Scalable Algorithms for Learning High-Dimensional Linear Mixed Models

Zilong Tan
Duke University
ztan@cs.duke.edu

Kimberly Roche
Duke University
kimberly.roche@duke.edu

Xiang Zhou
University of Michigan
xzhousph@umich.edu

Sayan Mukherjee
Duke University
sayan@stat.duke.edu

Abstract

Linear mixed models (LMMs) are used extensively to model observations that are not independent. Parameter estimation for LMMs can be computationally prohibitive on big data. State-of-the-art learning algorithms require computational complexity which depends at least linearly on the dimension p of the covariates, and often use heuristics that do not offer theoretical guarantees. We present scalable algorithms for learning high-dimensional LMMs with sublinear computational complexity dependence on p. Key to our approach are novel dual estimators which use only kernel functions of the data, and fast computational techniques based on the subsampled randomized Hadamard transform. We provide theoretical guarantees for our learning algorithms, demonstrating the robustness of parameter estimation. Finally, we complement the theory with experiments on large synthetic and real data.

1 INTRODUCTION

Linear mixed models (LMMs) are widely used in many real world applications ranging from longitudinal data analysis (Laird and Ware, 1982; Demidenko, 2013) and genome wide association studies (Kang et al., 2008; Lippert et al., 2011; Zhou, 2017) to recommender systems (Zhang et al., 2016). LMMs provide a flexible framework for modeling a wide range of data types, including clustered, longitudinal, and spatial data. Parameter estimation for LMMs is computationally prohibitive for big data, both for large sample size n (Zhou and Stephens, 2014; Darnell et al., 2017; Perry, 2017) and for high-dimensional covariates p (Schelldorfer et al., 2011). The main computational bottlenecks for parameter estimation arise from the non-convexity of the optimization problem (Kang et al., 2008; Perry, 2017) as well as the computational cost of matrix inversions (Zhou, 2017; Laird et al., 1987; Lindstrom and Bates, 1988; Bates et al., 2015). State-of-the-art methods for parameter estimation in LMMs require computational complexity that depends at least linearly on p: (i) $O\left(nkp\right)$ for the setting $n > p$ with a rank k covariance matrix (Zhou, 2017; Darnell et al., 2017); and (ii) $O\left(n^2p\right)$ per iteration for $p \gg n$ (Schelldorfer et al., 2011, 2014; Jakubík, 2015). In this paper, we present scalable algorithms with sublinear computational complexity in p, making the proposed approach useful for high-dimensional LMMs. In addition, we provide a theoretical analysis for our approach that states provable error guarantees between the estimated and ground-truth parameters.

Two sets of parameters are estimated in LMMs, the fixed-effects coefficients and the variances for the unobservable random effects and noise. The random-effects variance is generally assumed to have a certain structure, such as a block-diagonal matrix (Laird and Ware, 1982; Demidenko, 2013). To estimate both sets of parameters, an expectation maximization (EM) algorithm is typically used (Laird et al., 1987; Bates et al., 2015) to handle the latent random-effect variable. The M-step in the EM algorithm incurs high computational costs due to matrix inversions. Newton-Raphson has been used to reduce the number of iterations required for parameter estimates to converge (Lindstrom and Bates, 1988); however, each iteration is still costly due to matrix inversions. A recent research focus is to avoid matrix inversions at each iteration. For instance, when $n > p$ a spectral algorithm is available (Kang et al., 2008; Lippert et al., 2011; Patterson and Thompson, 1971). The state-of-the-art algorithm (Darnell et al., 2017) further improved the computational complexity of the spectral algorithm using randomized singular value decomposition (Darnell et al., 2017).

While approximate learning algorithms (Zhou, 2017;

Darnell et al., 2017) are efficient, few provide provable guarantees in terms of estimation accuracy. Recently, a guaranteed non-iterative algorithm was proposed in (Perry, 2017), which runs in $O\left(n\left(p+d\right)^4\right)$ time for d random effects. Inference with guarantees for high-dimensional LMMs, i.e., $p \gg n$, typically incurs greater computational complexity due to the regularization required to address high-dimensional data (Schelldorfer et al., 2011, 2014). In the high-dimensional setting, most algorithms perform block coordinate descent with an $O\left(n^2 p\right)$ per-iteration cost (Schelldorfer et al., 2011, 2014). In this paper, we show that efficiency and provable guarantees can be achieved simultaneously for learning high-dimensional LMMs.

There are two key ideas we use in our efficient algorithms. The first idea is to propose an approximate estimator that relies on an $n \times n$ kernel matrix (§ 3) which can be computed efficiently using the subsampled randomized Hadamard transform (SRHT) (Tropp, 2011). This reduces the linear complexity dependence on p. Unlike some other approximation algorithms (Lu et al., 2013), the proposed estimator also has the advantage of recovering the fixed-effects coefficients for all p dimensions as opposed to the reduced dimensions. This allows us to provide effect sizes in terms of the original covariates, a requirement in many applications. The second idea is the introduction of *approximate variance components* (AVCs) to replace variance components when estimating the fixed-effects coefficients. These AVCs have a closed-form expression and are fast to compute.

We apply our novel approach to LMMs with a both general covariances as well as a block-diagonal covariances for the random effects. The former can be viewed as a special case of the latter with a single block, and has been adopted in genome-wide association studies (Kang et al., 2008; Lippert et al., 2011; Zhou, 2017). LMMs with a block-diagonal covariance structure have been widely used for modeling repeated measures data (Laird and Ware, 1982). We propose a non-iterative algorithm for the general covariance setting and a fast EM variant for the block-diagonal setting.

Contribution Our main contribution is providing a class of approximation algorithms for parameter inference in high-dimensional LMMs with provable guarantees. In Table 1, we state the computational complexity for several standard and state-of-the-art parameter inference algorithms. In the table and in this paper, n is the sample size, p is the number of covariates, k is the rank of the covariance matrix, s are the number of subsamples, and ϵ is the approximation error. Our method is the only one that is sublinear in p, and can be a $n/\log p$ magnitude faster than the others (discussed in § 4.1). In addition to theoretical advantages, we demonstrate the empirical accuracy and speed of our method on both synthetic and real data in § 6.

Table 1: Computational complexity for parameter inference. † denotes that the estimator has provable guarantees.

REML (LIPPERT ET AL., 2011)	$O\left(n^2 p\right)$
†MOMENTS (PERRY, 2017)	$O\left(n\left(p+q\right)^4\right)$
SUBSAMPLING (ZHOU, 2017)	$O\left(ps^2\right)$
RSVD (DARNELL ET AL., 2017)	$O\left(pnk\right)$
†THIS WORK	$O\left(\frac{n^2(k+\log p)\log k}{\epsilon^2}\right)$

Notation We denote the maximum and minimum eigenvalues of a matrix $\boldsymbol{A}$ by $\lambda_{\max}\left(\boldsymbol{A}\right)$ and $\lambda_{\min}\left(\boldsymbol{A}\right)$, respectively. Similarly, we denote the maximum and minimum singular values respectively by $\sigma_{\max}\left(\boldsymbol{A}\right)$ and $\sigma_{\min}\left(\boldsymbol{A}\right)$. $\boldsymbol{A}^\dagger$ represents the Moore–Penrose pseudoinverse of $\boldsymbol{A}$, and $\kappa\left(\boldsymbol{A}\right)$ denotes the condition number of $\boldsymbol{A}$. The superscripted notation $\boldsymbol{y}^{(i)}$ refers to the copy of $\boldsymbol{y}$ for group i. We write the spectral norm of a matrix as $\|\cdot\|_2$, the Frobenius norm as $\|\cdot\|_F$, and the Ky Fan k-norm (the sum of the k largest singular values) as $\|\!|\cdot|\!\|_k$.

Organization Section 2 provides the background on standard LMMs. In section 3, we formulate the L_2-regularized LMMs and present approximate estimators based on a kernel matrix. Section 4 describes fast computational techniques for the approximate estimators. In section 5, we provide theoretical guarantees for our estimators. Section 6 reports empirical evidence of the speed and accuracy of our methods, and section 7 concludes this paper.

2 LINEAR MIXED MODELS

Consider a regression problem with n observations, where $\boldsymbol{y} \in \mathbb{R}^n$ denotes the response vector and $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ represents the covariate matrix with p covariates. The standard LMM is given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + c\mathbf{1} + \boldsymbol{e} \quad \text{with} \quad \begin{bmatrix}\boldsymbol{\gamma} \\ \boldsymbol{e}\end{bmatrix} \sim \text{MVN}\left(\mathbf{0}, \begin{bmatrix}\boldsymbol{\Lambda} & \mathbf{0} \\ \mathbf{0} & \sigma^2\boldsymbol{I}\end{bmatrix}\right), \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the fixed-effect coefficient vector, $\boldsymbol{Z} \in \mathbb{R}^{n \times q}$ is a full-rank random-effects design matrix, $\boldsymbol{\gamma} \in \mathbb{R}^q$ is the random-effect coefficient vector, c is the intercept, and $\boldsymbol{e} \in \mathbb{R}^n$ is the noise vector. The parameters to be estimated are the fixed-effects coefficients $\boldsymbol{\beta}$, and variance components $\boldsymbol{\Lambda}$ and σ^2.

In general, the variables $\boldsymbol{X}$, $\boldsymbol{y}$, $\boldsymbol{\gamma}$, and $\boldsymbol{e}$ in (1) correspond to observations from m classes, and are grouped by the following structure (Laird and Ware, 1982):

$$\begin{bmatrix} \boldsymbol{X}^{(1)} \\ \boldsymbol{X}^{(2)} \\ \vdots \\ \boldsymbol{X}^{(m)} \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{y}^{(1)} \\ \boldsymbol{y}^{(2)} \\ \vdots \\ \boldsymbol{y}^{(m)} \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{\gamma}^{(1)} \\ \boldsymbol{\gamma}^{(2)} \\ \vdots \\ \boldsymbol{\gamma}^{(m)} \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{e}^{(1)} \\ \boldsymbol{e}^{(2)} \\ \vdots \\ \boldsymbol{e}^{(m)} \end{bmatrix},$$

where $\cdot^{(i)}$ denote the variables specific to group i, whose dimensions are $\boldsymbol{X}^{(i)} \in \mathbb{R}^{n_i \times p}$, $\boldsymbol{\gamma}^{(i)} \in \mathbb{R}^{d}$, and $\boldsymbol{y}^{(i)}, \boldsymbol{e}^{(i)} \in \mathbb{R}^{n_i}$, $\sum_{i=1}^{m} n_i = n$. The LMM assumes that $\boldsymbol{\gamma}^{(i)}$ corresponding to distinct classes are independent. In particular, the random-effects design matrix $\boldsymbol{Z}$ and the random-effects covariance are block-diagonal

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}^{(1)} & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{Z}^{(m)} \end{bmatrix}, \quad \boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{H} & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & \boldsymbol{H} \end{bmatrix}$$

with $\boldsymbol{Z}^{(i)} \in \mathbb{R}^{n_i \times d}$, $\boldsymbol{H} \in \mathbb{R}^{d \times d}$, and $q = md$.

Computational challenges Parameter inference in LMMs aims to accurately recover $\mathcal{P} \coloneqq \{\boldsymbol{\beta}, \boldsymbol{\Lambda}, \sigma^2\}$ from $\{\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z}\}$. This is straightforward if $\boldsymbol{\Lambda}$ is given. When $\boldsymbol{\Lambda}$ is unknown, inference can be computationally challenging even in the standard setting where $n > p$ (Laird and Ware, 1982; Lippert et al., 2011; Zhou, 2017; Laird et al., 1987; Lindstrom and Bates, 1988; Patterson and Thompson, 1971; Zhang et al., 2011).

First, parameter estimation problem is non-convex for both maximum likelihood and restricted maximum likelihood (REML) (Laird et al., 1987; Patterson and Thompson, 1971; Harville, 1974). For instance, the methods using REML (Kang et al., 2008; Lippert et al., 2011) project the data onto two uncorrelated parts, and then estimate the fixed-effects and variance components separately on each part. This has the advantage of giving unbiased estimates of the variance components. However, the REML likelihood function is a non-convex function which involves the eigenvalues of the variance of the projected data (Patterson and Thompson, 1971).

Second, regularization is typically required to support the high-dimensional setting, which adds further computational overheads (Lippert et al., 2011; Zhou, 2017; Schelldorfer et al., 2011, 2014; Jakubík, 2015). To address these challenges, we develop novel approximate estimators that are efficient to compute (§ 4), and have provable accuracy guarantees (§ 5).

3 APPROXIMATE ESTIMATORS FOR HIGH-DIMENSIONAL LMMS

In this section, we consider an L_2-regularized LMM to support the high-dimensional setting $p > n$, and develop efficient approximate estimators for the parameters.

Standard parameter estimation algorithms for LMMs such as (Kang et al., 2008; Laird et al., 1987; Bates et al., 2015) do not support the high-dimensional setting $p > n$. We consider introducing the L_2 regularization on the fixed-effects coefficients, which can be viewed as adding the prior $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Phi})$. The L_2-regularized LMM has the following log-likelihood

$$\begin{aligned} &\log p(\boldsymbol{y}, \boldsymbol{\beta} \mid \boldsymbol{X}; \boldsymbol{V}) \\ &\propto -\frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{\Phi}^{-1}\boldsymbol{\beta} - \frac{1}{2}\log\det \boldsymbol{V} \\ &\quad -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - c\boldsymbol{1})\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - c\boldsymbol{1}) \end{aligned} \tag{2}$$

with the marginal variance $\boldsymbol{V} \coloneqq \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{Z}^\top + \sigma^2\boldsymbol{I}$.

Parameter estimation of an LMM is typically iterative and computationally prohibitive, especially in the high-dimensional setting (Darnell et al., 2017; Perry, 2017; Schelldorfer et al., 2011). To improve the computational efficiency, we propose dual as well as approximate estimators. These estimators are non-iterative and have reduced computational complexity, as we will show in § 4.

3.1 FIXED-EFFECT COEFFICIENTS

We first derive the estimators for the fixed-effects coefficients $\widehat{\boldsymbol{\beta}}$ and $\widehat{c}$, which are the maximizers of the log-likelihood (2). A dual estimator of $\boldsymbol{\beta}$ is then given for use in the high-dimensional setting. Using the partial derivatives, it is straightforward to show

$$\left(\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{\Phi}^{-1}\right)\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^\top\boldsymbol{V}^{-1}(\boldsymbol{y} - \widehat{c}\boldsymbol{1}) \tag{3}$$

$$\widehat{c} = \frac{\boldsymbol{1}^\top\boldsymbol{V}^{-1}\boldsymbol{y} - \boldsymbol{1}^\top\boldsymbol{V}^{-1}\boldsymbol{X}\widehat{\boldsymbol{\beta}}}{\boldsymbol{1}^\top\boldsymbol{V}^{-1}\boldsymbol{1}}. \tag{4}$$

Let $\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^\top\boldsymbol{V}^{-1}\left(\boldsymbol{1}^\top\boldsymbol{V}^{-1}\boldsymbol{1}\right)^{-1}$, we obtain

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{L}\boldsymbol{X} + \boldsymbol{\Phi}^{-1}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{L}\boldsymbol{y}. \tag{5}$$

The dual estimator using $\boldsymbol{X}\boldsymbol{\Phi}\boldsymbol{X}^\top$ was proposed in (Saunders et al., 1998) where the authors used Lagrange multipliers to obtain the following estimator for ridge regression:

$$\widehat{\boldsymbol{\beta}}_{\text{Dual}} = \boldsymbol{\Phi}\boldsymbol{X}^\top\left(\boldsymbol{V} + \boldsymbol{X}\boldsymbol{\Phi}\boldsymbol{X}^\top\right)^{-1}\boldsymbol{y}.$$

Here, $\boldsymbol{\Phi}$ is set to be diagonal, and the above estimator (6) can be evaluated in $O\left(n^2p\right)$ time, a significant improvement when $p \gg n$. However, the computational bottleneck becomes evaluating the *kernel matrix* $\boldsymbol{X}\boldsymbol{\Phi}\boldsymbol{X}^\top$.

For the zero intercept case $\widehat{c} = 0$, the dual estimator (6) is equivalent to (5) from the following variant of the Woodbury identity $\left(\boldsymbol{U}^{-1} + \boldsymbol{A}^\top \boldsymbol{V}^{-1} \boldsymbol{A}\right)^{-1} \boldsymbol{A}^\top \boldsymbol{V}^{-1} = \boldsymbol{U}\boldsymbol{A}^\top \left(\boldsymbol{A}\boldsymbol{U}\boldsymbol{A}^\top + \boldsymbol{V}\right)^{-1}$ for invertible matrices $\boldsymbol{U}$ and $\boldsymbol{V}$. The dual estimator can be generalized to any intercept,

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\Phi}\boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{L} \left(\boldsymbol{X}\boldsymbol{\Phi}\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{L} + \boldsymbol{I}\right)^{-1} \boldsymbol{y}. \quad (6)$$

Computing the dual estimator (6) takes $O\left(n^2 p\right)$ time as opposed to $O\left(p^3\right)$ time required by (5). This complexity will be further improved in § 4 for the setting $p \gg n$.

3.2 APPROXIMATE VARIANCE COMPONENTS

The variance components $\boldsymbol{\Lambda}$ and σ^2 are typically estimated using an iterative EM algorithm with a per-iteration cost $O\left(p^3\right)$ (Laird et al., 1987; Lindstrom and Bates, 1988) or an exhaustive grid search for the solution of a system of eigenvalue equations (Kang et al., 2008; Lippert et al., 2011). We consider an approximate non-iterative estimator based on the key observation that the optimization of the (2) has a simple closed-form solution if carried out with respect to $\boldsymbol{M} = \boldsymbol{V} + \boldsymbol{X}\boldsymbol{\Phi}\boldsymbol{X}^\top$. We will estimate $\boldsymbol{M}$ and use it as a proxy for estimating $\boldsymbol{\Lambda}$ as well as σ^2. The variance components inferred using $\boldsymbol{M}$ are referred to as the *approximate variance components* (AVCs). While AVCs may be used as variance components estimates under certain circumstances, the main purpose is to serve as fast replacements in estimating fixed-effects coefficients.

Proxy component estimation To perform the REML estimation of the variance components in terms of $\boldsymbol{M}$, we first rewrite the log-likelihood (2) as

$$\begin{aligned} l\left(\boldsymbol{\beta}, \boldsymbol{V}\right) = &-\frac{1}{2}\log\det \boldsymbol{V} \\ &-\frac{1}{2}\left(\boldsymbol{y} - \widehat{c}\boldsymbol{1}\right)^\top \boldsymbol{M}^{-1}\left(\boldsymbol{y} - \widehat{c}\boldsymbol{1}\right) \\ &-\frac{1}{2}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\left(\boldsymbol{V}\right)\right)^\top \boldsymbol{Q}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\left(\boldsymbol{V}\right)\right) \end{aligned} \quad (7)$$

where $\boldsymbol{Q} = \boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{\Phi}^{-1}$ and $\widehat{\boldsymbol{\beta}}\left(\boldsymbol{V}\right) = \left(\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X} + \boldsymbol{\Phi}^{-1}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{V}^{-1}\left(\boldsymbol{y} - \widehat{c}\boldsymbol{1}\right)$. Here, the estimate $\widehat{\boldsymbol{\beta}}$ depends on $\boldsymbol{V}$, and is consistent with the estimate given by (5). The $\widehat{c}$ in (7) can be set to the mean response, or estimated based on a prior distribution as in (Zhou et al., 2013).

Then, the REML estimator for the variance components is based on marginalizing the fixed effects $\boldsymbol{\beta}$ (Harville, 1974). It follows that

$$\begin{aligned} l_p\left(\boldsymbol{V}\right) &\propto \log \int_{\mathbb{R}^p} \exp\left(l\left(\boldsymbol{\beta}, \boldsymbol{V}\right)\right) d\boldsymbol{\beta} \\ &\propto -\frac{1}{2}\log\det \boldsymbol{V} - \frac{1}{2}\log\det \boldsymbol{Q} \\ &\quad -\frac{1}{2}\left(\boldsymbol{y} - \widehat{c}\boldsymbol{1}\right)^\top \boldsymbol{M}^{-1}\left(\boldsymbol{y} - \widehat{c}\boldsymbol{1}\right). \end{aligned}$$

From Sylvester's determinant theorem, one observes that $\det\left(\boldsymbol{M}\right) = \det\left(\boldsymbol{\Phi}\right)\det\left(\boldsymbol{V}\right)\det\left(\boldsymbol{Q}\right)$. Thus, we arrive at

$$\begin{aligned} l_p\left(\boldsymbol{V}\right) \propto &-\frac{1}{2}\log\det \boldsymbol{M} \\ &-\frac{1}{2}\left(\boldsymbol{y} - \widehat{c}\boldsymbol{1}\right)^\top \boldsymbol{M}^{-1}\left(\boldsymbol{y} - \widehat{c}\boldsymbol{1}\right). \end{aligned} \quad (8)$$

Now, what we have achieved through (8) is a simple closed-form REML estimate of $\boldsymbol{V}$, rather than the non-convex or iterative updates for $\widehat{\boldsymbol{\Lambda}}$ and $\widehat{\sigma}^2$ in state-of-the-art LMM parameter estimation algorithms. Unconstrained maximization of (8) with respect to $\boldsymbol{M}$ results in the closed-form equality

$$\boldsymbol{Z}\widehat{\boldsymbol{\Lambda}}\boldsymbol{Z}^\top + \widehat{\sigma}^2\boldsymbol{I} = \left(\boldsymbol{y} - \widehat{c}\boldsymbol{1}\right)\left(\boldsymbol{y} - \widehat{c}\boldsymbol{1}\right)^\top - \boldsymbol{X}\boldsymbol{\Phi}\boldsymbol{X}^\top, \quad (9)$$

for an optimal $\boldsymbol{M}$. Note that $\boldsymbol{Z}\widehat{\boldsymbol{\Lambda}}\boldsymbol{Z}^\top$ is positive semidefinite, whereas the right hand side has at most one positive eigenvalue. Thus, this optimal $\boldsymbol{M}$ may not be achievable and the unbiased estimate of $\boldsymbol{\Lambda}$ may possibly have negative eigenvalues. The issue of negative variance estimates in linear mixed models is an open problem (Demidenko, 2013) and beyond the scope of this paper. One resolution is to introduce a Gamma prior on $\boldsymbol{\Lambda}$ (Chung et al., 2013). For unbiased estimation, we allow $\boldsymbol{\Lambda}$ to have negative eigenvalues, and intuitively we refer to the variance estimators obtained this way as *approximate variance components*.

Approximate variance estimators Assume that $\boldsymbol{Z}$ has full column rank and let $\boldsymbol{S} = \left(\boldsymbol{y} - \widehat{c}\boldsymbol{1}\right)\left(\boldsymbol{y} - \widehat{c}\boldsymbol{1}\right)^\top - \boldsymbol{X}\boldsymbol{\Phi}\boldsymbol{X}^\top$. The approximate variance components $\widehat{\boldsymbol{\Lambda}}_{\text{AVC}}$ and $\widehat{\sigma}^2_{\text{AVC}}$ can be obtained via

$$\arg\min_{\boldsymbol{\Lambda}, \sigma^2} \left\|\boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{Z}^\top - \boldsymbol{S} + \sigma^2\boldsymbol{I}\right\|_F^2. \quad (10)$$

Optimizing with respect to $\boldsymbol{\Lambda}$ yields

$$\boldsymbol{\Lambda}_\star = \boldsymbol{Z}^\dagger\left(\boldsymbol{S} - \sigma^2\boldsymbol{I}\right)\boldsymbol{Z}^{\dagger\top}, \quad (11)$$

where $\boldsymbol{Z}^\dagger := \left(\boldsymbol{Z}^\top\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^\top$. The estimators are computed by substituting $\boldsymbol{\Lambda}_\star$ into (10) and optimizing with respect to σ^2:

$$\begin{aligned} \widehat{\sigma}^2_{\text{AVC}} &= \frac{\operatorname{tr}\left[\boldsymbol{S}\left(\boldsymbol{I} - \boldsymbol{Z}\boldsymbol{Z}^\dagger\right)\right]}{n - q} \\ \widehat{\boldsymbol{\Lambda}}_{\text{AVC}} &= \boldsymbol{Z}^\dagger\boldsymbol{S}\boldsymbol{Z}^{\dagger\top} - \widehat{\sigma}^2_{\text{AVC}}\left(\boldsymbol{Z}^\top\boldsymbol{Z}\right)^{-1}. \end{aligned} \quad (12)$$

Consider the parameterization $\boldsymbol{\Lambda} = \theta \boldsymbol{D}$ in (Kang et al., 2008; Lippert et al., 2011) with a fixed symmetric positive semi-definite $\boldsymbol{D}$, the solution to (10) is written as

$$\boldsymbol{\Lambda}_* = \frac{\operatorname{tr}\left(\boldsymbol{G}\left(\boldsymbol{S} - \sigma^2 \boldsymbol{I}\right)\right)}{\operatorname{tr}\left(\boldsymbol{G}^2\right)} \boldsymbol{D} \tag{13}$$

with $\boldsymbol{G} = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^\top$. Substituting into (10), we obtain

$$\widehat{\sigma}^2_{\text{AVC}} = \frac{1}{n-\alpha}\left[\operatorname{tr}(\boldsymbol{S}) - \frac{\operatorname{tr}(\boldsymbol{G}\boldsymbol{S})}{\operatorname{tr}\left(\boldsymbol{G}^2\right)}\right], \tag{14}$$

where $\alpha = \operatorname{tr}(\boldsymbol{G})^2 / \operatorname{tr}\left(\boldsymbol{G}^2\right)$. Combined with (13), we arrive at

$$\widehat{\boldsymbol{\Lambda}}_{\text{AVC}} = \frac{\operatorname{tr}\left(\boldsymbol{G}\left(\boldsymbol{S} - \widehat{\sigma}^2_{\text{AVC}} \boldsymbol{I}\right)\right)}{\operatorname{tr}\left(\boldsymbol{G}^2\right)} \boldsymbol{D}. \tag{15}$$

While AVCs may be used as variance components estimates under certain circumstances, the main purpose is to speed up estimating the fixed-effect coefficients. The complexity for computing the AVCs is $O\left(n^3\right)$, if $\boldsymbol{S}$ is given. Like the dual fixed-effects estimator (6), the computational bottleneck of AVCs also lies in evaluating $\boldsymbol{X}\boldsymbol{\Phi}\boldsymbol{X}^\top$.

4 FAST COMPUTATIONAL ALGORITHMS

In this section, we further improve the computational complexity $O\left(n^2 p\right)$ of the proposed approximate estimators in the high-dimensional setting $p \gg n$, where the computation bottleneck lies in evaluating the kernel $\boldsymbol{X}\boldsymbol{\Phi}\boldsymbol{X}^\top$. We adopt the subsampled randomized Hadamard transform (SRHT) (Tropp, 2011) to compute the kernel matrix efficiently. In particular, the high-dimensional data is first projected into lower dimensions using SRHT, and the parameters of the LMM are then estimated using the projected data. However, there are two main challenges involved: 1) the estimated parameter $\widehat{\boldsymbol{\beta}}$ now corresponds to the projected data of reduced dimensions, whereas the coefficients of the full original covariates are desired; and 2) the impact of applying the SRHT on the accuracy of parameter estimation needs to be justified. The techniques developed in this section recovers the coefficients to the full covariates from the SRHT projected data with high accuracy, as will be shown in § 5.

4.1 NON-ITERATIVE ALGORITHM FOR GENERAL LMMS

In this subsection, we provide a fast algorithm for parameter estimation in case of a general covariance matrix. Algorithm 1 takes as input the matrices $\boldsymbol{X}$ and $\boldsymbol{\Phi}$

Algorithm 1 Approximate kernel matrix computation.

Require: $\boldsymbol{X}$, $\boldsymbol{\Phi}$, and error tolerance ϵ.

1: Let $p' = 2^{\lceil \log_2 p \rceil}$, append $p' - p$ all zero columns to $\boldsymbol{X}$, and $p' - p$ all zero rows and columns to $\boldsymbol{\Phi}$. Compute a diagonal matrix $\boldsymbol{D}$ of dimension p' with Rademacher random diagonal elements.

2: Denote the fast Walsh-Hadamard transform by

$$\boldsymbol{W}_{p'} = \begin{bmatrix} \boldsymbol{W}_{p'/2} & \boldsymbol{W}_{p'/2} \\ \boldsymbol{W}_{p'/2} & -\boldsymbol{W}_{p'/2} \end{bmatrix} \quad \text{with} \quad \boldsymbol{W}_1 = 1.$$

Let r be the rank of $\boldsymbol{X}$ or $r = n$ for unknown rank, then define

$$s_\epsilon := \frac{6\left[\sqrt{r} + \sqrt{8\log(rp')}\right]^2 \log r}{\epsilon^2}.$$

Sample without replacement m rows of $\boldsymbol{W}_{p'}\boldsymbol{D}/\sqrt{s_\epsilon}$ to obtain the SRHT $\boldsymbol{\Pi}$. Compute $\boldsymbol{A} = \boldsymbol{X}\sqrt{\boldsymbol{\Phi}}\boldsymbol{\Pi}^\top$.

3: **return** the approximate kernel $\boldsymbol{A}\boldsymbol{A}^\top$, $\boldsymbol{A}$, and $\boldsymbol{\Pi}$.

(which will be typically diagonal) and an approximation error ϵ described in § 5. Both an approximation to the kernel matrix $\boldsymbol{X}\boldsymbol{\Phi}\boldsymbol{X}^\top$ and the SRHT matrix $\boldsymbol{\Pi}$ are computed. The computational efficiency of the algorithm is a result of replacing $\boldsymbol{X}$ with the smaller transform $\boldsymbol{A}$ in subsequent operations. Additionally, the structure of the SRHT allows for a divide-and-conquer scheme to compute $\boldsymbol{A} = \boldsymbol{X}\sqrt{\boldsymbol{\Phi}}\boldsymbol{\Pi}^\top$ in $O(np\log p)$ time. Note that the matrix $\boldsymbol{W}_{p'}$ is not formed explicitly. The computation $\boldsymbol{A}\boldsymbol{A}^\top$ requires $O\left(n^2 s_\epsilon\right)$ time, which becomes dominant setting $\epsilon \leq Cn\sqrt{\frac{\log n}{p\log p}}$ for some universal constant C. Thus, the overall runtime for the algorithm is $O\left(\frac{n^3 \log n}{\epsilon^2}\right)$ for dense full-rank $\boldsymbol{X}$, and will be faster if $\boldsymbol{X}$ is of low rank. The quality of the approximation depends on ϵ, which will be discussed in § 5.

Given the approximate kernel, it is straight forward to compute the AVCs $\boldsymbol{\Lambda}_{\text{AVC}}$ and σ^2_{AVC} via (12). The coefficients for the fixed-effects can also be computed efficiently using the following estimator

$$\widehat{\boldsymbol{\beta}} = \sqrt{\boldsymbol{\Phi}}\boldsymbol{\Pi}^\top \boldsymbol{A}^\top \widehat{\boldsymbol{V}}^{-1}\boldsymbol{L}\left(\boldsymbol{I} + \boldsymbol{A}\boldsymbol{A}^\top \widehat{\boldsymbol{V}}^{-1}\boldsymbol{L}\right)^{-1}\boldsymbol{y}. \tag{16}$$

Given the approximate kernel matrix and $\boldsymbol{A}$, computing $\boldsymbol{A}^\top \widehat{\boldsymbol{V}}^{-1}\boldsymbol{L}\left(\boldsymbol{I} + \boldsymbol{A}\boldsymbol{A}^\top \widehat{\boldsymbol{V}}^{-1}\boldsymbol{L}\right)^{-1}\boldsymbol{y}$ takes time $O\left(\max\left\{n^2 s_\epsilon, n^3\right\}\right)$ and multiplication of this vector by $\sqrt{\boldsymbol{\Phi}}\boldsymbol{\Pi}^\top$ is $O(p\log p)$ due to the structure of the SRHT matrix as well as the fact that $\sqrt{\boldsymbol{\Phi}}$ is diagonal. The resulting complexity in computing (16) is $O\left(\max\left\{n^2 s_\epsilon, n^3, p\log p\right\}\right)$.

Approximating the kernel matrix using the SRHT was proposed for ridge regression in (Lu et al., 2013), a special case of our setting. A method for estimating the full set of fixed-effects coefficients was not provided in (Lu et al., 2013). Instead, a reduced set of m fixed-effects coefficients corresponding to the transformed covariate matrix $\boldsymbol{X}\boldsymbol{\Pi}^\top$ was reported. For many applications, a major point of using an LMM is to estimate the effect-size of the fixed-effect coefficients, so computing $\widehat{\boldsymbol{\beta}}$ is essential to the problem.

4.2 FAST EM FOR MULTI-GROUP LMMS

For efficient parameter estimation in L_2-regularized LMMs with repeated measurements, we extend the EM algorithm for the low-dimensional setting $n \geq p$ (Laird et al., 1987) by combing the kernel estimators and Algorithm 1. While this high-dimensional EM variant is iterative, we show that the per-iteration computational cost is scalable in p.

The log-likelihood of the L_2-regularized LMM (17) can be rewritten in terms of class-specific variables as

$$\begin{aligned} &\log p\left(\boldsymbol{y}, \boldsymbol{\gamma}, \boldsymbol{\beta} \mid \boldsymbol{X}; \sigma^2, \boldsymbol{\Lambda}\right) \\ &\propto -\frac{1}{2}\boldsymbol{\beta}^\top \boldsymbol{\Phi}^{-1}\boldsymbol{\beta} - \frac{n}{2}\log\sigma^2 - \frac{m}{2}\log\det \boldsymbol{H} \\ &\quad - \frac{1}{2}\sum_{i=1}^{m} \boldsymbol{\gamma}^{(i)\top}\boldsymbol{H}^{-1}\boldsymbol{\gamma}^{(i)} - \frac{\boldsymbol{e}^\top \boldsymbol{e}}{2\sigma^2}, \end{aligned} \tag{17}$$

where $\boldsymbol{e} = \boldsymbol{y} - c\mathbf{1} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\gamma}$.

From the above log-likelihood, the posterior distribution of $\boldsymbol{\beta}$ conditioned on the data and parameter estimates $\widehat{\mathcal{P}} := \left\{\widehat{c}, \widehat{\sigma}^2, \widehat{\boldsymbol{H}}\right\}$ is multivariate normal with mean $\boldsymbol{\Phi}\boldsymbol{X}^\top\widehat{\boldsymbol{M}}^{-1}\left(\boldsymbol{y} - \widehat{c}\mathbf{1}\right)$ and covariance $\boldsymbol{\Phi} - \boldsymbol{\Phi}\boldsymbol{X}^\top\widehat{\boldsymbol{M}}^{-1}\boldsymbol{X}\boldsymbol{\Phi}$. Similarly, the posterior distribution of the vector of latent variables $\boldsymbol{\gamma}$ is multivariate normal with mean $\widehat{\boldsymbol{\Lambda}}\boldsymbol{Z}^\top\widehat{\boldsymbol{M}}^{-1}\left(\boldsymbol{y} - \widehat{c}\mathbf{1}\right)$ and covariance $\widehat{\boldsymbol{\Lambda}} - \widehat{\boldsymbol{\Lambda}}\boldsymbol{Z}^\top\widehat{\boldsymbol{M}}^{-1}\boldsymbol{Z}\widehat{\boldsymbol{\Lambda}}$. Denote by $\widehat{\boldsymbol{\gamma}}$ the mean of the posterior distribution of $\boldsymbol{\gamma}$, we also obtain the following posterior distributions of class-specific latent variable $\boldsymbol{\gamma}^{(i)}$:

$$\mathcal{N}\left(\widehat{\boldsymbol{\gamma}}^{(i)}, \widehat{\boldsymbol{H}} - \widehat{\boldsymbol{H}}\boldsymbol{Z}^{(i)\top}\left(\widehat{\boldsymbol{M}}^{-1}\right)^{(i)}\boldsymbol{Z}^{(i)}\widehat{\boldsymbol{H}}\right). \tag{18}$$

Note that $\cdot^{(i)}$ represents the block matrix corresponding to group i. These posteriors are used in the E-step, discussed next.

E-step In the E-step, we derive the expectation of the log-likelihood (17) with respect to the aforementioned posterior distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}^{(i)}$:

$$\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\gamma}\mid\boldsymbol{y},\widehat{\mathcal{P}}}\left[\log p\left(\boldsymbol{y}, \boldsymbol{\gamma}, \boldsymbol{\beta} \mid \boldsymbol{X}; \sigma^2, \boldsymbol{H}\right)\right].$$

We only need to consider terms in the expectation that involve c, σ^2, and $\boldsymbol{H}$. Denote by $\widehat{\Sigma}_{\boldsymbol{\gamma}^{(i)}}$ the variance of (18), the following holds $\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\gamma}\mid\boldsymbol{y},\widehat{\mathcal{P}}}\left(\boldsymbol{\gamma}^{(i)\top}\boldsymbol{H}^{-1}\boldsymbol{\gamma}^{(i)}\right) = \widehat{\boldsymbol{\gamma}}^{(i)\top}\boldsymbol{H}^{-1}\widehat{\boldsymbol{\gamma}}^{(i)} + \operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}^{(i)}}\boldsymbol{H}^{-1}\right)$. Using the previously derived posterior distributions, we get

$$\begin{aligned} \mathbb{E}\left(\boldsymbol{e} \mid \boldsymbol{y}, \widehat{\mathcal{P}}\right) &= \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\widehat{\boldsymbol{\gamma}} - \widehat{c}\mathbf{1} \\ \operatorname{cov}\left(\boldsymbol{e} \mid \boldsymbol{y}, \widehat{\mathcal{P}}\right) &= \widehat{\sigma}^2\boldsymbol{I} - \widehat{\sigma}^4\widehat{\boldsymbol{M}}^{-1}. \end{aligned}$$

Thus, we arrive at $\mathbb{E}_{\boldsymbol{\beta},\boldsymbol{\gamma}\mid\boldsymbol{y},\widehat{\mathcal{P}}}\left(\boldsymbol{e}^\top\boldsymbol{e}\right) = \widehat{\boldsymbol{e}}^\top\widehat{\boldsymbol{e}} + \widehat{\sigma}^2\boldsymbol{I} - \widehat{\sigma}^4\widehat{\boldsymbol{M}}^{-1}$, where $\widehat{\boldsymbol{e}} := \mathbb{E}\left(\boldsymbol{e} \mid \boldsymbol{y}, \widehat{\mathcal{P}}\right)$.

M-step We now update the parameter estimates with the maximizers of the expectation from the E-step. First, observe that the $\boldsymbol{\beta}$ estimate from the posterior distribution is the same as the the dual estimator developed in § 3. To maximize the expectation with respect to $\boldsymbol{H}$ and σ^2, we take the partial derivatives with respect to $\boldsymbol{H}^{-1}$ and σ^{-2}, and set them to zero. This gives the following M-step updates:

$$\begin{aligned} \widehat{\boldsymbol{H}} &\leftarrow \frac{1}{m}\sum_{i=1}^{m}\left(\widehat{\boldsymbol{\gamma}}^{(i)}\widehat{\boldsymbol{\gamma}}^{(i)\top} + \widehat{\Sigma}_{\boldsymbol{\gamma}^{(i)}}\right) \\ \widehat{\sigma}^2 &\leftarrow \widehat{\sigma}^2 + \frac{1}{n}\left[\widehat{\boldsymbol{e}}^\top\widehat{\boldsymbol{e}} - \widehat{\sigma}^4\operatorname{tr}\left(\widehat{\boldsymbol{M}}^{-1}\right)\right]. \end{aligned} \tag{19}$$

The fast version of the above EM variant uses Algorithm 1 for computing the kernel. Note that the original $\boldsymbol{X}$ is no longer needed after the SRHT projection. This provides additional space advantages as data $\boldsymbol{X}$ can be preprocessed, and the Hadamard transform in Step 1 requires a small constant amount of memory. Overall, the per-iteration computational complexity of the EM algorithm is $O\left(\max\left\{n^2 s_\epsilon, n^3\right\}\right)$.

5 THEORETICAL GUARANTEES

In this section, we provide an analysis of the difference in the parameters estimated via the approximate algorithms versus minimizing the L_2-regularized LMM. We are not proving consistency of our estimator—convergence of the parameter estimates to the population quantity. Consistency results for LMMs and-regularized LMMs were provided in (Schelldorfer et al., 2011; Cui et al., 2004; Hall and Yao, 2003). See the supplementary materials for proofs of the theorems in this section.

Theorem 1 (Fixed-effect norm error)**.** *Let $\widehat{\boldsymbol{\beta}}$ be the fixed-effect coefficients estimated by (5) and $\widehat{\boldsymbol{\beta}}'$ be the fixed-effect coefficients estimated by the approximate proce-*

dure in (16)*. Then, with probability at least* $1 - 3/n$

$$\frac{\left\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}'\right\|}{\left\|\widehat{\boldsymbol{\beta}}\right\|} \leq \frac{\epsilon}{1-\epsilon} \frac{\left\|\boldsymbol{\Phi}^{-1}\right\|_2 \kappa\left(\boldsymbol{\Gamma}\right)}{\frac{\|\boldsymbol{\Phi}\|_2^{-1}}{1+\sqrt{2/3\epsilon}} + \lambda_{min}\left(\boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{X}\right)}$$

with $\boldsymbol{\Gamma} \coloneqq \boldsymbol{\Phi}^{-1} + \boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{X}$*, or loosely*

$$\frac{\left\|\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}'\right\|}{\left\|\widehat{\boldsymbol{\beta}}\right\|} \leq \frac{\epsilon\left(1+\sqrt{2/3\epsilon}\right)}{1-\epsilon} \kappa\left(\boldsymbol{\Phi}\right) \kappa\left(\boldsymbol{\Gamma}\right)$$

for all $0 \leq \epsilon < 1$.

An intuitive interpretation of the theorem is that the fixed-effects coefficients estimator (16) has better accuracy when the predefined $\boldsymbol{\Phi}$ is better conditioned and has smaller spectral norm. One can certainly improve the accuracy by setting a smaller ϵ, which in turn uses more samples in Algorithm 1.

Theorem 2 (AVC approximation errors)**.** *Let* $\widehat{\sigma}^2_{AVC}$ *and* $\widehat{\boldsymbol{\Lambda}}_{AVC}$ *be computed using* (12)*. Let* $\widehat{\sigma}'^2_{AVC}$ *and* $\widehat{\boldsymbol{\Lambda}}'_{AVC}$ *be computed using the same equations but with approximate kernel from Algorithm 1. Then, the following two statements hold jointly with probability at least* $1 - 3/n$*:*

$$\left|\widehat{\sigma}^2_{AVC} - \widehat{\sigma}'^2_{AVC}\right| \leq \epsilon \cdot \frac{\left\|\left|\boldsymbol{X}\boldsymbol{\Phi}\boldsymbol{X}^\top\right|\right\|_{n-q}}{n-q} \quad \textit{and}$$

$$\left\|\widehat{\boldsymbol{\Lambda}}_{AVC} - \widehat{\boldsymbol{\Lambda}}'_{AVC}\right\|_2 \leq \frac{\epsilon}{\widehat{\sigma}_{min}\left(\boldsymbol{Z}\right)^2}\left(\left\|\boldsymbol{X}\boldsymbol{\Phi}\boldsymbol{X}^\top\right\|_2 + \frac{\left\|\left|\boldsymbol{X}\boldsymbol{\Phi}\boldsymbol{X}^\top\right|\right\|_{n-q}}{n-q}\right).$$

Note that the fraction of the Ky Fan norm does not exceed the spectral norm. A looser but more convenient bounds are $\left|\widehat{\sigma}^2_{\text{AVC}} - \widehat{\sigma}'^2_{\text{AVC}}\right| \leq \epsilon \left\|\boldsymbol{X}\boldsymbol{\Phi}\boldsymbol{X}^\top\right\|_2$ and $\left\|\widehat{\boldsymbol{\Lambda}}_{\text{AVC}} - \widehat{\boldsymbol{\Lambda}}'_{\text{AVC}}\right\|_2 \leq 2\epsilon\sigma_{\min}\left(\boldsymbol{Z}\right)^{-2}\left\|\boldsymbol{X}\boldsymbol{\Phi}\boldsymbol{X}^\top\right\|_2$.

6 EXPERIMENTS

In this section, we conduct a simulation study as well as numerical experiments on real data. The simulation study demonstrates the accuracy of parameter estimation using the proposed Approximate Ridge LMM (`arLMM`) methods. We also examined the results on a real data example from the Wellcome Trust Case Control Consortium (WTCCC) study (The Wellcome Trust Case Control Consortium, 2007), which include about 14,000 cases from seven common diseases and a total of about 450,000 SNPs.

The main finding of the experiments is that the proposed approximate inference algorithms enjoy similar predictive accuracy as state-of-the-art methods at a significantly reduced computation cost in practice. In particular, our Matlab prototype implementation is 6x faster than the optimized C implementation of the state-of-the-art `BSLMM` method for genome-wide association studies.

6.1 SIMULATION STUDIES

To evaluate parameter estimation, we consider two performance metrics. The first one is the correlation between the estimated and ground-truth fixed-effect coefficients. The second metric is the Negative Log Likelihood (NLL) of the standard LMM, which meaningfully reflects the quality of variance estimation.

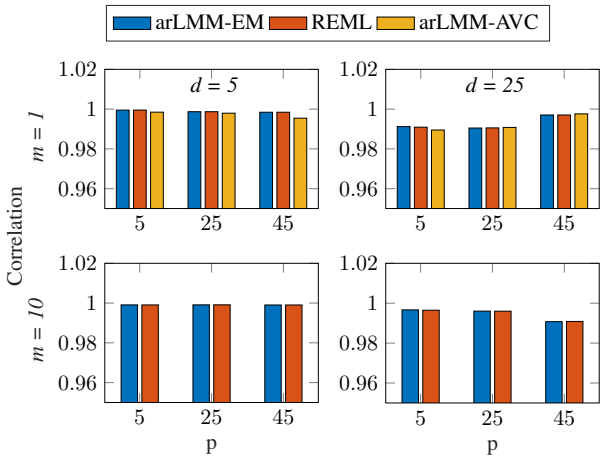
For the simulation, we compare the performance of our non-iterative algorithm `arLMM-AVC` based on (16) and (12), the proposed multi-group variant `arLMM-EM` based on (19), the standard `REML` (Bates et al., 2015), L_1-regularized LMM `lmmlasso` (Schelldorfer et al., 2011), and `CovexLasso` using both L_1- and L_2-regularization (Jakubík, 2015).

Synthetic data generation The simulation is based on synthetic training and validation sets sampled from a fixed LMM distribution. The design matrices as well as the parameters for the fixed LMM are randomly generated. Specifically,

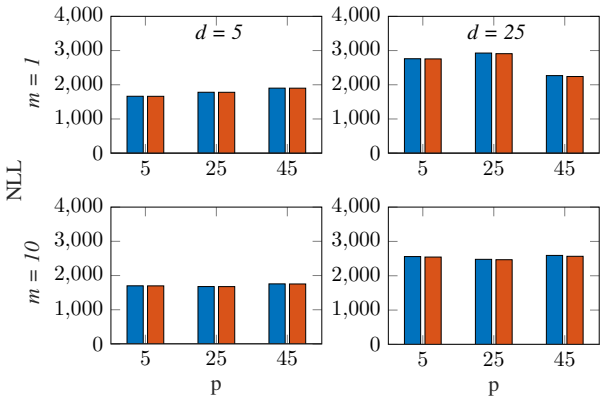
$$\begin{aligned} X_{ij} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0,1\right) & Z^{(k)}_{ij} &\overset{\text{i.i.d.}}{\sim} \mathcal{U}\left(0,1\right) \\ \boldsymbol{\gamma}^{(k)} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{K}^\top\boldsymbol{K}\right) & K_{ij} &\overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0,1\right) \\ \boldsymbol{\beta} &\sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}\right) & \sigma^2 &\sim \mathcal{U}\left(0,d\right). \end{aligned}$$

Note that there are d random-effect variables with covariance $\boldsymbol{K}^\top\boldsymbol{K}$. Thus, the random-effect design matrix $\boldsymbol{Z} \in \mathbb{R}^{n\times q}$, $q = md$, will be block-diagonal with diagonal blocks $\boldsymbol{Z}^{(k)}$. Given the number of observations n, we randomly sample n_k observations for each group k, where the fractions n_k/n are specified by the Dirichlet distribution with the concentration parameters $(1, 1, \cdots)^\top$.

Overdetermined settings Let us first consider the standard setting $n > p$, which are supported by many parameter estimation algorithms of LMMs. We evaluate the performance of `arLMM-AVC` and `arLMM-EM` in a variety of p, d, and m settings. The parameter estimates obtained using the proposed methods are compared with the estimates given by the standard REML (see e.g., (Kang et al., 2008; Lippert et al., 2011; Laird et al., 1987; Bates et al., 2015)) which is known to produce unbiased estimates.



(a) Correlation between estimated and true $\boldsymbol{\beta}$'s.



(b) NLL at convergence.

Figure 1: Comparing the performance of parameter estimation on synthetic data with $n = 1,000$ observations. Note that `arLMM-AVC` is only applicable to the single group setting. This figure shows that the `arLMM-EM` and `arLMM-AVC` achieve comparable estimation performance as `REML`.

Figure 1 shows the error for the fitted parameters using 1,000 observations sampled from the underlying LMM. The average results are reported over 10 runs on independently generated datasets. These generated datasets have the same number of observations $n = 1,000$ but different settings of p, d, and m.

As shown in Figure 1, `arLMM-EM` and `arLMM-AVC` exhibit comparable estimation accuracy as the standard `REML`. Note that `arLMM-AVC` is applicable only when $m = 1$ (the first row of Figure 1). Since `arLMM-AVC` is based on non-iterative approximation to the variance components, the error is slighted higher than the others as expected.

High-dimensional (underdetermined) setting We also examined the performance of our model in the high-dimensional setting where we are interested in variable selection based on the fixed-effects coefficients. In Table 2, we specify the three regimes for which we generate simulated data: an overdetermined LMM, a moderate-dimensional LMM, and a high-dimensional LMM. Each regime is characterized by n, p, d, and m, and an extra parameter s, the number of non-zeros in the ground-truth $\boldsymbol{\beta}_{\text{True}}$. Since $m > 1$ we did not apply `arLMM-AVC`.

Table 2: Regimes of data.

	(n, p, d, m, s)
LOW	$(100, 1000, 5, 3, 10)$
MOD	$(200, 10^4, 5, 3, 10)$
HIGH	$(10^4, 10^6, 10, 100, 100)$

Figure 2 reports variable selection results for `arLMM-EM`, `lmmlasso` (Schelldorfer et al., 2011), and `ConvexLasso`. All the settings in Table 2 have sparse ground-truth $\boldsymbol{\beta}_{\text{True}}$. Figure 2 shows the fraction of the signal (non-zeros in $\boldsymbol{\beta}_{\text{True}}$) recovered in the estimate $\widehat{\boldsymbol{\beta}}$. We varied the regularization parameters to obtain $\widehat{\boldsymbol{\beta}}$ with different sparsity $\left\|\widehat{\boldsymbol{\beta}}\right\|_0$. The entries with the largest magnitude of $\widehat{\boldsymbol{\beta}}$ is considered the signal in these evaluations. As can be seen, `arLMM-EM` delivers a competitive signal recovery ratio for $p = 10^3, 10^4$, and scales to considerably large dimensions $n = 10^4$ and $p = 10^6$, which the other two methods cannot handle.

6.2 GENOME WIDE ASSOCIATION STUDIES

LMMs have been used extensively for mapping traits in statistical genetics. The problem formulation is that of regressing a quantitative or categorical trait onto a high-dimensional vector of 450,000 single nucleotide polymorphisms (SNPs), or locations of discrete genetic variation, for each subject included in the study. The random effects are driven by population structure or the pairwise similarity or relatedness between individuals.

We compare our approximate estimator to the performance of a state-of-the-art estimator called `BSLMM` (Bayesian sparse linear mixed model) (Zhou et al., 2013). Specifically, we run BSLMM in its ridge-regression with mixed models setting, the fastest setting of the package for a fair comparison. In this setting `BSLMM` is computing the maximum a posteriori estimate of the regularized LMM. We compare performance on the Wellcome Trust Case Control Consortium (WTCCC) dataset of 14,000 cases of 7 diseases - bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D) - and 3,000 shared controls. This dataset characterizes over 450,000 single nu-

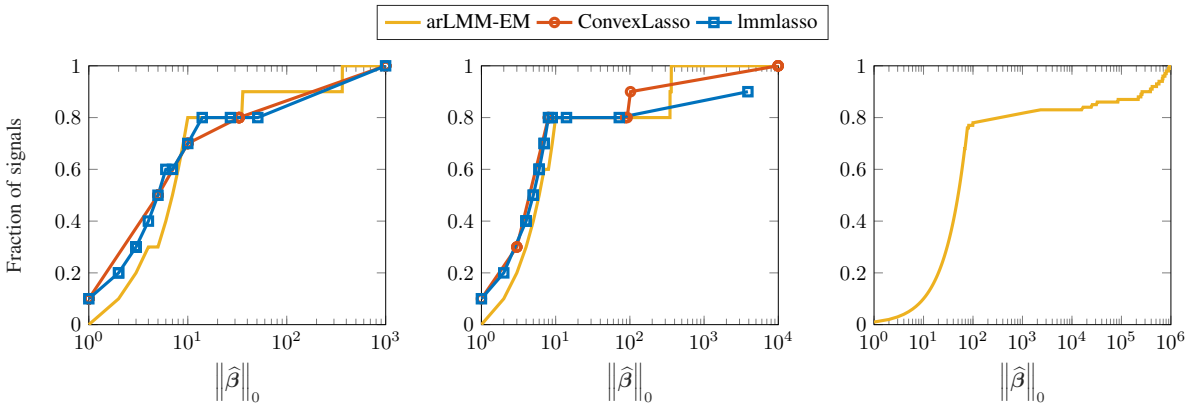


Figure 2: Fraction of signals captured by $\widehat{\beta}$. From left to right, the configurations are respectively LOW, MOD, and HIGH in Table 2. It shows that `arLMM-EM` performs competitively in variable selection.

Table 3: Comparing the prediction performance as well as the runtime of `BSLMM` and `arLMM-AVC` on the WTCCC dataset. Corr $\left(\widehat{\beta}_{\text{BSLMM}}, \widehat{\beta}_{\text{arLMM-AVC}}\right)$ denotes the correlation between the fixed-effect coefficient estimates given by `BSLMM` and `arLMM-AVC`.

DISEASE	TIME (MIN)		AUC		CORR $\left(\widehat{\beta}_{\text{BSLMM}}, \widehat{\beta}_{\text{arLMM-AVC}}\right)$
	BSLMM	ARLMM-AVC	BSLMM	ARLMM-AVC	
BD	115.8	25.1	0.6520	0.6461	0.9898
CAD	161.0	26.1	0.5899	0.5937	0.9776
CD	110.3	25.4	0.6260	0.6328	0.9862
HT	120.6	19.4	0.5956	0.6010	0.9766
RA	147.4	19.9	0.6173	0.6206	0.9834
T1D	120.0	20.4	0.6846	0.6840	0.9939
T2D	155.3	18.9	0.6003	0.5993	0.9783

cleotide polymorphisms (SNPs), or locations of discrete genetic variation, for each subject included in the study. Disease status is indicated as a binary response (1 for disease case, -1 for control). Each of the datasets had roughly equal numbers of cases and controls.

For this experiment, we adopted the same random-effect covariance parameterization used to control for population structure $\theta \boldsymbol{X}\boldsymbol{X}^\top/p$ as `BSLMM`, and used `arLMM-AVC` with AVCs (15) and (14). `arLMM-AVC` and `BSLMM` were run under identical conditions on each of the seven approximately 5,000-subject $\times$ 450,000-SNP datasets. This was the same experimental setup used to validate `BSLMM` in (Zhou et al., 2013).

Observed runtimes for each of the seven datasets are reported in Table 3. Correlation between the $\widehat{\beta}$ reported by `arLMM-AVC` and `BSLMM` in all cases was very high, 0.977 or greater.

We also compared disease status prediction by splitting each dataset into a training set comprised of 80% of subjects and a test set of the remaining 20%, selected at random. `arLMM-AVC` and `BSLMM` each estimated $\widehat{\beta}$ from the training set and attempted to predict disease status on the held-out set. We repeated this 20 times for each of the seven datasets and evaluated performance of prediction on the held-out set by area under the ROC curve (AUC). These results are also given in Table 3. Predictive performance by `arLMM-AVC` and `BSLMM` was almost identical. Predicting disease status from genetic markers is hard and it is well known that the effect sizes of genetic variants are individually small and that a great deal of variance in the response will also be driven by environmental factors.

7 CONCLUSIONS

State-of-the-art parameter inference in LMMs requires computational complexity which depends at least linearly on the number of covariates p and generally relies on heuristics. In this paper, we presented scalable learning algorithms which have sublinear computational complexity in p and provide theoretical guarantees for the accuracy of parameter estimation. Our approach combines novel approximate estimators that use a kernel matrix of the observations and the subsampled randomized Hadamard transform. Experiments on synthetic and real data corroborate the theory.

References

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

Y. Chung, S. Rabe-Hesketh, V. Dorie, A. Gelman, and J. Liu. A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, 78(4):685–709, 2013.

H. Cui, K. W. Ng, and L. Zhu. Estimation in mixed effects model with errors in variables. *Journal of Multivariate Analysis*, 91(1):53–73, 2004.

G. Darnell, S. Georgiev, S. Mukherjee, and B. E. Engelhardt. Adaptive randomized dimension reduction on massive data. *JMLR*, Apr. 2017.

E. Demidenko. *Mixed Models: Theory and applications with R*. Wiley, 2nd edition, 2013.

P. Hall and Q. Yao. Inference in components of variance models with low replication. *Annals of Statistics*, 31 (2):414–441, 2003.

D. A. Harville. Bayesian inference for variance components using only error contrasts. *Biometrika*, 61: 383–385, 1974.

J. Jakubík. Convex method for variable selection in high-dimensional linear mixed models. In *Proceedings of the 10th International Conference on Measurement*, pages 55–58, 2015.

H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178:1709–23, Mar 2008.

N. Laird, N. Lange, and D. Stram. Maximum likelihood computations with repeated measures: Application of the EM algorithm. *Journal of the American Statistical Association*, 82:97–105, 1987.

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, Dec. 1982.

M. J. Lindstrom and D. M. Bates. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83:1014–1022, 1988.

C. Lippert, J. Listgarten, Y. Liu, C. Kadie, R. Davidson, and D. Heckerman. Fast linear mixed models for genome-wide association studies. *Nature Methods*, 8 (10):833–835, Oct. 2011.

Y. Lu, P. Dhillon, D. P. Foster, and L. Ungar. Faster ridge regression via the subsampled randomized Hadamard transform. In *NIPS*, pages 369–377. 2013.

H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554, 1971.

P. O. Perry. Fast moment-based estimation for hierarchical models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):267–291, 2017.

C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *ICML*, pages 515–521, 1998.

J. Schelldorfer, P. Bühlmann, and S. V. De Geer. Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214, 2011.

J. Schelldorfer, L. Meier, and P. Bühlmann. GLMM-Lasso: An algorithm for high-dimensional generalized linear mixed models using ℓ_1-penalization. *Journal of Computational and Graphical Statistics*, 23(2):460–477, 2014.

The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447 (7145):661–678, June 2007.

J. A. Tropp. Improved analysis of the subsampled randomized Hadamard transform. *Advances in Adaptive Data Analysis*, 3(1-2):115–126, 2011.

X. Zhang, Y. Zhou, Y. Ma, B. Chen, L. Zhang, and D. Agarwal. GLMix: Generalized linear mixed models For large-scale response prediction. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 363–372, 2016.

Z. Zhang, G. Dai, and M. I. Jordan. Bayesian generalized kernel mixed models. *JMLR*, 12:111–139, Feb. 2011.

X. Zhou. A unified framework for variance component estimation with summary statistics in genome-wide association studies. *The Annals of Applied Statistics*, 11(4):2027–2051, 2017.

X. Zhou and M. Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, 2014.

X. Zhou, P. Carbonetto, and M. Stephens. Polygenic modeling with Bayesian sparse linear mixed models. *PLOS Genetics*, 9(2):1–14, 02 2013.