Multi-trait genomic selection via multivariate regression with structured regularization

Julien Chiquet[1,2], Stéphane Robin[2], and Tristan Mary-Huard[2]

[1]Laboratoire Statistique et Génome – UMR CNRS 8071/Université d'Évry, France
[2]Laboratoire MMIP – UMR INRA 518/AgroParisTech – Paris, France

Background. 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

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

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.

Structured regularization with underlying sparsity. The present work proposes a general multivariate regression framework with three purposes: *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 via a set of parameters $\boldsymbol{\Omega}$ 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 an estimator to achieve these three goals using the conditional GGM formulation proposed in [1], whose conditional likelihood L is penalized by two regularization terms: the first term accounts for sparsity of the direct effects $\boldsymbol{\Omega}$ and the second one accounts for linkage disequilibrium via a structuring matrix $\mathbf{L}$. The (convex) objective function writes

$$J(\boldsymbol{\Omega}, \mathbf{R}) = -\frac{1}{n} \log L(\boldsymbol{\Omega}, \mathbf{R}) + \frac{\lambda_2}{2} \mathrm{tr}\left({}^t\boldsymbol{\Omega}\mathbf{L}\boldsymbol{\Omega}\mathbf{R}\right) + \lambda_1 \|\boldsymbol{\Omega}\|_1.$$

This work comes with an accompanying optimization procedure to minimize J.

Genomic selection in Brassica napus. We illustrate our proposal on the study conducted by [2] where $n = 103$ lines of *Brassica napus* are considered, on which $p = 300$ genetic markers and $q = 8$ traits were recorded. Traits included are five percent winter survival for 1992, 1993, 1994, 1997 and 1999 and days to flowering after 0, 4 and 8 weeks vernalization (flower0, 4 and 8). The left panel of Figure 1 gives both the regression coefficients (top) and the direct effects (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

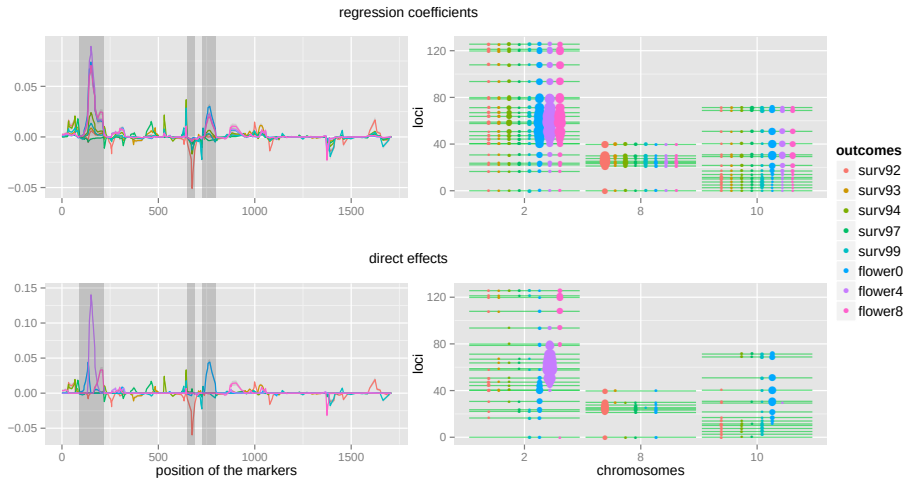

Fig. 1. Estimation of direct $\mathbf{\Omega}$ and indirect $\mathbf{B}$ genetic effects of the markers

considering direct effects rather than regression coefficients appears clearly on Figure 1, 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.

References

1. Sohn, K., Kim, S.: Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization. JMLR **W&CP**(22) (2012) 1081–1089
2. Ferreira, M., Satagopan, J., Yandell, B., Williams, P., Osborn, T.: Mapping loci controlling vernalization requirement and flowering time in brassica napus. Theor. Appl. Genet. **90** (1995) 727–732