Multi-trait genomic selection via multivariate regression with structured regularization

Julien Chiquet, Tristan Mary, Stephane Robin

Multi-trait genomic selection via multivariate regression with structured regularization

Julien Chiquet[1,2] JULIEN.CHIQUET@GENOPOLE.CNRS.FR
Stéphane Robin[2] STEPHANE.ROBIN@AGROPARISTECH.FR
Tristan Mary-Huard[2] TRISTAN.MARY-HUARD@AGROPARISTECH.FR
[1]Laboratoire Statistique et Génome – UMR CNRS 8071/Université d'Évry Val d'Essonne – Évry, France
[2]Laboratoire MMIP – UMR INRA 518/AgroParisTech – Paris, France

1. Introduction

Genomic selection aims at predicting one or several phenotypes on the basis of available genetic information (markers). In plant and animal genetics, deriving an accurate prediction of complex traits is of major importance for the early detection and selection of individuals of high genetic value. Due to the continued advancement of high-throughput genotyping and sequencing technologies, genetic information is now available at low cost in the form of thousands of markers. On the other hand, acquiring trait information remains expensive, and a typical experiment will only contain a few hundreds of phenotyped individuals. This leads to the classical "high dimensional setting" where the number of features is much higher than the number of observations available for training. Consequently regularization methods such as Ridge or Lasso regression or their Bayesian counterparts have been proposed since the very beginning of genomic selection (de los Campos et al., 2012).

In many application fields, statisticians have successfully made use of the ability of regularized methods to take into account distinctive features of the data (Schölkopf & Smola, 2002). In genomics for instance it is current practice to integrate biological prior knowledge such as gene networks, pathways or GO attributes to the regularization function to improve both the performance and the interpretability of the prediction function. In genomic selection regularized methods have mostly been used for their ability to handle high dimensional data, and little attention has been devoted to the development of penalty functions including prior knowledge. Moreover, while several traits are usually considered in a given experiment, most methods only perform single trait genomic selection, neglecting correlations between phenotypes and leading to poor performance for the prediction of traits with low heritability.

To circumvent these limitations, we consider the general linear model to simultaneously predict q responses (output variables) using the same set of p markers (input variables) based on a training sample $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1,\dots,n}$. One has

$$\mathbf{y}_i = \mathbf{B}^T \mathbf{x}_i + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{R}), \quad \forall i = 1, \dots, n, \quad (1)$$

where $\boldsymbol{\varepsilon}_i$ is a noise term with a q-dimensional unknown covariance matrix $\mathbf{R}$, and $\mathbf{B}$ is the $p \times q$ matrix of regression coefficients (we omit the intercept term and center the data for clarity). The purposes of the present work are the following: i) to account for the dependency structure between the outputs, i.e. to integrate the estimation of $\mathbf{R}$ in the inference process; ii) to integrate some prior information about linkage disequilibrium to account for the dependency structure between markers and evaluate its influence on the different phenotypes; iii) to induce sparsity on partial covariances rather than on the regression coefficients $\mathbf{B}$, since according to the Gaussian graphical models (GGM) direct effects are measured by partial covariances between predictors and responses. We present our estimator to achieve these three goals, and illustrate its behavior through an application to the *Brassica napus* dataset (Kole et al., 2002).

2. Model setup and learning

2.1. Convex parametrization of multivariate regression

The statistical framework arises from a different parametrization of (1), making a connection between multivariate regression and GGM that was first underlined in Sohn & Kim (2012). This amounts to investigate the joint probability distribution of $(\mathbf{x}_i, \mathbf{y}_i)$ in the Gaussian case, with the following block-wise decomposition of the covariance matrix $\boldsymbol{\Sigma}$ and its inverse $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}} \\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}} \end{pmatrix}, \quad \boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{\mathbf{xx}} & \boldsymbol{\Omega}_{\mathbf{xy}} \\ \boldsymbol{\Omega}_{\mathbf{yx}} & \boldsymbol{\Omega}_{\mathbf{yy}} \end{pmatrix}. \quad (2)$$

Back to the distribution of $\mathbf{y}_i$ conditional on $\mathbf{x}_i$, multivariate Gaussian analysis shows that, for centered $\mathbf{x}_i$ and $\mathbf{y}_i$,

$$\mathbf{y}_i | \mathbf{x}_i \sim \mathcal{N}\left(-\boldsymbol{\Omega}_{\mathbf{yy}}^{-1} \boldsymbol{\Omega}_{\mathbf{yx}} \mathbf{x}_i, \boldsymbol{\Omega}_{\mathbf{yy}}^{-1}\right). \quad (3)$$

Introducing the empirical matrices of covariance $\mathbf{S_{yy}} = n^{-1}\sum_{i=1}^n \mathbf{y}_i\mathbf{y}_i^T$, $\mathbf{S_{xx}} = n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T$, and $\mathbf{S_{yx}} = n^{-1}\sum_{i=1}^n \mathbf{y}_i\mathbf{x}_i^T$, the log-likelihood associated with these parameters – which is a conditional likelihood regarding the joint model (2) – verifies

$$-\frac{2}{n}\log L(\mathbf{\Omega_{xy}}, \mathbf{\Omega_{yy}}) = -\log|\mathbf{\Omega_{yy}}| + \mathrm{tr}\left(\mathbf{S_{yy}\Omega_{yy}}\right) + 2\mathrm{tr}\left(\mathbf{S_{xy}\Omega_{yx}}\right) + \mathrm{tr}(\mathbf{\Omega_{yx}S_{xx}\Omega_{xy}\Omega_{yy}^{-1}}) + \mathrm{cst}. \quad (4)$$

This has been referred to as a partial or conditional Gaussian Graphical Model in the recent literature (cGGM, see Yin & Li, 2011; Sohn & Kim, 2012; Lee & Liu, 2012; Yuan & Zhang, 2012). We notice by comparing the cGGM (3) to the multivariate regression model (1) that $\mathbf{\Omega_{yy}^{-1}} = \mathbf{R}$ and $\mathbf{B} = -\mathbf{\Omega_{xy}\Omega_{yy}^{-1}}$. Although equivalent to (1), parametrization (3) shows two important differences with several implications. First, the negative log likelihood (4) can be shown to be jointly convex in $(\mathbf{\Omega_{xy}}, \mathbf{\Omega_{yy}})$ while its counterpart in $(\mathbf{B}, \mathbf{R}^{-1})$ is only biconvex. Hence, minimization problems involving (4) are amenable to a global solution, which facilitates both optimization and theoretical analysis. The conditional negative log-likelihood (4) will thus serve as a building block for our learning criterion. Second, it unveils new interpretations for the relationships between input and output variables, as discussed in Sohn & Kim (2012): $\mathbf{\Omega_{xy}}$ describes the *direct* links between predictors and responses, the support of which we are looking for to select relevant interactions. On the other hand, $\mathbf{B}$ entails *both direct and indirect* influences, possibly due to strong correlations between the responses, described by the covariance matrix $\mathbf{R}$ (or equivalently its inverse $\mathbf{\Omega_{yy}}$).

2.2. Structured regularization with underlying sparsity

Our regularization scheme starts by considering some structural prior information: adjacent markers (inputs) on the sequence should have similar direct relationships with the phenotypic traits (outputs). This depicts a pattern on the predictors that acts along the rows of $\mathbf{B}$ or $\mathbf{\Omega_{xy}}$ as substantiated by the following Bayesian point of view.

Bayesian interpretation. Assume similarities between inputs can be encoded into a matrix $\mathbf{L}$. The Bayesian framework provides a convenient setup to define the way the structural information should be accounted for when learning the coefficients. In the single output case (see, e.g. Marin & Robert, 2007), the conjugate prior for $\boldsymbol{\beta}$ would be $\mathcal{N}(\mathbf{0}, \mathbf{L}^{-1})$. Combined with the covariance between the outputs, this gives $\mathrm{vec}(\mathbf{B}) \sim \mathcal{N}(\mathbf{0}, \mathbf{R}\otimes\mathbf{L}^{-1})$, where $\otimes$ is the Kronecker product. By properties of the vec operator this can be equivalently stated for the direct links as $\mathrm{vec}(\mathbf{\Omega_{xy}}) \sim \mathcal{N}(\mathbf{0}, \mathbf{R}^{-1}\otimes\mathbf{L}^{-1})$. Choosing such a prior results in

$$\log\mathbb{P}(\mathbf{\Omega_{xy}}|\mathbf{L},\mathbf{R}) = \frac{1}{2}\mathrm{tr}\left(\mathbf{\Omega_{xy}^T L\Omega_{xy}R}\right) + \mathrm{cst}.$$

LD-based regularization. Information about linkage disequilibrium between markers can be used to define the regularization matrix $\mathbf{L}$. Genetic maps provide the genetic distances between the markers. This distance is expressed in centi-Morgan (cM), meaning that the correlation between two markers distant from a distance of d cM is ρ^d, where $\rho = .98$[1] The covariance matrix $\mathbf{L}^{-1}$ can hence be defined as $\mathbf{L}_{ij}^{-1} = \rho^{d_{ij}}$. Because of the Markovian structure of linkage disequilibrium (the associated graphical model is a chain graph), $\mathbf{L}$ is tridiagonal with general elements

$$w_{i,i} = \frac{1-\rho^{2d_{i-1,i}+2d_{i,i+1}}}{(1-\rho^{2d_{i-1,i}})(1-\rho^{2d_{i,i+1}})},$$
$$w_{i,i+1} = \frac{-\rho^{d_{i,i+1}}}{1-\rho^{2d_{i,i+1}}}$$

and $w_{i,j} = 0$ if $|i-j| > 1$. For the first (resp. last) marker, the distance $d_{i-1,i}$ (resp. $d_{i,i+1}$) is infinite.

Link to structured regression methods. When $q = 1$, $\mathbf{R}$ turns to a variance σ^2 and $\mathbf{\Omega_{xy}}$ a vector $\boldsymbol{\omega}_{\mathbf{xy}}$, so as

$$\frac{1}{2}\mathrm{tr}\left(\mathbf{\Omega_{xy}^T L\Omega_{xy}R}\right) = \frac{\sigma^2}{2}\boldsymbol{\omega}_{\mathbf{xy}}^T\mathbf{L}\boldsymbol{\omega}_{\mathbf{xy}}.$$

When $\mathbf{L} = \mathbf{I}$, we recognize the regularization term of ridge regression. Coupled with a ℓ_1 norm, we meet back the Elastic-Net of Zou & Hastie (2005). If $\mathbf{L}$ is the Laplacian of a graph describing direct relationships between parameters this is the structured Elastic-Net of Slawski et al. (2010).

Criterion. By this argument, we propose the following criterion with two regularizing terms: a smooth trace term relying on the available structural information $\mathbf{L}$ and a ℓ_1 norm encouraging sparsity among the direct links. The optimization problem turns to the joint minimization of

$$J(\mathbf{\Omega_{xy}}, \mathbf{\Omega_{yy}}) = -\frac{1}{n}\log L(\mathbf{\Omega_{xy}}, \mathbf{\Omega_{yy}}) + \frac{\lambda_2}{2}\mathrm{tr}\left(\mathbf{\Omega_{yx}L\Omega_{xy}\Omega_{yy}^{-1}}\right) + \lambda_1\|\mathbf{\Omega_{xy}}\|_1. \quad (5)$$

Optimization. Thanks to the convexity of (4) combined with the trace and ℓ_1 norms, Problem (5) is jointly convex in $(\mathbf{\Omega_{xy}}, \mathbf{\Omega_{yy}})$ when (λ_1, λ_2) are fixed. An algorithm alternating optimization on $\mathbf{\Omega_{yy}}$ and $\mathbf{\Omega_{xy}}$ is guaranteed to converge to the minimum:

$$\hat{\mathbf{\Omega}}_{\mathbf{yy}}^{(k+1)} = \underset{\mathbf{\Omega_{yy}}\succ 0}{\arg\min}\, J_{\lambda_1\lambda_2}(\hat{\mathbf{\Omega}}_{\mathbf{xy}}^{(k)}, \mathbf{\Omega_{yy}}), \quad (6a)$$

$$\hat{\mathbf{\Omega}}_{\mathbf{xy}}^{(k+1)} = \underset{\mathbf{\Omega_{xy}}}{\arg\min}\, J_{\lambda_1\lambda_2}(\mathbf{\Omega_{xy}}, \hat{\mathbf{\Omega}}_{\mathbf{yy}}^{(k+1)}). \quad (6b)$$

We can solve analytically (6a) when $q < n$ based on simple matrix algebra, while (6b) can be recast as an Elastic-Net problem, for which very efficient algorithms are available.

[1] This value directly arises from the definition of the cM itself.

3. Genomic selection in Brassica napus

In the study conducted by Ferreira et al. (1995) and Kole et al. (2002), $n = 103$ lines of *Brassica napus* where considered, on which $p = 300$ genetic markers and $q = 8$ traits (responses) were recorded. Traits included are percent winter survival for 1992, 1993, 1994, 1997 and 1999 (surv92, surv93, surv94, surv97, surv99, respectively), and days to flowering after no vernalization (flower0), 4 weeks vernalization (flower4) or 8 weeks vernalization (flower8). We applied the proposed methodology to study the influence of each marker on the traits and compare its predictive performance with these of its competitors. Prediction error (PE) was estimated by randomly splitting the 103 samples into a test set and a training set with sizes 33 and 70. Before adjusting the models, we first scaled the outcomes on the training and test sets to facilitate interpretability. Five-fold cross-validation was used on the training set to choose the tuning parameters. The estimated PE is given in Table 1. All methods provide similar results although the one we propose provides the smallest error for five of the eight traits. It also shows that the survival traits have a larger residual variability than the flowering traits, suggesting a higher sensitivity to environmental conditions. A picture of the estimated between-response covariance matrix is given in Figure 1. In this context, this matrix reflects the correlation between the traits that are either explained by an unexplored part of the genotype, by the environment or by some interaction between the two. The residuals of the flowering times exhibit strong correlations, whereas the correlation between the survival rates are weak.

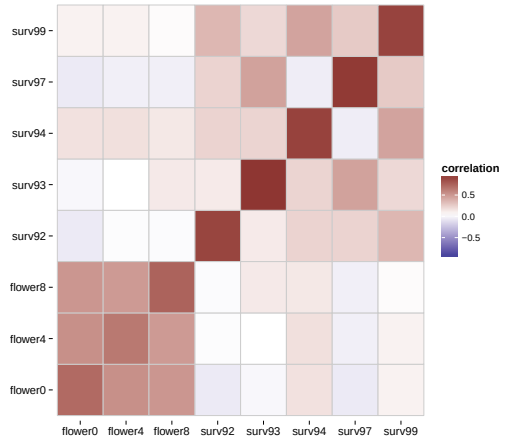

Figure 1. Brassica study: residual covariance estimation

We then turned to the effects of each marker on the different traits. The left panel of Figure 2 gives both the regression coefficients (top) and the direct effects[2] (bottom). The grey zones correspond to chromosomes 2, 8 and 10, respectively. The exact location of the markers within these chromosomes are displayed in the right panel, where the size of the dots reflects the absolute value of the regression coefficients (top) and of the direct effects (bottom). The interest of considering direct effects rather than regression coefficients appears clearly here, looking for example at chromosome 2. Three large overlapping regions are observed in the coefficient plot, for each flowering trait. A straightforward interpretation would suggest that the corresponding region controls the general flowering process. The direct effect plot allows to go deeper and shows that these three responses are actually controlled by separated sub-regions within this chromosome. The confusion in the coefficient plot only results from the strong correlations observed between the three flowering traits.

[2] we represent $-\hat{\mathbf{\Omega}}_{\mathbf{xy}}$ to facilitate the comparison with $\hat{\mathbf{B}}$.

References

de los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., and Calus, M.P.L. Whole genome regression and prediction methods applied to plant and animal breeding. *Genetics*, 2012.

Ferreira, ME, Satagopan, J., Yandell, BS, Williams, PH, and Osborn, TC. Mapping loci controlling vernalization requirement and flowering time in brassica napus. *Theor. Appl. Genet.*, 90: 727–732, 1995.

Kole, C, Thorman, CE, Karlsson, BH, Palta, JP, Gaffney, P, Yandell, BS, and Osborn, TC. Comparative mapping of loci controlling winter survival and related traits in oilseed brassica rapa and B. napus. *Mol. Breed.*, 1:201–210, 2002.

Lee, W. and Liu, Y. Simultaneous multiple response regression and inverse covariance matrix estimation via penalized gaussian maximum likelihood. *J. Multivar. Anal.*, 111:241–255, 2012.

Marin, J.-M. and Robert, Ch. P. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer-Verlag: New-York, 2007.

Schölkopf, B. and Smola, A.J. *Learning with Kernels, Support Vector Machines, Regularization, Optimization and Beyond*. The MIT Press, 2002.

Slawski, M., zu Castell, W., and Tutz, G. Feature selection guided by structural information. *Ann. Appl. Stat.*, 4:1056–1080, 2010.

Sohn, K.A and Kim, S. Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. *JMLR*, W&CP(22):1081–1089, 2012.

Yin, J. and Li, H. A sparse conditional gaussian graphical model for analysis of genetical genomics data. *Ann. Appl. Stat.*, 5: 2630–2650, 2011.

Yuan, X.-T. and Zhang, T. Partial gaussian graphical model estimation. Technical report, arXiv preprint, 2012.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B*, 67:301–320, 2005.

Method	*surv92*	*surv93*	*surv94*	*surv97*	*surv99*	*flower0*	*flower4*	*flower8*	Mean PE
LASSO	0.79	0.98	0.90	1.02	1.00	0.58	0.53	0.74	0.818
group-LASSO	0.90	1.00	0.92	**0.99**	**0.92**	0.59	0.55	0.74	0.825
E-net (no LD)	0.87	1.01	0.97	1.03	1.03	0.55	0.54	0.69	0.836
Str. E-net (with LD)	**0.75**	0.98	0.89	1.03	1.02	0.55	0.50	0.74	0.808
our proposal (with LD)	0.77	**0.96**	**0.84**	1.00	1.02	**0.48**	**0.46**	**0.68**	**0.77**

Table 1. Estimated prediction error for the *Brassica napus* data

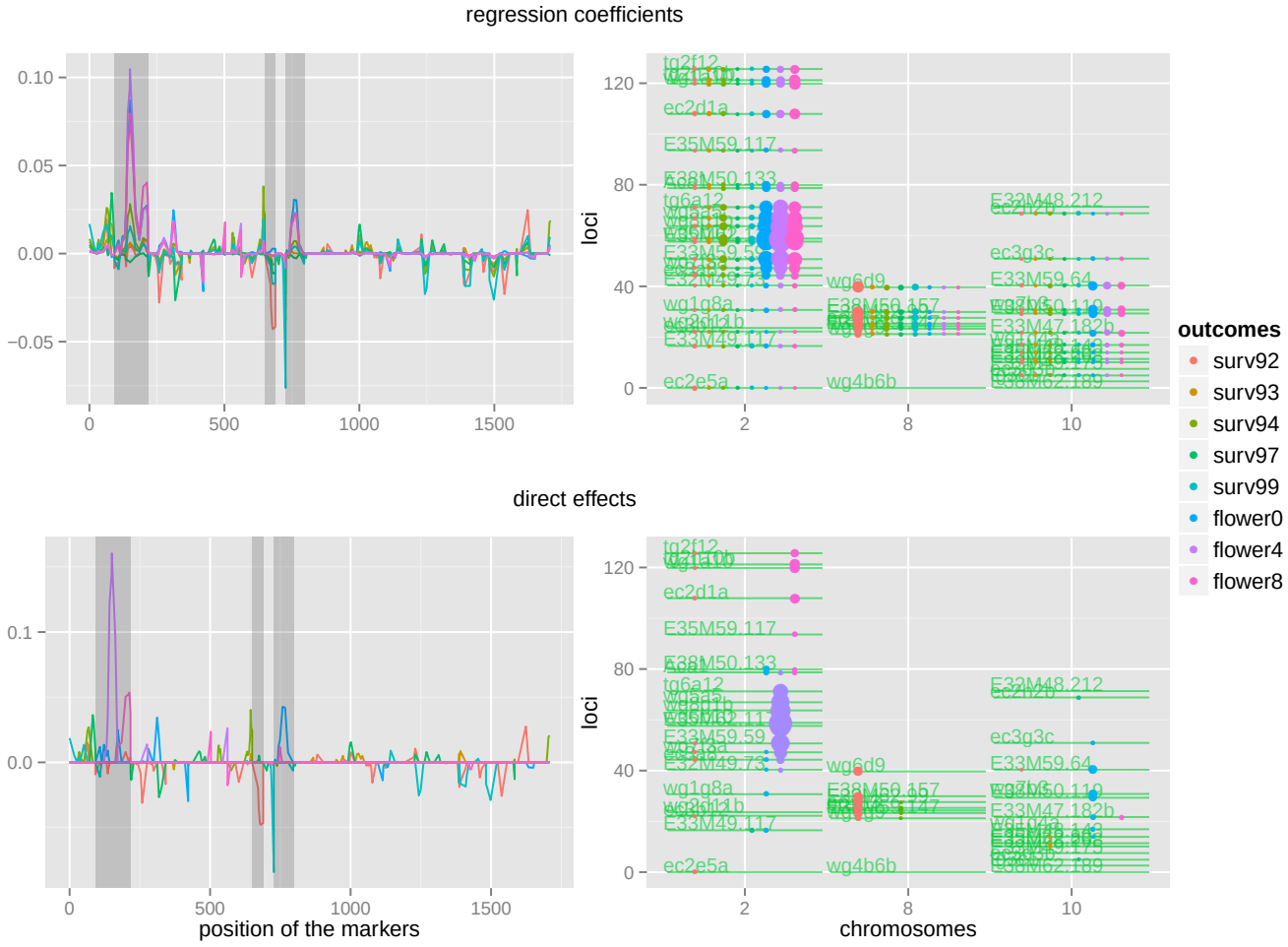

Figure 2. Brassica Study: estimation of direct and indirect genetic effects of the markers on the traits