Received: 21 July 2019 | Revised: 17 November 2019 | Accepted: 24 December 2019

DOI: 10.1002/sim.8477

Check for updates

RESEARCH ARTICLE

Statistics in Medicine WILEY

Multikernel linear mixed model with adaptive lasso for complex phenotype prediction

Yalu Wen[1] | **Qing Lu**[2]

[1]Department of Statistics, The University of Auckland, Auckland, New Zealand

[2]Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, Michigan

Correspondence
Yalu Wen, Department of Statistics, The University of Auckland, 38 Princes Street, Auckland Central, New Zealand.
Email: y.wen@auckland.ac.nz

Funding information
National Institute on Drug Abuse, Grant/Award Number: R01DA043501; National Natural Science Foundation of China, Grant/Award Number: 81502887; U.S. National Library of Medicine, Grant/Award Number: R01LM012848

Linear mixed models (LMMs) and their extensions have been widely used for high-dimensional genomic data analyses. While LMMs hold great promise for risk prediction research, the high dimensionality of the data and different effect sizes of genomic regions bring great analytical and computational challenges. In this work, we present a multikernel linear mixed model with adaptive lasso (KLMM-AL) to predict phenotypes using high-dimensional genomic data. We develop two algorithms for estimating parameters from our model and also establish the asymptotic properties of LMM with adaptive lasso when only one dependent observation is available. The proposed KLMM-AL can account for heterogeneous effect sizes from different genomic regions, capture both additive and nonadditive genetic effects, and adaptively and efficiently select predictive genomic regions and their corresponding effects. Through simulation studies, we demonstrate that KLMM-AL outperforms most of existing methods. Moreover, KLMM-AL achieves high sensitivity and specificity of selecting predictive genomic regions. KLMM-AL is further illustrated by an application to the sequencing dataset obtained from the Alzheimer's disease neuroimaging initiative.

KEYWORDS
adaptive lasso, high-dimensional sequencing data, linear mixed model, risk prediction

1 | INTRODUCTION

Accurate disease risk prediction is an essential step toward precision medicine, an emerging model of healthcare that tailors treatment strategies based on individuals' profiles.[1,2] The successes from genome-wide association studies have provided insights into the genetic etiology of complex diseases,[3,4] which has led to a growing interest in predicting phenotypes using genetic variants.[5] Although promising, most of the existing models can only explain a small proportion of disease heritability and thus lack sufficient accuracy for clinical use.[6,7]

Complex traits are influenced by multiple genetic variants through complex biological pathways, and thus progress toward accurately predicting phenotypes requires the development of analytical methods that can model all genetic variants jointly.[8-11] Best linear unbiased prediction (BLUP) within the linear mixed model (LMM) framework has long been considered the method of choice for predicting phenotypes when a large number of genetic variants are jointly considered.[8-10,12-16] It has gained tremendous popularities in recent years.[8,9] Instead of estimating the effect size for each genetic predictor, LMMs attempt to estimate their cumulative effects. At the core, LMM assumes that genetic similarity can lead to the phenotypic similarity, and it encodes genetic effects through a genomic similarity matrix (GSM).[8,9]

 © 2020 John Wiley & Sons, Ltd.

Specifically, GSM is used to specify the correlation structure of a random effect term in LMMs. Traditionally, in animal and plant breeding, GSM is estimated using kinship coefficient, and a single random effect term is used to model genome-wide additive effects.[14,15] With the development of high-throughput technologies, GSMs nowadays can be estimated empirically from genome-wide data.[8-10,12] The widely used genomic BLUP (gBLUP) specifies a single-random effect term in LMM with the correlation structure specified according to the GSM estimated directly from genome-wide data. The implicit assumption for gBLUP is that effect sizes for all genetic variants come from a common Gaussian distribution and they act in an additive manner.[10] MultiBLUP generalizes the gBLUP model by allowing genetic variants located at different genomic regions (eg, coding, intron, and eQTLs) having separate random effects, where the correlation structures are determined by GSMs calculated from each genomic region.[8] MKLMM further generalizes MultiBLUP by constructing kernel functions under the reproducing kernel Hilbert space (RKHS) to estimate GSMs for each genomic region, where potential interaction effects within each genomic region can be considered.[9]

LMM-based methods encode genetic effects from multiple variants through GSMs, which substantially reduces the data dimension and makes it possible to jointly consider the predictive effects of all genetic variants. However, for high-dimensional genomic data, most of the measured genetic variants are not related with phenotypes. As noted by Byrnes et al,[17] variable selection algorithms can substantially improve the prediction accuracy when good biological annotations are absent. Including GSMs estimated from all genomic regions can attenuate the effects of those predictive regions, and thus reduce the prediction accuracy. Moreover, when the number of random effects is large, the estimation involves a high-dimensional covariance matrix estimation that can increase computational instability.[18] Traditional variable selection methods (eg, Mallow C_p,[19] akaike information criterion [AIC],[20] and forward/backward/stepwise selection) suffer from lack of theoretical justification and statistical stability.[21] Bayesian information criteria (BIC) and generalized information criterion (GIC) are consistent variable selection procedures for fixed effects,[22-24] but they perform poorly for selecting random effects.[25] Recent work has focused on selecting random effects simultaneously with model estimation. Chen and Dunson[26] and Kinney et al[27] selected random effects through a hierarchical Bayesian approach. Bondell et al[28] utilized the reparameterized technique proposed by Chen et al and further developed an expectation-maximization (EM)-algorithm to select random effects based on a penalized likelihood function. Ahn et al[29] developed a moment-based method to select random effects, and Lin[30] proposed a two-stage method for random effect selection. However, none of these methods can be directly applied to LMMs used in genetic research. For standard LMM, there are multiple clusters (eg, m clusters). It assumes that the outcome vector for each cluster comes from a multivariate normal distribution (eg, $\boldsymbol{Y}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma_0}), \forall i \in (1, 2, \dots, m)$), and thus the resulting variance covariance matrix for $\boldsymbol{Y}$ is a block diagonal matrix with m blocks (ie, $\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\Sigma}_0)$). For LMMs used in genetic research,[8,9] the variance-covariance matrix for the outcome vector $\boldsymbol{Y}$ is of the form $\boldsymbol{\Sigma} = \sum_r^R \boldsymbol{K}^r \sigma_r^2 + \sigma_0^2 \boldsymbol{I}$, where $\boldsymbol{K}^r$ represents the GSM estimated directly from the rth genomic region. Since $\boldsymbol{K}^r$ is usually a dense matrix, the outcome vector $\boldsymbol{Y}$ is a single observation obtained from a multivariate normal distribution (ie, $m = 1$). Therefore, the standard asymptotic behaviors established when $m \to \infty$ are not applicable.

Mounting evidences have suggested that epistasis widely exists,[31,32] and thus it is crucial to capture the potential interaction effects when building prediction models.[9,33] However, the vast majority of LMMs assume that genetic variants influence phenotypes only in an additive manner.[8,15,16] A few studies that investigated the effects of interactions (eg, dominant effect,[34,35] two-way interactions,[36] and high-order interactions using kernels under RKHS[37-39]) have only achieved limited successes partially due to the exponentially large search space of interactions and the simple assumption of homogeneous effect sizes across the entire genome. The recent proposed MKLMM addresses these limitations by modeling the high-order interactions within each region using kernels under RKHS and accounting for heterogeneous effects by specifying multiple random effects for different genomic regions.[9] Although it improves the prediction accuracy, it is hard to prespecify kernels as the genetic architecture of complex diseases is unknown in advance. Moreover, similar to other LMMs, MKLMM also lacks theoretical justifications for selecting predictive regions, as the number of regions is usually determined empirically.

In this article, we develop a multikernel linear mixed model with adaptive lasso (KLMM-AL) to address these issues. The KLMM-AL (i) specifies multiple random effects to allow for heterogeneous effects for different genomic regions, (ii) allows multiple kernels per genomic region to account for various types of genetic effects, and (iii) establishes the theoretical justification for selecting predictive regions (ie, selecting random effects from LMM with only one dependent observation vector). Therefore, the KLMM-AL cannot only account for additive effects and various types of nonadditive effects, but also select predictive regions efficiently. In the following sections, we will first lay out the details of the proposed method and its theoretical properties. We will then compare its accuracy with existing widely used methods through simulation studies and further illustrate the KLMM-AL through an application to a whole-genome sequencing dataset obtained from the Alzheimer's disease neuroimaging initiative (ADNI).[40]

2 | METHODS

LMMs and their extensions have been widely used for prediction research with high-dimensional genomic data. For completeness, we first present the LMMs used for prediction research with genomic data. We will then (i) propose our method for selecting random effects from high-dimensional genomic data (ie, selecting random effects based on a single dependent observation), (ii) describe the computational algorithms, and (iii) derive the asymptotic properties of our estimators.

2.1 | Linear mixed model for risk prediction with high-dimensional genomic data

Utilizing a similar idea used in MultiBLUP[8] and MKLMM,[9] we first divide the genome into R regions and assume each region has its own effect size. We model the outcomes within the LMM framework as,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{r}^{R} \boldsymbol{g}^r + \boldsymbol{e}, \boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{I}_n\sigma_0^2), \tag{1}$$

where $\boldsymbol{Y}$ is a $n \times 1$ vector of the outcomes, $\boldsymbol{X}$ is a $n \times p$ matrix of demographic variables (eg, age and gender), $\boldsymbol{\beta}$ is the effect size of demographic variables, and $\boldsymbol{g}^r$ is the genetic effects from the rth genomic region with $\boldsymbol{g}^r \sim N(\boldsymbol{0}, \boldsymbol{K}^r\sigma_r^2)$.

The covariance matrix of $\boldsymbol{Y}$ is influenced by both $\boldsymbol{K}^r$ and σ_r^2 (ie, $\text{var}(\boldsymbol{Y}) = \sigma_0^2\boldsymbol{I}_n + \sum_r^R \boldsymbol{K}^r\sigma_r^2$), and $\boldsymbol{K}^r\sigma_r^2$ encodes the assumptions of genetic effects of the rth region on the outcome. For example, when $\boldsymbol{K}^r = \frac{\boldsymbol{Z}_r\boldsymbol{Z}_r^T}{p_r}$ with $\boldsymbol{Z}_r$ and p_r being genotypes and the number of genetic variants of the rth genomic region, it implicitly assumes that genetic variants located at the rth region have additive effects on the outcomes and their effect sizes follow a normal distribution (ie, $\boldsymbol{g}^r = \boldsymbol{Z}_r\boldsymbol{\gamma}_r, \gamma_r \sim N(0, \sigma_r^2)$). This assumption has been used by both gBLUP and MultiBLUP.[8,15] When $\boldsymbol{K}^r = \boldsymbol{K}_1^r \circ \boldsymbol{K}_1^r$ with $\boldsymbol{K}_1^r = \boldsymbol{Z}_r\boldsymbol{Z}_r^T$ and $\circ$ being the Hadamard product (ie, $\boldsymbol{g}^r \sim N(\boldsymbol{0}, \boldsymbol{K}_1^r \circ \boldsymbol{K}_1^r\sigma_r^2)$), it implicitly assumes that there are pairwise interactions among genetic variants on the rth genomic region. Indeed, $\boldsymbol{K}^r$ can be defined using various kernel functions to capture both linear and nonlinear effects. This idea is similar to that used in MKLMM.[9] However, different from MKLMM that assumes only one specific effect for each genomic region (eg, additive-only or pairwise-interaction-only effects), we allow the same genetic regions having multiple types of effects. For example, if rth genomic region has both the additive and pairwise interaction effects, then $\boldsymbol{g}^r \sim N(0, \boldsymbol{K}_1^r\sigma_{r1}^2 + \boldsymbol{K}_2^r\sigma_{r2}^2)$. This makes our model much more flexible.

2.2 | Penalized maximum likelihood function

The genetic causes for most of complex diseases are unknown in advance, and thus it is quite likely that a substantial amount of genomic regions and their types of effects (eg, additive) included in the analyses are not disease-related. While we focus on region selection in the following sections, the same rule can be applied to select the type of effects for each genomic region. Under model 1, if the rth genetic region is not predictive, and then σ_r^2 is expected to be zero. Selecting predictive regions are equivalent to determine which σ_r^2 is not zero, and thus a natural choice for our model is to use L_1 penalty to select random effects.

Let $\boldsymbol{\theta}_R = (\sigma_1^2, \ldots, \sigma_R^2)^T$, $\boldsymbol{\theta} = (\sigma_0^2, \boldsymbol{\theta}_R^T)^T$, and $\boldsymbol{\phi} = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$. The log-likelihood function for Equation (1) is

$$l(\boldsymbol{\phi}) = -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \boldsymbol{\Sigma} = \boldsymbol{I}_n\sigma_0^2 + \sum_r^R \boldsymbol{K}^r\sigma_r^2. \tag{2}$$

The corresponding penalized log-likelihood function with L_1 penalty is,

$$l_p(\boldsymbol{\phi}) = l(\boldsymbol{\phi}) - \lambda \sum_{i=1}^{p+R+1} \omega_i(|\phi_i|), \quad \text{with } \omega_i = \begin{cases} \omega_i, \text{for } \phi_i \in \boldsymbol{\theta}_R \\ 0, \text{otherwise,} \end{cases} \tag{3}$$

where λ is a nonnegative regularization parameter, ω_i are adaptive weights, typically $\omega_i = 1/|\tilde{\phi}_i|$, with $\tilde{\phi}_i$ denoting an initial $\sqrt{n}$ consistent estimator of $\boldsymbol{\phi}$ (eg, the maximum likelihood estimators). It is worth noting that the penalties are only

put on random effects (ie, $\boldsymbol{\theta_R}$) as the focus is on the selection of predictive genomic regions. Maximizing $l_p(\boldsymbol{\phi})$ can enable variable selection and parameter estimation simultaneously, as the effects of less important factors are shrunk to zeros under the L_1-penalty. The regularization parameters ($\lambda\omega_i$) are allowed to vary with the genetic effects, which is similar to the idea of adaptive lasso.

2.3 | Computation of penalized maximum likelihood estimator

2.3.1 | EM method

The Cholesky decomposition has been used extensively in LMMs to estimate random effect parameters, but it cannot be directly used for our model. Cholesky decomposition does not allow for the elimination of random effects, and thus is incapable of selecting predictive genomic regions. Moreover, kernel matrices $\boldsymbol{K}^r$ used to encode various types of genetic effects are only guaranteed to be positive semidefinite. To address these challenges, we utilize the same idea used in Chen and Dunson,[26] and factorize $\boldsymbol{K}^r$ of the random effect $\boldsymbol{g}^r$ as $\boldsymbol{K}^r = (\boldsymbol{F_r D_r})(\boldsymbol{F_r D_r})^T$, where $\boldsymbol{D}_r$ is a diagonal matrix and $\boldsymbol{F}_r$ is a lower triangular matrix with 1s on its diagonal. Both $\boldsymbol{D}_r$ and $\boldsymbol{F}_r$ are unique. Let $\boldsymbol{g}_r \sim N(0, \sigma_0^2 \boldsymbol{I}_n)$, the model in Equation (1) can be reparameterized as,

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \sum_r^R d_r \boldsymbol{L_r g_r} + \boldsymbol{e}, \boldsymbol{e} \sim N(0, \sigma_0^2 \boldsymbol{I}_n), \tag{4}$$

where $d_r = \sqrt{\sigma_r^2/\sigma_0^2}$ and $\boldsymbol{L}_r = \boldsymbol{F}_r \boldsymbol{D}_r$.

EM algorithms can be used to estimate parameters in model (4),[28] where the complete data comprised of the observed outcomes ($\boldsymbol{Y}$) and the unobserved random genetic effects ($\boldsymbol{g} = (\boldsymbol{g}_1^T, \boldsymbol{g}_2^T, \dots, \boldsymbol{g}_R^T)^T$). Let $\boldsymbol{d} = (d_1, d_2, \dots, d_R)^T$ and $\boldsymbol{\eta} = (\boldsymbol{\beta}^T, \boldsymbol{d}^T)^T$. Dropping constant terms, the complete data log-likelihood function can be written as,

$$l_c(\boldsymbol{\eta}|\boldsymbol{Y}, \boldsymbol{g}) = -\frac{n(R+1)}{2} \log(\sigma_0^2) - \frac{1}{2\sigma_0^2} \left(\|\boldsymbol{Y} - \boldsymbol{X\beta} - \sum_r^R d_r \boldsymbol{L}_r \boldsymbol{g}_r\|^2 + \boldsymbol{g}^T \boldsymbol{g} \right), \tag{5}$$

where $\|A\|^2$ is the L_2 norm of A. In the E-step, the conditional expectation (denoted as $E^p_{\boldsymbol{g}|\boldsymbol{Y},\boldsymbol{\eta}}$) is computed as $E_{\boldsymbol{g}|\boldsymbol{Y},\boldsymbol{\eta}}(l_c(\boldsymbol{\eta}|\boldsymbol{Y}, \boldsymbol{g})) + \lambda \sum_r^R \omega_r' |d_r|$, where $\omega_r' = 1/\tilde{d}_r$ and $\tilde{d}_r$ is a $\sqrt{n}$ consistent estimator of d_r. In the M-step, $E^p_{\boldsymbol{g}|\boldsymbol{Y},\boldsymbol{\eta}}$ is the maximized with respective to parameters, which is equivalent to minimize Equation (6).

$$Q_c(\boldsymbol{\eta}|\boldsymbol{Y}, \boldsymbol{g}) = E_{\boldsymbol{g}|\boldsymbol{Y},\boldsymbol{\eta}} \left(\|\boldsymbol{Y} - \boldsymbol{X\beta} - \sum_r d_r \boldsymbol{L}_r \boldsymbol{g}_r\|^2 \right) + \lambda \sum_r^R \omega_r' |d_r|. \tag{6}$$

At iteration step t, $\boldsymbol{g}|\boldsymbol{Y}, \hat{\boldsymbol{\eta}^{(t)}} \sim N(\hat{\boldsymbol{g}^{(t)}}, \hat{\boldsymbol{U}^{(t)}})$ with mean and variances are given by

$$\begin{aligned} \hat{\boldsymbol{g}}^{(t)} &= (\boldsymbol{I} + \boldsymbol{B}^{(t)} \boldsymbol{L}^T \boldsymbol{L} \boldsymbol{B}^{(t)})^{-1} (\boldsymbol{L}\boldsymbol{B}^{(t)})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}) \\ \hat{\boldsymbol{U}}^{(t)} &= (\boldsymbol{I} + \boldsymbol{B}^{(t)} \boldsymbol{L}^T \boldsymbol{L} \boldsymbol{B}^{(t)})^{-1} \sigma_0^{2(t)} \\ \sigma_0^{2(t)} &= \sqrt{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)})^T (\boldsymbol{L}\boldsymbol{B}^{(t)} \boldsymbol{B}^{(t)} \boldsymbol{L}^T + \boldsymbol{I})^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)})/N}, \end{aligned} \tag{7}$$

where $\boldsymbol{L} = [\boldsymbol{L}_1, \boldsymbol{L}_2, \dots, \boldsymbol{L}_R]$ is a $n \times (Rn)$ matrix and $\boldsymbol{B}^{(t)}$ is a $(Rn) \times (Rn)$ block diagonal matrix with the rth block equal to $d_r^{(t)} \boldsymbol{I}_n$. For high-dimensional data, inverting a $(Rn) \times (Rn)$ matrix is computationally intensive. However, as shown in Appendix B, $\boldsymbol{g_r}|\boldsymbol{Y}, \boldsymbol{\eta}$ is also normally distributed, and the mean and variances are given by,

$$\begin{aligned} \hat{\boldsymbol{g}}_r^{(t)} &= d_r^{(t)} \boldsymbol{L}_r^T \left(\sum_i^M \boldsymbol{K}^i d_i^{2(t)} + \boldsymbol{I}_N \right)^{-1} \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(t)} \right) \\ \hat{\boldsymbol{U}}_r^{(t)} &= \left(\boldsymbol{I}_n - d_r^{2(t)} \boldsymbol{L}_r^T \left(\sum_i^M \boldsymbol{K}^i d_i^{2(t)} + \boldsymbol{I}_N \right)^{-1} \boldsymbol{L}_r \right) \sigma_0^{2(t)}. \end{aligned} \tag{8}$$

Clearly, for those noise regions (ie, $d_r = 0$), $\boldsymbol{g}_r = 0$ and $\boldsymbol{U}_r = \boldsymbol{I}_n\sigma_0^2$. As most of the genomic regions are not predictive, instead of directly inverting a $(Rn) \times (Rn)$ matrix (Equation (7)), we use Equation (8) to compute the conditional distributions. Therefore, for E-step at iteration step t, the conditional expectation (ie, Equation (6)) is calculated as,

$$Q(\boldsymbol{\eta}|\boldsymbol{\eta}^{(t)}) = E_{\boldsymbol{g}_r|\boldsymbol{Y},\boldsymbol{\eta}^{(t)}}\left(\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \sum_r d_r\boldsymbol{L}_r\boldsymbol{g}_r\|^2\right) + \lambda\sum_r^R \omega_r'|d_r|. \tag{9}$$

For M-step at step t, Equation (9) is minimized with respect to $\boldsymbol{\eta}$, which can be achieved using the quadratic programming. The details for computing the conditional expectation and performing the M-step are in Appendix B. This EM algorithm is designed for a fixed value of λ. To choose the tuning parameter λ, we use a BIC,

$$\text{BIC}_\lambda = -2l(\hat{\boldsymbol{\phi}}) + \log(n) \times (df_\lambda), \tag{10}$$

where df_λ is the number of nonzero coefficient in $\hat{\boldsymbol{\eta}}$.

2.3.2 | Approximate penalized maximum likelihood estimator

The maximization algorithms for linear mixed models are usually based on two basic approaches: the EM and Newton-Raphson (NR) method.[41,42] While the above EM algorithm can be used to estimate parameters, it can be computationally intensive when dealing with high-dimensional data, especially when the tuning parameter λ also needs to be selected. To optimize Equation (6), the NR algorithm whose convergence rate is quadratic with good initial values can also be used. Motivated by the idea used in References 42, and 44, we propose to locally approximate the penalized log-likelihood function as,

$$l_p(\boldsymbol{\phi}) \approx l(\boldsymbol{\phi}^{(0)}) + l'(\boldsymbol{\phi}^{(0)})^T(\boldsymbol{\phi} - \boldsymbol{\phi}^{(0)}) + \frac{1}{2}(\boldsymbol{\phi} - \boldsymbol{\phi}^{(0)})^T l''(\boldsymbol{\phi}^{(0)})(\boldsymbol{\phi} - \boldsymbol{\phi}^{(0)}) - \lambda\sum_r^R \omega_r|\phi_r|. \tag{11}$$

We set $\boldsymbol{\phi}^{(0)}$ to be the maximum likelihood estimator of $l(\boldsymbol{\phi})$ that can be obtained by the existing software (eg, MultiBLUP). It can be shown that the maximizers for Equation (11) can be attained equivalently by using

$$\hat{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi}} \left\{\frac{1}{2}(\boldsymbol{\phi} - \boldsymbol{\phi}^{(0)})^T(-l''(\boldsymbol{\phi}^{(0)}))(\boldsymbol{\phi} - \boldsymbol{\phi}^{(0)}) + \lambda\sum_r^R \omega_r|\phi_r|\right\}. \tag{12}$$

Clearly, Equation (12) can be efficiently solved by the least angel regression (LAR) algorithm that allows computing of the entire regularization path very efficiently.[45] The regularization parameter λ is determined according to the BIC type of criterion specified in Equation (10).

2.4 | Asymptotic properties

The asymptotic properties for the maximum likelihood estimators for the traditional LMMs have been well established under the settings where the number of clusters goes to infinity. However, these results cannot be directly applied to our model. $\boldsymbol{K}^r$ is usually a dense matrix. Therefore, the outcome vector in our model is a single observation from a multivariate normal distribution (ie, $\boldsymbol{Y}_n \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma} = \boldsymbol{I}_n\sigma_0^2 + \sum_r^R \boldsymbol{K}^r\sigma_r^2)$), and the number of clusters is 1 by design in our model. Sweeting[46] and Mardia and Marshall[47] have derived a general result of weak consistency and uniform asymptotic normality for maximum likelihood estimators based on dependent observations. In our work, we consider the same framework established by Mardia and Marshal[47] and use a similar idea introduced by Kyung et al[48] and Chu et al[49] to investigate the asymptotic properties of our estimators. The assumptions of our model are listed in Appendix A.1. The assumptions (S.1) to (S.6) are similar to those used in Mardia et al[47] (detailed proof is shown in Appendix A.2). Together with the assumption (S.7), they yield a central limit theorem for $l'(\boldsymbol{\phi})$ and convergence in probability of $l''(\boldsymbol{\phi})$. To be specific, under the assumptions (S.1) to (S.7), for any $\boldsymbol{\phi} \in \mathbb{R}^p \times \Theta$, as $n \to \infty$, we have $n^{-1/2}l'(\boldsymbol{\phi}) \to_d N(\boldsymbol{0}, \boldsymbol{J}(\boldsymbol{\phi}))$ and $n^{-1}l''(\boldsymbol{\phi}) \to_p -\boldsymbol{J}(\boldsymbol{\phi})$, where $\boldsymbol{J}(\boldsymbol{\phi}) = \text{diag}\{\boldsymbol{J}(\boldsymbol{\beta}), \boldsymbol{J}(\boldsymbol{\theta})\}$ (The detailed proof is shown in Appendix A.3). This result demonstrates the asymptotic

behavior of the first and second derivatives of the log-likelihood function. It also shows that the maximum likelihood estimator is $\sqrt{n}$ consistent and asymptotically normal.

Let $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{10}^T, \boldsymbol{\theta}_{20}^T)^T$ denote the true values of $\boldsymbol{\theta}$. Without loss of generality, $\boldsymbol{\theta_{10}}$ is a $s \times 1$ vector whose components are not zero and $\boldsymbol{\theta_{20}}$ is the $(R+1-s)$ remaining components of $\boldsymbol{\theta_0}$, so that $\boldsymbol{\theta_{20}} = \mathbf{0}$. Let $\boldsymbol{\beta_0}$ denote the true values of $\boldsymbol{\beta}$. Therefore, the true values of $\boldsymbol{\phi}$ can be written as $\boldsymbol{\phi}_0 = (\boldsymbol{\phi}_{10}^T, \boldsymbol{\phi}_{20}^T)^T$, where $\boldsymbol{\phi}_{10} = (\boldsymbol{\beta}_0^T, \boldsymbol{\theta}_{10}^T)^T$ and $\boldsymbol{\phi}_{20} = \boldsymbol{\theta_{20}} = \mathbf{0}$. In a similar manner, $\boldsymbol{\phi}$ can be decomposed as $\boldsymbol{\phi} = (\boldsymbol{\phi}_1^T, \boldsymbol{\phi}_2^T)^T = (\boldsymbol{\beta}^T, \boldsymbol{\theta}_1^T, \boldsymbol{\phi}_2^T)^T = (\boldsymbol{\beta}^T, \boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$. For the penalized log-likelihood function given in Equation (3), let $l(\boldsymbol{\phi}_1) \equiv l\left\{(\boldsymbol{\phi}_1, \mathbf{0})^T\right\}$ and $l_p(\boldsymbol{\phi}_1) \equiv l_p\left\{(\boldsymbol{\phi}_1, \mathbf{0})^T\right\}$ denote the log-likelihood and the penalized log-likelihood of the first s components of $\boldsymbol{\phi}$ (ie, by letting $\boldsymbol{\phi}_2 = \mathbf{0}$ and $\boldsymbol{\phi} = (\boldsymbol{\phi}_1^T, \mathbf{0}^T)^T$), respectively.

In the web appendix A, we showed that the penalized estimators enjoy the oracle property and are asymptotically normally distributed.[52] In particular, we first showed that there exists a local maximizer in a $\sqrt{n}$ neighborhood with $\hat{\boldsymbol{\phi}}_2 = 0$, suggesting that the penalized likelihood estimator can identify the true model with probability tending to 1. To be specific, under the assumptions (S.1) to (S.8) given in Appendix A.1, we have (i) there exists a local maximizer $\hat{\boldsymbol{\phi}} = \left(\hat{\boldsymbol{\phi}}_1, \mathbf{0}\right)^T$ of $l_p(\boldsymbol{\phi}_1)$ such that $\hat{\boldsymbol{\phi}}_1$ is $\sqrt{n}$ consistent for $\boldsymbol{\phi}_{10}$ and (ii) $l_p\left((\boldsymbol{\phi}_1^T, \mathbf{0})^T\right) = \max_{\|\boldsymbol{\phi}_2\| \le Mn^{-1/2}} l_p\left((\boldsymbol{\phi}_1^T, \boldsymbol{\phi}_2^T)^T\right)$ for any $\boldsymbol{\phi}_1$ satisfying $\|\boldsymbol{\phi_1} - \boldsymbol{\phi_{10}}\| \le Mn^{-1/2}$ and some constant $M > 0$. We further showed that the penalized maximum likelihood estimators for those nonzero parameters are asymptotically normally distributed. To be specific, under the assumptions (S.1) to (S.8) given in Appendix A.1, we have

a. $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to_d N(0, J(\boldsymbol{\beta}_0)^{-1})$
b. $\sqrt{n}J(\boldsymbol{\theta}_{10})\left[\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10} + \frac{\lambda_n}{n}J(\boldsymbol{\theta}_{10})^{-1}h(\boldsymbol{\theta}_{10})\right] \to_d N(0, J(\boldsymbol{\theta}_{10}))$,

where $\boldsymbol{J}(\boldsymbol{\theta_{10}})$ consists of the first $s \times s$ upper-left submatrix of $\boldsymbol{J}(\boldsymbol{\theta})$ and $h(\boldsymbol{\theta}_{10}) = (\omega_{p+1}\text{sgn}(\theta_{10}^1), \omega_{p+2}\text{sgn}(\theta_{10}^2), \ldots, \omega_{p+s}\text{sgn}(\theta_{10}^s))$ with θ_{10}^j being the jth element in vector $\boldsymbol{\theta_{10}}$. It is straightforward to see that $\sqrt{n}J(\boldsymbol{\theta}_{10})(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}) \to_d N(0, J(\boldsymbol{\theta}_{10}))$, to the first order.

3 | SIMULATIONS

Simulation studies are conducted to evaluate the performance of KLMM-AL, where for each genomic region, we consider two kernels (ie, the linear kernel and the polynomial kernel with degree 2). We further compare the performances of KLMM-AL with other commonly used methods (ie, gBLUP,[15] MultiBLUP,[8] and MKLMM[9]). In the first scenario, we compare the performance of KLMM-AL with parameters estimated by both EM (denoted by KLMM-AL-EM) and NR with local approximation (denoted by KLMM-AL-NR). In the second scenario, we compare the performance of our KLMM-AL-NR method with the other three existing methods by increasing the number of noise genetic regions. In the third scenario, we evaluate the performance of our method when epistasis (ie, interaction effects) are present. For both MultiBLUP and MKLMM, we use the default settings. As the number of regions for MKLMM are determined empirically, we considered three regions (denoted by MKLMM3) and eight regions (denoted by MKLMM8) for MKLMM in our simulations. For all the simulation studies, we consider two analytical techniques for KLMM-AL method: (i) only additive effects (ie, the linear kernels) are considered (denoted by KLMM-AL-Lin) and (ii) both additive effects (ie, the linear kernels) and interaction effects (ie, polynomial kernel with degree 2) are considered (denoted by KLMM-AL-Adapt).

In all simulation studies, we use the training samples to build predictive models and use the testing samples to evaluate their performance. The Pearson correlation and the mean square error (MSE) in the testing samples are calculated for KLMM-AL, gBLUP, MultiBLUP, and MKLMM. For the KLMM-AL method, we also calculate the chances of selecting predictive and nonpredictive genomic regions. To mimic the distribution of minor allele frequencies and linkage disequilibrium in the real human genome, in all simulations described below the genomic data are drawn from chromosome 1 of the 1000 Genome Project. In particular, we first cut the genome into regions with each being 75 Kb, and then randomly select these genomic regions for each replicate. For all the simulations considered below, three regions are selected as causal for each type of effects. Within each causal region, 20% of the genetic variants are set causal.

3.1 | Simulation I: The comparison between KLMM-AL-NR and KLMM-AL-EM

In this section, we compare the performance of KLMM-AL-NR with KLMM-AL-EM to assess whether the local approximation can achieve similar accuracy as the EM algorithm. In particular, we want to evaluate the impact of sample

size on the local approximation. For this set of simulations, we only consider the additive effects and simulate the phenotypes as:

$$Y_i = \sum_{r=1}^{3} \sum_{j=1}^{N_r} Z_{rj}\beta_{rj}I_{rj} + \epsilon_i, \epsilon_i \sim N(0, \sigma^2), \beta_{rj} \sim N(0, \sigma_r^2), \tag{13}$$

where N_r is the total number of genetic variants on the rth causal region, and Z_{rj} and β_{rj}, respectively, represent the genotype and its effect of the jth genetic variant on the rth causal region. I_{rj} is an indicator with $I_{rj} = 1$ if the jth marker on the rth causal region is causal and $I_{rj} = 0$ otherwise. We set $P(I_{rj}) = 20\%$. The first, second, and third causal regions account for 6.7%, 13.3%, and 20% of the heritability, respectively. In total, all causal genetic variants account for 40% of the heritability.

While keeping the testing sample size being 100, we gradually change the training sample size from 50 to 500. For each sample size setting, we consider two and seven noise regions (ie, total number of regions is five and 10, respectively). We generate 500 replicates for each setting. For each replicate, we use training samples to train the model and use the testing data to assess its performance. We calculate the Pearson correlations and MSEs between the predicted values calculated from KLMM-AL-NR and those from KLMM-AL-EM to assess the consistency between these methods. We further report the selection consistency between these two methods in terms of the chances of correctly identifying causal and noncausal genomic regions.

The consistencies between the predicted values for KLMM-AL-NR and KLMM-AL-EM are summarized in Figure 1. As the training sample size increases, the consistencies between the two methods increase. Indeed, when the training sample size equal to 500, regardless of the number of noise regions, the mean of Pearson correlation between the predicted values derived from these two methods is almost 1, and the mean of MSEs is very close to zero. The selection consistencies

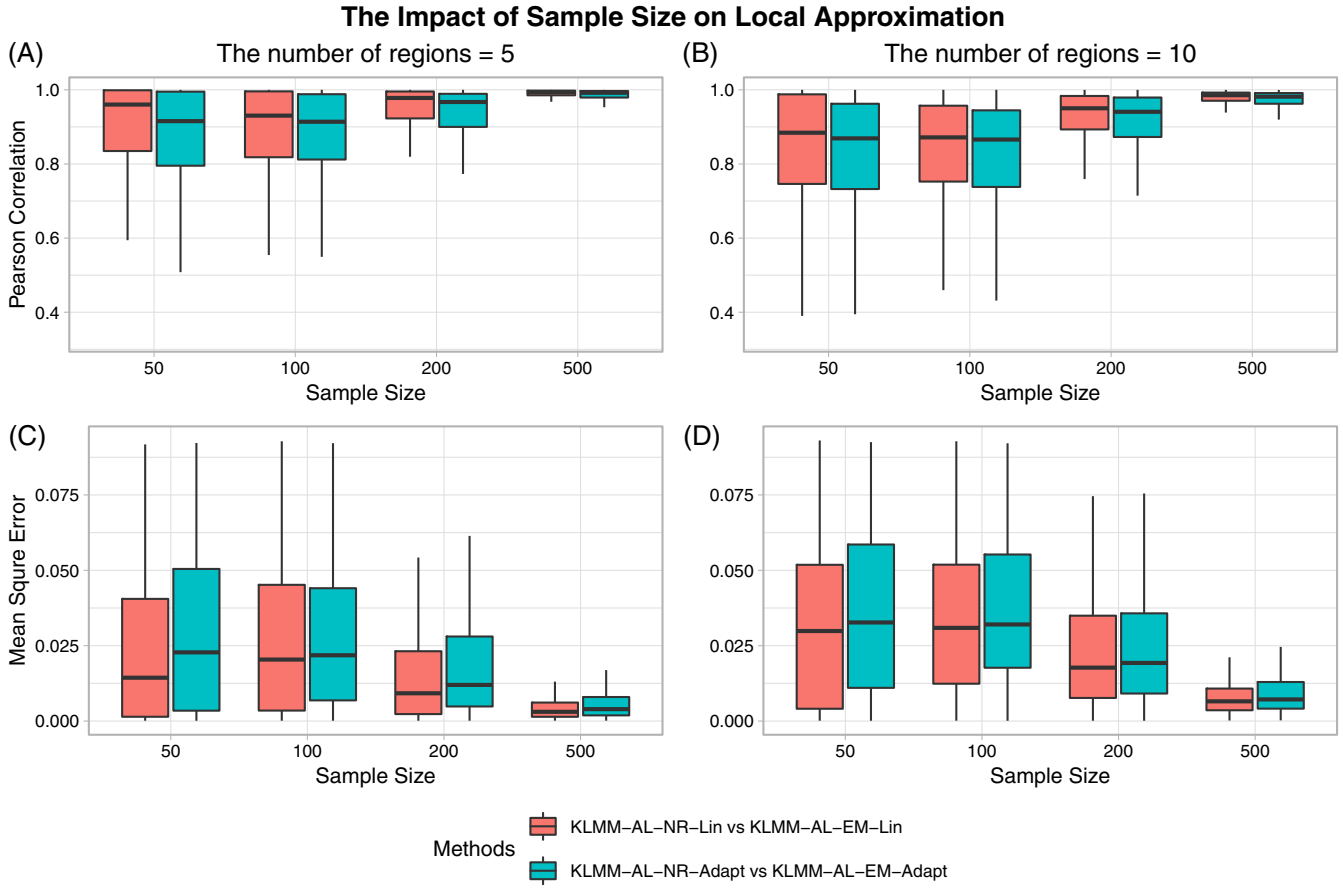

FIGURE 1 The impact of sample size on local approximation for the KLMM-AL method [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 The consistency of selection between KLMM-AL-EM and KLMM-AL-NR

	KLMM-AL-Lin		KLMM-AL-Adapt	
No. Sample	**All[a]**	**True[b]**	**All[a]**	**True[b]**
The number of regions = 5				
50	0.812	0.785	0.821	0.925
100	0.791	0.762	0.827	0.911
200	0.811	0.844	0.805	0.915
500	0.863	0.962	0.841	0.955
The number of regions = 10				
50	0.840	0.768	0.871	0.936
100	0.825	0.727	0.871	0.889
200	0.814	0.829	0.859	0.915
500	0.851	0.955	0.873	0.957

[a]The chances of KLMM-AL-EM and KLMM-AL-NR selecting the same regions.
[b]The chances of KLMM-AL-EM and KLMM-AL-NR selecting the same causal regions.

between the two methods are summarized in Table 1. On average, the chances of selecting the same regions from both methods is 84% and the chances of selecting the same causal regions from both methods is 88%, indicating the selection consistency between KLMM-AL-EM and KLMM-AL-NR is relatively high. Indeed, when the training sample size is 500, the chances of both methods selecting the same causal regions are above 95% (Table 1). This suggests that when the training sample size is sufficiently large, KLMM-AL-NR performs very similar to KLMM-AL-EM with regard to both the predicted values and the selected regions. KLMM-AL-EM can be computationally demanding when the sample size is large due to the selection of tuning parameters (ie, λ). Because for each value of λ, an EM algorithm is used to estimate the parameters as detailed in Section 2.3.1. The KLMM-AL-NR, on the other hand, can efficiently calculate the entire regularization pathways using the LAR algorithm (Section 2.3.2), which can substantially improve the computational efficiencies, especially when both the sample size and the number of genomic regions are large. KLMM-AL-NR can asymptotically achieve similar performance as KLMM-AL-EM, and thus we recommend to use KLMM-AL-NR when the training sample size is relatively large.

3.2 | Simulation II: The impact of the number of noise regions

In this set of simulations, we compare the performance of KLMM-AL with three commonly used methods (ie, gBLUP, MultiBLUP, and MKLMM) by gradually increasing the number of noise genomic regions from two (ie, the total number of regions is five) to 97 (ie, the total number of regions is 100). As shown in Section 3.1, KLMM-AL-NR and KLMM-AL-EM achieve similar performance. Because KLMM-AL-EM requires a substantial amount of computational time, we only focus on KLMM-AL-NR in this set of simulations. Similar to Section 3.1, we only consider additive effects and simulate the phenotypes using Equation (13). For each given number of noise regions, we vary the total heritability and allow different regions contributing differently to the total heritability. Specifically, for the rth causal genomic region, it accounts for $rh/6$ of the heritability, where $r = 1, 2, 3$, and h changes from 20% to 80%. The sample sizes for training samples and testing samples are all set to be 500. Based on 500 Monte Carlo replicates, we calculate the Pearson correlations and MSEs from the testing samples and report the proportions of correctly identifying causal and noncausal genomic regions for the KLMM-AL method.

The Pearson correlations and MSEs are shown in Figure 2. The computational time is shown in Appendix Figure S1. Among all the scenarios considered, KLMM-AL performs better than the other methods. As the number of noise regions increases, the performance of gBLUP drops significantly. While MultiBLUP and MKLMM tend to be more robust compared with gBLUP by allowing for different effect sizes of genetic regions, their performances also drop as the number of noise regions increases, especially when the heritability is high. For the KLMM-AL methods, regardless of whether

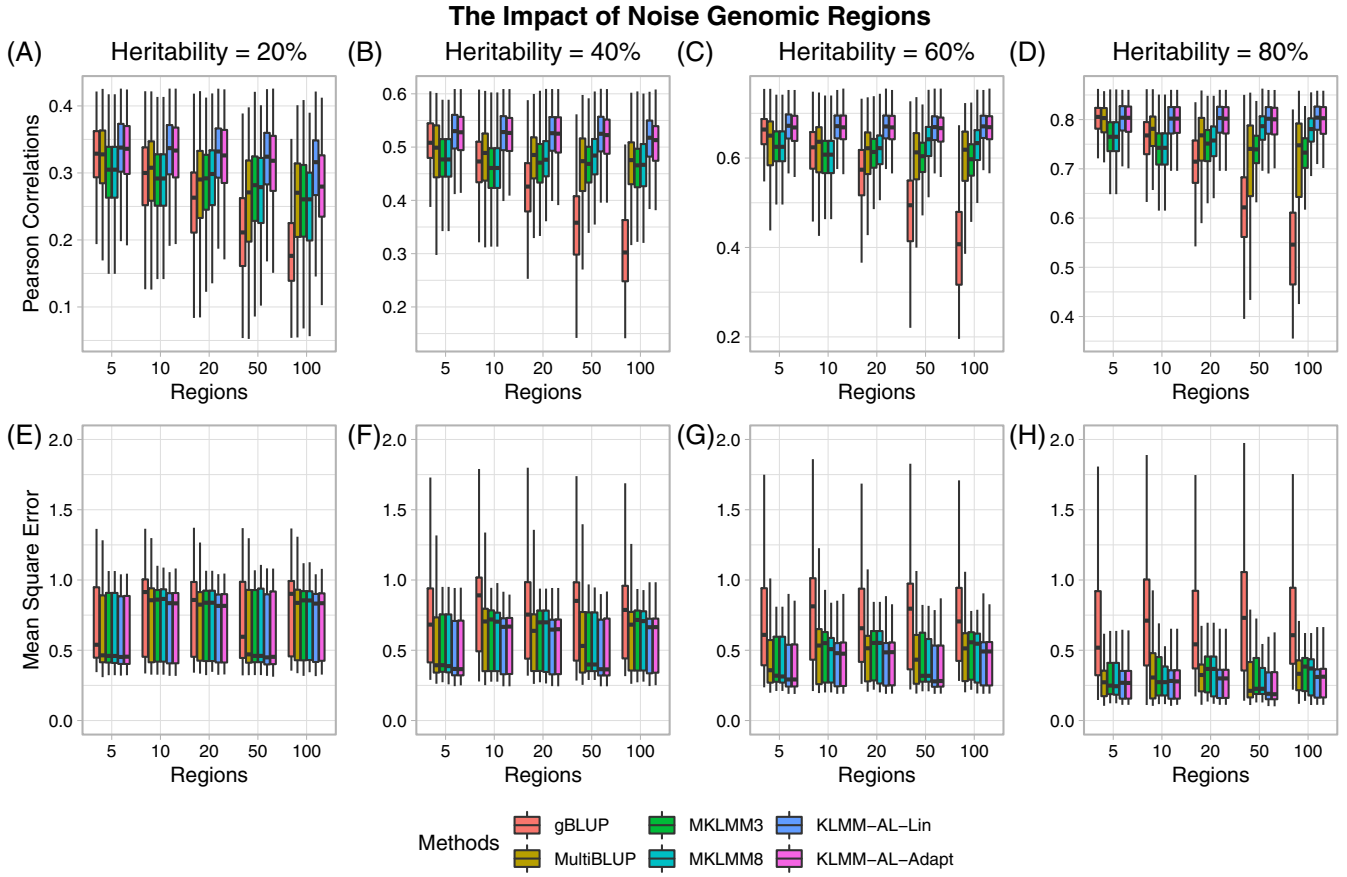

FIGURE 2 The impact of the number of noise genomic regions on Pearson correlations and mean square errors calculated from the testing samples [Colour figure can be viewed at wileyonlinelibrary.com]

we use one kernel (ie, KLMM-AL-Lin) or two kernels per region (ie, KLMM-AL-Adapt), the performances are relatively robust as the number of noise regions increases. This indicates excluding noise regions cannot only improve prediction accuracy, but also improve the robustness of the prediction model.

In practice, the underlying disease model is usually unknown in advance, and thus a model that can adaptively choose the right kernels and achieve accurate prediction is preferred. With regard to variable selection, KLMM-AL-Lin has high sensitivity and specificity (Table 2). KLMM-AL-Adapt may misclassify the predictive effects of the variants located on those causal regions (ie, by selecting the polynomial kernel rather than the linear kernel), but this misclassification rate becomes negligible as the heritability increases. Nevertheless, KLMM-AL-Adapt has a similar proportion of correctly identifying the predictive regions as KLMM-AL-Lin, and its specificity is also very similar to that of KLMM-AL-Lin (Table 2). Moreover, while it is expected that the KLMM-AL-Lin performs better than KLMM-AL-Adapt as the kernel it uses represents the true disease model, the differences of prediction accuracies between these two methods are very small (Figure 2).

3.3 | Simulation III: The impact of epistasis

In this set of simulations, we evaluate the performance of KLMM-AL-NR when interaction effects are present. We further compare its performance with gBLUP, MultiBLUP, and MKLMM. When simulating phenotypes, we consider both the additive effects (denoted by $\boldsymbol{Y}_{\text{av}}$) and the interaction effects (denoted by $\boldsymbol{Y}_{\text{int}}$). The phenotypes $\boldsymbol{Y}$ are simulated as a weighted linear combination of these genetic effects,

$$\boldsymbol{Y} = \boldsymbol{Y}_{\text{av}} + \boldsymbol{Y}_{\text{int}} + \boldsymbol{e}, \quad \boldsymbol{e} \sim N(0, \sigma^2 \boldsymbol{I}). \tag{14}$$

TABLE 2 The chances of selecting predictive/noise regions as the number of noise regions increases

	KLMM-AL-Lin		KLMM-AL-Adapt				
No. Regions	**TP[a]**	**FP[b]**	**TP-Lin[c]**	**FP-Lin[d]**	**TP-Int[e]**	**FP-Int[f]**	**TP-Both[g]**
The heritability = 20%							
5	0.798	0.023	0.594	0.013	0.206	0.013	0.786
10	0.769	0.011	0.564	0.008	0.201	0.007	0.757
20	0.745	0.011	0.570	0.009	0.176	0.007	0.736
50	0.692	0.007	0.519	0.006	0.172	0.006	0.685
100	0.641	0.005	0.450	0.012	0.144	0.007	0.582
The heritability = 40%							
5	0.944	0.007	0.818	0.003	0.130	0.003	0.939
10	0.940	0.008	0.812	0.006	0.138	0.003	0.935
20	0.933	0.006	0.820	0.006	0.130	0.002	0.934
50	0.926	0.006	0.812	0.004	0.130	0.003	0.922
100	0.903	0.003	0.765	0.011	0.143	0.003	0.870
The heritability = 60%							
5	0.937	0.000	0.886	0.000	0.044	0.000	0.922
10	0.936	0.002	0.897	0.001	0.045	0.000	0.932
20	0.936	0.002	0.897	0.002	0.048	0.000	0.934
50	0.934	0.003	0.881	0.003	0.057	0.001	0.921
100	0.950	0.002	0.890	0.004	0.081	0.001	0.932
The heritability = 80%							
5	0.919	0.002	0.910	0.002	0.010	0.000	0.915
10	0.909	0.000	0.907	0.000	0.012	0.000	0.913
20	0.908	0.000	0.902	0.000	0.011	0.000	0.907
50	0.899	0.000	0.893	0.000	0.016	0.000	0.899
100	0.939	0.000	0.907	0.001	0.022	0.000	0.914

[a]The chances of selecting causal regions for KLMM-AL-Lin.
[b]The chances of selecting noise regions for KLMM-AL-Lin.
[c]The chances of selecting causal regions with the additive effects by the linear kernel for KLMM-AL-Adapt.
[d]The chances of selecting noise regions by the linear kernel for KLMM-AL-Adapt.
[e]The chances of selecting causal regions with the additive effects by the polynomial kernel for KLMM-AL-Adapt.
[f]The chances of selecting noise regions by the polynomial kernel for KLMM-AL-Adapt.
[g]The chances of selecting causal regions by any kernels for KLMM-AL-Adapt.

Similar to the above simulations, we simulate three causal regions for each effect. We use $\boldsymbol{Z}_{r,\text{av}}$ ($p_{r,\text{av}}$) and $\boldsymbol{Z}_{r,\text{int}}$ ($p_{r,\text{int}}$) to respectively denote the causal variants (the number of causal variants) on the rth causal region for the additive and interaction effects. We use h_{av} and h_{int} to denote the heritability accounted by the additive and interaction effects, respectively.

For the additive effects, Equation (13) is used. We can show that Equation (13) is equivalent to

$$\boldsymbol{Y}_{\text{av}} \sim N(\boldsymbol{0}, \sum_{r}^{3} \boldsymbol{K}_{r,\text{av}}\sigma^2_{r,\text{av}}), \quad \text{where } \boldsymbol{K}_{r,\text{av}} = \frac{\boldsymbol{Z}_{r,\text{av}}\boldsymbol{Z}^T_{r,\text{av}}}{p_{r,\text{av}}}. \tag{15}$$

Therefore, Equation (15) is used to generate the additive effects. To allow various regions having different effect sizes, the rth causal region ($r = 1, 2, 3$) with the additive effects accounts for $r \times h_{\text{av}}/6$ of the heritability.

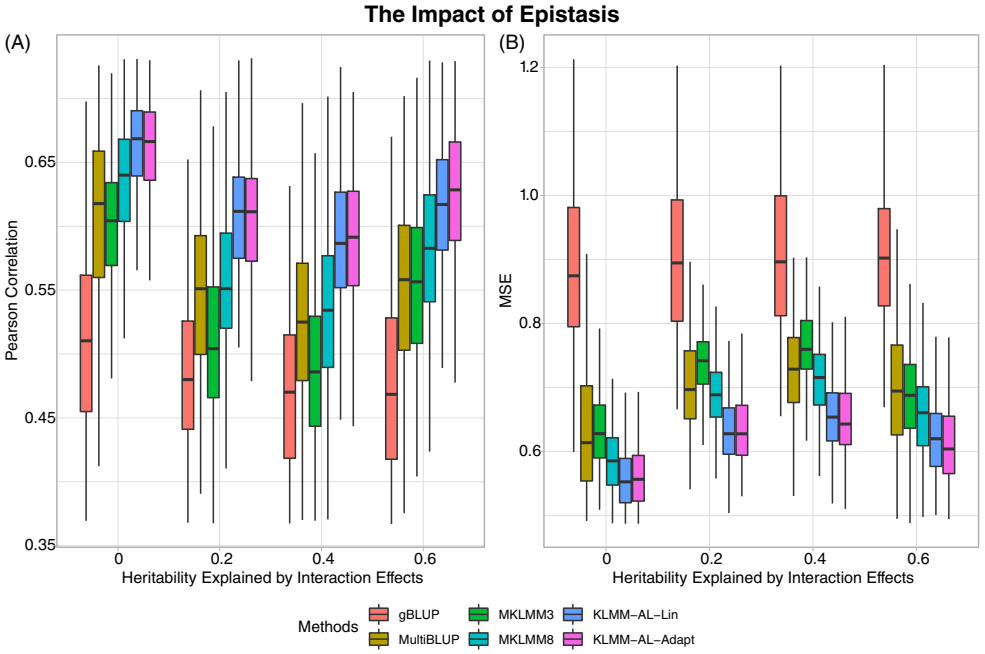

FIGURE 3 The impact of epistasis: the total heritability is 60% [Colour figure can be viewed at wileyonlinelibrary.com]

For the interaction effect, we only consider the local pairwise interactions (ie, interactions within the causal region), and simulate the interaction effects as $\boldsymbol{Y}_{\text{int}} \sim N(\boldsymbol{0}, \sum_r^3 \boldsymbol{K}_{r,\text{int}}\sigma^2_{r,\text{int}})$, where $\boldsymbol{K}_{r,\text{int}}$ is used to capture the interaction effects from the rth causal region. For the pairwise interactions, we set $\boldsymbol{K}_{r,\text{int}} = \boldsymbol{K}_{r,\text{av}} \circ \boldsymbol{K}_{r,\text{av}}$ with $\circ$ being Hadamard product. To allow different effect sizes, the rth causal region ($r = 1, 2, 3$) with the interaction effects accounts for $r \times h_{\text{int}}/6$ of the heritability.

For all the simulations, we fix the total heritability to be 60% (ie, $h_{\text{av}} + h_{\text{int}} = 0.6$) and gradually change the proportion of heritability accounted by the interaction effects (ie, h_{int} increases from 0 to 0.6). We also vary the total number of noise regions (ie, 2, 7, 17, 47, and 97). We set both the training and testing sample sizes being 500. We generate 500 replicates for each setting and evaluate the performances based on Pearson correlations and MSEs calculated from the testing samples.

The prediction accuracy when the total number of regions is 50 is summarized in Figure 3. The remaining results are summarized in Figure S2. KLMM-AL works better than the other methods under all the settings considered, and this indicates teasing out the noise genomic regions can improve prediction accuracy. Comparing KLMM-AL-Adapt with KLMM-AL-Lin, as expected, the KLMM-AL-Adapt performs better than KLMM-AL-Lin when the interaction effects account for a substantial amount of heritability. What we have noticed is that even when the interaction effects are absent, KLMM-AL-Adapt achieves very similar performance to that of KLMM-AL-Lin. With regard to variable selection, KLMM-AL-Lin has high sensitivity and specificity for selecting the prediction genomic regions (on average, sensitivity = 88% and specificity = 99.4%). Although the chances of correctly selecting the genomic regions with the interaction (additive) effects vary with their effect sizes and KLMM-AL-Adapt may misclassify different kinds of effects (Table 3), the sensitivity and specificity of selecting the prediction genomic regions for KLMM-AL-Adapt remains high (on average, sensitivity = 87% and specificity = 99.4%). We consider the robust performance of KLMM-AL-Adapt with respect to both prediction accuracy and region selection is important. This is because the underlying disease model is unknown in advance, and an algorithm that can adaptively choose the kernel functions close to the underlying genetic effects has the potential to improve the prediction accuracy. Indeed, as shown in Figure 3, KLMM-AL-Adapt attains better performance than all the other methods and outperforms KLMM-AL-Lin when the interaction effects are large.

4 | REAL-DATA APPLICATION

We analyze the whole-genome sequencing data from the ADNI using the proposed method with both the linear kernel and the polynomial kernel of degree 2 to capture both the additive and interaction effects (ie, KLMM-AL-Adapt). We further compare our method with commonly used methods, including gBLUP,[15] MultiBLUP,[8] and MKLMM.[9] For MultiBLUP

TABLE 3 The chances of selecting predictive/noise genomic regions when epistasis is present

	KLMM-AL-Lin		KLMM-AL-Adapt						
Her[a]	TP[b]	FP[c]	TP-Lin[d]	FP-Lin[e]	TP-Int[f]	FP-Int[g]	TP-E[h]	TP[i]	FP[j]
The total number of noise regions = 3									
0.6	0.949	0.008	N/A	0.003	0.664	0.002	0.664	0.892	0.005
0.4	0.848	0.010	0.657	0.002	0.597	0.005	0.627	0.817	0.008
0.2	0.817	0.008	0.827	0.005	0.370	0.000	0.598	0.789	0.005
0.0	0.937	0.000	0.886	0.000	N/A	0.000	0.886	0.922	0.000
The total number of noise regions = 7									
0.6	0.950	0.007	N/A	0.004	0.667	0.001	0.667	0.899	0.004
0.4	0.847	0.010	0.666	0.007	0.600	0.003	0.633	0.841	0.010
0.2	0.819	0.008	0.839	0.006	0.381	0.003	0.610	0.807	0.009
0.0	0.936	0.002	0.897	0.001	N/A	0.000	0.897	0.932	0.002
The total number of noise regions = 17									
0.6	0.943	0.003	N/A	0.002	0.666	0.001	0.666	0.902	0.003
0.4	0.843	0.008	0.652	0.006	0.606	0.003	0.629	0.841	0.009
0.2	0.821	0.008	0.835	0.005	0.400	0.003	0.618	0.806	0.008
0.0	0.936	0.002	0.897	0.002	N/A	0.000	0.897	0.934	0.003
The total number of noise regions = 47									
0.6	0.938	0.004	N/A	0.003	0.667	0.001	0.667	0.908	0.004
0.4	0.844	0.008	0.628	0.007	0.629	0.002	0.628	0.846	0.009
0.2	0.809	0.004	0.824	0.005	0.405	0.002	0.614	0.813	0.006
0.0	0.934	0.003	0.881	0.003	N/A	0.001	0.881	0.921	0.003
The total number of noise regions = 97									
0.6	0.932	0.003	N/A	0.006	0.686	0.002	0.686	0.902	0.008
0.4	0.803	0.005	0.621	0.009	0.608	0.002	0.614	0.829	0.011
0.2	0.817	0.005	0.786	0.007	0.370	0.002	0.578	0.784	0.008
0.0	0.950	0.002	0.890	0.004	N/A	0.001	0.890	0.932	0.005

[a]Heritability explained by the interaction effects.
[b]The chances of selecting causal regions for KLMM-AL-Lin.
[c]The chances of selecting noise regions for KLMM-AL-Lin.
[d]The chances of selecting causal regions with the additive effects by the linear kernel for KLMM-AL-Adapt.
[e]The chances of selecting noise regions by the linear kernel for KLMM-AL-Adapt.
[f]The chances of selecting causal regions with the interaction effects by the polynomial kernel for KLMM-AL-Adapt.
[g]The chances of selecting noise regions by the polynomial kernel for KLMM-AL-Adapt.
[h]The chances of selecting causal regions by kernels representing the underlying effects for KLMM-AL-Adapt.
[i]The chances of selecting causal regions by any kernels for KLMM-AL-Adapt.
[j]The chances of selecting noise regions by any kernels for KLMM-AL-Adapt.

and MKLMM, we use the default settings. As the number of regions is preselected for MKLMM, we use both three and eight regions for these analyses.

The ADNI is a longitudinal study designed to assess clinical, imaging, genetic, and biomarkers through the process of normal aging to Alzheimer's disease (AD).[40] Study participants were followed and assessed over time to investigate the pathology of AD. DNA samples were obtained and analyzed using Illumina's non-CLIA whole-genome sequencing. Imaging data (eg, MR imaging and PET imaging) and clinical data (eg, cognitive tests) were also collected at each visit. For our analyses, we are interested in using sequencing data to predict PET-imaging outcomes, FDG, and AV-45 scans, which were performed on all newly enrolled subjects within 2 weeks of the baseline in-clinic assessments.

We annotated the genetic variants based on GRch37 assembly and included a total of 310 genes that have been previously reported to be associated with AD. The complete genes included in our analyses are listed in Table S1. In total, 344,337 single-nucleotide variants are included in the final analyses and the distribution of the minor allele frequencies for these variants are shown in Figure S3. To avoid overfitting, we randomly selected 80 subjects to serve as the testing samples and used the remaining samples to build the model (ie, 422 and 551 samples for AV-45 and FDG, respectively). We calculated the Pearson correlations and the MSEs based on the testing samples. To avoid the chance findings, we repeated this process 100 times.

The results for AV-45 and FDG are shown in Figure 4 and Figure 5, respectively. For both AV-45 and FDG, the Pearson correlations of the KLMM-AL are higher and MSEs of KLMM-AL are smaller than the other methods, suggesting KLMM-AL achieves better prediction accuracy than the existing methods. This indicates that excluding noise genes from the prediction can improve prediction accuracy. The proportion of each gene being selected by KLMM-AL for AV-45 and

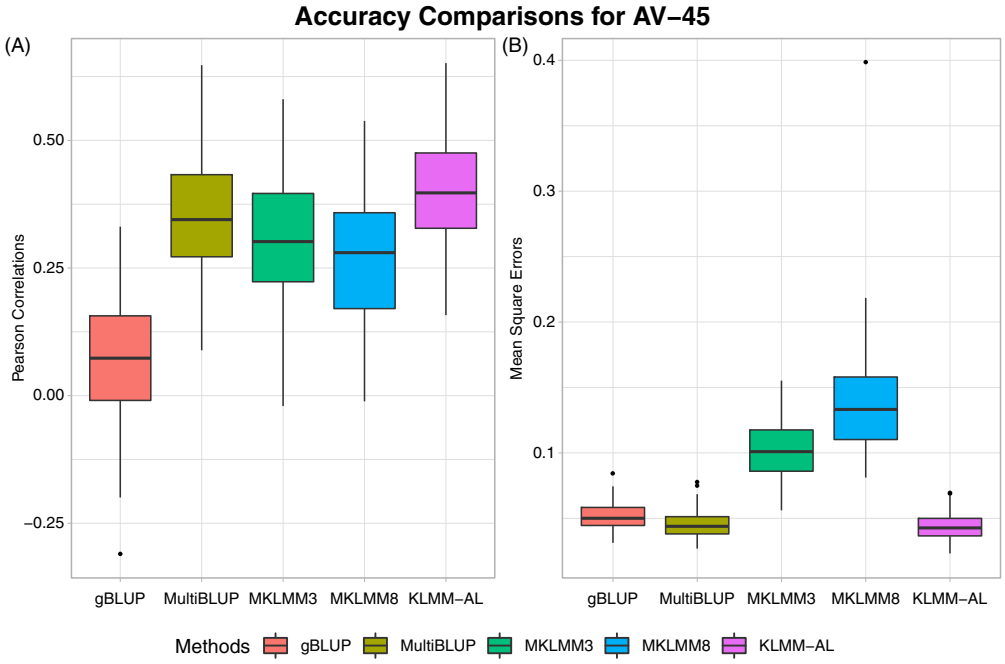

FIGURE 4 Pearson correlations and mean square errors calculated from the testing samples for AV-45 [Colour figure can be viewed at wileyonlinelibrary.com]

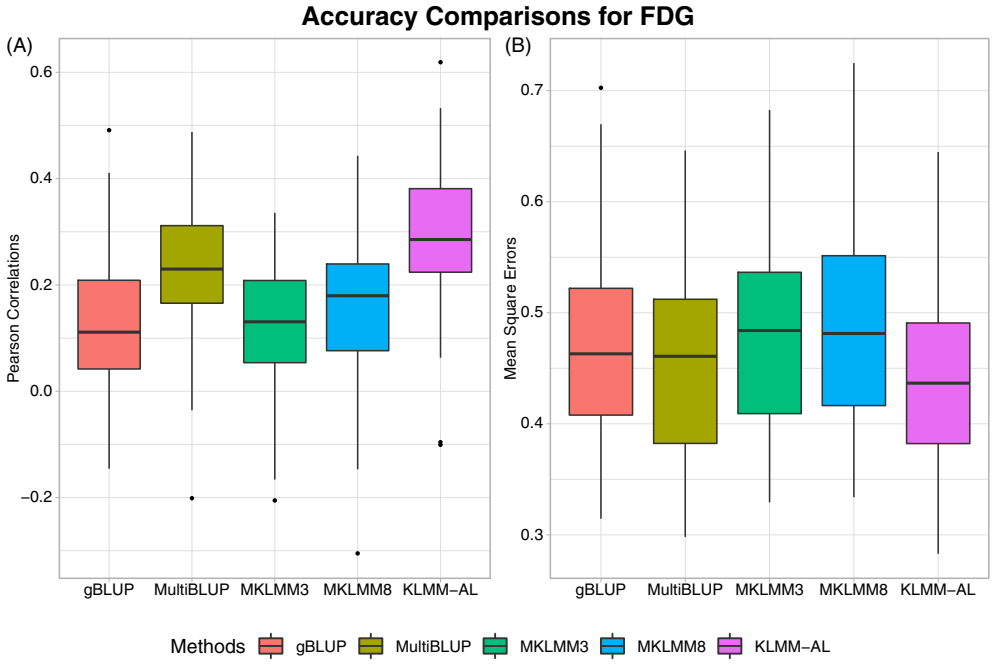

FIGURE 5 Pearson correlations and mean square errors calculated from the testing samples for FDG [Colour figure can be viewed at wileyonlinelibrary.com]

FDG is summarized in Table S1. The KLMM-AL achieves robust performance with regard to the variable selection. The *APOE* gene on chromosome 19, a well-known risk factor for AD, has been selected 100% for both AV-45 and FDG. Only the linear kernel is chosen for *APOE*, suggesting the variants on *APOE* having the additive effects. No other genes are selected for AV-45. For FDG, among the 310 genes, in addition to *APOE*, only four genes have been selected. The fibroblast growth factor (*FGF-1*) located on chromosome 5 has been selected about 63% times, and only the linear kernel (ie, the additive effect) has been selected. *FGF-1* promotes the survival of neurons, and it was reported that serum *FGF-1* in patients with AD was higher than in patients without AD.[50] It was also reported that variants within *FGF-1* were associated with AD in Chinese Han population.[51] The *ADRA1A* gene located on chromosome 8 has been selected 15%, and only the additive effects have been selected. It has been shown that mutations in *ADRA1A* can lead to early onset of AD.[53] The *NTRK1* gene located at chromosome 1 has been selected 24%, and only the interaction effects have been selected. Counts et al[54] found that *NTRK1* expression was reduced in basal forebrain cholinergic neurons through the postmortem examination of the brains of patients with early stage AD. It has been found that rs6336 on *NTRK1* is associated with early-onset AD in Italian population.[55] The gene *CHRNA4* located at chromosome 20 has been selected 14% with both the linear effects and interaction effects. It was reported that genetic polymorphisms in the *CHRNA4* gene were associated with AD.[56,57]

5 | DISCUSSION

We have proposed a multikernel linear mixed model with adaptive lasso for predicting phenotypes using high-dimensional genomic data. We have developed two algorithms to estimate the parameters for our model, and have further established the asymptotic properties of the estimators (ie, the asymptotic behavior of the maximum penalized likelihood estimators under LMM when the outcome vector is a single observation obtained from a multivariate normal distribution). The KLMM-AL can (i) account for heterogeneous effect sizes for different genomic regions by specifying multiple random effects, (ii) capture various types of genetic effects (eg, the additive and pairwise interactions) by using multiple kernel functions per genomic region, and (iii) adaptively and efficiently select predictive regions using the theory built from this work. The software implementing this algorithm can be downloaded from https://github.com/YaluWen/KLMMALPackage.

Through simulation studies, we have demonstrated that the computationally efficient algorithm (ie, KLMM-AL-NR) designed to obtain the approximate penalized maximum likelihood estimators can have similar performance as the exact method (ie, KLMM-AL-EM) with respect to both prediction accuracy and predictive region selection. This makes our model easily scale up to large-scale genetic studies. We also demonstrated that the KLMM-AL is robust against the number of noise regions (Figure 2), whereas the prediction accuracies for other methods drop to various degrees as the number of noise regions increases. Although both MKLMM and MultiBLUP allow different regions contributing differently to the outcomes and build an empirical selection process (ie, selecting the regions based on empirical criteria), neither of them can achieve the same level of prediction accuracy as the KLMM-AL partially due to the lack of theoretical justification of selecting predictive regions. Indeed, MKLMM lets the users to determine the number of regions and its performance depends on the users' decision (Figures 2-5). From simulations, we also show that the data adaptive version of KLMM-AL (ie, KLMM-AL-Adapt) can capture potential interaction effects (Figures 3 and S2) and has relatively high sensitivities and specificities of selecting predictive genomic regions (Tables 2 and 3). Moreover, KLMM-AL-Adpat can also provide some insights into the types of genetic effects (eg, the additvie or pairwise interactions). Although we demonstrate the KLMM-AL-Adapt with only two kernels per region (ie, the linear kernel and the polynomial kernel), it can easily incorporate other kernels to capture more complicated interactions. For example, the saturate pathway kernel[9] can also be implemented into KLMM-AL-Adapt. Through the real-data application, we further demonstrate that the selection of our algorithm is consistent (Table S1) and it can achieve better prediction performance than the existing methods (Figures 4 and 5).

The work introduced in this article focuses on the continuous phenotype with normal distribution. The analysis of binary outcomes within mixed effect model framework can be challenging, as the parameter inference is intractable.[9] Several recent studies have demonstrated that treating binary outcomes as if they were continuous using LMM can achieve reasonable predictions.[8,9] Though easy to implement, it would be interesting to study, within the framework of generalized LMM, other link functions (eg, logit and log) for the prediction of outcomes with various distributions (eg, binary and Poisson) in the future. Similar to many existing parametric models,[8,9] our method depends on the distributional assumptions that can be violated in practice (eg, model misspecification and outcomes from heavy tailed distributions). It could

of great importance to incorporate robust modeling and variable selection methods (an extensive review of such methods can be found in Wu et al[58]) into our proposed framework, and this will be a future direction of our research.

ACKNOWLEDGEMENTS

The project was supported by the National Natural Science Foundation of China (Award No. 81502887), the Faculty Research Development Funds from the University of Auckland, the National Institute on Drug Abuse (Award No. R01DA043501), and the National Library of Medicine (Award No. R01LM012848). The authors wish to acknowledge the contribution of NeSI high-performance computing facilities to the results of this research. NZs national facilities are provided by the NZ eScience Infrastructure and funded jointly by NeSIs collaborator institutions and through the Ministry of Business, Innovation & Employment's Research Infrastructure program.

Author Contributions

Y.W. and Q.L. developed the research idea. Y.W. designed the algorithm and evaluated its theoretical performance. Y.W. also conducted simulation studies and performed analyses for the real example. Y.W. wrote the article and revised following the comments and revisions from Q.L.

Conflict of Interest

The authors declare no potential conflict of interests.

ORCID

Yalu Wen https://orcid.org/0000-0002-0071-5917

REFERENCES

1. Ashley EA. The precision medicine initiative: a new national effort. *JAMA.* 2015;313(21):2119-2120. https://doi.org/10.1001/jama.2015.3595.
2. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med.* 2015;372(9):793-795. https://doi.org/10.1056/NEJMp1500523.
3. Afshari NA, Jr I, Morris NJ, et al. Genome-wide association study identifies three novel loci in Fuchs endothelial corneal dystrophy. *Nat Commun.* 2017;8:14898. https://doi.org/10.1038/ncomms14898.
4. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.* 2007;447(7145):661-678. https://doi.org/10.1038/nature05911.
5. Makowsky R, Pajewski NM, Klimentidis YC, et al. Beyond missing heritability: prediction of complex traits. *PLoS Genet.* 2011;7(4):e1002051. https://doi.org/10.1371/journal.pgen.1002051.
6. Kim H, Grueneberg A, Vazquez AI, Hsu S, Los CG. Will big data close the missing heritability gap? *Genetics.* 2017;207(3):1135-1145. https://doi.org/10.1534/genetics.117.300271.
7. Speed D, Cai N, Consortium Ucleb, Johnson MR, Nejentsev S, Balding DJ. Reevaluation of SNP heritability in complex human traits. *Nat Genet*. 2017;49(7):986-992. https://doi.org/10.1038/ng.3865.
8. Speed D, Balding DJ. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* 2014;24(9):1550-1557. https://doi.org/10.1101/gr.169375.113.
9. Weissbrod O, Geiger D, Rosset S. Multikernel linear mixed models for complex phenotype prediction. *Genome Res.* 2016;26(7):969-979. https://doi.org/10.1101/gr.201996.115.
10. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet.* 2010;42(7):565-569. https://doi.org/10.1038/ng.608.
11. Zeng P, Zhou X. Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nat Commun.* 2017;8(1):456. https://doi.org/10.1038/s41467-017-00470-2.
12. Los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D. Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet.* 2013;9(7):e1003608. https://doi.org/10.1371/journal.pgen.1003608.
13. Los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics.* 2013;193(2):327-345. https://doi.org/10.1534/genetics.112.143313.
14. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. *Genetics.* 2001;157(4):1819-1829.
15. VanRaden PM. Efficient methods to compute genomic predictions. *J Dairy Sci.* 2008;91(11):4414-4423. https://doi.org/10.3168/jds.2007-0980.

16. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet.* 2012;44(7):821-824. https://doi.org/10.1038/ng.2310.
17. Byrnes AE, Wu MC, Wright FA, Li M, Li Y. The value of statistical or bioinformatics annotation for rare variant association with quantitative trait. *Genet Epidemiol.* 2013;37(7):666-674. https://doi.org/10.1002/gepi.21747.
18. Peng H, Lu Y. Model selection in linear mixed effect models. *J Multivar Anal.* 2012;109:109-129. https://doi.org/10.1016/j.jmva.2012.02.005.
19. Mallows CL. Some comments on C_P. *Technometrics.* 1973;15(4):661-675. https://doi.org/10.2307/1267380.
20. Hirotogu A. Information Theory and an Extension of the Maximum Likelihood Principle. In: Parzen E, Tanabe K, Kitagawa G, eds. *Selected Papers of Hirotugu Akaike. Springer Series in Statistics (Perspectives in Statistics).* New York, NY: Springer; 1998:199-213.
21. Pan J, Huang C. Random effects selection in generalized linear mixed models via shrinkage penalty function. *Stat Comput.* 2013;24(5):725-738. https://doi.org/10.1007/s11222-013-9398-0.
22. Nishii R. Asymptotic properties of criteria for selection of variables in multiple regression. *Ann Stat.* 1984;12(2):758-765.
23. Rao R, Wu Y. A strongly consistent procedure for model selection in a regression problem. *Biometrika.* 1989;76(2):369-374. https://doi.org/10.1093/biomet/76.2.369.
24. Schwarz G. Estimating the dimension of a model. *Ann Stat.* 1978;6(2):461-464.
25. Pu W, Niu X-F. Selecting mixed-effects models based on a generalized information criterion. *J Multivar Anal.* 2006;97(3):733-758. https://doi.org/10.1016/j.jmva.2005.05.009.
26. Chen Z, Dunson DB. Random effects selection in linear mixed models. *Biometrics.* 2003;59(4):762-769. https://doi.org/10.1111/j.0006-341X.2003.00089.x.
27. Kinney SK, Dunson DB. Fixed and random effects selection in linear and logistic models. *Biometrics.* 2007;63(3):690-698. https://doi.org/10.1111/j.1541-0420.2007.00771.x.
28. Bondell HD, Krishna A, Ghosh SK. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics.* 2010;66(4):1069-1077. https://doi.org/10.1111/j.1541-0420.2010.01391.x.
29. Ahn M, Zhang HH, Lu W. Moment-based method for random effects selection in linear mixed models. *Stat Sin.* 2012;22(4):1539-1562. https://doi.org/10.5705/ss.2011.054.
30. Lin X. Estimation using penalized quasilikelihood and quasi-pseudo-likelihood in Poisson mixed models. *Lifetime Data Anal.* 2007;13(4):533-544. https://doi.org/10.1007/s10985-007-9071-z.
31. Buil A, Brown AA, Lappalainen T, et al. Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat Genet.* 2015;47(1):88-91. https://doi.org/10.1038/ng.3162.
32. Moore JH, Williams SM. Epistasis and its implications for personal genetics. *Am J Hum Genet.* 2009;85(3):309-320. https://doi.org/10.1016/j.ajhg.2009.08.006.
33. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet.* 2009;10(6):392-404. https://doi.org/10.1038/nrg2579.
34. Vitezica ZG, Varona L, Legarra A. On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics.* 2013;195(4):1223-1230. https://doi.org/10.1534/genetics.113.155176.
35. Zhu Z, Bakshi A, Vinkhuyzen AA, et al. Dominance genetic variation contributes little to the missing heritability for human complex traits. *Am J Hum Genet.* 2015;96(3):377-385. https://doi.org/10.1016/j.ajhg.2015.01.001.
36. Munoz PR, Jr R, Gezan SA, et al. Unraveling additive from nonadditive effects using genomic relationship matrices. *Genetics.* 2014;198(4):1759-1768. https://doi.org/10.1534/genetics.114.171322.
37. Li SY, Cui YH. Gene-centric gene-gene interaction: a model-based kernel machine method. *Ann Appl Stat.* 2012;6(3):1134-1161.
38. Akdemir D, Jannink JL. Locally epistatic genomic relationship matrices for genomic association and prediction. *Genetics.* 2015;199(3):857-871. https://doi.org/10.1534/genetics.114.173658.
39. Ober U, Erbe M, Long N, Porcu E, Schlather M, Simianer H. Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. *Genetics.* 2011;188(3):695-708. https://doi.org/10.1534/genetics.111.128694.
40. Saykin AJ, Shen L, Foroud TM, et al. Alzheimer's disease neuroimaging initiative biomarkers as quantitative phenotypes: genetics core aims, progress, and plans. *Alzheimers Dement.* 2010;6(3):265-273. https://doi.org/10.1016/j.jalz.2010.03.013.
41. Lindstrom MJ, Bates DM. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J Am Stat Assoc.* 1988;83(404):1014-1022.
42. Lin B, Pang Z, Jiang J. Fixed and random effects selection by REML and pathwise coordinate optimization. *J Comput Graph Stat.* 2013;22(2):341-355. https://doi.org/10.1080/10618600.2012.681219.
43. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc.* 2001;96(456):1348-1360. https://doi.org/10.1198/016214501753382273.
44. Fan Y, Li R. Variable selection in linear mixed effects models. *Annals Stat.* 2012;40(4):2043-2068. https://doi.org/10.1214/12-AOS1028.
45. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Ann Stat.* 2004;32(2):407-451.
46. Sweeting TJ. Uniform asymptotic normality of the maximum likelihood estimator. *Ann Stat.* 1980;8(6):1375-1381.
47. Mardia KV, Marshall RJ. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika.* 1984;71(1):135-146. https://doi.org/10.1093/biomet/71.1.135.
48. Kyung M, Ghosh SK. Maximum likelihood estimation for directional conditionally autoregressive models. *J Stat Plan Infer.* 2010;140(11):3160-3179. https://doi.org/10.1016/j.jspi.2010.04.012.

49. Chu T, Zhu J, Wang H. Penalized maximum likelihood estimation and variable selection in geostatistics. *Ann Stat.* 2011;39(5):2607-2625. https://doi.org/10.1214/11-AOS919.
50. Mashayekhi F, Hadavi M, Vaziri HR, Naji M. Increased acidic fibroblast growth factor concentrations in the serum and cerebrospinal fluid of patients with Alzheimer's disease. *J Clin Neurosci.* 2010;17(3):357-359.
51. Tao QQ, Sun YM, Liu ZJ, et al. A variant within FGF1 is associated with Alzheimer's disease in the Han Chinese population. *Am J Med Genet B Neuropsychiatr Genet.* 2014;165B(2):131-136.
52. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc.* 2006;101(476):1418-1429.
53. Kunkle Brian W, Vardarajan Badri N, Naj Adam C, et al. Identification of novel candidate genes for early-onset Alzheimer's Disease through integrated whole-exome sequencing and exome chip array association analysis. *Alzheimer's Dement*. 2016;12(7):P177-P178. https://doi.org/10.1016/j.jalz.2016.06.306.
54. Counts SE, Mufson EJ. The role of nerve growth factor receptors in cholinergic basal forebrain degeneration in prodromal Alzheimer disease. *J Neuropathol Exp Neurol.* 2005;64(4):263-272.
55. Cozza A, Melissari E, Iacopetti P, et al. SNPs in neurotrophin system genes and Alzheimer's disease in an Italian population. *J Alzheimers Dis.* 2008;15(1):61-70.
56. Kawamata J, Shimohama S. Association of novel and established polymorphisms in neuronal nicotinic acetylcholine receptors with sporadic Alzheimer's disease. *J Alzheimers Dis.* 2002;4(2):71-76.
57. Dorszewska J, Florczak J, Rozycka A, Jaroszewska-Kolecka J, Trzeciak WH, Kozubski W. Polymorphisms of the CHRNA4 gene encoding the alpha4 subunit of nicotinic acetylcholine receptor as related to the oxidative DNA damage and the level of apoptotic proteins in lymphocytes of the patients with Alzheimer's disease. *DNA Cell Biol.* 2005;24(12):786-794. https://doi.org/10.1089/dna.2005.24.786.
58. Wu C, Ma S. A selective review of robust variable selection with applications in bioinformatics. *Brief Bioinform.* 2015;16(5):873-883. https://doi.org/10.1093/bib/bbu046.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Wen Y, Lu Q. Multikernel linear mixed model with adaptive lasso for complex phenotype prediction. *Statistics in Medicine*. 2020;39:1311–1327. https://doi.org/10.1002/sim.8477