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., 2009). 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) (Van Raden, 2008), BayesA and BayesB (Meuwissen et al., 2001), BayesC $\pi$ (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., 2019), 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 Jamink, 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 forges 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 (e.g. biomass) (Westhues et al., 2017) is defined as, the centered multiple output regression 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 L2 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., 2013) 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 L2,1-norm joint, a novel multivariate method that models the response variables jointly in the penalize likelihood framework using the L2,1-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 L2,1-norm. Finally, we recall the statistical formulation of some regression models that will be used to compare our proposed method, i.e. the L2,1-norm regression for GS (L2,1-fs), the centered multiple output regression (cMOR), and RR. Throughout the rest of this paper, for a matrix $V = (v_{ij})$, we denote by $v'$ and $v_j$ its $r$th row and $j$th column respectively. The symbols $\mathbf{I}$ and $\mathbf{0}$ stand for trace and vectorization operators respectively. $V^{-1}$ is the inverse and $V^t$ the transpose of $V$. The $n$-dimensional identity matrix is denoted $I_n$ and the $p$-norm of a vector $v \in \mathbb{R}^n$ is defined as,

$$\|v\|_p = \left( \sum_{i=1}^{n} |v_i|^p \right)^{1/p},$$

where $v_i$ represents the $i$th element of $v$.

2.1 L2,1-norm

First introduced in (Ding et al., 2006), the L2,1-norm of a matrix $V \in \mathbb{R}^{n \times m}$ is defined by,

$$\|V\|_{2,1} = \sum_{j=1}^{m} \|v_j\|_2 = \sum_{j=1}^{m} \left( \sum_{i=1}^{n} |v_{ij}|^2 \right)^{1/2},$$

It has been shown that the L2,1-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:

$$\|VR\|_{2,1} = \|V\|_{2,1}. \quad (3)$$

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 $q_{ii} = \frac{1}{\|v_i\|_2}$.

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

Let $Y = [y_1, y_2, \ldots, y_p] \in \mathbb{R}^{n \times p}, X = [x_1, x_2, \ldots, x_p] \in \mathbb{R}^{n \times p}, B = [b_1, b_2, \ldots, b_p] \in \mathbb{R}^{p \times p}$ and $E = [e_1, e_2, \ldots, e_p] \in \mathbb{R}^{n \times p}$ 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 $y_1, y_2, \ldots, y_p$ and $p$ predictor variables $x_1, x_2, \ldots, x_p$, so that, if the $i$th observation of the response, the $p$th value of the predictor variables and the $r$th unobserved random vector are respectively defined by $y_i = (y_{i1}, y_{i2}, \ldots, y_{ip})$, $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $e_i = (e_{i1}, e_{i2}, \ldots, e_{ip})$, then the linear regression model takes the following matrix representation:

$$Y = XB + E. \quad (4)$$

We also assume that $e_i$ are independent and have identical multivariate normal distribution with mean vector $0$ and covariance matrix $\Sigma$. 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

$$J(B, \Omega) = tr \left( \frac{1}{\Omega} (Y - XB) \Omega^{-1} (Y - XB)^t \right) - \log \det \Omega, \quad (5)$$

with maximum likelihood estimate (MLE) for $B$ that does not depend on $\Omega$

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

$\hat{B}_{mle}$ 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 $\hat{B}_{mle}$ 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,

$$\min_{B} ||Y - XB||_{2.1} + \lambda_2||B||_{2.1}.\quad(7)$$

The solution of this optimization problem is given by:

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

where $D$ is the diagonal matrix with the $r$th entry $d_r = \frac{1}{A_{rr}}$.

$$A = [X, l_0] \in \mathbb{R}^{n \times o}, \quad W = [B, 1] \in \mathbb{R}^{m \times o}, \quad \text{and} \quad m = p + n.$$  

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 $L_1$-norm penalty on the regression coefficient matrix $B$. Among several penalty functions, 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 $x -$ master regulator ($MR_x$) if the corresponding row in the estimated sparse regression coefficient matrix $\hat{B} \in \mathbb{R}^{p \times o}$ is $x -$ dense, i.e. at least an $x -$ fraction, $0.5 \leq x \leq 1$, of the entries in the corresponding row are non-zero. Moreover, for the purpose of this study, we only consider the case $x = 1$ and use $MR_1$ to define the proportion of rows that are $MR_1$.

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

$$\hat{B}(\lambda) = \arg\min_{B} ||Y - XB||_2 + \lambda ||B||_2,$$

with solution:

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

In contrast, LASSO solves

$$\hat{B}(\lambda) = \arg\min_{B} ||Y - XB||_2 + \lambda_1 ||B||_1.$$

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)

$$\hat{B}(\lambda) = \arg\min_{B} ||Y - \Phi(X)B||_2 + \lambda \sqrt{\lambda} ||B||_1,$$

where $\Phi$ is the kernel function. Finally, the multiple output regression solves

$$(\hat{B}, \hat{\Omega}, \hat{\Sigma}) = \arg\min_{(B, \Omega, \Sigma)} \text{tr}(Y'Y - XB\Omega^{-1}(Y'XB)') - n \log \Omega^{-1} - \lambda_1 \text{tr}(BB') + \lambda_2 \text{tr}(\Sigma^{-1}B'B) - p \log |\Sigma^{-1}|$$

$$+ \lambda_3 \text{tr}(\Omega^{-1}) + \lambda_4 \text{tr}(\Sigma^{-1}).\quad(13)$$

$\Omega^{-1}$ and $\Sigma^{-1}$ 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:

$$K(B, \Omega) = \text{tr} \left[ \frac{1}{s^2} (Y' - B'X')' \Omega^{-1} (Y' - B'X') \right] - \log |\Omega|.\quad(14)$$

The $L_1$ penalty is then applied on the precision matrix $\Omega$ 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 $B$ which, in turn, can aid the interpretation of the fitted model. Our model then provides the estimates $B$ and $\Omega$ by solving the following optimization problem:

$$f(B, \Omega) = \arg\min_{B, \Omega} \left\{ \text{tr} \left[ \frac{1}{s^2} (Y' - B'X')' (Y' - B'X') \Omega \right] - \log |\Omega| + \lambda_2 ||B||_{2.1} \right\}.\quad(15)$$

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 $\Omega$ at a chosen point $\Omega_0$ is equivalent to optimizing

$$\hat{B}(\Omega_0) = \arg\min_{B} \left\{ \text{tr} \left[ \frac{1}{s^2} (Y' - B'X')' (Y' - B'X') \Omega_0 \right] - \log |\Omega_0| + \lambda_2 ||B||_{2.1} \right\}.\quad(16)$$

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

$$B = \left[ X'\Omega_0 X + \frac{\lambda_2}{s^2} C \right]^{-1} X'\Omega_0 Y.\quad(17)$$

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:

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

where $C$ is the diagonal matrix with $r$th entry $c_r = \frac{1}{A_{rr}}$. 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 $C = I_n$, we have the RR estimate, and, finally, when $C = I_n$, we have the $L_{2.1}$-norm based variable selection.

Solving Eq. (15) for $\Omega$ 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 $\hat{\Omega}$ in the model given in Eq. (19), below:

$$\hat{\Omega}(B_0) = \arg\min_{\Omega} \left\{ \text{tr} \left[ \frac{1}{s^2} (Y' - B_0'X')' (Y' - B_0'X') \Omega \right] - \log |\Omega| + \lambda_1 ||\Omega||_1 \right\}.\quad(19)$$
Here, $Y^t$ and $X^t$ are respectively the $k^t$-response and predictor matrices, while $B_{ij}^{0:k}$ is the regression coefficient matrix estimated out of the $k^t$-fold for $l_1$ and $l_2$. In addition, seq(3, 12, 1) and $2^{seq(5, 2, 1)}$ are used as search grids to obtain the optimal $l_1$ and $l_2$ 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 $B$. These are given respectively by the proportion of non-zero entries in the true coefficient $B$ identified correctly by the estimate $\hat{B}$ and the proportion of zero entries in $B$ that $\hat{B}$ 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 MR1 by computing mr1, the proportion of rows with non-zero entries in $B$ correctly identify by $\hat{B}$. Therefore, mr1 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 $(\Sigma_0)_{ij} = \gamma_{ij}$, so that rows of the design matrix $X \in \mathbb{R}^{n \times p}$ are independently generated from the multivariate normal distribution $N(0, \Sigma)$. For the genomic prediction application, different coding for the genotypes (predictors) matrix $X$ can be obtained. For instance, all absolute values in the intervals $[0, 5], [5, 1]$ and $[1, \infty]$ can respectively be coded as 0, 1 and 2, which is an alternative to randomly sample the genotype matrix $X$ from $[0, 1, 2]$. For the error matrix $E \in \mathbb{R}^{n \times q}$, an autoregressive covariance structure of order 1, AR(1), is considered, implying that rows of $E$ are independently drawn from the multivariate normal distribution $N(0, \Sigma)$, with $(\Sigma)_{ij} = \rho^{j-i}$ and $\rho$ taking values (.1, .5, .9). Using the matrix element wise product $B = W \times Q + K \times W$, a sparse regression coefficient matrix is obtained. With the modification, we further obtain some rows in $B$ that are non-zero so that the proportion of correctly identified master regulators can be computed. In this setting, each entry of $W$ 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 $\rho$ independent Bernoulli draws with $s_2$ probability of success. Following Eq(4), 30 trials 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 assess 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 (.7and.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 55K SNP array (Cook et al., 2012) are used to simulate 12 highly correlated traits controlled by 80 MR1. 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 $L_2,1$-norm 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

---

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

**Input:** $\lambda_1, \lambda_2$, $X \in \mathbb{R}^{p \times x}$, $Y \in \mathbb{R}^{n \times x}$

**Output:** $\Omega \in \mathbb{R}^{p \times n}$, $B \in \mathbb{R}^{p \times x}$

1. Set $t = 0$ and initialize
   - The diagonal matrix $C_1 \in \mathbb{R}^{p \times p}$ as identity,
   - $\Omega_l$ as the average of the ridge covariance matrix
   - $\hat{B}$ as $\hat{B}(\hat{\Omega}_t)$ solving Eq. (18)
   - $\Omega$ as $\Omega_l(\hat{B})$ by solving Eq. (19) using GLASSO

2. repeat
   - Update $C_0$ by computing $C_{t+1}$ with the $j^{th}$ diagonal entry
     $$c_{jj} = \frac{1}{2\|b_j^t\|^2}$$
   - Update $\Omega$ by computing $\hat{\Omega}_{t+1} = \hat{\Omega}(\hat{B}_{t+1})$ by solving Eq. (19) using GLASSO
   - Update $\hat{B}$ by computing $\hat{B}_{t+1} = \hat{B}(\hat{\Omega}_{t+1})$ 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 $B$. In addition, we use the sample covariance matrix of the RR residual to scale the convergence of the precision matrix ($\text{Chen et al., 2014}$). This implies that the convergence criteria for $B$ and $\Omega$ are met when $\sum_{j} |b_{ij}| - \tilde{b}_{ij} | < \varepsilon_1 \sum_{j} |b_{ij}|$ and $\sum_{j} |\hat{b}_{ij}^{(t+1)} - \hat{b}_{ij}^{(t)}| < \varepsilon_2 \sum_{j} |\hat{b}_{ij}^{(t)}|$, respectively. Here, $\varepsilon_1$ and $\varepsilon_2$ are the tolerance parameters that we set to $10^{-5}$. Moreover, because our objective function is convex in $B$ when the other parameter is fixed and monotonically decreasing in each iteration, another convergence criteria one can use is given by $\|\hat{B}_{t+1}\|_{2,1} \geq \|B^{(t)}\|_{2,1}$ 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:

$$RV(Y, \hat{Y}) = \frac{\text{tr}(YY^T \hat{Y} \hat{Y}^T)}{\sqrt{\text{tr}(YY^T) \text{tr}(\hat{Y} \hat{Y}^T)}}$$

and

$$MSE(Y, \hat{Y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$
predictability, and fixed p = 800 for sparsity and MR1 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 80 markers set as RR cMOR L2-1-fs L2-1-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 (0.06) 0.81 (0.07) 0.80, 0.02
50,300,30 0.71 (0.07) 0.72 (0.07) 0.47 (0.01) 0.72 (0.07) 0.69 (0.07) 0.65 (0.07) 0.70 (0.07) 0.70 (0.06) 0.80
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) 0.80

(B) Recovery rate of MR1

For methods with ability to correctly identify non-zero entries in the true regression coefficient matrix, all metrics are averaged over 20 replicates with AR(1) parameter 51.2 • 0.3 joint RR mLASSO Elastic-Net cMOR remMAP MBayesB GBLUP
50,800,30 0.1 – – 38.8 80.3 0 – 0 0 0.27 0.63 (0.06) 0.61 (0.07) 0.25 (0.05) 0.42 (0.05) 0.42 (0.05) 0.22 (0.06) 0.03 (0.04)
50,100,30 0.5 – – 38.9 80.4 0 – 0 0 0.24 (0.05) 0.64 (0.05) 0.64 (0.04) 0.24 (0.05) 0.24 (0.05) 0.64 (0.05) 0.64 (0.04) 0.24 (0.05) 0.24 (0.05)
0.9 – – 38.8 80.02 0 – 0 0 0.23 (0.04) 0.58 (0.05) 0.59 (0.06) 0.22 (0.06) 0.03 (0.04) 0.64 (0.05) 0.64 (0.04) 0.24 (0.05) 0.24 (0.05)

Note: The dataset consists of s = 30 simulated phenotypes, n = 50 observations and varying number of predictors p ∈ [100, 300, 800] to see their impact on predictability, and fixed p = 800 for sparsity and MR1 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 MR1, (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 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 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.

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

(A) Predictability

(B) Recovery rate of MR1

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 known MR1. 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.

(B) Recovery rate of MR1

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 known MR1. 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.

(B) Recovery rate of MR1

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 known MR1. 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.
Table 3. Comparison of model performance on Brassica napus data

<table>
<thead>
<tr>
<th>Model</th>
<th>RV-Coeff</th>
<th>mr₁</th>
<th>flower 0</th>
<th>flower 4</th>
<th>flower 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>RR</td>
<td>0.46</td>
<td>–</td>
<td>2.40 (0.69)</td>
<td>1.98 (0.57)</td>
<td>1.81 (0.52)</td>
</tr>
<tr>
<td>cMOR</td>
<td>0.46</td>
<td>–</td>
<td>2.35 (0.67)</td>
<td>2.00 (0.57)</td>
<td>1.84 (0.53)</td>
</tr>
<tr>
<td>L₂,₁-fs</td>
<td>0.43</td>
<td>33%</td>
<td>2.51 (0.72)</td>
<td>1.95 (0.56)</td>
<td>1.75 (0.50)</td>
</tr>
<tr>
<td>L₂,₁-joint</td>
<td>0.46</td>
<td>23%</td>
<td>2.48 (0.71)</td>
<td>1.95 (0.55)</td>
<td>1.73 (0.49)</td>
</tr>
<tr>
<td>kmLASSO</td>
<td>0.44</td>
<td>–</td>
<td>4.54 (1.31)</td>
<td>3.28 (0.94)</td>
<td>1.85 (0.53)</td>
</tr>
<tr>
<td>mLASSO</td>
<td>0.30</td>
<td>–</td>
<td>2.51 (0.73)</td>
<td>1.95 (0.57)</td>
<td>1.72 (0.50)</td>
</tr>
<tr>
<td>MRCE</td>
<td>0.10</td>
<td>97%</td>
<td>2.22 (0.65)</td>
<td>1.76 (0.51)</td>
<td>1.62 (0.45)</td>
</tr>
<tr>
<td>GBLUP</td>
<td>0.49</td>
<td>–</td>
<td>2.46 (0.73)</td>
<td>1.92 (0.54)</td>
<td>1.74 (0.53)</td>
</tr>
<tr>
<td>Elastic-Net</td>
<td>0.36</td>
<td>–</td>
<td>2.46 (0.72)</td>
<td>1.91 (0.62)</td>
<td>1.73 (0.49)</td>
</tr>
<tr>
<td>MBayesB</td>
<td>0.57</td>
<td>–</td>
<td>2.50 (0.74)</td>
<td>1.95 (0.56)</td>
<td>1.74 (0.57)</td>
</tr>
<tr>
<td>remMAP</td>
<td>0.28</td>
<td>4%</td>
<td>2.51 (0.71)</td>
<td>1.94 (0.53)</td>
<td>1.73 (0.50)</td>
</tr>
</tbody>
</table>

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.

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 > 0.78) 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 L₂,₁-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 L₂,₁-joint is the third best performing, after MBayesB and GBLUP, for
formed metabolomics profiles for 94 primary metabolites.

Here, we compare the performance of L2

3.3 Comparative analysis with wheat dataset

We removed all SNPs with less than 5% minimum allele frequency (MAF), leaving us with 200 180 to build the L2

Table 4. Comparison of model performance on wheat data

<table>
<thead>
<tr>
<th>Model</th>
<th>RV-Coeff</th>
<th>mr1</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>trait 1</td>
</tr>
<tr>
<td>RR</td>
<td>0.10</td>
<td>–</td>
<td>3.1 (0.7)</td>
</tr>
<tr>
<td>cMOR</td>
<td>0.09</td>
<td>–</td>
<td>3.7 (0.8)</td>
</tr>
<tr>
<td>L2_1-fs</td>
<td>0.10</td>
<td>70%</td>
<td>2.6 (0.5)</td>
</tr>
<tr>
<td>L2_1-joint</td>
<td>0.13</td>
<td>21%</td>
<td>1.9 (0.4)</td>
</tr>
<tr>
<td>kMLASSO</td>
<td>0.008</td>
<td>–</td>
<td>2 (0.4)</td>
</tr>
<tr>
<td>mLASSO</td>
<td>0.08</td>
<td>–</td>
<td>1.7 (0.3)</td>
</tr>
<tr>
<td>GBLUP</td>
<td>0.14</td>
<td>–</td>
<td>1.8 (0.4)</td>
</tr>
<tr>
<td>Elastic-Net</td>
<td>0.11</td>
<td>1%</td>
<td>1.9 (0.4)</td>
</tr>
<tr>
<td>MBayesB</td>
<td>0.15</td>
<td>53%</td>
<td>1.9 (0.4)</td>
</tr>
<tr>
<td>remMAP</td>
<td>0.13</td>
<td>–</td>
<td>1.8 (0.4)</td>
</tr>
</tbody>
</table>

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. With mod-els allowing the identification of master regulators (remMAP and MBayesB), we see that L2_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 L2_1-joint we also identify that 23% of SNPs are found as MR1, 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 L2_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 (Perez 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, L2_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 L2_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 L2_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 MR1, we can say that, for moderate number of predictors, L2_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 0.78 and only few values above 0.5. 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 L2_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, L2_1-joint and L2_1-fs.

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 L2_1-joint and L2_1-fs models are superior. Given the observation that our L2_1-joint model outperforms L2_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 L2_1-joint achieve equal predictability as quantified by the number of time each method outperforms the contender, with L2_1-joint exhibiting high performance when simultaneously considering predictability and variable selection with respect to the competitors.

<table>
<thead>
<tr>
<th>Model</th>
<th>RV-Coeff</th>
<th>Selected variables</th>
<th>mr1</th>
</tr>
</thead>
<tbody>
<tr>
<td>RR</td>
<td>0.26</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>cMOR</td>
<td>0.26</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>L2_1-fs</td>
<td>0.27</td>
<td>55960 (27.9%)</td>
<td>4249 (2.12%)</td>
</tr>
<tr>
<td>L2_1-joint</td>
<td>0.26</td>
<td>135597 (67.73%)</td>
<td>30819 (15.39%)</td>
</tr>
</tbody>
</table>

Note: RV coefficient for predicting metabolites levels across the 40 testing lines, features selection and identification of MR1 for L2_1-joint, L2_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.
Anhydro-beta-D-glucose attained by RR. The desirable property of L2,1-norm regularized multivariate genomic prediction 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 BREDCAFS [GA No. 727934] and PlantaSYS [FPA No. 664620].

Data availability
The data underlying this article are publicly available and their corresponding references provided within the article.

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

References


