Bioinformatics, 37(18), 2021, 2896–2904
doi: 10.1093/bioinformatics/btab212
Advance Access Publication Date: 28 March 2021
Original Paper

OXFORD

Genetics and population analysis

$L_{2,1}$-norm regularized multivariate regression model with applications to genomic prediction

Alain J. Mbebi [1,2], Hao Tong[1,2,3] and Zoran Nikoloski[1,2,3,*]

[1]Systems Biology and Mathematical Modeling Group, Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam-Golm, Germany, [2]Bioinformatics Group, Institute of Biochemistry and Biology, University of Potsdam, 14476 Potsdam-Golm, Germany and [3]Center for Plant Systems Biology and Biotechnology, Ruski 139, 4000 Tsentar, Plovdiv, Bulgaria

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on June 16, 2020; revised on March 16, 2021; editorial decision on March 22, 2021; accepted on March 26, 2021

Abstract

Motivation: Genomic selection (GS) is currently deemed the most effective approach to speed up breeding of agricultural varieties. It has been recognized that consideration of multiple traits in GS can improve accuracy of prediction for traits of low heritability. However, since GS forgoes statistical testing with the idea of improving predictions, it does not facilitate mechanistic understanding of the contribution of particular single nucleotide polymorphisms (SNP).

Results: Here, we propose a $L_{2,1}$-norm regularized multivariate regression model and devise a fast and efficient iterative optimization algorithm, called $L_{2,1}$-joint, applicable in multi-trait GS. The usage of the $L_{2,1}$-norm facilitates variable selection in a penalized multivariate regression that considers the relation between individuals, when the number of SNPs is much larger than the number of individuals. The capacity for variable selection allows us to define master regulators that can be used in a multi-trait GS setting to dissect the genetic architecture of the analyzed traits. Our comparative analyses demonstrate that the proposed model is a favorable candidate compared to existing state-of-the-art approaches. Prediction and variable selection with datasets from *Brassica napus*, wheat and *Arabidopsis thaliana* diversity panels are conducted to further showcase the performance of the proposed model.

Availability and implementation: : The model is implemented using R programming language and the code is freely available from https://github.com/alainmbebi/L21-norm-GS.

Contact: nikoloski@mpimp-golm.mpg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

First introduced in Meuwissen *et al.* (2001), genomic selection (GS) is considered the most promising breeding method to speed up the development and release of improved genotypes (Crossa *et al.*, 2017). It uses machine learning approaches to integrate phenotypic data of a given trait with molecular markers [i.e. single nucleotide polymorphisms (SNPs)] in a statistical model for a training population. The model is then used to predict genomic estimated breeding values for the trait of genotypes in a testing population, which have been genotyped but not phenotyped (Hayes *et al.*, 2001). The predictions for unseen genotypes can be used for selection without any further phenotyping. Therefore, an increase in GS accuracy for agronomically important traits can accelerate genetic gain by shortening the breeding cycles (Heffner *et al.*, 2010).

Early applications of GS used diverse machine learning approaches to predict individual traits in a setting where the number of SNPs is larger than the size of the training population. Most widely applied approaches include regularized mixed effect models, such as: the ridge regression best linear unbiased prediction (rrBLUP) (Henderson, 1975), its variant genomic-BLUP (GBLUP) (VanRaden, 2008), BayesA and BayesB (Meuwissen *et al.*, 2001), BayesC π (Habier *et al.*, 2011) and the BayesLASSO (Park and Casella, 2008), to mention a few. In addition to the Bayesian regression family, that induces model sparseness by an appropriate prior density (e.g. Student-*t*) for regression coefficients, regularized high-dimensional regressions have also been used, including: the ridge regression (RR) (Hoerl and Kennard, 1970; Ogutu *et al.*, 2012), the LASSO (Usai *et al.*, 2009) and the elastic-net (Wang *et al.*, 2019).

Experience from breeding programs indicates that genetic correlations between traits are quite common, and can thereby be exploited since one trait carries information about others (Jia and Jannink, 2012). Several studies already proposed multi-trait GS models and tested their effects on data from simulations and crop breeding programs. These models account for the genetic (co)variance between the traits, and their applications have shown that

predictability for low-heritability traits can be increased by multi-trait GS (Calus and Veerkamp, 2011; Karaman *et al.*, 2018). These approaches rely on vectorizing the matrix of traits (i.e. responses) and fitting the BLUP models. However, the Bayesian family of approaches in multi-trait setting, while statistically sound, can quickly become computationally expensive because of the Markov Chain Monte Carlo (MCMC) steps required to achieve convergence during parameter estimation.

In addition, multi-trait GS, like the classical approach, does not provide features selection capability and forgoes statistical testing of the effects of the SNPs to improve predictability for the studied traits. Therefore, multi-trait GS approaches have not been exploited to simultaneously provide sparse estimates and determine master regulators, i.e. markers which can simultaneously explain a large proportion in the majority of traits. Insights in master regulators may help narrow down the search for key genes underlying multiple traits, and will thus leverage the pleiotropy in the analyzed traits. This is particularly relevant when studying gene regulation and metabolism, for which the transcriptomic and metabolic phenotype arise due to the interconnection of thousands of genes and metabolites, shown to be jointly predictive of agronomically relevant traits (e.g. biomass) (Westhues *et al.*, 2017). In this sense, the multi-trait Bayesian approaches often do not result in sparse estimates for the model parameters, rendering it difficult to specify such master regulators.

Another means to develop multi-trait GS, with the aim of identifying master regulators, is to cast it in the framework of multi-output or multi-response regression that accounts for sparsity. For instance, a classical approach in this area is the Curds & Whey (Breiman and Friedman, 1997) that is only suitable for low dimension settings. Another approach is given by the simultaneous variable selection (Turlach *et al.*, 2005), an extension of the LASSO where the L_∞ norm penalty is imposed on the regression coefficient matrix. Although this norm results in sparsity of the selected predictors, it can lead to bias in model estimation. Finally, one can also jointly estimate the regression coefficient and the precision matrices. For instance, the multiple-output regression (Cai *et al.*, 2014; He *et al.*, 2016) incorporates both the covariances between traits (i.e. responses) and between errors in the model to improve the regression coefficient estimate, while the multivariate penalized likelihood (Lee and Liu, 2012; Rothman *et al.*, 2010) utilizes the covariance between the responses or the errors. However, these approaches are computationally challenging, since in the setting where the number of markers (i.e. SNPs) used as predictors is larger than the number of genotypes (i.e. observations) their maximum likelihood estimate of the precision matrix usually do not converge (Lee and Liu, 2012).

To improve selection of markers, while not compromising estimation and predictability, we assume that the responses are multivariate Gaussian and propose the $L_{2,1}$-joint, a novel multivariate method that models the response variables jointly in the penalize likelihood framework using the $L_{2,1}$-norm penalty. We propose a fast and efficient optimization algorithm that simultaneously constructs sparse estimates of the regression coefficients along with the precision matrix. Comparative analyses with simulated and real-world metabolomics data show that the proposed approach is a competitive candidate solution to the contenders.

2 Materials and methods

First, we introduce the matrix formulation of the multivariate linear regression model, and then briefly review the $L_{2,1}$-norm. Finally, we recall the statistical formulation of some regression models that will be used to compare our proposed method, i.e. the $L_{2,1}$-norm regression for GS ($L_{2,1}$-fs), the centered multiple output regression (cMOR), and RR. Throughout the rest of this paper, for a matrix $V = (v_{ij})$, we denote by $\boldsymbol{v^i}$ and $\boldsymbol{v_j}$ its ith row and jth column respectively. The symbols **tr** and **vec** stand for trace and vectorization operators respectively. V^{-1} is the inverse and V' the transpose of V. The n-dimensional identity matrix is denoted I_n and the L_p-norm of a vector $\boldsymbol{v} \in \mathbb{R}^n$ is defined as,

$$\|v\|_p = (\sum_{i=1}^{n} ||v_i||^p)^{\frac{1}{p}}, \tag{1}$$

where v_i represents the ith element of $\boldsymbol{v}$.

2.1 $L_{2,1}$-norm

First introduced in (Ding *et al.*, 2006), the $L_{2,1}$-norm of a matrix $V \in \mathbb{R}^{n\times m}$ is defined by,

$$||V||_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} v_{ij}^2} = \sum_{i=1}^{n} ||\boldsymbol{v^i}||_2. \tag{2}$$

It has been shown that the $L_{2,1}$-norm is rotation-invariant with respect to the rows, i.e. for any rotational matrix R of conformable size, the equality in Eq. (3), below, holds:

$$||VR||_{2,1} = ||V||_{2,1}. \tag{3}$$

An important notion that is used to solve the optimization problem in Eq. (15) is the partial derivative of $||V||_{2,1}$, defined as $\frac{\partial}{\partial V}||V||_{2,1} = QV$, with $Q \in \mathbb{R}^{n\times n}$ the diagonal matrix with entries $q_{ii} = \frac{1}{2||\boldsymbol{v^i}||_2}$.

2.2 Multivariate linear regression and the maximum likelihood estimate (MLE)

Let $Y = [\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y_s}] \in \mathbb{R}^{n\times s}$, $X = [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_p}] \in \mathbb{R}^{n\times p}$, $B = [\boldsymbol{b_1}, \boldsymbol{b_2}, \ldots, \boldsymbol{b_s}] \in \mathbb{R}^{p\times s}$ and $E = [\boldsymbol{\varepsilon_1}, \boldsymbol{\varepsilon_2}, \ldots, \boldsymbol{\varepsilon_s}] \in \mathbb{R}^{n\times s}$ represent matrices of observed responses, predictors, unknown regression coefficients and errors respectively. Statistical analysis using a multivariate linear regression model models the relationship between s response variables $\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y_s}$ and p predictor variables $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_p}$, so that, if the ith observation of the response, the i^{th} value of the predictor variables and the ith unobserved random vector are respectively defined by $\boldsymbol{y_i} = (y_{i1}, y_{i2}, \ldots, y_{is})'$, $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ and $\boldsymbol{\varepsilon_i} = (x_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{is})'$, then the linear regression model takes the following matrix representation:

$$Y = XB + E. \tag{4}$$

We also assume that $\boldsymbol{\varepsilon_i}$ are independent and have identical multivariate normal distribution with mean vector **0** and covariance matrix Σ. This model aims to predict multiple responses with a single set of predictors. For simplicity and without loss of generality, columns of X and Y are assumed centered so that the intercept term can be omitted. Then, up to a constant not dependent on the regression coefficient matrix B and the precision matrix $\Omega = \Sigma^{-1}$, the negative log-likelihood is

$$J(B, \Omega) = \mathbf{tr}\left[\frac{1}{n}(Y - XB)\Omega(Y - XB)'\right] - \log|\Omega|, \tag{5}$$

with maximum likelihood estimate (MLE) for B that does not depends on Ω

$$\hat{B}_{\text{mle}} = (X'X)^{-1}X'Y. \tag{6}$$

$\hat{B}_{\text{mle}}$ is the same estimate obtained by regressing separately each response on the same set of predictors, which is exactly the ordinary least squares (OLS) estimate and does not take into account the possible shared information among the responses. Furthermore, in the context of high dimensional data and large-p with small-n regression where X is not full rank, deriving $\hat{B}_{\text{mle}}$ using directly Eq. (6) is not possible.

In the following, we assume that we have n genotypes across each of which we measured s traits and identified p SNPs, so that Y and X represent the traits (e.g. metabolite profiles) and the SNPs matrices respectively.

2.3 GS with $L_{2,1}$-norm variable selection model and contending approaches

GS based on the $L_{2,1}$-norm solves,

$$\underset{B}{\operatorname{argmin}} ||Y - XB||_{2,1} + \lambda||B||_{2,1}. \tag{7}$$

The solution of this optimization problem is given by:

$$W = D^{-1}A'(AD^{-1}A')^{-1}Y, \tag{8}$$

where D is the diagonal matrix with the ith entry $d_{ii} = \frac{1}{2||w^i||_2}$,

$$A = [X, \lambda I_n] \in \mathbb{R}^{n\times m}, \quad W = \begin{bmatrix} B \\ P \end{bmatrix} \in \mathbb{R}^{m\times s}, \quad \text{and} \quad m = p + n.$$

Detailed explanations regarding the computation steps can be found in (Nie *et al.*, 2010).

From Eq. (2), it becomes apparent that the L_2-norm of each row in the $L_{2,1}$-norm penalty plays a specific role. As explained in (Sun *et al.*, 2009), the $L_{2,1}$-norm quantifies the effect of the ith predictor with the L_2-norm, while performing summation over all data points with the L_1-norm. This gives the $L_{2,1}$-norm the ability to induce row sparsity in the regression coefficient matrix B. Among several potential penalties, we opted for the $L_{2,1}$-norm since it also penalizes all the entries in the coefficient matrix and addresses one of our aims to identify master regulators.

To make it precise, we say that a column of the predictor matrix $X \in \mathbb{R}^{n\times p}$ is an $\alpha-$ master regulator ($\mathbf{MR}_\alpha$) if the corresponding row in the estimated sparse regression coefficient matrix $\hat{B} \in \mathbb{R}^{p\times s}$ is $\alpha-$ dense, i.e. at least an $\alpha-$ fraction, $0.5 \leq \alpha \leq 1$, of the entries in the corresponding row are non-zero. Moreover, for the purpose of this study, we only consider the case $\alpha = 1$ and use $\mathbf{mr}_1$ to define the proportion of rows that are $\mathbf{MR}_1$.

For completeness, we recall the RR optimization problem, given in Eq. (9)

$$\hat{B}(\lambda) = \underset{B}{\operatorname{argmin}} ||Y - XB||_2 + \lambda||B||_2, \tag{9}$$

with solution:

$$\hat{B}(\lambda) = (X'X + \lambda I_p)^{-1}X'Y. \tag{10}$$

In contrast, LASSO solves

$$\hat{B}(\lambda) = \underset{B}{\operatorname{argmin}} ||Y - XB||_2 + \lambda||B||_1. \tag{11}$$

The kernel LASSO that aims to account for possible non-linear dependence between the response and predictor, extends LASSO by using some suitable basis functions (kernel) as predictor and solves the optimization problem described in Eq. (12)

$$\hat{B}(\lambda) = \underset{B}{\operatorname{argmin}} ||Y - \Phi(X)B||_2 + \sqrt{\lambda}||B||_1, \tag{12}$$

where Φ is the kernel function. Finally, the multiple output regression solves:

$$\begin{aligned}(\hat{B}, \hat{\Omega}, \hat{\Sigma}) = \underset{(B,\Omega^{-1},\Sigma^{-1})}{\operatorname{argmin}} \; & \mathbf{tr}[(Y - XB)\Omega^{-1}(Y - XB)'] \\ & - n\log|\Omega^{-1}| + \lambda_1\mathbf{tr}(BB') + \lambda_2\mathbf{tr}(B\Sigma^{-1}B') - p\log|\Sigma^{-1}| \\ & + \lambda_3\mathbf{tr}(\Omega^{-1}) + \lambda_4\mathbf{tr}(\Sigma^{-1}).\end{aligned} \tag{13}$$

Ω^{-1} and Σ^{-1} represent the inverse covariances for the error and response respectively. We note that the optimization problem in Eq. (13) is not convex when all variables are considered jointly, and is convex for each individual variable when all others are kept constant. An iterative algorithm is then used to solve the convex problem (He *et al.*, 2016).

2.4 $L_{2,1}$-norm regularized multivariate regression and covariance estimation

Here, our aim is to design a multivariate regression model for GS that exploits the correlation between genotypes to obtain marker effects estimates along with variable selection. Applying the transpose operator on Eq. (4) yields the following negative log-likelihood function:

$$K(B, \Omega) = \mathbf{tr}\left[\frac{1}{s}(Y' - B'X')\Omega(Y' - B'X')'\right] - \log|\Omega|. \tag{14}$$

The L_1 penalty is then applied on the precision matrix Ω to reduce the number of parameters to be estimated when the number of responses variables (i.e. traits) is large (Rothman *et al.*, 2008) and to ensure the existence of an optimal solution with finite value of the objective function, in the situation where one has more responses than samples (Rothman *et al.*, 2010). In addition, the $L_{2,1}$ penalty is imposed on the regression coefficient matrix B to provide sparse $\hat{B}$ which, in turn, can aid the interpretation of the fitted model. Our model then provides the estimates $\hat{B}$ and $\hat{\Omega}$ by solving the following optimization problem:

$$f(B, \Omega) = \underset{B,\Omega}{\operatorname{argmin}}\{K(B, \Omega) + \lambda_1\|\Omega\|_1 + \lambda_2\|B\|_{2,1}\}, \tag{15}$$

with tuning parameters $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ to be determined from the data.

However, solving Eq. (15) is challenging since the optimization problem is not convex and the $L_{2,1}$-norm is not smooth. We overcome the challenge by iteratively solving for one parameter while keeping the other one constant. In doing so, we transform Eq. (15) into a convex optimization problem and ensure that the problem has a global optimum. Solving Eq. (15) for B with constant Ω at a chosen point Ω_0 is equivalent to optimizing

$$\hat{B}(\Omega_0) = \underset{B}{\operatorname{argmin}}\left\{a\mathbf{tr}\left[\frac{1}{s}(Y' - B'X')'(Y' - B'X')\Omega_0\right] - \log|\Omega_0| + \lambda_2||B||_{2,1}\right\}. \tag{16}$$

Taking the partial derivative with respect to B and equating to zero yields

$$\hat{B} = \left[X'\Omega_0X + \frac{s\lambda_2}{2}C\right]^{-1}X'\Omega_0Y. \tag{17}$$

Using the Woodbury matrix identity (Riedel, 1992) in the case where $\lambda_2 \neq 0$, we obtain the formulation in Eq. (18) that is the core of our algorithm. More specifically, the inversion of the $p \times p$ matrix is avoided and we, instead, invert an $n \times n$ matrix in the following:

$$\hat{B} = \frac{2}{s\lambda_2}C^{-1}X'\Omega_0\left[Y - \frac{2}{s\lambda_2}\left(I_n + \frac{2}{s\lambda_2}XC^{-1}X'\Omega_0\right)^{-1} \times XC^{-1}X'\Omega_0Y\right], \tag{18}$$

where C is the diagonal matrix with ith entry $c_{ii} = \frac{1}{2||b^i||_2}$. A close look at Eq. (17) reveals the generality of our estimate: When $\lambda_2 = 0$, and $\Omega_0 = I_n$, we obtain the OLS estimate. When $\Omega_0 = I_n$ and $C = I_p$ we have the RR estimate, and, finally, when $C = I_p$ we have the $L_{2,1}$-norm based variable selection.

Solving Eq. (15) for Ω with fixed B at a chosen point B_0 corresponds to the L_1-penalized covariance estimation problem (Yuan and Lin, 2006) and the well-known efficient solution given by the graphical lasso (GLASSO) of (Friedman *et al.*, 2008). We make use of GLASSO to estimate Ω in the model given in Eq. (19), below:

$$\hat{\Omega}(B_0) = \operatorname{argmin}_\Omega\left\{a\mathbf{tr}\left[\frac{1}{s}(Y' - B_0'X')'(Y' - B_0'X')\Omega\right] - \log|\Omega| + \lambda_1\|\Omega\|_1\right\}. \tag{19}$$

Algorithm 1: $L_{2,1}$**-joint**

Input: $\lambda_1, \lambda_2, \mathbf{X} \in \mathbb{R}^{p\times s}, \mathbf{Y} \in \mathbb{R}^{n\times s}$,
Output: $\boldsymbol{\Omega} \in \mathbb{R}^{n\times n}, \mathbf{B} \in \mathbb{R}^{p\times s}$
1: Set $t = 0$ and initialize
- The diagonal matrix $\mathbf{C}_t \in \mathbb{R}^{p\times p}$ as identity,
- $\widehat{\boldsymbol{\Omega}}_t$ as the inverse of the ridge covariance matrix
- $\widehat{\mathbf{B}}_t$ as $\widehat{\mathbf{B}}_t(\widehat{\boldsymbol{\Omega}}_t)$ solving Eq. (18)
- $\widehat{\boldsymbol{\Omega}}_t$ as $\widehat{\boldsymbol{\Omega}}_t(\widehat{\mathbf{B}}_t)$ by solving Eq. (19) using GLASSO
2: **repeat**
- Update $\mathbf{C}$ by computing $\mathbf{C}_{t+1}$ with the j^{th} diagonal entry $c_{jj} = \frac{1}{2\|\mathbf{b}_t^j\|_2}$
- Update $\widehat{\boldsymbol{\Omega}}$ by computing $\widehat{\boldsymbol{\Omega}}_{t+1} = \widehat{\boldsymbol{\Omega}}(\widehat{\mathbf{B}}_{t+1})$ by solving Eq. (19) using GLASSO
- Update $\widehat{\mathbf{B}}$ by computing $\widehat{\mathbf{B}}_{t+1} = \widehat{\mathbf{B}}(\widehat{\boldsymbol{\Omega}}_{t+1})$ using Eq. (18)
$t = t + 1$
until *Convergence*;

The following Algorithm (1), referred to as $L_{2,1}$-joint, summarizes the computational steps for optimizing our model in Eq. (15).

2.5 Convergence criteria

Because the RR estimate is well-defined, including the case when the predictors are collinear, we use its L_1-norm to scale the convergence criterion for our regression coefficient matrix $\hat{\mathbf{B}}$. In addition, we use the sample covariance matrix of the RR residual to scale the convergence of the precision matrix (Chen *et al.*, 2014). This implies that the convergence criteria for $\hat{\mathbf{B}}$ and $\hat{\boldsymbol{\Omega}}$ are met when $\sum_{i,j}|\hat{b}_{ij}^{(t+1)} - \hat{b}_{ij}^{(t)}| < \varepsilon_1 \sum_{ij}|\hat{b}_{ij}^{\text{ridge}}|$ and $\sum_{i,j}|\hat{\omega}_{ij}^{(t+1)} - \hat{\omega}_{ij}^{(t)}| < \varepsilon_2 \sum_{ij}|\hat{\omega}_{ij}^{\text{ridge}}|$, respectively. Here, ε_1 and ε_2 are the tolerance parameters that we set to 10^{-5}. Moreover, because our objective function is convex in $\mathbf{B}$ when the other parameter is fixed and monotonically decreasing in each iteration, another convergence criteria one can use is given by $||\hat{\mathbf{B}}^{(t+1)}||_{2,1} \geq ||\hat{\mathbf{B}}^{(t)}||_{2,1}$ or when an a priori set maximum number of iterations is reached.

2.6 Model evaluation and hyper-parameters

To evaluate the predictability we use the RV coefficient (Escoufier, 1973) that measures the relationship between two sets of variables (measured and predicted) and the multi-output extension of the mean squared error (MSE), respectively defined by Eqs. (20) and (21) below:

$$\text{RV}(\mathrm{Y}, \tilde{\mathrm{Y}}) = \frac{\text{tr}(\mathrm{Y}\mathrm{Y}'\tilde{\mathrm{Y}}\tilde{\mathrm{Y}}')}{\sqrt{\text{tr}((\mathrm{Y}\mathrm{Y}')^2)\text{tr}((\tilde{\mathrm{Y}}\tilde{\mathrm{Y}}')^2)}}, \tag{20}$$

and

$$\text{MSE}(\mathrm{Y}, \hat{\mathrm{Y}}) = \frac{1}{n}\frac{1}{s}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{21}$$

with y_i and $\hat{y}_i$ denoting the observed and predicted (or estimated) s output of $\mathrm{Y} \in \mathbb{R}^{n\times s}$, respectively. The degree of penalization that can be imposed on the model is not fixed; thus, for each penalty level a different solution path is founds. It is therefore of great importance to select the best estimator based on the optimal penalty level, which we determine by cross-validation (CV). To this end, we split the entire dataset into K non-overlapping subsets of nearly equal size. Using K-fold cross-validation, we select the optimal $\hat{\lambda}_1$ and $\hat{\lambda}_2$ by solving

$$(\hat{\lambda}_1, \hat{\lambda}_2) = argmin_{\lambda_1\lambda_2}\sum_{k=1}^{\mathrm{K}}||\mathrm{Y}^k - \mathrm{X}^k\mathrm{B}_{\lambda_1\lambda_2}^{(-k)}||_2. \tag{22}$$

Here, Y^k and X^k are respectively the k^{th}-fold response and predictor matrices, while. $\mathrm{B}_{\lambda_1\lambda_2}^{(-k)}$ is the regression coefficient matrix estimated out of the k^{th}-fold for λ_1 and λ_2. In addition, seq(3, 12, 1) and $2^{\text{seq}(-5,-2,1)}$ are used as search grids to obtain the optimal λ_1 and λ_2 respectively.

We also use the true positive rate (TPR) and the true negative rate (TNR) to quantify the degree of sparsity recognition by the estimate of the regression coefficient matrix $\hat{\mathrm{B}}$. These are given respectively by the proportion of non-zero entries in the true coefficient B identified correctly by the estimate $\hat{\mathrm{B}}$ and the proportion of zero entries in B that $\hat{\mathrm{B}}$ matched correctly. Since from the simulation design we know exactly what the master regulators are, we also evaluate the ability of all models to correctly identify the true MR_1 by computing mr_1, the proportion of rows with non-zero entries in B correctly identify by $\hat{\mathrm{B}}$. Therefore, mr_1 corresponds to the proportion of master regulators.

3 Results and discussion

3.1 Comparative analysis with synthetic data

To quantify the performance of the proposed method, we devise a series of two synthetic datasets. (1) By modifying a previously studied simulation design (Yuan *et al.*, 2007). We set $(\Sigma_{\mathrm{X}})_{ij} = .7^{|i-j|}$, so that rows of the design matrix $\mathrm{X} \in \mathbb{R}^{n\times p}$, are independently generated from the multivariate normal distribution $N_p(0, \Sigma_{\mathrm{X}})$. For the genomic prediction application, different coding for the genotypes (predictors) matrix $\mathbf{X}$ can be obtained. For instance, all absolute values in the intervals $[0, .5]$, $]0.5, 1]$ and $]1, \infty[$ can respectively be coded as 0, 1 and 2, which is an alternative to randomly sample the genotype matrix $\mathbf{X}$ from $\{0, 1, 2\}$. For the error matrix $\mathbf{E} \in \mathbb{R}^{n\times s}$, an autoregressive covariance structure of order 1, AR(1), is considered, implying that rows of E are independently drawn from the multivariate normal distribution $N_s(0, \Sigma)$, with $(\Sigma)_{ij} = \rho^{|i-j|}$ and ρ taking values (.1,.5,.9). Using the matrix element wise product $\mathbf{B} = \mathbf{W} * \mathbf{Q} + \mathbf{K} * \mathbf{W}$, a sparse regression coefficient matrix is obtained. With the modification, we further obtain some rows in $\mathbf{B}$ that are non-zero so that the proportion of correctly identified master regulators can be computed. In this setting, each entry of $\mathbf{W}$ is an independent draw from $N(0, 1)$, the entries of K are independent realization from a Bernoulli distribution with s_1 probability of success. Each row of Q is either a vector of ones or zeros, the rows of all one are determined based on p independent Bernoulli draws with s_2 probability of success. Following Eq.(4), 30 traits were simulated and their heritabilities are provided in Supplementary Table S1. For each data generation process, 20 replicates are drawn and we consider a test dataset of sample size 20 to assess the predictability. (2) To further asses the predictability of the proposed model, in the second synthetic dataset, a pleiotropy architecture under low (.1 and.2), mild (.4 and.5) and high (.7 and.8) heritability scenario is considered. The R package simplePHENOTYPES (Fernandes and Lipka, 2020) and the included genotypic data composed of 282 inbred maize association panel using the 55 K SNP array (Cook *et al.*, 2012) are used to simulate 12 highly correlated traits controlled by 80 MR_1. Note that, for the purpose of this study, we only used 2000 SNPs and 80 lines were always keep for testing during the CV.

In what follows, the performance of our proposed $L_{2,1}$-norm regularized multivariate regression and covariance estimation is assessed and compared on the synthetic dataset with eight contenders: (1) the efficient and robust feature selection via joint $L_{2,1}$-norms minimization ($L_{2,1}$-fs) (Nie *et al.*, 2010), (2) the recent centered multiple output regression (cMOR) (He *et al.*, 2016) which showed that centering of the predictor matrix improves prediction performance, (3) GBLUP, (4) the Elastic-Net (Zou and Hastie, 2005), (5) the Regularized multivariate regression for identifying master predictors (remMAP) (Peng *et al.*, 2010), (6) the multiple-trait Bayesian regression (MBayesB) (Cheng *et al.*, 2018) implemented with BGLR package in R (Pérez and de Los Campos, 2014) with the proportion of influential SNPs estimated rather than chosen, (7) the LASSO, and

Table 1. Comparison of model performance on synthetic data

(A) Predictability

n, p, s	RR	cMOR	$L_{2,1}$-fs	$L_{2,1}$-joint	mLASSO	MBayesB	remMAP	Elastic-Net
50,100,30	0.84 (0.07)	0.85 (0.07)	0.69 (0.01)	0.85 (0.06)	0.79 (0.08)	0.78 (0.08)	0.85 (.06)	0.81 (0.07)
50,300,30	0.71 (0.07)	0.72 (0.07)	0.47 (.01)	0.72 (0.07)	0.69 (0.07)	0.65 (.07)	0.70 (0.07)	0.70 (0.06)
50,800,30	0.64 (0.09)	0.64 (0.09)	0.37 (0.03)	0.62 (0.08)	0.58 (0.02)	0.58 (0.01)	0.63 (0.09)	0.62 (0.08)

(B) Recovery rate of MR_1

n, p, s	ρ	RR	cMOR	$L_{2,1}$-fs	$L_{2,1}$-joint	mLASSO	MBayesB	remMAP	Elastic-Net
50,800,30	0.1	–	–	38.8	80.3	0	–	0	0
	0.5	–	–	38.9	80.4	0	–	0	0
	0.9	–	–	38.8	80.02	0	–	0	0

(C) Sparsity recovery TPR/TNR

n, p, s	ρ	RR	cMOR	$L_{2,1}$-fs	$L_{2,1}$-joint	mLASSO	MBayesB	remMAP	Elastic-Net
50,800,30	0.1	–/0	–/0	51.8/52.9	89.01/13.26	1.05/99.1	–/0	5.2/95.6	2.7/97.7
	0.5	–/0	–/0	51.7/52.9	89.02/13.2	1/99.2	–/0	8.7/92.6	2.7/97.7
	0.9	–/0	–/0	51.7/53.02	89.04/13.25	1.15/99.07	–/0	5.31/95.5	2.58/97.8

Note: The dataset consists of $s = 30$ simulated phenotypes, $n = 50$ observations and varying number of predictors $p \in \{100, 300, 800\}$ to see their impact on predictability, and fixed $p = 800$ for sparsity and MR_1 analysis. (A) The predictability assessed as the RV coefficient between the true and predicted responses in the unseen data with standard errors in parentheses. (B) The true positive rate (in %) for master regulator recovery, which determines the ability of each model to correctly identify the known MR_1 (the non-zero rows in the true regression coefficient matrix). (C) The sparsity recovery quantified by the true positive rate/true negative rate (in %) for the regression coefficient matrix estimate $\hat{\mathbf{B}}$, specifying the potential of each model to correctly identify non-zero entries in the true coefficient matrix. All metrics are averaged over 20 replicates with AR(1) parameter ρ and for all models the tuning parameters were selected using 5-fold CV. The symbol '–' denotes the fact that all entries in the estimated regression coefficients were non-zero and hence could not be used to quantify the parameter of interest.

Table 2. Comparison of model performance on simulated phenotypes at different levels of heritability based on SNP data from maize

(A) Predictability

Traits	H^2	$L_{2,1}$-fs	$L_{2,1}$-joint	RR	mLASSO	Elastic-Net	cMOR	remMAP	MBayesB	GBLUP
1	0.1	0.08 (0.05)	0.21 (0.05)	0.01 (0.04)	0.03 (0.06)	0.02 (0.06)	0.08 (0.03)	**0.24 (0.05)**	0.22 (0.06)	0.03 (0.04)
2	0.4	0.48 (0.06)	**0.64 (0.06)**	0.27 (0.05)	0.63 (0.06)	0.63 (0.07)	0.23 (0.04)	0.58 (0.05)	0.59 (0.06)	0.25 (0.05)
3	0.7	0.69 (0.06)	**0.84 (0.06)**	0.38 (0.05)	0.81 (0.07)	0.81 (0.05)	0.31 (0.03)	0.81 (0.04)	0.82 (0.08)	0.37 (0.05)
4	0.2	0.20 (0.06)	0.36 (0.05)	0.11 (0.04)	0.37 (0.06)	0.37 (0.04)	0.10 (0.03)	**0.40 (0.04)**	0.29 (0.07)	0.19 (0.04)
5	0.5	0.57 (0.08)	**0.73 (0.07)**	0.37 (0.06)	0.73 (0.08)	0.73 (0.05)	0.32 (0.05)	0.67 (0.06)	0.71 (0.07)	0.37 (0.06)
6	0.8	0.76 (0.05)	**0.87 (0.05)**	0.42 (0.03)	0.83 (0.05)	0.83 (0.05)	0.39 (0.03)	0.83 (0.05)	0.86 (0.05)	0.43 (0.03)
7	0.1	0.12 (0.05)	0.22 (0.06)	0.01 (0.04)	0.19 (0.06)	0.18 (0.04)	0.04 (0.02)	**0.27 (0.05)**	0.01 (0.06)	0.01 (0.04)
8	0.4	0.49 (0.06)	**0.65 (0.05)**	0.32 (0.03)	0.64 (0.05)	0.64 (0.04)	0.24 (0.03)	0.58 (0.05)	0.61 (0.07)	0.30 (0.03)
9	0.7	0.68 (0.06)	**0.83 (0.06)**	0.33 (0.04)	0.81 (0.06)	0.81 (0.04)	0.31 (0.03)	0.81 (0.06)	0.81 (0.06)	0.33 (0.04)
10	0.2	0.24 (0.06)	0.44 (0.05)	0.06 (0.03)	0.44 (0.06)	0.44 (0.06)	0.07 (0.03)	**0.46 (0.05)**	0.38 (0.07)	0.06 (0.04)
11	0.5	0.54 (0.07)	**0.67 (0.06)**	0.38 (0.05)	0.68 (0.07)	0.69 (0.07)	0.39 (0.04)	0.62 (0.05)	0.66 (0.06)	0.38 (0.06)
12	0.8	0.81 (0.05)	**0.91 (0.06)**	0.42 (0.03)	0.86 (0.06)	0.86 (0.06)	0.37 (0.02)	0.86 (0.05)	0.89 (0.06)	0.42 (0.03)

(B) Recovery rate of MR_1

$L_{2,1}$-fs	$L_{2,1}$-joint	RR	mLASSO	Elastic-Net	cMOR	remMAP	MBayesB	GBLUP
51.2	40.3	–	0	0	–	0	–	–

Note: (A) The predictability of each trait assessed by the correlation coefficient between the true and predicted trait in the unseen data with standard errors in parentheses. (B) The true positive rate (in %) for master regulator recovery, which determines the ability of each model to correctly identify the 80 markers set as MR_1. The metrics are averaged over 20 replicates and the tuning parameters were selected using 3-fold CV. The symbol '–' denotes the fact that all entries in the estimated regression coefficients were non-zero and hence could not be used to quantify the parameter of interest.

(8) the RR estimate that is included due to its quality in term of predictability.

Let us consider the sparsity parameter $s_1 = .5$ and three choices of the AR(1) parameter $\rho = \{0.1, 0.5, 0.9\}$. In terms of predictability depicted in Table 1A, the proposed $L_{2,1}$-joint, cMOR, remMAP and the RR achieve equal performance as quantified by the average RV coefficient between the validation sample and the corresponding predicted values. However, from Table 1B, the $L_{2,1}$-joint correctly identified on average 80% of MR_1, while remMAP failed to identify any, cMOR and RR do not achieve variable selection. Moreover, among the methods with variable selection capability, $L_{2,1}$-joint outperforms the $L_{2,1}$-fs, multivariate LASSO (mLASSO) and Elastic-Net and achieve equal performance with remMAP, for low and high correlation. In contrast to $L_{2,1}$-joint, the classical $L_{2,1}$-fs shows a lower recovery rate of MR_1 by identifying correctly on average 39% of MR_1 over all replicates. In addition, for methods with ability to reveal loci that can regulate more than one trait, the MR_1 recovery rate computed in Table 1B shows the superiority of $L_{2,1}$-joint with

Table 3. Comparison of model performance on *Brassica napus* data

Model	RV-Coef	$\mathbf{mr}_1$	MSE		
			flower 0	flower 4	flower 8
RR	0.46	–	2.40 (0.69)	1.98 (0.57)	1.81 (0.52)
cMOR	0.46	–	2.35 (0.67)	2.00 (0.57)	1.84 (0.53)
$L_{2,1}$-fs	0.43	33%	2.51 (0.72)	1.95 (0.56)	1.75 (0.50)
$L_{2,1}$-joint	0.46	23%	2.48 (0.71)	1.95 (0.55)	1.73 (0.49)
kmLASSO	0.44	–	4.54 (1.31)	3.28 (0.94)	1.85 (0.53)
mLASSO	0.30	–	2.51 (0.73)	1.95 (0.57)	1.72 (0.50)
MRCE	0.10	97%	**2.22** (0.65)	**1.76** (0.51)	**1.62** (0.45)
GBLUP	0.49	–	2.46 (0.73)	1.92 (0.54)	1.74 (0.53)
Elastic-Net	0.36	–	2.46 (0.72)	1.91 (0.62)	1.73 (0.49)
MBayesB	**0.57**	–	2.50 (0.74)	1.95 (0.56)	1.74 (0.57)
remMAP	0.28	4%	2.51 (0.71)	1.94 (0.53)	1.73 (0.50)

Note: Predictability measured by the RV coefficient between the observed and predicted values for all three traits and the proportion of SNPs found to be master regulators by a specific GS model for the *Brassica napus* dataset. The estimated prediction error for all traits along with the minimum mean squared error (MSE) value to each trait highlighted in bold for specific GS model, and standard errors in parentheses. The symbol '–' denotes the fact that all entries in the estimated regression coefficients were non-zero and hence could not be used to quantify the parameter of interest.

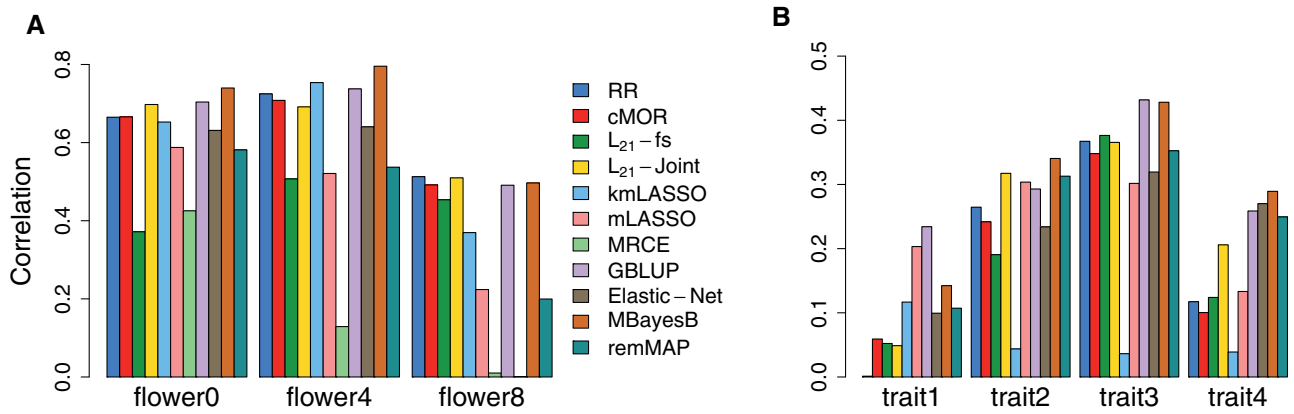

Fig. 1. Predictability for (**A**) *Brassica napus* and (**B**) wheat traits, computed as the correlation coefficient between the observed phenotypes and predicted breeding values for individual traits in the validation set

respect to $L_{2,1}$-fs, remMAP and MBayesB. One may argue that $L_{2,1}$-fs performs better than $L_{2,1}$-joint simply because it includes fewer MR_1. However, as shown in Table 1B, summarizing the ability of each model to correctly identify the true MR_1, we find that the $L_{2,1}$-fs fails on average 61% of time to correctly identify non-zero rows in the true regression coefficient matrix compared to the $L_{2,1}$-joint with an average failure rate of 20%. Since it has been shown that RR provides accurate prediction in the context $p \gg n$, we conclude that the proposed model maintain the desired properties of RR while achieving variable selection that helps for model interpretability. For the sparsity recognition measured by the TPR and TNR, Table 1C shows that for all correlation patterns, the $L_{2,1}$-joint and $L_{2,1}$-fs dominate the RR and cMOR and exhibit comparable performance in correctly identifying the non-zero entries in the true regression coefficient matrix. The $L_{2,1}$-joint is, however, superior when it comes to identifying the true zero entries in B. Predictability presented in Table 2A, for highly correlated traits simulated using maize maker data, reveals that the proposed $L_{2,1}$-joint outperforms the contenders for 8 traits out of 12 for mild and high heritability and is the second best after remMAP for low heritability. Overall, models capable to reveal master regulators are more accurate when predicting traits with low heritability. These findings are in line with the theoretical considerations, in the sense that traits with high heritability are highly predictable andpredictability in genomic prediction can increase when simultaneously considering correlated traits of lower heritability. For the MR_1 recovery rate shown in Table 2B, we observe that the proposed $L_{2,1}$-joint is second best performing after the $L_{2,1}$-fs with respectively 40.3% and 51.2% out of 80 master regulators correctly identified.

3.2 Comparative analysis with *Brassica napus* data

Some of the multi-trait genomic selection (MTGS) models are only tractable for small or moderate number of markers (p), such as: (1) The sparse multivariate regression with covariance estimation (Rothman *et al.*, 2010) (MRCE), (2) the multivariate LASSO (mLASSO) and (3) the kernelized multivariate LASSO (Xu and Yin, 2013) (kmLASSO), implemented in the MTGS package in R (Budhlakoti *et al.*, 2019). Although not a multiple output regression, GBLUP method is also included because of its reputation in GS.

In our comparative analysis, we used a dataset from *Brassica napus* (rapeseed) (Kole *et al.*, 2002), provided as a part of MTGS package. The data consists of 3 highly correlated (correlation >.78) traits, associated to days of flowering at different weeks (flower 0, flower 4, flower 8) and 50 lines obtained from two cultivars (Stellar and Major) and genotyped for 100 markers. The first 40 lines are used in 5-fold CV to build the training and validation samples and the remaining (testing set) put aside for prediction assessment. In this setting, comparison of all selected multi-trait approaches can be carried out, due to the limited number of modeled traits.

Our findings show that MBayesB, GBLUP, RR, cMOR and the $L_{2,1}$-joint capture the largest part of the linear relationship between the responses and predictors, as assessed by the RV coefficient in Table 3. Focusing on individual traits Figure 1A, we see that $L_{2,1}$-joint is the third best performing, after MBayesB and GBLUP, for

Table 4. Comparison of model performance on wheat data

Model	RV-Coef	mr_1	MSE			
			trait 1	trait 2	trait 3	trait 4
RR	0.10	–	3.1 (0.7)	1.4 (0.3)	1.9 (0.4)	2.8(0.6)
cMOR	0.09	–	3.7 (0.8)	1.9 (0.4)	2.4 (0.5)	3.2 (0.7)
$L_{2,1}$-fs	0.10	70%	2.6 (0.5)	1.1 (0.2)	1.6 (0.3)	2.4 (0.5)
$L_{2,1}$-joint	0.13	21%	1.9 (0.4)	0.6 (0.1)	**1** (0.2)	1.7 (0.4)
kmLASSO	0.008	–	2 (0.4)	0.6 (0.1)	1.29 (0.2)	2 (0.4)
mLASSO	0.08	–	**1.7** (0.3)	0.9 (0.2)	1.1 (0.2)	2.1 (0.4)
GBLUP	0.14	–	1.8 (0.4)	**0.5** (0.1)	1.1 (0.2)	1.8 (0.4)
Elastic-Net	0.11	1%	1.9 (0.4)	0.6 (0.1)	1.1 (0.2)	1.7 (0.3)
MBayesB	0.15	55%	1.9 (0.4)	0.6 (0.1)	**1**(0.2)	**1.6** (0.3)
remMAP	0.13	–	1.8 (0.4)	0.7 (0.1)	1.2 (0.2)	1.9 (0.4)

Note: Predictability quantified by the RV coefficient between the observed and the predicted values for all four traits in the wheat dataset. Also shown is the proportion of SNPs identified as master regulators and the estimated prediction error for all traits, and standard errors in parentheses. The minimum mean squared error (MSE) value corresponding to each trait is highlighted in bold for specific GS model. The symbol '–' denotes the fact that all entries in the estimated regression coefficients were non-zero and hence could not be used to quantify the parameter of interest.

flower 0, and second best performing, after RR, for flower 8. Withing models allowing the identification of master regulators (remMAP and MBayesB), we see that $L_{2,1}$-joint outperforms the contender for flower 8 and is the second best performing after MBayesB for the other two traits. However, in the case of $L_{2,1}$-joint we also identify that 23% of SNPs are found as MR_1, which provides additional information that could not be obtained by MBayesB based on the estimated π values. This is due to the strong relationship between regression coefficients estimated from MBayesB and the choice of π. Some π values may actually provide sparse estimates and facilitate master regulators identification.

3.3 Comparative analysis with wheat dataset

Here, we compare the performance of $L_{2,1}$-joint against other models for a moderate number of predictors. This is done using a collection of 599 historical wheat lines from the international maize and wheat improvement center (CIMMYT) global wheat breeding program. Part of the BGLR R package (Pérez and de Los Campos, 2014), the dataset comprises, 4 phenotypic traits representing the average grain yield of the 599 evaluated lines in four environments. Altogether, 1279 markers were retained for analysis, and we use the first 500 lines in 5-fold CV to build the training and validation samples and the remaining used as unseen data to evaluate the predictability.

Table 4 shows that the predicted values by MBayesB, $L_{2,1}$-joint, GBLUP and remMAP are the closest to the measured phenotypic values in the test sample, as assessed by the RV coefficient. At the individual trait level, we can see that $L_{2,1}$-joint, MBayesB, GBLUP and mLASSO achieve the smallest prediction error for one out of four traits. A further analysis of the correlation between measured and predicted individual traits as shown in Figure 1B, ranks $L_{2,1}$-joint as the best performing for trait2 and second best performing after MBayesB, for trait3 and trait4 among methods allowing master regulator identification. With its additional variable selection property evidenced here by the identification of 21% of SNPs as MR_1, we can say that, for moderate number of predictors, $L_{2,1}$-joint exhibits high performance when simultaneously considering predictability and variable selection with respect to the competitors.

3.4 Comparative analysis with *Arabidopsis thaliana* data

To further test our methodology on real-world datasets, we consider the gas chromatography mass spectrometry (GC-MS) log-transformed metabolomics profiles for 94 primary metabolites from leaves in a natural *Arabidopsis thaliana* population consisting of 312 accessions, used already in genome-wide association analyses of primary metabolites (Wu *et al.*, 2016). The correlation analysis on the metabolomic data reveals a maximum correlation of.78 and only few values above.5. In addition, we used 214 051 SNPs obtained using AffymetrixGeneChip Array 6.0 (Horton *et al.*, 2012). We removed all SNPs with less than 5% minimum allele frequency (MAF), leaving us with 200 180 to build the $L_{2,1}$-joint model. In such a setting, the usage of the Bayesian multi-trait approaches is prohibitive, due to the large number (94) of modeled traits. As a result, the comparison includes only four approaches, namely: RR, cMOR, $L_{2,1}$-joint and $L_{2,1}$-fs.

Table 5. Predictability on *A. thaliana* data

Model	RV-Coef	Selected variables	mr_1
RR	0.26	–	–
cMOR	0.26	–	–
$L_{2,1}$-fs	0.27	55960 (27.9%)	4249 (2.12%)
$L_{2,1}$-joint	0.26	135597 (67.73%)	30819 (15.39%)

Note: RV coefficient for predicting metabolites levels across the 40 testing lines, features selection and identification of MR_1 for $L_{2,1}$-joint, $L_{2,1}$-fs, cMOR and RR. Tuning parameters are selected by 5-fold CV. The symbol '–' denotes the fact that all entries in the estimated regression coefficients were non-zero and hence could not be used to quantify the parameter of interest.

In terms of the predictability for the full metabolomic profile determined by the RV coefficient between the measured metabolite levels and the predicted breeding values in the test sample (i.e. the last 40 lines, the unseen data), Table 5 shows that, all considered models achieve almost similar results. However, a look at the number of markers entering the models demonstrate that the $L_{2,1}$-joint and $L_{2,1}$-fs models are superior. Given the observation that our $L_{2,1}$-joint model outperforms $L_{2,1}$-fs with respect to identification of master regulators, it finally shows the suitability of the proposed solution in high-dimensional setting and when more than four traits are considered.

Further correlation analysis between measured and predicted individual traits for known metabolite classes in the selection candidates, (see Supplementary Figs S1–S3), shows that: (i) For the 26 organic acids metabolites, RR and $L_{2,1}$-joint achieve equal predictability as quantified by the number of time each method outperforms the contender, with $L_{2,1}$-joint achieving the maximum correlation of.48 on citric acid. (ii) Regarding the ability to predict the levels of the 26 amino acids, we find that both models are superior half of the time and achieve equal maximum correlation of .49 on isoleucine and serine for $L_{2,1}$-joint and RR, respectively. (iii) Concerning the predictability of the 17 sugars, we observe another split, as both models are superior on 8 counts, achieve equal performance on Glucose, and the maximum correlation of.5 on 1,6-

Anhydro-beta-D-glucose attained by RR. The desirable property of $L_{2,1}$-joint to perform variable selection suggest the proposed model as a better candidate.

We quantify the effect of a given SNP on metabolites by the sum of absolute values of the corresponding row in the estimated regression coefficient matrix. Using this approach, rows contribution of $\hat{B}$ were ranked and the most relevant SNPs identified. Even though the majority of high ranked SNPs were also master regulators, in the following we focus only on those which are master regulators. Since linkage disequilibrium (LD) decays on average within 10 kb in *Arabidopsis thaliana* (Kim *et al.*, 2007), we used a 10 kb window for genes search (i.e. 5 kb left and right for the considered SNP). Using this procedure, a subset of the 20 most prominent (in decreasing order) SNPs fulfilling the MR_1 conditions were singled out (see Supplementary Table S1). These include: (i) the lead SNP m22901 on chromosome 1, at locus AT4G36240, encoding GATA transcription factor 7, involved in cell differentiation, circadian rhythm and response to light stimulus (Manfield *et al.*, 2007), (ii) SNP m50264, on chromosome 2, implicating four loci, one of which, AT2G03500, Early flowering MYB protein, directly represses flowering locus T expression in the leaf vasculature (Yan *et al.*, 2014) and acts as transcriptional activators in abscisic acid signal transduction pathway (Abe *et al.*, 2003), (iii) SNP m105589, on chromosome 3, at locus AT3G47290, encoding ATPLC8, reported to be involved in seedling growth and endoplasmic reticulum stress responses (Kanehara *et al.*, 2015), (iv) SNP m120899 on chromosome 4, implicating three genes, of which AT4G04720, encoding CPK21, is involved in plant growth regulation and abiotic stress responses (Shi *et al.*, 2018).

A further exploration of marker effects shows that some of these MR_1 fall within a dense region (i.e. interval with five or more consecutive SNPs with high effect). For instance: (i) On chromosome 4, a 2.5 kb window (position 8298588–8296004), exhibits SNPs m132532, m132541 and m132544 as MR_1. This is a smaller interval of the region where AT4G14400, also known as accelerated cell death 6 (ACD6) gene, is located. (ii) On chromosome 5, a 1.5 kb interval (positions 13637852–13636269), SNPs m173856, m173857, m173858, m173861 and m173862 are also MR_1. In this interval, AT5G35410 is found, a regulatory component controlling plant potassium uptake and involved in the response to salt stress, protein phosphorylation and intracellular signal transduction (Wang *et al.*, 2018). On a more general note, we observe that the MR_1 form hot spots, in the sense that once an MR_1 is identified, it is more likely to find another one in its vicinity. This can be further visualized when looking along chromosomes for the dataset at hand. We observe that the average distance between two MR_1 is 874 bp compare to 134 bp the average distance between two SNPs. The presented models do not consider the effect of environment nor the interaction between genotypes and environments, although the latter are particularly relevant for selection of genotypes that are better performing in specific environments. Future efforts will be directed toward incorporation of environment in covariates and consideration of weighted variants of the used $L_{2,1}$-norm. Further, we note that the current formulation of the model assume same variance for all SNPs used. Therefore, future work will focus on using weighted variants of the $L_{2,1}$-norm to begin to investigate generalizations in which variance is not-equal across the SNPs. Moreover, it will be important to investigate the extent to which MR_1 obtained from our approach agree with results from classical approaches for genome-wide associations, which determine the effect of individual markers. Such efforts will highlight the usage of the developed prediction models for inference of underlying molecular mechanisms.

4 Conclusion

Despite the use of a rather strong assumption that a locus simultaneously affects all the traits or none of them, standard multi-trait GS methods greatly improved the accuracy of genomic prediction. In this work, we relaxed this assumption by putting no restriction on the fraction of traits on which a marker can be causal, thus opening the possibility to identify master regulators. Using simulated and real-world data, we demonstrated the effectiveness of the $L_{2,1}$-norm as a tool for variable selection and master regulators identification in a penalized multivariate regression when the number of SNPs, as predictors, is much larger than the number of genotypes.

Acknowledgements

A.J.M. and Z.N. thank Dr. Marcus McHale from Galway University, Ireland, for fruitful discussions during his research sojourn in Z.N.'s group at the Max Planck Institute of Molecular Plant Physiology.

Author Contributions

Z.N. conceived the project, A.J.M and Z.N. designed the model, H.T. prepared the *Arabidopsis thaliana* dataset, A.J.M., H.T., Z.N. analyzed the data, A.J.M. and Z.N. prepared the manuscript. All authors read and approved the final manuscript.

Funding

This project was funded by the European Union's Horizon 2020 research and innovation programme projects BREEDCAFS [GA No. 727934] and PlantaSYST [FPA No. 664620].

Data availability

The data underlying this article are publicly available and their corresponding references provided withing the article.

Conflict of Interest: The Authors declare that there is no conflict of interest.

References

Abe,H. *et al.* (2003) Arabidopsis atmyc2 (bhlh) and atmyb2 (myb) function as transcriptional activators in abscisic acid signaling. *Plant Cell*, **15**, 63–78.

Breiman,L. and Friedman,J.H. (1997) Predicting multivariate responses in multiple linear regression. *J. R. Stat. Soc. Ser. B (Methodological)*, **59**, 3–54.

Budhlakoti,N. *et al.* (2019) MTGS: Multiple traits genomic selection. R package version 0.1.0 — For new features, see the 'Changelog' file (in the package source). *Journal of Computational Biology*, **26**, 1100–1112.

Cai,H. *et al.* (2014) Multi-output regression with tag correlation analysis for effective image tagging. In: *International Conference on Database Systems for Advanced Applications*. Springer, Bali, Indonesia. pp. 31–46.

Calus,M.P. and Veerkamp,R.F. (2011) Accuracy of multi-trait genomic selection using different methods. *Genet. Select. Evol.*, **43**, 26.

Chen,L. *et al.* (2014) Regularized multivariate regression models with skew-t error distributions. *J. Stat. Plann. Inference*, **149**, 125–139.

Cheng,H. *et al.* (2018) Genomic prediction from multiple-trait Bayesian regression methods using mixture priors. *Genetics*, **209**, 89–103.

Cook,J.P. *et al.* (2012) Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. *Plant Physiol.*, **158**, 824–834.

Crossa,J. *et al.* (2017) Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.*, **22**, 961–975.

Ding,C. *et al.* (2006) R 1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization. In *Proceedings of the 23rd International Conference on Machine Learning*. ACM, Pittsburgh, Pennsylvania, USA, pp. 281–288.

Escoufier,Y. (1973) Le traitement des variables vectorielles. *Biometrics*, **29**, 751–760.

Fernandes,S.B. and Lipka,A.E. (2020) simplephenotypes: simulation of pleiotropic, linked and epistatic phenotypes. *BMC Bioinformatics*, **21**, 1–10.

Friedman,J. *et al.* (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.

Habier,D. *et al.* (2011) Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, **12**, 186.

Hayes,B. *et al.* (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819–1829.

He,D. *et al.* (2016) Novel applications of multitask learning and multiple output regression to multiple genetic trait prediction. *Bioinformatics*, **32**, i37–i43.

Heffner,E.L. *et al.* (2010) Plant breeding with genomic selection: gain per unit time and cost. *Crop Sci.*, **50**, 1681–1690.

Henderson,C.R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics*, **31**, 423–447.

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

Horton,M.W. *et al.* (2012) Genome-wide patterns of genetic variation in worldwide arabidopsis thaliana accessions from the regmap panel. *Nat. Genet.*, **44**, 212–216.

Jia,Y. and Jannink,J.-L. (2012) Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics*, **192**, 1513–1522.

Kanehara,K. *et al.* (2015) Arabidopsis atplc2 is a primary phosphoinositide-specific phospholipase c in phosphoinositide metabolism and the endoplasmic reticulum stress response. *PLoS Genet.*, **11**, e1005511.

Karaman,E. *et al.* (2018) Genomic prediction using multi-trait weighted gblup accounting for heterogeneous variances and covariances across the genome. *Genes Genomes Genet.*, **8**, 3549–3558.

Kim,S. *et al.* (2007) Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat. Genet.*, **39**, 1151–1155.

Kole,C. *et al.* (2002) Comparative mapping of loci controlling winter survival and related traits in oilseed *Brassica rapa* and *B. napus*. *Mol. Breed.*, **9**, 201–210.

Lee,W. and Liu,Y. (2012) Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *J. Multivariate Anal.*, **111**, 241–255.

Manfield,I.W. *et al.* (2007) Conservation, convergence, and divergence of light-responsive, circadian-regulated, and tissue-specific expression patterns during evolution of the arabidopsis gata gene family. *Plant Physiol.*, **143**, 941–958.

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

Nie,F. *et al.* (2010) Efficient and robust feature selection via joint *l*2,1-norms minimization. *Adv. Neural Inf. Process. Syst.*, **23**, 1813–1821.

Ogutu,J.O. *et al.* (2012) Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. In *BMC Proceedings*, Vol. 6. Springer, Rennes, France.

Park,T. and Casella,G. (2008) The Bayesian lasso. *J. Am. Stat. Assoc.*, **103**, 681–686.

Peng,J. *et al.* (2010) Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Ann. Appl. Stat.*, **4**, 53–77.

Pérez,P. and de Los Campos,G. (2014) Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, **198**, 483–495.

Riedel,K.S. (1992) A Sherman-Morrison-Woodbury identity for rank augmenting matrices with application to centering. *SIAM J. Matrix Anal. Appl.*, **13**, 659–662.

Rothman,A.J. *et al.* (2008) Sparse permutation invariant covariance estimation. *Electronic J. Stat.*, **2**, 494–515.

Rothman,A.J. *et al.* (2010) Sparse multivariate regression with covariance estimation. *J. Comput. Graph. Stat.*, **19**, 947–962.

Shi,S. *et al.* (2018) The Arabidopsis calcium-dependent protein kinases (CDPKS) and their roles in plant growth regulation and abiotic stress responses. *Int. J. Mol. Sci.*, **19**, 1900.

Sun,L. *et al.* (2009) Efficient recovery of jointly sparse vectors. In: Bengio,Y. *et al.* (eds.) *Advances in Neural Information Processing Systems*, Vol. 22. Curran Associates, Inc., Vancouver, British Columbia, Canada. pp. 1812–1820

Turlach,B.A. *et al.* (2005) Simultaneous variable selection. *Technometrics*, **47**, 349–363.

Usai,M.G. *et al.* (2009) Lasso with cross-validation for genomic selection. *Genet. Res.*, **91**, 427–436.

VanRaden,P.M. (2008) Efficient methods to compute genomic predictions. *J. Dairy Sci.*, **91**, 4414–4423.

Wang,C. *et al.* (2018) Sip1, a novel sos2 interaction protein, is involved in salt-stress tolerance in Arabidopsis. *Plant Physiol. Biochem.*, **124**, 167–174.

Wang,X. *et al.* (2019) Evaluation of gblup, bayesb and elastic net for genomic prediction in Chinese Simmental beef cattle. *PLoS One*, **14**, e0210442.

Westhues,M. *et al.* (2017) Omics-based hybrid prediction in maize. *Theor. Appl. Genet.*, **130**, 1927–1939.

Wu,S. *et al.* (2016) Combined use of genome-wide association data and correlation networks unravels key regulators of primary metabolism in *Arabidopsis thaliana*. *PLoS Genet.*, **12**, e1006363.

Xu,J. and Yin,J. (2013) Kernel least absolute shrinkage and selection operator regression classifier for pattern classification. *IET Comput. Vis.*, **7**, 48–55.

Yan,Y. *et al.* (2014) A MYB-domain protein EFM mediates flowering responses to environmental cues in Arabidopsis. *Dev. Cell*, **30**, 437–448.

Yuan,M. and Lin,Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **68**, 49–67.

Yuan,M. *et al.* (2007) Dimension reduction and coefficient estimation in multivariate linear regression. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **69**, 329–346.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **67**, 301–320.