definite and $\sigma_\epsilon^2 > 0$. Therefore,

$$\mathbf{z}_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \begin{pmatrix} \boldsymbol{\beta}^\top \Sigma \boldsymbol{\beta} + \sigma_\epsilon^2 & \boldsymbol{\beta}^\top \Sigma \\ \Sigma \boldsymbol{\beta} & \Sigma \end{pmatrix}\right).$$

With only mild conditions (without Gaussian assumption) on the design distribution, bounds of the non-asymptotic risk of OLS, ridge regression, and misspecified OLS are derived by Hsu et al. (2011). However, exact asymptotic risks are challenging without multivariate Gaussian assumption on $\mathbf{z}_i$. Similar Gaussian assumption on F has been made in linear model risk studies such as Breiman and Freedman (1983); Leeb et al. (2009); Dicker (2013). Lemma 3 and 4 depends on the closed-form expectation of trace of inverse Wishart distribution. Moreover, proofs of these lemmas rely on orthogonal invariance of Gaussian noise distribution.

4.2.2 Generalization of ordinary least squares estimator

For online-retailing data, dimensionality of covariates is approximately the same as number of observations; therefore, we are still interested in the case where $\rho < 1$. OLS method is the classical estimator when $p < n$, and its asymptotic risk can be computed under assumption (4.4).

Lemma 3. *Let $n, p \to \infty$ and $p/n \to \rho \in (0, 1)$, assume condition (4.4) holds, then*

$$\lim_{n,p\to\infty} \|\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}\|_\Sigma^2 = \frac{\sigma_\epsilon^2 \rho}{1 - \rho}.$$

The proof of Lemma 3 is contained in Section 4.7, and is consistent with results derived by Dicker (2013). As ρ approaches to 1 from the left, OLS is expected to gain tremendous out-of-sample error. Therefore, it is reasonable to consider other high-dimensional methods in this senario. Risk for other linear shrinkage estimators such as oracle solution of ridge regression is also derived similarly under Gaussian assumption (4.4) (Dicker, 2013, 2016). Asymptotic squared-error risk of LASSO is proofed based on analysis of approximate message passing algorithm discussed in Chapter 2 under assumption (4.4) and $\Sigma = I$ (Bayati and Montanari, 2012).

In the e-commerce data, predicted demands are further analyzed for the purpose of inventory planning. Since the cost of failing to fulfill cosumers' demand outweigh cost of storing extra inventory, we may also be interested in other loss functions such as quantile loss and absolute loss. Exact risk is only available for squared-error loss, but in practice similar phenomena seem to hold for these other loss functions.

4.3 Projected least squares estimation

For the purpose of out-of-sample prediction, we propose an OLS-type estimator, which regresses $\mathbf{y}$ on linearly transformed predictors XC, where $C \in \mathbb{R}^{p\times k}$ $(k/p \to 0)$. This approach provides a sharp estimate of the coefficients while introducing some misspecification error.

The transformation matrix C is given a priori. For example, in online-retailing data, suppose there are m products sold online, then C could correspond to pooling the data across products and create a common univariate time-series/regression model for all products. The linearly transformed predictors provide information of average overall demand for each previous week. For multi-polygenic score approach in genetic risk prediction, the weights in C are noisy estimations of each parameter acquired in a large repetition sample (Selzam et al., 2017). The projected design summarizes all genetic variants associated with a given risk factor by weighted sum over all genotypes.

In many other problems where projection directions are not obvious, it is important to investigate the suitable transformations. The directions can be found by dimension reduction techniques, including classical principal component analysis and partial least squares noted previously. In addition, these directions derived can be converted into a coarser group mapping by projected-clustering method such as k-means. The clustering pattern might also lie within subspace of the data; however this approach might not be as effective due to the curse of dimensionality (Kriegel et al., 2009). However, for out-of-sample prediction, the best choice of finite dimensional projection C should maximize its corresponding C-heritability.

4.3.1 Asymptotic risk

Similar to the argument on (3.17) from Chapter 3, with random-design assumption, we can decompose $\mathbf{y}$ into the following form

$$\begin{aligned}
\mathbf{y} &= X\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
&= Z\Sigma^{1/2}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
&= Z\Sigma^{1/2}C(C^\top \Sigma C)^{-1}C^\top \Sigma\boldsymbol{\beta} + Z(I - \Sigma^{1/2}C^\top(C^\top\Sigma C)^{-1}C^\top\Sigma^{1/2})\Sigma^{1/2}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
&= XC\boldsymbol{\mu} + X\boldsymbol{\gamma} + \boldsymbol{\epsilon}
\end{aligned}$$

where

$$\begin{aligned}
Z &:= X\Sigma^{-1/2}, \\
\boldsymbol{\mu} &:= (C^\top\Sigma C)^{-1}C^\top\Sigma\boldsymbol{\beta} \\
\boldsymbol{\gamma} &:= \boldsymbol{\beta} - C\boldsymbol{\mu}.
\end{aligned}$$

The estimand of projected least squares $\boldsymbol{\mu}$ is a linear transformation of $\boldsymbol{\beta}$. Consider the new noise being $\boldsymbol{\epsilon}' = (\epsilon'_1, \ldots, \epsilon'_n)^\top = X\boldsymbol{\gamma} + \boldsymbol{\epsilon}$, under assumption (4.4),

$$\epsilon'_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \boldsymbol{\gamma}^\top\Sigma\boldsymbol{\gamma} + \sigma_\epsilon^2) \quad \text{and independent of } C^\top\mathbf{x}_i.$$

Then we take the least squares approach to estimate $C\boldsymbol{\mu}$. The least squares estimate provides a sharp estimate on $C\boldsymbol{\mu}$ as $k/p \to 0$, while introducing some bias because of model misspecification. Such bias-variance tradeoff is shown in asymptotic out-of-sample error of projected least squares below.

Lemma 4. *Assume that the linear model* (4.1) *holds, that* $\boldsymbol{\beta} \in \mathbb{R}^p$ *is a fixed vector. Suppose that* $\mathbf{z}_i = (y_i, X_{i\cdot})^\top$ *and*

$$\mathbf{z}_i \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \begin{pmatrix} \boldsymbol{\beta}^\top \Sigma \boldsymbol{\beta} + \sigma_\epsilon^2 & \boldsymbol{\beta}^\top \Sigma \\ \Sigma \boldsymbol{\beta} & \Sigma \end{pmatrix}\right).$$

Let $C \in \mathbb{R}^{p \times k}$, $k/p \to 0$, *and define*

$$\hat{\boldsymbol{\beta}}_{proj} := C\hat{\boldsymbol{\mu}}_{proj} = C(C^\top X^\top X C)^{-1} C^\top X^\top \mathbf{y}.$$

Then,

$$\lim_{n\to\infty} \|\hat{\boldsymbol{\beta}}_{proj} - \boldsymbol{\beta}\|_\Sigma^2 = \boldsymbol{\gamma}^\top \Sigma \boldsymbol{\gamma}.$$

This lemma is derived in Section 4.7. In the large limit, all of out-of-sample error of projected least squares comes from the misspecified error. Moreover, from the heritability point of view, consider new response be $\boldsymbol{\epsilon}'$, then $\boldsymbol{\gamma}^\top \Sigma \boldsymbol{\gamma}$ measures the total amount of variation with in $\boldsymbol{\epsilon}'$ that is attributable to predictors X. In other words, the portion of explained variation for projected least squares achieves its upper bound (C-heritability) in the large limit.

4.4 Risk estimation and model evaluation

4.4.1 Random-effects assumption

In Chapter 3, we have discussed the maximum likelihood approach for fixed-effects variance component estimation. When $\Sigma = I$, the random-effects Gaussian maximum likelihood method estimate is consistent in estimating $\boldsymbol{\gamma}^\top \Sigma \boldsymbol{\gamma}$ (Dicker and Erdogdu, 2016b). When Σ is known, the data can be reduced to the iid design case after transforming the data $(\mathbf{y}, X) \mapsto (\mathbf{y}, X\Sigma^{-1/2})$. Therefore, for risk estimation purpose, $\Sigma^{1/2}\boldsymbol{\gamma} \in \mathbb{R}^p$ is treated as iid random effects to quantify the magnitude of lost signal. Hence, assume that

$$\boldsymbol{\beta} = C\boldsymbol{\mu} + \boldsymbol{\gamma},\ C \in \mathbb{R}^{p\times k}, \boldsymbol{\mu} \in \mathbb{R}^k \text{ and } \Sigma^{1/2}\boldsymbol{\gamma} \sim \mathcal{N}\left(0, \frac{\tau^2}{p} I_p\right). \tag{4.5}$$

Under the assumption above, the out-of-sample error of projected least squares is equivalent to τ^2. Moreover, it has been previous noted in Chapter 3 that the popular linear kernel-based maximum likelihood approach is potentially biased in estimating $\boldsymbol{\gamma}^\top \Sigma \boldsymbol{\gamma}$ for fixed $\boldsymbol{\gamma}$, and the whitening transformation step $(\mathbf{y}, X) \mapsto (y, X\Sigma^{-1/2})$ is crucial in avoiding biased estimation results. Therefore, in this section, we assume that the structure of Σ is known.

As noted previously, τ^2 is the amount of signal variation within $\boldsymbol{\epsilon}'$, and variance component estimation requires $\boldsymbol{\epsilon}'$. As an immediate result of Lemma 4 ,

$$\lim_{n,p\to\infty} \|C\boldsymbol{\mu} - \hat{\boldsymbol{\beta}}_{proj}\|_\Sigma^2 \to 0.$$

Let

$$\tilde{\mathbf{y}} \;:=\; \mathbf{y} - X\hat{\boldsymbol{\beta}}_{proj}. \tag{4.6}$$

Then, the residual $\tilde{\mathbf{y}}$ is a proxy of $\boldsymbol{\epsilon}'$ in the large limit.

4.4.2 Heritability estimation

In prediction problems, the optimal approach is selected by minimizing the risk. In this subsection, we would like to discuss how to compare out-of-sample error of projected least squares estimation to that of alternative methods. Let $h^2 := \frac{\tau^2}{\tau^2+\sigma_\epsilon^2}$, which measures the variation in residual $\boldsymbol{\epsilon}'$ that is attributable to predictors. Then h^2 indirectly measures risk of projected least squares, and the optimal approach to the problem can be measured by h^2 and heritabilities of other approaches.

First of all, when $\rho < 1$, h^2 provides a rule of thumb for deciding whether projected least squares is superior than OLS. Since in the online retailing data, $n \approx p$, we are still interested in the performance of projected least squares when $\rho < 1$. The decision rule is shown as follows:

Corollary 1. *As $n, p \to \infty$ and $p/n \to \rho \in (0, 1)$, under assumption (4.5), $l\left(\hat{\boldsymbol{\beta}}_{proj}\right) < l\left(\hat{\boldsymbol{\beta}}_{ols}\right)$ is equivalent to*

$$\frac{\tau^2}{\tau^2+\sigma_\epsilon^2} < \rho. \tag{4.7}$$

The corrollary is an immediate result of Lemma 3 and 4. The left hand side of the inequality h^2 compares the squared bias with noise variance for a model where variance of the estimator converges to zero in probability. Such risk evaluation can also be applied to residual of other high-dimensional methods with variance converges to zero.

Heritability h^2 can be measured by maximum likelihood estimation with Mahalanobis kernel. Motivated by random-effects assumption 4.5, let $\eta^2 = \tau^2/\sigma_\epsilon^2$ and the maximum likelihood estimator (MLE) for variance components $(\sigma_\epsilon^2, \eta^2)$ is

$$(\hat{\sigma}_\epsilon^2, \hat{\eta}^2) = \underset{\sigma_\epsilon^2, \eta^2 > 0}{\operatorname{argmax}} \; l(\sigma_\epsilon^2, \eta^2), \tag{4.8}$$

where

$$\begin{aligned} l(\sigma_\epsilon^2, \eta^2; X\Sigma^{-1/2}, \tilde{\mathbf{y}}) \quad = \quad & -\frac{1}{2}\log(\sigma_\epsilon^2) - \frac{1}{2n}\log\det(\eta^2/pX\Sigma^{-1}X^\top + I) \\ & -\frac{1}{2n\sigma_\epsilon^2}\tilde{\mathbf{y}}^\top(\eta^2/pX\Sigma^{-1}X^\top + I)^{-1}\tilde{\mathbf{y}} \end{aligned}$$

is the log-likelihood of $\tilde{\mathbf{y}}|X\Sigma^{-1/2}$. The corresponding MLE of h^2 is

$$\hat{h}^2 = \frac{\hat{\eta}^2}{1+\hat{\eta}^2}.$$

We refer statistical properties (i.e. consistency and asymptotic normality) of $\hat{h}^2$ to Proposition 2 in Chapter 3. Comparing to ordinary least squares, the projected least squares is preferred when $\hat{h}^2 < \rho$, a standard Wald test can be applied for the decision of whether to perform the projected least squares method.

When comparing out-of-sample error of projected least squares with that of other sharp estimation approach (e.g. projected least squares with other candidate projections), the preference can also be carried out by variance component estimation (i.e. the projection matrix is prefered when its associated C-heritability is larger). Suppose $\hat{\boldsymbol{\beta}}_{alt}$ is the estimate of an alternative method such that

$$\lim_{n,p\to\infty} \left\| \hat{\boldsymbol{\beta}}_{alt} - \mathbb{E}\left(\hat{\boldsymbol{\beta}}_{alt}\right) \right\|_\Sigma^2 \to 0.$$

Then let the residual of the alternative method be $\tilde{\mathbf{y}}' = \mathbf{y} - X\hat{\boldsymbol{\beta}}_{alt}$ and assume $\tilde{\mathbf{y}}' \to X\boldsymbol{\gamma}' + \boldsymbol{\epsilon}$ where $\boldsymbol{\gamma}' = \boldsymbol{\beta} - \mathbb{E}(\hat{\boldsymbol{\beta}}_{alt})$. Let $\eta'^2 = \tau'^2/\sigma_\epsilon^2$, then $\hat{\eta}'^2$ is an estimator of η'^2 via Gaussian variance component MLE (4.8). Then the original projected least squares method has a smaller misspecified error when Wald test rejects the null hypothesis $H_0^{(a)} : \eta^2 - \eta'^2 > 0$.

For other estimators that are less sharp, direct analysis on variance components in residual could potentially cause biased estimation results. However, if the asymptotic risk has an explicit form in terms of quadratic form-based signal-to-noise ratio, we can also use similar technique to estimate the risk.

Take ridge regression as an example. Provided $\text{Cov}(\mathbf{x}) = \Sigma$, define the ridge regression estimator corresponds to the regularization estimator $t^2 \in [0, \infty]$ be

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{ridge}(t^2) &:= \Sigma^{-1/2}(\Sigma^{-1/2}X^\top X\Sigma^{-1/2} + p/t^2)^{-1}\Sigma^{-1/2}X^\top \mathbf{y}, \\ &= (X^\top X + p/t^2\Sigma)^{-1}X^\top \mathbf{y}.\end{aligned}$$

Let $\kappa^2 = \boldsymbol{\beta}^\top \Sigma \boldsymbol{\beta}/\sigma_\epsilon^2$ be the signal-to-noise ratio (SNR) of linear model (4.1), then define the optimal ridge regression estimator be

$$\hat{\boldsymbol{\beta}}_{ridge}(\kappa^2) := (X^\top X + p/\kappa^2\Sigma)^{-1}X^\top \mathbf{y}. \tag{4.9}$$

$\hat{\boldsymbol{\beta}}_{ridge}(\kappa^2)$ is "oracle" in the sense that

$$l\left(\hat{\boldsymbol{\beta}}_{ridge}(\kappa^2)\right) = \inf_{t^2 \in [0,\infty]} l\left(\hat{\boldsymbol{\beta}}_{ridge}(t^2)\right).$$

Moreover, assuming $\boldsymbol{\beta}^\top \Sigma \boldsymbol{\beta} < \infty$, we have

$$\lim_{n,p\to\infty} \sup_{\boldsymbol{\beta}\in\mathbb{R}^p} \left| l\left(\hat{\boldsymbol{\beta}}_{ridge}(\kappa^2)\right) - \sigma_\epsilon^2 R_{ridge}(\kappa^2, \rho)\right| = 0,$$

where

$$R_{ridge}(\kappa^2, \rho) = \frac{1}{2\rho}\left[\kappa^2(\rho - 1) - \rho + \sqrt{(\kappa^2\rho - \kappa^2 - \rho)^2 + 4\kappa^2\rho^2}\right].$$

The asymptotic risk of oracle ridge estimator is derived in (Dicker, 2013, 2016). Similar to residual variance component estimation (4.8) discussed earlier, assuming that $\Sigma^{1/2}\boldsymbol{\beta}$ is iid Gaussian, the MLE of κ^2 is $\hat{\kappa}^2$, which estimates the SNR within the original linear model (4.1). Then we claim that projected least squares is expected to have smaller out-of-sample error when Wald test rejects $H_0^{(b)} : \eta^2 - R_{ridge}(\kappa^2, \rho) > 0$.

4.5 Application: genetic risk prediction

4.5.1 Projected least squares in genetics

In animal and plant breeding, and even in human genetics, genetic merit could be predicted based on genome-wide association studies (GWAS) data. Since (Meuwissen and Goddard, 2001), many literature assume a high-dimensional linear model, such that

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.10}$$

where $\mathbf{y} \in \mathbb{R}^n$ is the vector of quantitative traits, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the effect-sizes, and $\boldsymbol{\epsilon}$ is additive environmental noise follows $\mathcal{N}(0, \sigma_\epsilon^2)$. The i^{th} row of X is a centered and standardized corresponding p-dimensional vector of predictors for individual i indicating the genotyped SNP. The genotype j for individual i is trinary and depends on minor allele frequency (MAF) of SNP j (Yang et al., 2010; Zaitlen and Kraft, 2012).

Since quantitative traits are usually affected by numerous SNPs with small effects (Buckler et al., 2009), comparing to limited number of observations, we generally consider the setting where $n << p$. A classical method in genetic prediction is best linear unbiased prediction (BLUP) proposed by (Henderson, 1950, 1984). In the ideal case when there is no linkage disequilibrium (i.e. $\text{Cov}(\mathbf{x}) = I$), and assuming that $\beta_1, \ldots, \beta_p \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\beta^2/p)$, The BLUP is equivalent to the oracle ridge regression estimator, such that

$$\hat{\boldsymbol{\beta}}_{blup} := (X^\top X + p\sigma_\epsilon^2/\sigma_\beta^2 I)^{-1} X^\top \mathbf{y}.$$

The ridge regression effectively controls model complexity and is commonly used in predicting breeding values (de Vlaming and Groenen, 2015).

Many common complex traits have been discovered to be associated with finitely many intermediate risk factors. Although most of SNPs genotyped are insignificant, they are jointly associated with the trait (Consortium et al., 2009). Multi-polygenic score method allows all SNPs to be included for predicting the quantitative phenotype of these complex traits (Krapohl et al., 2017). A polygenic score summarizes all genetic variants associated with a given risk factor by weighted sum over all genotypes, and the phenotype is then predicted by a function of polygenic scores. In practice, the weighted sum is acquired from a replication sample by simple regression analysis (Selzam et al., 2017). In other words, weights associated to a risk factor is a noisy estimation of the effects. Therefore, we further assume that $\boldsymbol{\beta}$ follows assumption 4.5. Although the individual effect-sizes are small and insignificant, the signal is significant after aggregating the genotypes with known weight matrix C, where each parameter of the projected design matrix quantifies the impact of a specific risk factor.

4.5.2 Numerical results

The rest of this section includes simulated quantitative trait prediction. Consider $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\mathbf{y} \in \mathbb{R}^n$ is quantitative trait and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$. In addition, we let

(i) $n = 1000$, $p = 2000$, $k = 10$.

(ii) $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma)$ where ij-th entry of Σ is $0.5^{|i-j|}$. Although entries in $\mathbf{x}$ are trinary and depends on MAF of each SNP, we approximate the predictors with Gaussian design, similar to Chapter 3 .

(iii) Let

$$C = \begin{pmatrix} \mathbf{c}_1 & 0 & \ldots & 0 \\ 0 & \mathbf{c}_2 & 0 & \ldots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & \mathbf{c}_k \end{pmatrix}$$

where $\mathbf{c}_i \in \mathbb{R}^{p/k}$ represents computed weights for SNPs associated with risk factor i. Coordinates of $\mathbf{c}_i$ are iid and follow $(k/p)\chi_1^2$.

(iv) $\boldsymbol{\mu} \sim \sqrt{5}\mathcal{N}(0, I)$.

(v) $\sigma_\epsilon^2 = 0.1$.

Let

$$\boldsymbol{\beta} = \sqrt{1-\omega}C\boldsymbol{\mu} + \boldsymbol{\gamma}, \tag{4.11}$$

where $\boldsymbol{\gamma} \sim \mathcal{N}(0, \sigma_\gamma^2/pI)$ and $\sigma_\gamma^2 = \omega\boldsymbol{\mu}^\top C^\top \Sigma C\boldsymbol{\mu}$. Fixing C and $\boldsymbol{\mu}$, we vary $\omega = 0.2, 0.3, \ldots, 0.9$. For each setting, we simulate 50 independent datasets ($\mathbf{y}$,X), and compute the projected least squares estimator provided C and oracle ridge regression estimator (4.9) provided $\kappa^2 = \boldsymbol{\beta}^\top\Sigma\boldsymbol{\beta}/\sigma_\epsilon^2$. Risk of each estimator is computed by Gaussian variance component MLE (4.8). Mean of risk estimators are shown in Table 4.1.

When ω varies, r_{proj}^2 increase linearly in ω. Signal-to-noise ratio is expected to hold the same when ω varies, and the fluctuations in r_{blup}^2 for each ω in Table 4.1 is due to randomness in $\boldsymbol{\gamma}$. Estimated risk for projected least squares $\hat{r}_{proj}^2$ shows downward

Table 4.1: Mean estimates of out-of-sample error of projected least squares and the BLUP estimator. r^2_{proj} and r^2_{blup} are the expected out-of-sample error of projected least squares and the BLUP estimator respectively. Their average risk estimates are $\hat{r}^2_{proj}$ and $\hat{r}^2_{blup}$, follows the Mahalanobis kernel maximum likelihood estimator from Chapter 3 . $\tilde{r}^2_{proj}$ is the finite sample corrected estimate for $\hat{r}^2_{proj}$.

ω	r^2_{proj}	$\hat{r}^2_{proj}$	$\tilde{r}^2_{proj}$	r^2_{blup}	$\hat{r}^2_{blup}$
0.2	0.154	0.136	0.140	0.465	0.461
0.3	0.242	0.229	0.234	0.470	0.470
0.4	0.313	0.312	0.319	0.454	0.456
0.5	0.385	0.381	0.389	0.450	0.449
0.6	0.423	0.410	0.419	0.420	0.408
0.7	0.508	0.504	0.515	0.429	0.430
0.8	0.560	0.547	0.559	0.434	0.431
0.9	0.639	0.631	0.644	0.419	0.422

bias, especially when ω is large, this is because the difference between r^2_{proj} and actual estimand of $\hat{r}^2_{proj}$ is

$$\iota^2 = \frac{k-1}{n-k} r^2_{proj} + \frac{k}{n-k-1}(\sigma^2_\gamma + \sigma^2_\epsilon),$$

and we expect the difference to be vanished as $n \to \infty$. Let $\tilde{r}^2_{proj} = \hat{r}^2_{proj} + \hat{\iota}^2$ to be the finite sample corrected estimate of r^2_{proj}. After adjusting the non-asymptotic bias for projected least squares, the Mahalanobis MLE is consistent in estimating the risk for both methods. We refer to Chapter 3 for asymptotic variance the Mahalanobis estimator.

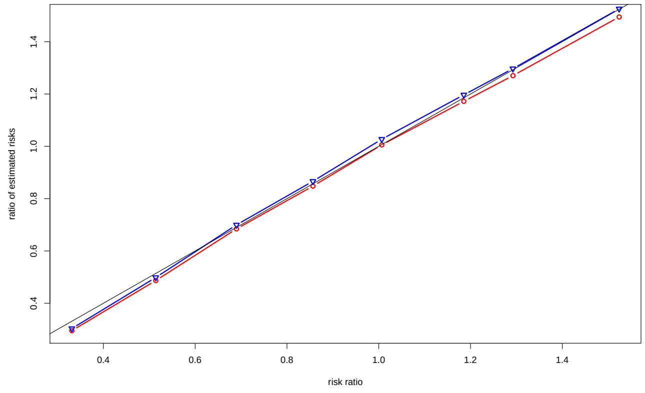

Figure 4.1: For various ω, ratio of risks between projected least squares and BLUP r^2_{proj}/r^2_{blup} is plotted against the ratio of average risk estimates $\hat{r}^2_{proj}/\hat{r}^2_{blup}$ in red line with dots. r^2_{proj}/r^2_{blup} is plotted against $\tilde{r}^2_{proj}/\hat{r}^2_{blup}$ in blue line with dots. The black line is a diagonal reference line.

Figure 4.1 shows the ratio of risks between projected least squares and BLUP against the ratio of mean estimated risk for various ω. A ratio of $r^2_{proj}/r^2_{blup} > 1$ means that projected least squares given C is preferred, and the black diagonal reference line indicates the situation where estimated ratio perfectly identifies the underlying risk ratio. It is shown that the unadjusted ratio usually underestimates the underlying ratio as expected. The finite sample corrected estimated ratio oscillates around the reference diagonal line, suggesting that the relationship between projected least squares and BLUP is well-estimated.

In the next experiment, we fix $\omega = 0.6$ in (4.11) and vary the undersampling ratio

$$n/p = 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7$$

given $p = 2000$. Keeping all other settings the same, we simulate 50 independent datasets with fixed $\boldsymbol{\beta}$. The distributional statistics are summarized in Table 4.2.

Table 4.2: Mean estimates of out-of-sample error of projected least squares the BLUP estimator. r^2_{proj} and r^2_{blup} are the expected out-of-sample error of projected least squares and the BLUP estimator respectively. Their average risk estimates are $\hat{r}^2_{proj}$ and $\hat{r}^2_{blup}$, follows the Mahalanobis kernel maximum likelihood estimator. The adjusted projected least squares risk estimator for non-asymptotic bias is $\tilde{r}^2_{proj}$.

n	r^2_{proj}	$\hat{r}^2_{proj}$	$\tilde{r}^2_{proj}$	r^2_{blup}	$\hat{r}^2_{blup}$
700	0.479	0.474	0.488	0.569	0.576
800	0.478	0.464	0.477	0.539	0.537
900	0.478	0.465	0.476	0.507	0.507
1000	0.476	0.464	0.473	0.478	0.475
1100	0.475	0.470	0.479	0.446	0.447
1200	0.475	0.461	0.470	0.420	0.415
1300	0.475	0.471	0.479	0.391	0.393
1400	0.474	0.456	0.463	0.364	0.366

r^2_{proj} is fixed because of fixed $\boldsymbol{\beta}$ within this simulation. In Table 4.2, we again observed that the estimated risk for projected least squares are biased because of finite sample sizes. After adjusting for the non-asymptotic bias, the gap between average estimated risk and r^2_{proj} is eliminated. In Figure 4.2, estimated ratio between two method is biased without finite sample correction while the adjusted estimated ratio oscillates around the referece line.

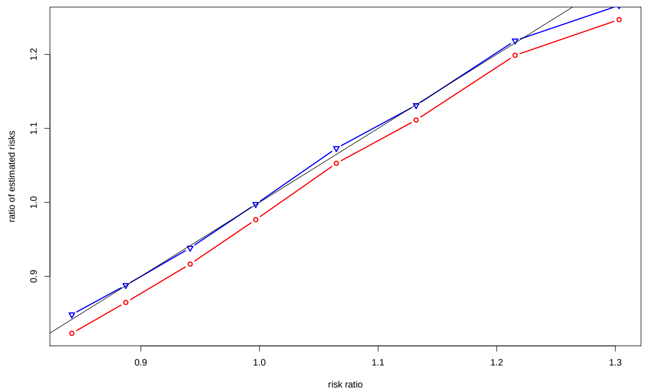

Figure 4.2: For various n, ratio of risk between projected least squares and BLUP r^2_{proj}/r^2_{blup} is plotted against the ratio of average risk estimates $\hat{r}^2_{proj}/\hat{r}^2_{blup}$ in red line with dots. r^2_{proj}/r^2_{blup} is plotted against $\tilde{r}^2_{proj}/\hat{r}^2_{blup}$ in blue line with dots. The black line is a diagonal reference line.

4.6 Application: demand forecasting

4.6.1 Banded vector autoregressive model

The forecasting team in online-retailing industry is responsible to forecast the weekly demand for all $m \to \infty$ products sold online, that is to estimate an extremely high-dimensional time series. Let $\boldsymbol{d}_t = (d_{1,t}, \ldots, d_{m,t})^\top \in \mathbb{R}^m$ be the demand vector for week t. We could assume that $\boldsymbol{d}_t$ follows the vector autoregressive (VAR) model, such that

$$\boldsymbol{d}_t = \sum_{j=1}^{q} B_j \boldsymbol{d}_{t-j} + \boldsymbol{e}_t, \tag{4.12}$$

where $q << m$, B_j is a sparse coefficient matrix, $\boldsymbol{e}_t$ is independent of $\boldsymbol{d}_{t-j}$ $\forall j < t$, with $\mathbb{E}(\boldsymbol{e}_t) = 0$, and $\mathrm{Var}(\boldsymbol{e}_t) = \Omega \in \mathbb{R}^{m \times m}$. Estimating the full coefficient matrices is extremely challenging, therefore works have been developed on model selection techniques on VAR model (e.g. Hsu et al., 2008; Haufe et al., 2008).

Suggested by Guo et al. (2016), after re-arranging the order of products, we can assume that B_j is banded in the sense that $(B_j)_{i,i'} = 0$ if $|i - i'| > s_0$. Provided training data $\boldsymbol{d}_t$, $t = 1, \ldots, T$, let $p_1 = q(2s_0 + 1)$ and $n_1 = T - q$, this data set can be written in linear models. For almost every product $a \in [m]$, the individual linear model is $\mathbf{y}_a = X_a \boldsymbol{\beta}_a + \boldsymbol{\epsilon}_a$, where $\boldsymbol{\beta}_a \in \mathbb{R}^{p_1}$ and the design matrix is $n_1 \times p_1$ matrix with full rank and $n_1 > p_1$ (despite $2k_0$ design matrices among them have less than p_1, i.e. when

$a \leq s_0$ or $a \geq m - s_0 + 1$). Similar to the random-design assumption for generic design matrix, suppose that

$$(X_a)_{i\cdot} \sim \mathcal{N}(0, \Sigma_1) \ \forall a \in [m] \text{ and } i \in [n_1]; \tag{4.13}$$

$$\boldsymbol{\beta}_a = \boldsymbol{\mu} + \boldsymbol{\gamma}_a, \text{where } \boldsymbol{\mu} \in \mathbb{R}^k \text{ and } \Sigma_1^{1/2}\boldsymbol{\gamma}_a \sim \mathcal{N}\left(0, \frac{\sigma_\beta^2}{p} I_{p_1}\right). \tag{4.14}$$

Let $\Sigma_* = \text{diag}(\Sigma_1, \ldots, \Sigma_1)$, let $n = mn_1$, $p = mp_1$ and $n/p \to \rho \in (0, 1)$. In a stacked form, the model is

$$\mathbf{y}_* = X_* \boldsymbol{\beta}_* + \boldsymbol{\epsilon}_*, \tag{4.15}$$

where $\mathbf{y}_* = (\mathbf{y}_1^\top, ..., \mathbf{y}_m^\top)^\top$, $X_* = \text{diag}(X_1, X_2, \cdots, X_m)$, $\boldsymbol{\beta}_* = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_m^\top)^\top$ and $\boldsymbol{\epsilon}_* = (\boldsymbol{\epsilon}_1^\top, \ldots, \boldsymbol{\epsilon}_m^\top)^\top$.

Guo et al. (2016) focuses on finding optimal s_0 and estimate $\boldsymbol{\beta}_*$ with OLS estimator. Although $\boldsymbol{\epsilon}_*$ is correlated, the OLS provides consistent estimation results. However, it is worth mentioning that the design matrix X_* is extremely sparse, such that only the m $n_1 \times p_1$ block matrices along the diagonal in the design matrix are non-sparse. Due to the sparsity and high dimensionality ($n \approx p$), the OLS estimator usually generalizes poorly, we would like to improve the OLS estimator by performing dimension reduction as well as reducing sparsity within the design matrix X_*. Let the transformation matrix be $I_* := (I_p, \ldots, I_p)^\top$, then I_* can be interpreted as the grouping of features, i.e. j^{th} feature of $\boldsymbol{\beta}_a$ is assigned to group j $\forall a \in [m]$.

4.6.2 Asymptotic risks of projected least squares and OLS

In the following subsection, we assume $\boldsymbol{\epsilon}_* \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$, the uncorrelated error is unrealistic for time series, however it is essential for computing the asymptotic risk.

The risk of $\hat{\boldsymbol{\beta}}_*^{(ols)} = \left(\hat{\boldsymbol{\beta}}_1^{(ols)}, \ldots, \hat{\boldsymbol{\beta}}_m^{(ols)}\right)^\top$ is calculated as follows, the risk is multiplied by $1/m$ to control it to a finite value.

$$\begin{aligned} \frac{1}{m}\|\hat{\boldsymbol{\beta}}_*^{(ols)} - \boldsymbol{\beta}_*\|_{\Sigma_*}^2 &= \frac{1}{m}\sum_{a=1}^m \|\boldsymbol{\beta}_a^{(ols)} - \boldsymbol{\beta}_a\|_{\Sigma_1}^2 \\ &= \frac{1}{m}\sum_{a=1}^m \|(X_a^\top X_a)^{-1} X_a^\top \mathbf{y}_a - \boldsymbol{\beta}_a\|_{\Sigma_1}^2 \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{m}\sum_{a=1}^{m} \|(X_a^\top X_a)^{-1} X_a^\top \boldsymbol{\epsilon}_a\|_{\Sigma_1}^2 \\
&= \frac{1}{m}\sum_{a=1}^{m} \|(Z_a^\top Z_a)^{-1} Z_a^\top \boldsymbol{\epsilon}_a\|_2^2,
\end{aligned}
$$

where $Z_a = X_a \Sigma_1^{-1/2}$ is iid $n_1 \times p_1$ Gaussian unitary random matrix.

In the large limit, by law of large numbers,

$$
\lim_{m\to\infty} \frac{1}{m}\sum_{a=1}^{m} \|(Z_a^\top Z_a)^{-1} Z_a^\top \boldsymbol{\epsilon}_a\|_2^2 \to \mathbb{E}\|(Z_a^\top Z_a)^{-1} Z_a^\top \boldsymbol{\epsilon}_a\|_2^2,
$$

where

$$
\mathbb{E}\|(Z_a^\top Z_a)^{-1} Z_a^\top \boldsymbol{\epsilon}_a\|_2^2 = \mathbb{E}_Z \mathbb{E}_{\epsilon|Z} \|(Z_a^\top Z_a)^{-1} Z_a^\top \boldsymbol{\epsilon}_a\|_2^2 = \frac{p_1 \sigma_\epsilon^2}{n_1 - p_1 - 1}
$$

by properties of inverse-Wishart distributions.

In this problem, the projection direction can be viewed as a way of grouping the features. The j-th feature of every $\boldsymbol{\beta}_a$ is mapped to group j, i.e. letting the projection direction be $I_* := (I_p, \ldots, I_p)^\top$. From the parameter perspective, the projection direction utilizes the knowledge of grouping in effects, assuming that effects in the same group have small variance. For example, a projection direction of I_* indicates that the j^{th} feature of every $\boldsymbol{\beta}_a$ are the same. For this special case with sparse design matrix, from the design matrix perspective, the projection direction reduces the dimensionality and sparsity within the design matrix, and can be treated as a way of compressing the signal in the design matrix. Define

$$
\hat{\boldsymbol{\beta}}_*^{(proj)} := I_* (I_*^\top X_*^\top X_* I_*)^{-1} I_*^\top X^\top \mathbf{y}_*
$$

The assumption (4.13) also makes it possible to compute the risk of projected least squares estimator.

Corollary 2. *Assume (4.13) holds, let $m \to \infty$ and $n_1/p_1 = \rho \in (0,1)$, then*

$$
\lim_{m\to\infty} \frac{1}{m} \|\hat{\boldsymbol{\beta}}_*^{(proj)} - \boldsymbol{\beta}_*\|_{\Sigma_*}^2 = \frac{\sigma_\beta^2 \mathrm{tr}(\Sigma_1)}{m p_1}.
$$

With assumption that X_a's share the same Σ_1, this corrollary is an immediate result of Theorem 1. Then similar heritability estimation technique and Wald test in Section 4.4 can be applied to this problem.

4.6.3 Numerical results

Consider linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ assuming uncorrelated error $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$. We first investigate how projected least squares react with block diagonal design matrix. Let

(i) $n = 2000$, $p = 1000$, $k = 10$.

(ii) $X_* = \text{diag}(X_1, X_2, \cdots, X_m)$;
for $a = 1, \ldots, m$, $(X_a)_{i\cdot} \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma_1)$ where $(\Sigma_1)_{ij} = 0.2^{|i-j|}$.

(iii) Let $\mathbf{1}$ is be length-p/k vector of 1s, assume

$$C = \begin{pmatrix} \mathbf{1} & 0 & \ldots & 0 \\ 0 & \mathbf{1} & 0 & \ldots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & 0 & \mathbf{1} \end{pmatrix} \tag{4.16}$$

(iv) $\boldsymbol{\mu} \sim 1/\sqrt{10}\mathcal{N}(0, I)$.

(v) $\sigma_\epsilon^2 = 1$.

Let $\boldsymbol{\beta} = \sqrt{1-\omega}C\boldsymbol{\mu} + \boldsymbol{\gamma}$, where $\boldsymbol{\gamma} \sim \mathcal{N}(0, \sigma_\gamma^2/pI)$ and $\sigma_\gamma^2 = \omega\boldsymbol{\mu}^\top C^\top \Sigma C\boldsymbol{\mu}$. Fixing C and $\boldsymbol{\mu}$, we vary $\omega = 0.2, 0.3, \ldots, 0.9$. For each setting, we simulate 50 independent datasets, and compute projected least squares and ordinary least squares.

Table 4.3: Mean estimates of out-of-sample error of projected least squares and ordinary least squares. r_{proj}^2 and r_{ols}^2 are the expected out-of-sample error of projected least squares and OLS respectively. Their average risk estimates are $\hat{r}_{proj}^2$ and $\hat{r}_{ols}^2$, follows the Mahalanobis kernel maximum likelihood estimator. $\tilde{r}_{proj}^2$ is the finite sample corrected estimate for $\hat{r}_{proj}^2$.

ω	r_{proj}^2	$\hat{r}_{proj}^2$	$\tilde{r}_{proj}^2$	r_{ols}^2	$\hat{r}_{ols}^2$
0.2	22.1	20.1	21.7	111.4	112.4
0.3	42.8	41.5	42.4	110.8	110.8
0.4	49.7	48.4	49.4	111.8	111.5
0.5	61.7	61.3	62.4	109.5	110.3
0.6	71.0	69.3	70.5	110.3	110.2
0.7	84.4	81.8	83.1	109.7	112.3
0.8	99.7	98.7	100.2	110.9	110.6
0.9	117.1	114.0	115.6	109.5	110.8

Mean of risk estimators computed by Gaussian variance component MLE (4.8) are shown in Table 4.3 and the risk ratio plot in Figure 4.3. Although the design is a sparse block diagonal matrix, Table 4.3 and Figure 4.3 show similar results as in genetic risk prediction (with non-sparse design).

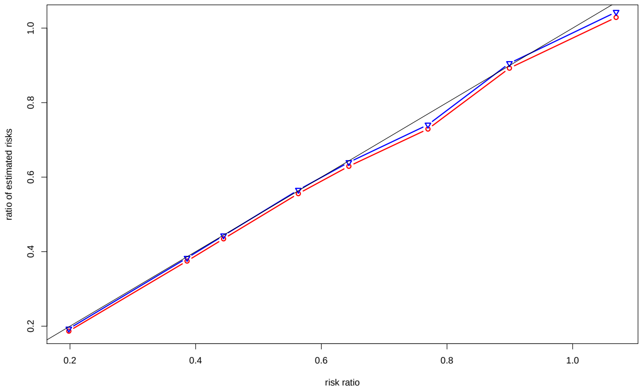

Figure 4.3: For various ω, ratio of risk between projected least squares and ordinary least squares r^2_{proj}/r^2_{ols} is plotted against the ratio of average risk estimates $\hat{r}^2_{proj}/\hat{r}^2_{ols}$ in red line with dots. r^2_{proj}/r^2_{ols} is plotted against $\tilde{r}^2_{proj}/\hat{r}^2_{ols}$ in blue line with dots. The black line is a diagonal reference line.

For the next experiment, we fix $\omega = 0.3$ and $\sigma_e = 0.5$, and vary the number of observations from 1600 to 3000 in increments of 200s. Keeping all other settings the same, we simulate 50 independent datasets with fixed $\boldsymbol{\beta}$. Similar patterns are observed in Table 4.4 and Figure 4.4 below.

Table 4.4: Mean estimates of out-of-sample error of projected least squares and ordinary least squares. r^2_{proj} and r^2_{ols} are the expected out-of-sample error of projected least squares and OLS respectively. Their average risk estimates are $\hat{r}^2_{proj}$ and $\hat{r}^2_{ols}$, follows the Mahalanobis kernel maximum likelihood estimator. $\tilde{r}^2_{proj}$ is the finite sample corrected estimate for $\hat{r}^2_{proj}$.

n	r^2_{proj}	$\hat{r}^2_{proj}$	$\tilde{r}^2_{proj}$	r^2_{ols}	$\hat{r}^2_{ols}$
1600	38.2	37.1	37.9	98.9	100.7
1800	38.1	36.5	37.1	71.1	71.2
2000	38.1	37.8	38.4	55.1	54.5
2200	38.0	36.8	37.3	44.6	45.5
2400	38.0	37.6	38.1	38.1	38.0
2600	38.0	36.7	37.2	32.8	33.2
2800	37.9	37.5	38.0	29.7	29.7
3000	37.9	37.6	38.0	26.0	26.3

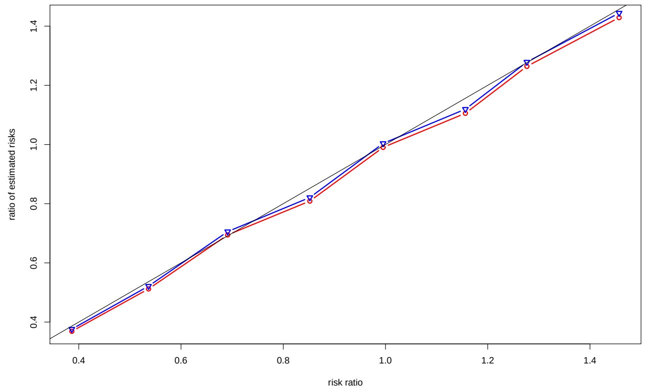

Figure 4.4: For various n, ratio of risk between projected least squares and ordinary least squares r^2_{proj}/r^2_{ols} is plotted against the ratio of average risk estimates $\hat{r}^2_{proj}/\hat{r}^2_{ols}$ in red line with dots. r^2_{proj}/r^2_{ols} is plotted against $\tilde{r}^2_{proj}/\hat{r}^2_{ols}$ in blue line with dots. The black line is a diagonal reference line.

In the following section, we would like to implement projected least squares for with times series data. We simulate the e-commerce data according to the following VAR model. For $t = k+1, \ldots, T$, let

$$\boldsymbol{d}_t = \sum_{j=1}^{k} \boldsymbol{b}_{t-j}^\top \boldsymbol{d}_{t-j} + \boldsymbol{e}_t, \quad k < T \tag{4.17}$$

where $\boldsymbol{d}_t$ is the m-dimensional demand vector at time t and $\boldsymbol{b}_j$ is a coefficient vector. We assume that $\boldsymbol{e}_t \sim \mathcal{N}(0, \sigma_e^2 I)$ are noise vectors for $t = 1, \ldots, T$. Let $n = T - k$, $D = (\boldsymbol{d}_1, \ldots, \boldsymbol{d}_T)$, $B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_k)^\top$ and $E = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)^\top$, then we can write an individual linear model for each product $a \in [m]$. Let $\mathbf{y}_a = X_a \boldsymbol{\beta}_a + \boldsymbol{\epsilon}_a$, where

$$\begin{aligned}
\mathbf{y}_a &= (D_{k+1,a}, D_{k+2,a}, \ldots, D_{n,a})^\top \\
X_a &= \begin{pmatrix} D_{1,a} & \ldots & D_{k,a} \\ D_{2,a} & \ldots & D_{k+1,a} \\ \vdots & \ddots & \vdots \\ D_{n,a} & \ldots & D_{n+k-1,a} \end{pmatrix} \\
\boldsymbol{\beta}_a &= B_{a\cdot} \\
\boldsymbol{\epsilon}_a &= E_{a\cdot}
\end{aligned}$$

In a stacked form, (4.17) is equivalent to

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.18}$$

with

$$\begin{aligned}
\mathbf{y} &= (\mathbf{y}_1^\top, \ldots, \mathbf{y}_m^\top)^\top, \\
X &= \begin{pmatrix} X_1 & 0 & \ldots & 0 \\ 0 & X_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & \ldots & X_m \end{pmatrix}, \\
\boldsymbol{\beta} &= (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_m^\top)^\top, \\
\boldsymbol{\epsilon} &= (\boldsymbol{\epsilon}_1^\top, \ldots, \boldsymbol{\epsilon}_m^\top).
\end{aligned}$$

In the experiment, we let

(i) $m = 100$, $n = 20$, $k = 10$.

(ii) $\mu_1, \ldots, \mu_k \overset{\text{iid}}{\sim} \frac{1}{k}\text{Unif}(0.3, 1)$.

(iii) $\sigma_e^2 = 1$.

For $i = 1, \ldots, k$, let $\boldsymbol{b}_i = \sqrt{1-\omega}\boldsymbol{\mu} + \boldsymbol{\gamma}_i$, where $\boldsymbol{\gamma}_i \sim \mathcal{N}(0, \sigma_\gamma^2/pI)$ and $\sigma_\gamma^2 = m\omega\boldsymbol{\mu}^\top\boldsymbol{\mu}$. We vary $\omega = 0.2, 0.3, \ldots, 0.9$. Projected least squares given (4.16) and ordinary least squares are fitted for the linear model (4.18). Estimated asymptotic risks for two approaches are listed in Table 4.5.

Table 4.5: Mean estimates of out-of-sample error of projected least squares and ordinary least squares. r^2_{proj} and r^2_{ols} are the expected out-of-sample error of projected least squares and OLS respectively. Their average risk estimates (assuming additive white Gaussian noise) are $\hat{r}^2_{proj}$ and $\hat{r}^2_{ols}$, follows the Mahalanobis kernel maximum likelihood estimator. $\tilde{r}^2_{proj}$ is the finite sample corrected estimate for $\hat{r}^2_{proj}$.

ω	r^2_{proj}	$\hat{r}^2_{proj}$	$\tilde{r}^2_{proj}$	r^2_{ols}	$\hat{r}^2_{ols}$
0.2	1.57	0.98	1.48	76.32	105.60
0.3	2.17	1.12	1.63	77.26	107.15
0.4	2.59	1.60	2.12	77.51	107.21
0.5	3.65	2.64	3.17	76.97	106.71
0.6	3.51	2.62	2.79	76.22	108.75
0.7	4.05	2.83	3.37	77.01	109.18
0.8	4.59	3.52	4.06	77.06	110.33
0.9	5.31	4.15	4.70	76.96	108.96

In Table 4.5, both r^2_{proj} and r^2_{ols} are the average out-of-sample error of 50 replicates. Because both methods belong to the class of ordinary least squares estimation, they are misspecified when responses are correlated. As a result, the correlation in noise causes the difference between empirical risk and estimand of risk estimators, and all estimators show bias in estiating risk of least squares with correlated error. Although Gaussian variance component MLE assumes additive white Gaussian noise, the estimated σ_e^2 and σ_γ^2 remain unbiased. r^2_{ols} has an unexpected large upward bias, we suspect it is due to the limited sample size.

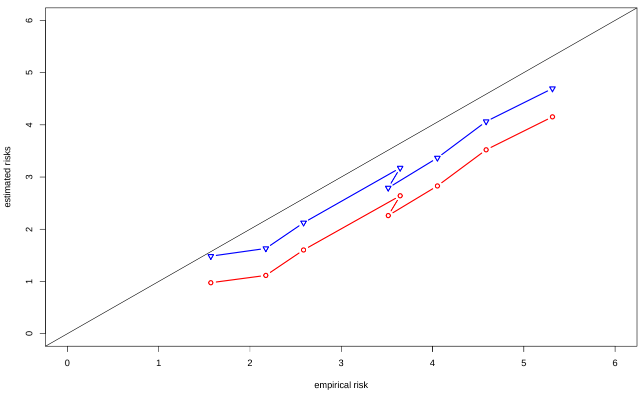

Figure 4.5: For various ω, empirical risk of projected least squares r^2_{proj} is plotted against the average risk estimates $\hat{r}^2_{proj}$in red line with dots, and plotted against $\tilde{r}^2_{proj}$ in blue line with dots. The black line is a diagonal reference line.

Figure 4.5 plots empirical and estimated risks for projected least squares. We omit ordinary least squares since it is unreliable with limited sample size. We observe that the portion of bias is relative stable for various ω.

The final experiment assumes $\omega = 0.3$ and vary n from 1600 to 3000 increments of 200s. The distributional statistics of estimated risks are reported in Table 4.6.

From Table 4.6, we observed that empirical risk r^2_{proj} decreases gradually as n increases. Moreover, the difference (unusualy pattern) between r^2_{ols} and $\hat{r}^2_{ols}$ is smaller as sample size increase. We conjecture that empirical risk for OLS-type estimation is very sensitive to sample size when responses are actually correlated, which suggests that the unexpected pattern for OLS we observed in the previous setting is due to insufficient sample size.

In real e-commerce data analysis, m is usually in units of millions. By stacking the predictors, we could essentially create an ideal situation where $k/n \to 0$. We expect that

Table 4.6: Mean estimates of out-of-sample error of projected least squares and ordinary least squares. r^2_{proj} and r^2_{ols} are the expected out-of-sample error of projected least squares and OLS respectively. Their average risk estimates are $\hat{r}^2_{proj}$ and $\hat{r}^2_{ols}$, follows the Mahalanobis kernel maximum likelihood estimator. $\tilde{r}^2_{proj}$ is the finite sample corrected estimate for $\hat{r}^2_{proj}$.

n	r^2_{proj}	$\hat{r}^2_{proj}$	$\tilde{r}^2_{proj}$	r^2_{ols}	$\hat{r}^2_{ols}$
1600	2.28	1.35	2.00	123.70	191.92
1800	2.24	1.17	1.75	94.52	137.42
2000	2.11	1.03	1.54	77.79	106.60
2200	2.08	0.92	1.39	65.38	87.76
2400	2.08	1.18	1.61	57.38	78.88
2600	2.02	1.19	1.59	50.92	64.87
2800	2.02	1.45	1.82	47.17	57.32
3000	1.99	1.07	1.41	42.04	51.21

in real data analysis the risk estimation for projection least squares becomes unbiased because of sufficiently large sample size.

4.7 Appendix

4.7.1 Proof of Lemma 3

Lemma 3. *Let $n, p \to \infty$ and $p/n \to \rho \in (0,1)$, assume condition (4.4) holds, then*

$$\lim_{n\to\infty} \|\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta}\|_\Sigma^2 = \frac{\sigma_\epsilon^2 \rho}{1-\rho}.$$

Proof. Let $Z = X\Sigma^{-1/2}$ and $Z = U\Lambda V^\top$ be its singular value decomposition, then

$$\begin{aligned} l\left(\hat{\boldsymbol{\beta}}_{ols}\right) &= \|(X^\top X)^{-1}X^\top \mathbf{y} - \boldsymbol{\beta}\|_\Sigma^2 \\ &= \|(X^\top X)^{-1}X^\top (X\boldsymbol{\beta} + \boldsymbol{\epsilon}) - \boldsymbol{\beta}\|_\Sigma^2 \\ &= \|(X^\top X)^{-1}X^\top \boldsymbol{\epsilon}\|_\Sigma^2 \\ &= \|(\Sigma^{1/2}Z^\top Z\Sigma^{1/2})^{-1}\Sigma^{1/2}Z^\top \boldsymbol{\epsilon}\|_\Sigma^2 \\ &= \|(Z^\top Z)^{-1}Z^\top \boldsymbol{\epsilon}\|_2^2 \\ &= \boldsymbol{\epsilon}^\top U\Lambda V^\top (V\Lambda^2 V^\top)^{-2} V\Lambda U^\top \boldsymbol{\epsilon} \\ &= \boldsymbol{\epsilon}^\top U\Lambda^{-2}U^\top \boldsymbol{\epsilon} \end{aligned}$$

Let $\tilde{\boldsymbol{\epsilon}} = (\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_p)^\top = U^\top \boldsymbol{\epsilon}$, then $\tilde{\epsilon}_j \overset{i.i.d.}{\sim} N(0, \sigma_\epsilon^2)$ because of orthogonal invariance of multivariate Gaussian distribution. Let $\tilde{\Lambda} = \mathrm{diag}(\tilde{\lambda}_1, \ldots, \tilde{\lambda}_p) = \Lambda^{-2}$. Then

$$l\left(\hat{\boldsymbol{\beta}}^{(ols)}\right) = \sum_{j=1}^p \tilde{\lambda}_j \tilde{\epsilon}_j^2$$

By (Etemadi, 2006),

$$\frac{1}{\sum_{j=1}^p \tilde{\lambda}_j} \sum_{j=1}^p \tilde{\lambda}_j \tilde{\epsilon}_j^2 \overset{a.s.}{\to} \mathbb{E}(\tilde{\epsilon}_j^2).$$

Let $W = Z^\top Z$, then $W \sim \text{Wishart}(n, I_p)$ and $\mathrm{tr}(W^{-1}) = \sum_{j=1}^p \tilde{\lambda}_j$. By (von Rosen, 1988),

$$\mathbb{E}\mathrm{tr}(W^{-1}) = \mathrm{tr}(\mathbb{E}(W^{-1})) = \mathrm{tr}\left(\frac{1}{n-p-1} I_p\right) = \frac{p}{n-p-1}$$

$$\begin{aligned} \mathbb{E}(\mathrm{tr}(W^{-1})^2) &= \mathrm{tr}(\mathbb{E}(\mathrm{tr}(W^{-1})W^{-1})) \\ &= \mathrm{tr}(c_1 \mathrm{tr}(I_p) I_p + 2c_2 I_p) \\ &= c_1 p^2 + 2c_2 p \end{aligned}$$

where $c_1 = (n-p-2)c_2$ and $c_2^{-1} = (n-p)(n-p-1)(n-p-3)$. Then

$$\begin{aligned}
\mathrm{Var}(\mathrm{tr}(W^{-1})) &= c_1p^2 + 2c_2p - \frac{p^2}{(n-p-1)^2} \\
&= \frac{(n-p-1)(n-p-2)p^2 + 2p(n-p-1) - p^2(n-p)(n-p-3)}{(n-p)(n-p-1)^2(n-p-3)} \\
&= \frac{2p^2 + 2p(n-p-1)}{(n-p)(n-p-1)^2(n-p-3)} \\
&= \frac{2p(n-1)}{(n-p)(n-p-1)^2(n-p-3)}.
\end{aligned}$$

Then in the large limit,

$$\lim_{n,p\to\infty} \mathrm{Var}(\mathrm{tr}(W^{-1})) = \frac{2np}{(n-p)^4} = \frac{2\rho}{n^2(1-\rho)^4}$$

Since $\rho < 1$, $1/(1-\rho) < \infty$, we thus have $\frac{2np}{(n-p)^4} \to 0$. Therefore, we have shown that $\mathbb{E}(\mathrm{tr}(W^{-1}) - \mathbb{E}(\mathrm{tr}(W^{-1})))^2 \to 0$, then

$$\mathrm{tr}(W^{-1}) \to \frac{\rho}{1-\rho}$$

Therefore, by Slutsky's theorem,

$$\boldsymbol{\epsilon}^\top U \Lambda^{-2} U^\top \boldsymbol{\epsilon} \to \frac{\sigma_\epsilon^2 \rho}{1-\rho}.$$

□

4.7.2 Proof of Lemma 4

Lemma 4. *Assume that the linear model* (4.1) *holds, that* $\boldsymbol{\beta} \in \mathbb{R}^p$ *is a fixed vector. Suppose that* $\mathbf{z}_i = (y_i, X_{i\cdot})^\top$ *and*

$$\mathbf{z}_i \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \begin{pmatrix} \boldsymbol{\beta}^\top\Sigma\boldsymbol{\beta} + \sigma_\epsilon^2 & \boldsymbol{\beta}^\top\Sigma \\ \Sigma\boldsymbol{\beta} & \Sigma \end{pmatrix}\right).$$

Let $C \in \mathbb{R}^{p\times k}$*,* $k/p \to 0$*, and define*

$$\hat{\boldsymbol{\beta}}_{proj} := C\hat{\boldsymbol{\mu}}_{proj} = C(C^\top X^\top XC)^{-1}C^\top X^\top \mathbf{y}.$$

Then,

$$\lim_{n\to\infty} \|\boldsymbol{\beta}_{proj} - \boldsymbol{\beta}\|_\Sigma^2 = \boldsymbol{\gamma}^\top\Sigma\boldsymbol{\gamma}.$$

Proof. The loss of $\hat{\boldsymbol{\beta}}_{proj}$ is

$$\begin{aligned} l\left(\hat{\boldsymbol{\beta}}_{proj}\right) &= \|C(C^\top X^\top XC)^{-1}C^\top X^\top \mathbf{y} - \boldsymbol{\beta}\|_\Sigma^2 \\ &= \|C(C^\top X^\top XC)^{-1}C^\top X^\top (XC\boldsymbol{\mu} + X\boldsymbol{\gamma} + \boldsymbol{\epsilon}) - C\boldsymbol{\mu} - \boldsymbol{\gamma}\|_\Sigma^2 \\ &= \|C(C^\top X^\top XC)^{-1}C^\top X^\top (X\boldsymbol{\gamma} + \boldsymbol{\epsilon}) - \boldsymbol{\gamma}\|_\Sigma^2 \end{aligned}$$

Let P_C be a $p \times k$ matrix with orthonormal columns such that $\Sigma^{1/2}C(C^\top \Sigma C)^{-1}C^\top \Sigma^{1/2} = P_C P_C^\top$. Let $P_{C^\perp}$ be a corresponding $p \times (p-k)$ matrix with orthonormal columns satisyfing $P_C^\top P_{C^\perp} = 0$ and $\mathbf{I} = P_C P_C^\top + P_{C^\perp} P_{C^\perp}^\top$. Then

$$\begin{aligned} & \|\hat{\boldsymbol{\beta}}_{proj} - \boldsymbol{\beta}\|_\Sigma^2 \\ = \; & \|C(C^\top X^\top XC)^{-1}C^\top X^\top (X\Sigma^{-1/2}(P_C P_C^\top + P_{C^\perp} P_{C^\perp}^\top)\Sigma^{1/2}\boldsymbol{\gamma} + \boldsymbol{\epsilon}) - \boldsymbol{\gamma}\|_\Sigma^2 \\ = \; & \|C(C^\top X^\top XC)^{-1}C^\top X^\top (X\Sigma^{-1/2}(P_{C^\perp} P_{C^\perp}^\top)\Sigma^{1/2}\boldsymbol{\gamma} + \boldsymbol{\epsilon}) + C(C^\top \Sigma C)^{-1}\Sigma\boldsymbol{\gamma} - \boldsymbol{\gamma}\|_\Sigma^2 \\ = \; & \|\Sigma^{1/2}C(C^\top X^\top XC)^{-1}C^\top X^\top (X\Sigma^{-1/2}(P_{C^\perp} P_{C^\perp}^\top)\Sigma^{1/2}\boldsymbol{\gamma} + \boldsymbol{\epsilon})\|_2^2 + \|P_{C^\perp}^\top \Sigma^{1/2}\boldsymbol{\gamma}\|_2^2. \end{aligned}$$

Here the first term is noise variation in estimating $\boldsymbol{\beta}$, the second term is caused by the bias we introduced when projecting the design matrix, the cross-term is zero because $\|P_{C^\perp}^\top \Sigma^{1/2} C\|_F^2 = 0$. We further decompose the noise variation:

$$\begin{aligned} & \|\Sigma^{1/2}C(C^\top X^\top XC)^{-1}C^\top X^\top (X\Sigma^{-1/2}P_{C^\perp} P_{C^\perp}^\top \Sigma^{1/2}\boldsymbol{\gamma} + \boldsymbol{\epsilon})\|_2^2 \\ = \; & \|\Sigma^{1/2}C(C^\top \Sigma C)^{-1/2}(Z_C^\top Z_C)^{-1}Z_C\boldsymbol{\epsilon}'\|_2^2 \\ = \; & \|(Z_C^\top Z_C)^{-1}Z_C\boldsymbol{\epsilon}'\|_2^2. \end{aligned}$$

where $Z_C = XC(C^\top \Sigma C)^{-1/2}$ is a Gaussian random matrix, and $\boldsymbol{\epsilon}' = X\Sigma^{-1/2}P_{C^\perp} P_{C^\perp}^\top \Sigma^{1/2}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$. We have Z_C independent of $\boldsymbol{\epsilon}'$ because $\|P_{C^\perp}^\top \Sigma^{1/2}C\|_F^2 = 0$. We can apply previous technique after finding the first two moments of $\boldsymbol{\epsilon}'$. Since $\boldsymbol{\epsilon}_i' \overset{i.i.d.}{\sim} \mathcal{N}(0, \|P_{C^\perp}^\top \Sigma^{1/2}\boldsymbol{\gamma}\|_2^2 + \sigma_\epsilon^2)$, we have

$$\|(Z_C^\top Z_C)^{-1}Z_C\boldsymbol{\epsilon}_{C^\perp}\|_2^2 \to \frac{k}{n-k-1}(\|P_{C^\perp}^\top \Sigma^{1/2}\boldsymbol{\gamma}\|_2^2 + \sigma_\epsilon^2).$$

Therefore,

$$\begin{aligned} \lim_{n\to\infty} \|\hat{\boldsymbol{\beta}}_{proj} - \boldsymbol{\beta}\|_\Sigma^2 &= \lim_{n\to\infty} \frac{n-1}{n-k}\|P_{C^\perp}^\top \Sigma^{1/2}\boldsymbol{\gamma}\|_2^2 + \frac{k}{n-k-1}\sigma_\epsilon^2 \\ &= \|P_{C^\perp}^\top \Sigma^{1/2}\boldsymbol{\gamma}\|_2^2 \end{aligned}$$

Moreover,

$$\begin{aligned}
\|P_{C^\perp}^\top \Sigma^{1/2}\boldsymbol{\gamma}\|_2^2 &= \boldsymbol{\gamma}^\top \Sigma \boldsymbol{\gamma} - \boldsymbol{\gamma}^\top \Sigma^{1/2} P_C P_C^\top \Sigma^{1/2} \boldsymbol{\gamma} \\
&= \boldsymbol{\gamma}^\top \Sigma \boldsymbol{\gamma} - \|P_C^\top \Sigma^{1/2}(\Sigma^{-1/2} P_{C^\perp} P_{C^\perp}^\top \Sigma^{1/2} \boldsymbol{\beta})\|_2^2 \\
&= \boldsymbol{\gamma}^\top \Sigma \boldsymbol{\gamma}.
\end{aligned}$$

□

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control* **19** 716–723.

Akdemir, D. and Jannink, J.-L. (2015). Locally epistatic genomic relationship matrices for genomic association and prediction. *Genetics* **199** 857–871.

Bai, Z., Miao, B., Pan, G. et al. (2007). On asymptotics of eigenvectors of large sample covariance matrix. *The Annals of Probability* **35** 1532–1572.

Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., Julier, C., Morahan, G., Nerup, J., Nierras, C. et al. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature genetics* **41** 703.

Bayati, M., Lelarge, M. and Montanari, A. (2012). Universality in polytope phase transitions and iterative algorithms. In *Proc. IEEE Int. Symp. Inform. Thy., (Cambridge, MA)*. IEEE.

Bayati, M. and Montanari, A. (2012). The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory* **58** 1997–2017.

Bonnet, A., Gassiat, E., Lévy-Leduc, C. et al. (2015). Heritability estimation in high dimensional sparse linear mixed models. *Electronic Journal of Statistics* **9** 2099–2129.

Boulesteix, A.-L. and Strimmer, K. (2006). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in bioinformatics* **8** 32–44.

Breiman, L. and Freedman, D. (1983). How many variables should be entered in a regression equation? *Journal of the American Statistical Association* **78** 131–136.

Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J. C. et al. (2009). The genetic architecture of maize flowering time. *Science* **325** 714–718.

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., Neale, B. M., of the Psychiatric Genomics Consortium, S. W. G. et al. (2015). Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics* **47** 291–295.

Consortium, I. S. et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460** 748.

Davis, L. K., Yu, D., Keenan, C. L., Gamazon, E. R., Konkashbaev, A. I., Derks, E. M., Neale, B. M., Yang, J., Lee, S. H., Evans, P. et al. (2013). Partitioning the heritability of tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. *PLoS Genet* **9** e1003864.

de Vlaming, R. and Groenen, P. J. (2015). The current and future use of ridge regression for prediction in quantitative genetics. *BioMed research international* **2015**.

Dicker, L. H. (2013). Optimal equivariant prediction for high-dimensional linear models with arbitrary predictor covariance. *Electronic Journal of Statistics* **7** 1806–1834.

Dicker, L. H. (2014). Variance estimation in high-dimensional linear models. *Biometrika* **101** 269–284.

Dicker, L. H. (2016). Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli* **22** 1–37.

Dicker, L. H. and Erdogdu, M. A. (2016a). Flexible results for quadratic forms with applications to variance components estimation. *Ann. Stat.* To appear.

Dicker, L. H. and Erdogdu, M. A. (2016b). Maximum likelihood for variance estimation in high-dimensional linear models. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics.*

Dicker, L. H. and Zhao, S. D. (2016). High-dimensional classification via nonparametric empirical Bayes and maximum likelihood inference. *Biometrika* **103** 21–34.

Donoho, D. L., Maleki, A. and Montanari, A. (2009). Message-passing algorithms for compressed sensing. *P. Natl. Acad. Sci. USA* **106** 18914–18919.

Donoho, D. L., Maleki, A. and Montanari, A. (2010a). Message passing algorithms for compressed sensing: I. motivation and construction. In *Proc. Inform. Theory Workshop (Cairo, Egypt)*. IEEE.

Donoho, D. L., Maleki, A. and Montanari, A. (2010b). Message passing algorithms for compressed sensing: Ii. analysis and validation. In *Proc. Inform. Theory Workshop (Cairo, Egypt)*. IEEE.

Donoho, D. L., Maleki, A. and Montanari, A. (2011). The noise-sensitivity phase transition in compressed sensing. *IEEE T. Inform. Theory* **57** 6920–6941.

Etemadi, N. (2006). Convergence of weighted averages of random variables revisited. *Proceedings of the American Mathematical Society* **134** 2739–2744.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2007). *Regression*. Springer.

Falconer, D. S. (1960). *Introduction to quantitative genetics*. Oliver And Boyd; Edinburgh; London.

Feng, L., Ma, R. and Dicker, L. H. (2017). Nonparametric maximum likelihood approximate message passing. In *Information Sciences and Systems (CISS), 2017 51st Annual Conference on*. IEEE.

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K. et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics* **47** 1228–1235.

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nature Reviews Genetics* **13** 135.

Gilmour, A. R., Thompson, R. and Cullis, B. R. (1995). Average information reml: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* 1440–1450.

Golan, D., Lander, E. S. and Rosset, S. (2014). Measuring missing heritability: inferring the contribution of common variants. *Proceedings of the National Academy of Sciences* **111** E5272–E5281.

Guo, D. and Wang, C.-C. (2006). Asymptotic mean-square optimality of belief propagation for sparse linear systems. In *Proc. Inform. Theory Workshop (Chengdu, China)*. IEEE.

Guo, D. and Wang, C.-C. (2007). Random sparse linear systems observed via arbitrary channels: A decoupling principle. In *Proc. IEEE Int. Symp. Inform. Thy., (Nice, France)*. IEEE.

Guo, S., Wang, Y. and Yao, Q. (2016). High-dimensional and banded vector autoregressions. *Biometrika* asw046.

Gusev, A., Bhatia, G., Zaitlen, N., Vilhjalmsson, B. J., Diogo, D., Stahl, E. A., Gregersen, P. K., Worthington, J., Klareskog, L., Raychaudhuri, S. et al. (2013). Quantifying missing heritability at known gwas loci. *PLoS genetics* **9** e1003993.

Gusev, A., Lee, S. H., Trynka, G., Finucane, H., Vilhjálmsson, B. J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E. et al. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *The American Journal of Human Genetics* **95** 535–552.

Haseman, J. K. and Elston, R. C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2** 3–19.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. Springer.

Haufe, S., Müller, K.-R., Nolte, G. and Krämer, N. (2008). Sparse causal discovery in multivariate time series. In *Proceedings of the 2008th International Conference on Causality: Objectives and Assessment-Volume 6*. JMLR. org.

Hayes, B. J., Visscher, P. M. and Goddard, M. E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics research* **91** 47–60.

Henderson, C. (1984). Applications of linear models in animal breeding .

Henderson, C. R. (1950). Abstract: Estimation of genetic parameters. *Ann. Math. Stat.* **21** 309–310.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **106** 9362–9367.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

Hotelling, H. (1957). The relations of the newer multivariate statistical methods to factor analysis. *British Journal of Mathematical and Statistical Psychology* **10** 69–79.

Hsu, D., Kakade, S. M. and Zhang, T. (2011). An analysis of random design linear regression. In *Proc. COLT*.

Hsu, N.-J., Hung, H.-L. and Chang, Y.-M. (2008). Subset selection for vector autoregressive processes using lasso. *Computational Statistics & Data Analysis* **52** 3645–3657.

Janson, L., Barber, R. F. and Candes, E. (2017). Eigenprism: inference for high dimensional signal-to-noise ratios. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79** 1037–1065.

Jiang, J. et al. (1996). REML estimation: Asymptotic behavior and related topics. *Ann. Stat.* **24** 255–286.

Jiang, W. and Zhang, C.-H. (2010). Empirical Bayes in-season prediction of baseball batting averages. In *Borrowing Strength: Theory Powering Applications – A Festschrift for Lawrence D. Brown*. Institute of Mathematical Statistics, 263–273.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Applied Statistics* 300–303.

Kang, H. M., Sul, J. H., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C., Eskin, E. et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* **42** 348.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Stat.* **27** 887–906.

Koenker, R. and Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Am. Stat. Assoc.* **109** 674–685.

Kostem, E. and Eskin, E. (2013). Improving the accuracy and efficiency of partitioning heritability into the contributions of genomic regions. *The American Journal of Human Genetics* **92** 558–564.

Krapohl, E., Patel, H., Newhouse, S., Curtis, C., von Stumm, S., Dale, P., Zabaneh, D., Breen, G., O'reilly, P. and Plomin, R. (2017). Multi-polygenic score approach to trait prediction. *Molecular psychiatry* .

Kriegel, H.-P., Kröger, P. and Zimek, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)* **3** 1.

Lee, P. H., Baker, J. T., Holmes, A. J., Jahanshad, N., Ge, T., Jung, J.-Y., Cruz, Y., Manoach, D. S., Hibar, D. P., Faskowitz, J. et al. (2016). Partitioning heritability analysis reveals a shared genetic basis of brain anatomy and schizophrenia. *Molecular psychiatry* **21** 1680.

Leeb, H. et al. (2009). Conditional predictive inference post model selection. *The Annals of Statistics* **37** 2838–2876.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry, and Applications*. IMS.

Lynch, M., Walsh, B. et al. (1998). *Genetics and analysis of quantitative traits*, vol. 1. Sinauer Sunderland, MA.

Majumdar, A., Witte, J. S. and Ghosh, S. (2015). Semiparametric allelic tests for mapping multiple phenotypes: binomial regression and mahalanobis distance. *Genetic epidemiology* **39** 635–650.

Mark, N. C. and Sul, D. (2011). When are pooled panel-data regression forecasts of exchange rates more accurate than the time-series regression forecasts? *Handbook of Exchange Rates* 265–281.

Meuwissen, H. B., T.H.E. and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157** 1819–1829.

Montanari, A. (2012). Graphical models concepts in compressed sensing. In *Compressed Sensing: Theory and Applications*. Cambridge University Press, 394–438.

Nasrabadi, N. M. (2007). Pattern recognition and machine learning. *Journal of electronic imaging* **16** 049901.

Patterson, N., Price, A. L. and Reich, D. (2006). Population structure and eigenanalysis. *PLoS genetics* **2** e190.

Powell, J. E., Visscher, P. M. and Goddard, M. E. (2010). Reconciling the analysis of ibd and ibs in complex trait studies. *Nature Reviews Genetics* **11** 800.

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *The American Journal of Human Genetics* **69** 124–137.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J. et al. (2007). Plink:

a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81** 559–575.

Rangan, S. (2010). Estimation with random linear mixing, belief propagation and compressed sensing. In *Proc. Conf. Inform. Science & Syst., (Princeton, NJ)*. IEEE.

Rangan, S. (2011). Generalized approximate message passing for estimation with random linear mixing. In *Proc. IEEE Int. Symp. Inform. Thy., (Saint Petersburg, Russia)*. IEEE.

Rangan, S., Schniter, P. and Fletcher, A. (2014). On the convergence of approximate message passing with arbitrary matrices. In *Proc. IEEE Int. Symp. Inform. Thy., (Honolulu, USA)*. IEEE.

Rangan, S., Schniter, P. and Fletcher, A. K. (2017). Vector approximate message passing. In *Information Theory (ISIT), 2017 IEEE International Symposium on*. IEEE.

Robbins, H. E. (1950). A generalization of the method of maximum likelihood: Estimating a mixing distribution (abstract). *Ann. Math. Stat.* **21** 314–315.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics* **6** 461–464.

Selzam, S., Krapohl, E., von Stumm, S., O'reilly, P., Rimfeld, K., Kovas, Y., Dale, P. S., Lee, J. and Plomin, R. (2017). Predicting educational achievement from dna. *Molecular psychiatry* **22** 267.

Speed, D. and Balding, D. J. (2015). Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics* **16** 33.

Speed, D., Hemani, G., Johnson, M. R. and Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91** 1011–1021.

Stahl, E. A., Raychaudhuri, S., Remmers, E. F., Xie, G., Eyre, S., Thomson, B. P., Li, Y., Kurreeman, F. A., Zhernakova, A., Hinks, A. et al. (2010).

Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nature genetics* **42** 508.

Stahl, E. A., Wegmann, D., Trynka, G., Gutierrez-Achury, J., Do, R., Voight, B. F., Kraft, P., Chen, R., Kallberg, H. J., Kurreeman, F. A. et al. (2012). Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature genetics* **44** 483.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B Met.* **58** 267–288.

Tikhonov, N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.* **5** 1035–1038.

Vila, J. P. and Schniter, P. (2013). Expectation-maximization gaussian-mixture approximate message passing. *IEEE T. Signal Proces.* **61** 4658–4672.

von Rosen, D. (1988). Moments for the inverted wishart distribution. *Scandinavian Journal of Statistics* 97–109.

Wold, H. (1975). Soft modelling by latent variables: the non-linear iterative partial least squares (nipals) approach. *Journal of Applied Probability* **12** 117–142.

Wold, S., Ruhe, A., Wold, H. and Dunn, W., III (1984). The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing* **5** 735–743.

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E. and Visscher, P. M. (2013). Pitfalls of predicting complex traits from snps. *Nature Reviews Genetics* **14** 507.

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A., Lee, S. H., Robinson, M. R., Perry, J. R., Nolte, I. M., van Vliet-Ostaptchouk, J. V. et al. (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature genetics* **47** 1114.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E. and Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42** 565–569.

Yang, J., Lee, S. H., Goddard, M. E. and Visscher, P. M. (2011a). Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88** 76–82.

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., De Andrade, M., Feenstra, B., Feingold, E., Hayes, M. G. et al. (2011b). Genome partitioning of genetic variation for complex traits using common snps. *Nature genetics* **43** 519.

Zaitlen, N. A. and Kraft, P. (2012). Heritability in the genome-wide association era. *Hum. Genet.* **131** 1655–1664.

Zhang, C., Bengio, S., Hardt, M., Recht, B. and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530* .