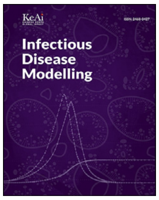
Infectious Disease Modelling 8 (2023) 1079–1087

Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm



High-dimensional supervised classification in a context of non-independence of observations to identify the determining SNPs in a phenotype

Aboubacry Gaye [a, b, *], Abdou Ka Diongue [a], Lionel Nanguep Komen [c], Amadou Diallo [b], Seydou Nourou Sylla [d], Maryam Diarra [b], Cheikh Talla [b], Cheikh Loucoubar [b]

[a] *Laboratory for Studies and Research in Statistics and Development, Gaston Berger University of Saint Louis, Senegal*
[b] *Epidemiology, Clinical Research and Data Science Unit, Institute Pasteur de Dakar, 220, Dakar, Senegal*
[c] *African Institute for Mathematical Sciences (AIMS), Senegal*
[d] *Information and Communication Technologies for Development, Alioune Diop University of Bambey, Senegal*

ARTICLE INFO

Article history:
Received 26 June 2023
Received in revised form 29 August 2023
Accepted 3 September 2023
Available online 9 September 2023
Handling Editor: Dr Daihai He

Keywords:
Non independence of observations
Correlated variables
High-dimensional supervised classification
SNP
Phenotype

ABSTRACT

This work addresses the problem of supervised classification for highly correlated high-dimensional data describing non-independent observations to identify SNPs related to a phenotype. We use a general penalized linear mixed model with a single random effect that performs simultaneous SNP selection and population structure adjustment in high-dimensional prediction models. Specifically, the model simultaneously selects variables and estimates their effects, taking into account correlations between individuals.

Single nucleotide polymorphisms (SNPs) are a type of genetic variation and each SNP represents a difference in a single DNA building block, namely a nucleotide. Previous research has shown that SNPs can be used to identify the correct source population of an individual and can act in isolation or simultaneously to impact a phenotype. In this regard, the study of the contribution of genetics in infectious disease phenotypes is of great importance.

In this study, we used uncorrelated variables from the construction of blocks of correlated variables done in a previous work to describe the most related observations of the dataset. The model was trained with 90% of the observations and tested with the remaining 10%. The best model obtained with the generalized information criterion (GIC) identified the SNP named rs2493311 located on the first chromosome of the gene called PRDM16 ((PR/SET domain 16)) as the most decisive factor in malaria attacks.

* Corresponding author. Laboratory for Studies and Research in Statistics and Development, Gaston Berger University of Saint Louis, Senegal.

E-mail addresses: aboubacry.gaye@pasteur.sn (A. Gaye), abdou.diongue@ugb.edu.sn (A.K. Diongue), lionel.n.komen@aims-senegal.org (L.N. Komen), amadou.diallo@pasteur.sn (A. Diallo), nourou03@gmail.com (S.N. Sylla), maryam.diarra@pasteur.sn (M. Diarra), cheikh.talla@pasteur.sn (C. Talla), cheikh.loucoubar@pasteur.sn (C. Loucoubar).

Peer review under responsibility of KeAi Communications Co., Ltd.

https://doi.org/10.1016/j.idm.2023.09.002

1. Introduction

GWAS has become the standard method for analyzing genetic datasets because of its success in identifying thousands of genetic variants associated with complex diseases. However, the discovered markers could only explain a small proportion of the phenotypic variance. According to (Yang et al., 2010) there are many causal variants that each explain a small amount of variation with small effect sizes. Methods such as GWAS, which test each variant or single nucleotide polymorphism (SNP) independently, may miss these true associations because of the strict significance thresholds required to reduce the number of false positives (Manolio et al., 2009). Another major problem to overcome is confounding due to geographic population structure, family, and/or cryptic kinship that can lead to false associations (Astle & Balding, 2009). Studies that separate their sample by ethnicity to address this confounding suffer from a loss of statistical power due to the decrease in sample size (Bhatnagar et al., 2020).

To address the first problem, multivariate regression methods have been proposed that simultaneously fit many SNPs in a single model (Hoggart et al., 2008; Wang et al., 2010). Studies such as (Lippert et al., 2011; Kang et al., 2010; Yu et al., 2006; Eu-Ahsunthornwattana et al., 2014) have attempted to address confounding by population structure. Today there are two main approaches to accounting for the relationship between subjects: the principal components adjustment (PC) method and the linear mixed model (LMM).

The first includes the upper CPs of the genome-wide SNP genotypes as additional covariates in the model (Price et al., 2006). The second uses a covariance matrix estimated from the genotypes of individuals and includes this information as a random effect (Astle & Balding, 2009).)

Other studies have recently focused on the use of penalized linear mixed models, which constrain the magnitude of effect sizes while controlling for confounding factors such as population structure. Examples include the LMM-lasso (Rakitsch et al., 2013) that places a Laplace prior on all main effects and the adaptive mixed lasso (Wang et al., 2011) that uses the L_1 penalty (Tibshirani, 1996) with adaptively chosen weights (Zou, 2006) to allow for differential shrinkage among variables in the model. Another method applied a combination of lasso and group lasso penalties to select variants within a gene most associated with the response (Ding et al., 2014). However, methods such as LMM-lasso are normally performed in two steps. First, the variance components are estimated once from an LMM with a single random effect. These LMMs normally use the covariance matrix estimated from the genotypes of the individuals to account for relatedness, but assume no main effect of SNP (i.e., a null model). The residuals from this null model with a single random effect can be treated as independent observations because the relationship has been effectively removed from the original response. In the second step, these residuals are used as the response in any high-dimensional model that assumes uncorrelated errors. This approach has both computational and practical advantages since existing penalized regression software such as glmnet (Friedman et al., 2010) and gglasso (Yang & Zou, 2015), which assumes independent observations, can be applied directly to the residuals. However, recent work has shown that there may be a loss of power if a causal variant is included in the computation of the covariance matrix because its effect will have been removed in the first step (Oualkacha et al., 2013; Yang et al., 2014).

In this work, we present a general penalized LMM developed in (Bhatnagar et al., 2020) and called ggmix that simultaneously selects variables and estimates their effects, taking into account correlations between individuals. It is a block coordinate descent algorithm with automatic selection of tuning parameters that is highly scalable, computationally efficient and has theoretical guarantees of convergence. The method can handle several sparsity-inducing penalties such as lasso (Tibshirani, 1996) and Elastic Network (Zou & Hastie, 2005). It works well even in the presence of highly correlated markers and when causal SNPs are included in the relatedness matrix. This method allowed us to identify the most important determinants of malaria access in related populations described by uncorrelated SNPs.

2. Materials and methods

2.1. Model configuration

Let $i = 1, \ldots, N$ be a grouping index, $j = 1, \ldots, n_i$ the observation index within a group and $N_T = \sum_{i=1}^{N} n_i$ the total number of observations. For each group either the observed vector of responses or phenotypes, X_i a $n_i \times (p+1)$ design matrix (with the column of 1 for the intercept), b_i a group-specific random effects vector of length n_i and $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})$ the individual error terms. Denote the stacked vectors $Y = (y_i, \ldots, y_N)^T \in \mathbf{R}^{(p+1)\times 1}$ as a vector of fixed-effects regression coefficients corresponding to X. We consider the following single random effect linear mixed model (Pirinen et al., 2013, pp. 369–390):

$$Y = X\beta + b + \epsilon$$

where the random effect b and the error variance ϵ are assigned to the distributions

$$b \sim \mathbf{N}(0, \eta\sigma^2\Phi) \qquad \epsilon \sim \mathbf{N}(0, (1-\eta)\sigma^2\mathbf{I})$$

here, $\Phi_{N_T \times N_T}$ is a known positive symmetric semidefinite covariance or relatedness matrix computed from SNPs sampled across the genome, $\mathbf{I}_{N_T \times N_T}$ is the identity matrix, and the parameters σ^2 and $\eta \in [0, 1]$ determine how the variance is divided between b and ϵ. Note that η is also the heritability in the strict sense (h^2), defined as the proportion of phenotypic variance attributable to additive genetic factors (Manolio et al., 2009). The joint density of Y is thus multivariate normal:

$$Y|(\beta,\eta,\sigma^2) \sim \mathbf{N}(X\beta, \eta\sigma^2\Phi + (1-\eta)\sigma^2\mathbf{I}) \tag{1}$$

The LMM-Lasso method (Rakitsch et al., 2013) considers an alternative but equivalent parameterization given by:

$$Y|(\beta,\eta,\sigma^2) \sim \mathbf{N}(X\beta, \sigma_g^2(\Phi + \delta\mathbf{I})) \tag{2}$$

where $\delta = \sigma_e^2/\sigma_g^2, \sigma_g^2$ is the genetic variance and σ_e^2 is the residual variance. We consider instead the parameterization in equation (1) since maximization is easier on the compact set $\eta \in [0,1]$ than on the unbounded interval $\delta \in [0, \infty)$ (Pirinen et al., 2013, pp. 369–390]. We define the full parameter vector as $\Theta = (\beta, \eta, \delta^2)$. The negative log-likelihood for equation (1) is given by

$$-l(\Theta) \propto \frac{N_T}{2} log(\sigma^2) + \frac{1}{2} log(det(V)) + \frac{1}{2\sigma^2}(Y - X\beta)^T V^{-1}(Y - X\beta) \tag{3}$$

where $V = \eta\Phi + (1-\eta)\mathbf{I}$ and det($\mathbf{V}$) is the determinant of $\mathbf{V}$.

Let $\Phi = \mathbf{UDU}^T$ be the eigen (spectral) decomposition of the kinship matrix Φ where $\mathbf{U}_{N_T \times N_T}$ is an orthonormal matrix of eigenvectors (i.e. $\mathbf{UU}^T = \mathbf{I}$) and $\mathbf{D}_{N_T \times N_T}$ is a diagonal matrix of eigenvalues $\wedge_i$. $\mathbf{V}$ can then be further simplified (Pirinen et al., 2013, pp. 369–390)

$$\begin{aligned} \mathbf{V} &= \eta\Phi + (1-\eta)\mathbf{I} \\ &= \eta\mathbf{UDU}^T + (1-\eta)\mathbf{UIU}^T \\ &= \mathbf{U}\eta\mathbf{DU}^T + \mathbf{U}(1-\eta)\mathbf{IU}^T \\ &= \mathbf{U}(\eta\mathbf{D} + (1-\eta)\mathbf{I})\mathbf{U}^T \\ &= \mathbf{U}\widetilde{\mathbf{D}}\mathbf{U}^T, \end{aligned} \tag{4}$$

where

$$\begin{aligned} \widetilde{D} &= \eta\mathbf{D} + (1-\eta)\mathbf{I} \\ &= \eta \begin{bmatrix} \wedge_1 & & & \\ & \wedge_2 & & \\ & & \ddots & \\ & & & \wedge_{N_T} \end{bmatrix} + (1-\eta) \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1+\eta(\wedge_1 - 1) & & & \\ & 1+\eta(\wedge_2 - 1) & & \\ & & \ddots & \\ & & & 1+\eta(\wedge_{N_T} - 1) \end{bmatrix} \\ &= diag\{1+\eta(\wedge_1 - 1), 1+\eta(\wedge_2 - 1), \ldots, 1+\eta(\wedge_{N_T} - 1)\}. \end{aligned} \tag{5}$$

Since equation (5) is a diagonal matrix, its inverse is also a diagonal matrix

$$\widetilde{D}^{-1} = diag\left\{\frac{1}{1+\eta(\wedge_1 - 1)}, \frac{1}{1+\eta(\wedge_2 - 1)}, \ldots, \frac{1}{1+\eta(\wedge_{N_T} - 1)}\right\} \tag{6}$$

From equations (4) and (5), $log(det(\mathbf{V}))$ simplifies to

$$\begin{aligned} log(det(\mathbf{V})) &= log(det(\mathbf{U})det(\widetilde{\mathbf{D}})det(\mathbf{U}^T)) \\ &= \log\left\{\prod_{i=1}^{N_T}(1+\eta(\wedge_i - 1))\right\} \\ &= \sum_{i=1}^{N_T} log(1+\eta(\wedge_i - 1)), \end{aligned} \tag{7}$$

since $det(\mathbf{U}) = 1$. It also follows from equation (4) that

$$\begin{aligned}\mathbf{V}^{-1} &= (\mathbf{U}\widetilde{\mathbf{D}}\mathbf{U}^{T})^{-1}\\ &= (\mathbf{U}^{T})^{-1}(\widetilde{\mathbf{D}})^{-1}\Big(\mathbf{U}^{-1}\\ &= \mathbf{U}\widetilde{\mathbf{D}}^{-1}\mathbf{U}^{T},\end{aligned} \tag{8}$$

since for an orthonormal matrix $U^{-1} = U^{T}$. By substituting equation (6), 7 and 8 in equation (3), the negative log-likelihood becomes

$$\begin{aligned}-l(\Theta) &\propto \frac{N_T}{2}log(\sigma^2) + \frac{1}{2}\sum_{i=1}^{N_T} log(1+\eta(\wedge_i - 1)) + \frac{1}{2\sigma^2}(Y - X\beta)^T\mathbf{U}\widetilde{\mathbf{D}}^{-1}\mathbf{U}^T(Y - X\beta)\\ &= \frac{N_T}{2}log(\sigma^2) + \frac{1}{2}\sum_{i=1}^{N_T} log(1+\eta(\wedge_i - 1)) + \frac{1}{2\sigma^2}(\mathbf{U}^TY - \mathbf{U}^TX\beta)^T\widetilde{\mathbf{D}}^{-1}(\mathbf{U}^TY - \mathbf{U}^TX\beta)\\ &= \frac{N_T}{2}log(\sigma^2) + \frac{1}{2}\sum_{i=1}^{N_T} log(1+\eta(\wedge_i - 1)) + \frac{1}{2\sigma^2}(\widetilde{Y} - \widetilde{X}\beta)^T\widetilde{\mathbf{D}}^{-1}(\widetilde{Y} - \widetilde{X}\beta)\\ &= \frac{N_T}{2}log(\sigma^2) + \frac{1}{2}\sum_{i=1}^{N_T} log(1+\eta(\wedge_i - 1)) + \frac{1}{2\sigma^2}\sum_{i=1}^{N_T}\frac{\left(\widetilde{Y}_i - \sum_{j=0}^{p}\widetilde{X}_{ij+1}\beta_j\right)^2}{1+\eta(\wedge_i - 1)},\end{aligned} \tag{9}$$

where $\widetilde{Y} = \mathbf{U^{TY}}$, $\widetilde{X} = \mathbf{U^{TX}}$, $\widetilde{Y}_i$ denotes the i_{th} element of $\widetilde{Y}$, $\widetilde{X}_{ij}$ is the i,j^{th} entry of $\widetilde{X}$ and $\mathbf{1}$ is a column vector of N_T units.

2.2. Likelihood estimator

We define the vector of length $p+4$ of parameters $(\Theta_0, \Theta_1, \ldots, \Theta_{p+1}, \Theta_{p+2}, \Theta_{p+3}) = (\beta, \eta, \sigma^2)$ where $\beta \in \mathbf{R}^{p+1}$, $\eta \in (0,1]$, $\sigma^2 > 0$. In what follows, $p+2$ and $p+3$ are the indices in Θ for η and σ^2, respectively. In light of the objectives of selecting variables associated with the response in high-dimensional data, a constraint placed on the magnitude of the regression coefficients is proposed. This can be achieved by adding a penalty term to the likelihood function 9. The penalty term is a necessary constraint because in our applications, the sample size is much smaller than the number of predictors. We define the following objective function:

$$\mathcal{Q}\lambda(\Theta) = f(\Theta) + \lambda\sum_{j\neq 0} v_j P_j(\beta_j),$$

where $f(\Theta) = -l(\Theta)$ is defined in equation (9) $P_j(\bullet)$ is a penalty term on the fixed regression coefficients $\beta_1, \ldots, \beta_{p+1}$ (we do not penalize the intercept) controlled by the nonnegative regularization parameter λ, and v_j is the penalty factor for the j_{th} covariate. These penalty factors allow the parameters to be penalized differently. Note that η or σ^2 are not penalized. An estimate of the regression parameters $\widetilde{\eta}_\lambda$ is obtained by

$$\widetilde{\eta}_\lambda = argmin_\Theta \mathcal{Q}_\lambda(\Theta). \tag{10}$$

3. Experiments and results

We consider a large genomic dataset consisting of 445 individuals: 235 malaria attack patients and 210 non-attack patients living in the villages of Dielmo and Ndiop. Individuals were genotyped using the Illumina microarray specific to African populations. Genotype data were generated for 719,656 SNPs (Single Nucleotide Polymorphism). For quality control, we excluded from the analysis SNPs with a MAF (Minor Allele Frequency) lower than 10%, or a call rate (% of genotyped individuals for the SNP) lower than 95% or a P-value lower than 10^{-4} for the Hardy-Weinberg Equilibrium test. We then applied the high LD block construction method using the remaining 699083 SNPs that met the quality control parameters. We assume an additive genetic model where the modalities of our variables (0, 1, and 2) count the number of minor alleles present on the SNP. The block partition method based on interval graph modeling investigated in a previous study partitioned the 699083 SNPs into 54150 blocks of high LD SNPs. In this study we described the 30 most related individuals according to their family identifiers with the 54150 representatives of the constituted blocks. Then we applied the ggmix on this dataset. We calculated the kinship matrix with the following formula:

$$\Phi = \frac{1}{p-1} X_{kinship} X_{kinship}^{T}$$

where $X_{kinship}$ is a genotype normalization matrix $n \times p$. The training of our model was done with 90% of observations randomly chosen and whose choice was rotated 100 times. We tested the model on the remaining 10%. The model simulates 100 values of λ, at each value it estimates the β coefficients of the SNPs, calculates the corresponding generalized information criterion (GIC) and selects the significant SNPs. We have the first 11 simulations of λ in 1. At the end of the simulation, the best model is the one with the minimal GIC value, in our case it is the 11^th simulation (last line of Table 1). The table of all the simulations is given in the appendix.

The best model selected only one significant SNP out of the 10000 entered. Fig. 1 shows the evolution of the coefficients of the SNPs over the simulations of λ. On this figure we notice that almost all the coefficients remain close and therefore not significant, except for one which decreases considerably over the simulations. This curve represents the coefficients of only one SNP which is strongly linked to our phenotype Y.

Fig. 2 represents the values of the GIC obtained with the simulations of λ. The value which gives the best model is indicated by the vertical dotted line.

4. Discussion

Most classical methods assume that the observations are independent and identically distributed, however this assumption is not always verified with real data. In this chapter we have tried to perform a supervised classification in high dimension in a context of non-independence of the observations. To do this we used a general penalized mixed linear model with a single random effect called ggmix which performs simultaneous SNP selection and population structure adjustment in high dimensional prediction models. It is a block coordinate descent algorithm with automatic selection of tuning parameters that is highly scalable, computationally efficient and has theoretical guarantees of convergence. In practice, the model simultaneously selects variables and estimates their effects, taking into account the correlations between individuals. We used uncorrelated variables obtained in the previous chapter to describe the most related observations of the data set. The model was trained with 90% of the observations and tested with the remaining 10%. The best model obtained with the generalized information criterion identified a SNP that is strongly related to malaria acces. The negative coefficients of this SNP show that it is a factor that favors malaria access. This SNP is rs2493311 located on chromosome 1 of the PRDM16 gene (PR/SET domain 16). Finding a single significant SNP in malaria access in a population is not new in the literature, as in the Gambia study where only the SNP rs334 in the coding region of HBB on chromosome 11 was identified. Furthermore, in (Fan & Tang, 2013) the generalized information criterion (GIC) selected only one gene where the BIC selected four genes and the AIC selected seven genes in the search for gene expression in acute lymphoblastic leukemia. The association of SNPs in even severe malaria attacks was demonstrated in a study carried out in Mali (Toure et al., 2012), Senegal's neighboring country, where allelic testing revealed potential associations of the HbS polymorphism (rs334, HBB gene), blood group O (and its components rs8176746 and rs8176719) and rs1126535 (CD40L+220) with severe malaria. Furthermore, sickle-cell (HbS) and ABO polymorphisms (rs8176746, rs8176719) have been shown to be significantly associated with severe malaria. The study conducted on sera, DNA samples and clinical data collected from 13,299 individuals at ten sites in Senegal, Mali, Burkina Faso, Sudan, Kenya, Tanzania and Sri Lanka using standardized methods revealed that homozygous recessives for CD36 (rs321198) had significantly lower levels of antimalarial antibodies against MSP2 (merozoite surface protein 2)(Shelton et al., 2015).

We can conclude that in the populations of Dielmo and Ndiop, in central Senegal, the SNP rs2493311 is a determining factor in malaria. The challenge of this method is to perform supervised classification on highly correlated high-dimensional data in a context of non-independence of observations. In this paper, we assume that the observations are not independent, in contrast to the classical model which assumes that the observations are independent and equally distributed. In addition, to solve the problem of uncorrelated SNPs in related populations, we use a well-established penalizing linear mixture model.

Table 1
The first 11 simulations of *lambda* with the corresponding coefficients and GIC.

Λ	loglik	(Intercept)	rs2493311	σ^2	GIC
0.648420700	−49.7910049	0.6231568	0.00000000	2.3429541484	121.5516
0.618948979	−49.1129300	0.7837084	−0.09037159	2.2281793593	131.1803
0.590816793	−48.4576983	0.9385106	−0.17742977	2.1226156764	129.8698
0.563963258	−47.8319564	1.0861968	−0.26049107	2.0264751202	128.6183
0.538330257	−47.2354014	1.2271647	−0.33977427	1.9388763958	127.4252
0.513862316	−46.6677609	1.3617561	−0.41546925	1.8590417819	126.2899
0.490506481	−46.1290651	1.4901850	−0.48770125	1.7863202829	125.2125
0.468212204	−45.6185458	1.6128347	−0.55667920	1.7200296799	124.1915
0.446931237	−45.1358958	1.7299222	−0.62252813	1.6596217166	123.2262
0.426617522	−44.6804051	1.8417334	−0.68540692	1.6045601105	122.3152
0.407227098	−44.2517994	1.9483952	−0.74539408	1.5544178120	121.4580

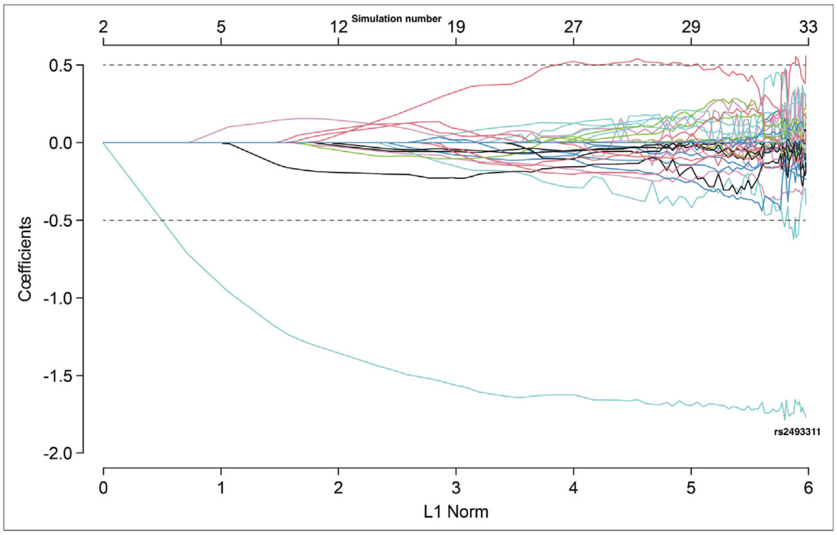


Fig. 1. Coefficients of SNPs according to the 100 simulated λ values.

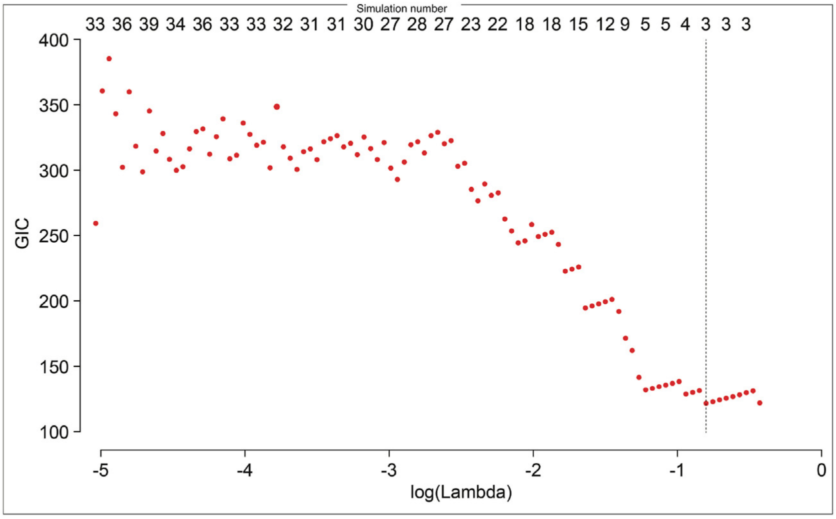


Fig. 2. GIC according to the 100 simulated *lambda* values.

This represents a significant departure from classical methods. Recent works such as (Mieth et al., 2016), which proposes a new, principled, reliable and reproducible methodology for identifying significant SNP-phenotype associations, have addressed the same problem as we have, but their method does not take into account the non-independence of observations.

In 2008 the authors of (Liang & Kelemen, 2008) presented a review of recent statistical advances and challenges related to the analysis of high-dimensional correlated SNP data in genomic association studies for complex diseases, but most of these methods do not address the non-independence of observations.

However we have noted that an increase in the number of observations and variables degrades considerably the kinship matrix and leads to a non convergence of the model. In the future we will try to overcome this limitation by exploring other methods of calculating the parentage matrix.

Declaration of competing interest

Authors declare no conflict of interest.

Appendix

Simulation	lambda	(Intercept)	rs2493311_1	eta	sigma2
s1	0.650798455858028	0.719307791844186	0	0.01	2.33796862512915
s2	0.621218661637185	0.91437808672394	−0.105402093458432	0.01	2.20484485246131
s3	0.592983314715305	1.10210559217338	−0.206764417955178	0.01	2.08261845332914
s4	0.566031307887713	1.28133743785218	−0.30353718491118	0.01	1.9712281262122
s5	0.540304311366496	1.45235489941225	−0.395879142765924	0.01	1.86976663303092
s6	0.515746646542624	1.61577422517146	−0.484107152025397	0.01	1.77723053567375
s7	0.492305165485778	1.77158368856428	−0.5682384414176	0.01	1.69300081332361
s8	0.469929135921098	1.92036950362605	−0.648573470554722	0.01	1.61622514808378
s9	0.448570131433913	2.06246475979266	−0.725291371558906	0.01	1.54623794731717
s10	0.428181926664838	2.1470850045096	−0.788386808994394	0.01	1.46905592186024
s11	0.408720397268412	2.22467272719357	−0.847907720261307	0.01	1.3980533366455
s12	0.390143424418774	2.2987330111569	−0.904722983531617	0.01	1.33335808175022
s13	0.372410803655704	2.37914510574842	−0.95912566352525	0.01	1.27243142132112
s14	0.355484157873745	2.50205702343823	−1.00954405190728	0.01	1.20755344280533
s15	0.339326854266115	2.61927766439498	−1.05764607135933	0.01	1.14845354567158
s16	0.323903925043636	2.73124367755361	−1.10356847217273	0.01	1.09459589265507
s17	0.309181991757113	2.83792015068815	−1.14738095357451	0.01	1.04554331237901
s18	0.295129193059384	2.93985942145303	−1.18921367918938	0.01	1.00083666885614
s19	0.281715115750687	3.0037739854799	−1.22709481569511	0.01	0.955323517376495
s20	0.268910728958128	3.02227655888651	−1.26076564456469	0.01	0.90521159454549
s21	0.25668832130679	3.07618306899497	−1.29010159706772	0.01	0.853287847744785
s22	0.245021440946514	3.12348209396986	−1.31729076139299	0.01	0.802313126612538
s23	0.233884838304553	3.15543942235609	−1.34299908486159	0.01	0.751973039401129
s24	0.223254411440214	3.19683648573182	−1.36840866485208	0.01	0.705849446653076
s25	0.213107153883202	3.23558630878388	−1.39271471874433	0.01	0.66379935129776
s26	0.203421104842807	3.27129656584732	−1.41595642066043	0.01	0.625458122872588
s27	0.194175301680146	3.30343693718036	−1.43800068219231	0.01	0.590564288556273
s28	0.18534973454063	3.35999583687332	−1.45704383554702	0.01	0.55661407683881
s29	0.176925303048439	3.42614259048904	−1.47554876262015	0.01	0.524695810436239
s30	0.168883774969314	3.48912121710709	−1.49319229971486	0.01	0.495668422385932
s31	0.161207746752184	3.55219098031935	−1.50515919745695	0.01	0.467459924370625
s32	0.153880605864229	3.6225654294685	−1.51664042399897	0.01	0.439180433146044
s33	0.146886494837887	3.71237087573108	−1.53167051334806	0.01	0.41278095499759
s34	0.140210276951972	3.96489056823753	−1.55200820449596	0.01	0.386575443169847
s35	0.133837503472633	4.12353056217201	−1.56543689714811	0.01	0.363780285607834
s36	0.127754382383275	4.2408842574395	−1.58250608830419	0.01	0.343855330830184
s37	0.121947748535741	4.37251401825539	−1.60207869286134	0.01	0.324964403250948
s38	0.116405035158177	4.41918124515789	−1.6143521678223	0.01	0.303892847312529
s39	0.111114246657904	4.44418356418651	−1.62505868275677	0.01	0.283058326959088
s40	0.106063932660445	4.47133898946828	−1.63637926585448	0.01	0.26367296522505
s41	0.101243163228512	4.55307553457703	−1.64150483760283	0.01	0.24508281276557
s42	0.0966415052073385	4.57018796379146	−1.63253062551628	0.01	0.226491139206482
s43	0.0922489996451415	4.57785468098522	−1.62541327715187	0.01	0.209876090803177
s44	0.0880561402398679	4.51728330046298	−1.62181616642271	0.01	0.194091966861866
s45	0.0840538527655639	4.45196843129541	−1.62359231963495	0.01	0.179273623218792
s46	0.0802334754338501	4.371691072325	−1.62236171539781	0.01	0.16497031343858
s47	0.0765867401479968	4.2920932956583	−1.63750404508623	0.01	0.151354328675346
s48	0.0731057546090312	4.32659941783817	−1.65635039086076	0.01	0.138490659258711
s49	0.0697829852351495	4.29947696379294	−1.66001896180604	0.01	0.126848351504806
s50	0.0666112408574676	4.22697839826831	−1.65621194588633	0.01	0.116278982947415
s51	0.0635836571568253	4.19592288565156	−1.65771782498483	0.01	0.106638002881586
s52	0.0606936818079628	4.17700349870347	−1.6618883727196	0.01	0.0977660281200947
s53	0.0579350602989153	4.12891539524687	−1.65458434886023	0.01	0.0898066674103416
s54	0.0553018223949399	4.23150385847896	−1.67761284321345	0.01	0.0824123386884932
s55	0.0527882692176768	4.3107693799249	−1.68569240582706	0.01	0.0757332020801086
s56	0.0503889609115827	4.20121182226923	−1.68759706889439	0.01	0.0695160339047503
s57	0.0480987048709448	3.93285596487975	−1.67160387409436	0.01	0.0634702809793338
s58	0.0459125445019934	4.17488084615611	−1.6950949874999	0.01	0.0587694506692449
s59	0.0438257484957956	3.79482766138991	−1.66005338906438	0.01	0.0535051037168795
s60	0.0418338005887114	3.87953553225343	−1.68753195154084	0.01	0.049326333354842
s61	0.0399323897882533	3.82281283862449	−1.69257812057704	0.01	0.0450209958042109
s62	0.038117401043196	3.77154741724061	−1.66952775581002	0.01	0.0413202724265934
s63	0.0363849063377429	3.9958181238604	−1.69945774075509	0.01	0.0374023937215729
s64	0.0347311561904779	3.75975503944738	−1.68073165817969	0.01	0.0341527411675211
s65	0.0331525715397018	3.67146028374316	−1.68249978845243	0.01	0.031279239338519
s66	0.0316457359975934	3.55578427654636	−1.68809322441416	0.01	0.0289476348832512

(*continued on next page*)

(*continued*)

Simulation	lambda	(Intercept)	rs2493311_1	eta	sigma2
s67	0.0302073884564306	3.73732707634953	−1.69293124127442	0.01	0.026052896059556
s68	0.0288344160308703	4.11769988781948	−1.71374056983016	0.01	0.0234887351597365
s69	0.0275238473210125	3.814802594525	−1.69996487871997	0.01	0.0217285756653297
s70	0.0262728459816685	3.96389929857223	−1.70471769067508	0.01	0.019729160006076
s71	0.0250787045839157	4.12500962132885	−1.71637992813799	0.01	0.0180597138831629
s72	0.0239388387556548	4.21981719029377	−1.72127424830992	0.01	0.0165542802961482
s73	0.0228507815884868	3.97531968055072	−1.69558874663295	0.01	0.0152196260687555
s74	0.0218121782988067	4.13989173546537	−1.72150373718836	0.01	0.0137985440010616
s75	0.0208207811315587	4.19768889568811	−1.71198314786973	0.01	0.0126229441637552
s76	0.0198744444956232	4.17297408733024	−1.71447339393066	0.01	0.0115155103778064
s77	0.0189711203203084	4.16078226948635	−1.66976035517302	0.01	0.0104258829465843
s78	0.0181088536228963	3.89678876369223	−1.74219393634947	0.01	0.00955164480346669
s79	0.0172857782776508	3.91342294109001	−1.74015282165734	0.01	0.00869896568274796
s80	0.0165001129771302	3.90713869568212	−1.73407996557493	0.01	0.00789143894094838
s81	0.0157501573770655	3.7494652953537	−1.71403339375951	0.01	0.00719626704387584
s82	0.0150342884164589	3.85839221997726	−1.73027793332792	0.01	0.00654786503082193
s83	0.0143509568049397	3.85971351116912	−1.72718160676676	0.01	0.00596537554459938
s84	0.0136986836697759	3.83965164547688	−1.70968281425755	0.01	0.00541531154191684
s85	0.0130760573552832	3.43219153199911	−1.67333992967833	0.01	0.00502137490147179
s86	0.0124817303677072	3.45185931417515	−1.78917795293387	0.01	0.00461877194970396
s87	0.0119144164589641	3.31114874457819	−1.65860235332315	0.01	0.00417361990880932
s88	0.0113728878429305	4.40894129806171	−1.73300991717024	0.01	0.00378173935171943
s89	0.0108559725382575	4.50859497738778	−1.76384420672811	0.01	0.0034296696429756
s90	0.0103625518319569	4.44711819917868	−1.74087352580782	0.01	0.00316477176769599
s91	0.00989155785827272	4.36075970515912	−1.74455255038913	0.01	0.00289306088952141
s92	0.00944197128759546	3.8446367070032	−1.65239123460242	0.01	0.00264907398314267
s93	0.0090128191204196	4.12646759517198	−1.6701359379359	0.01	0.00241130994998773
s94	0.00860317258156879	3.74286829229794	−1.687319661676	0.01	0.00220269364945199
s95	0.0082121451101319	3.91978864884732	−1.66372679886098	0.01	0.00201265598355357
s96	0.00783889044075943	3.19625252475614	−1.7349733740993	0.01	0.00182470736391412
s97	0.00748260077216812	3.2197432875659	−1.70247717783561	0.01	0.00164074790505021
s98	0.00714250501888972	3.64823453615657	−1.74149600799325	0.01	0.00151731470315982
s99	0.00681786714248059	3.58843909999135	−1.74499782912473	0.01	0.0013833824617226
s100	0.00650798455858028	3.61200168739519	−1.76478388829111	0.01	0.00126377640632813

References

Astle, W., & Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science, 24*(4), 451–471.

Bhatnagar, S. R., Yang, Y., Lu, T., Schurr, E., Loredo-Osti, J., Forest, M., Oualkacha, K., & Greenwood, C. M. (2020). Simultaneous snp selection and adjustment for population structure in high dimensional prediction models. *PLoS Genetics, 16*(5), Article e1008766.

Ding, X., Su, S., Nandakumar, K., Wang, X., & Fardo, D. W. (2014). A 2-step penalized regression method for family-based next-generation sequencing association studies. *BMC Proceedings, 8*, 1–6 (BioMed Central).

Eu-Ahsunthornwattana, J., Miller, E. N., Fakiola, M., W. T. C. C. C, Jeronimo, S. M., Blackwell, J. M., & Cordell, H. J. (2014). Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genetics, 10*(7), Article e1004445.

Fan, Y., & Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B, 75*(3), 531–552.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, 33*(1), 1.

Hoggart, C. J., Whittaker, J. C., De Iorio, M., & Balding, D. J. (2008). Simultaneous analysis of all snps in genome-wide and re-sequencing association studies. *PLoS Genetics, 4*(7), Article e1000130.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C., & Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics, 42*(4), 348–354.

Liang, Y., & Kelemen, A. (2008). *Statistical advances and challenges for analyzing correlated high dimensional snp data in genomic study for complex diseases*.

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

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747–753.

Mieth, B., Kloft, M., Rodríguez, J. A., Sonnenburg, S., Vobruba, R., Morcillo-Suárez, C., Farré, X., Marigorta, U. M., Fehr, E., Dickhaus, T., et al. (2016). Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. *Scientific Reports, 6*(1), Article 36671.

Oualkacha, K., Dastani, Z., Li, R., Cingolani, P. E., Spector, T. D., Hammond, C. J., Richards, J. B., Ciampi, A., & Greenwood, C. M. (2013). Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genetic Epidemiology, 37*(4), 366–376.

Pirinen, M., Donnelly, P., & Spencer, C. C. (2013). *Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies*. The Annals of Applied Statistics.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics, 38*(8), 904–909.

Rakitsch, B., Lippert, C., Stegle, O., & Borgwardt, K. (2013). A lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics, 29*(2), 206–214.

Shelton, J. M., Corran, P., Risley, P., Silva, N., Hubbart, C., Jeffreys, A., Rowlands, K., Craik, R., Cornelius, V., Hensmann, M., et al. (2015). Genetic determinants of anti-malarial acquired immunity in a large multi-centre study. *Malaria Journal, 14*, 1–18.

Tibshirani, 1996 Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B, 58*(1), 267–288.

Toure, O., Konate, S., Sissoko, S., Niangaly, A., Barry, A., Sall, A. H., Diarra, E., Poudiougou, B., Sepulveda, N., Campino, S., et al. (2012). *Candidate polymorphisms and severe malaria in a malian population*.

Wang, D., Eskridge, K. M., & Crossa, J. (2011). Identifying qtls and epistasis in structured plant populations using adaptive mixed lasso. *Journal of Agricultural, Biological, and Environmental Statistics, 16*(2), 170–184.

Wang, K., Li, M., & Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics, 11*(12), 843–854.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., et al. (2010). Common snps explain a large proportion of the heritability for human height. *Nature Genetics, 42*(7), 565–569.

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., & Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics, 46*(2), 100–106.

Yang, Y., & Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing, 25*(6), 1129–1141.

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics, 38*(2), 203–208.

Zou, 2006 Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B, 67*(2), 301–320.