Model Selection for GWAS

Jacob Antonsson
Dept. of Automatic Control, Lund University, Lund, Sweden
International Group for Data Analysis, Institut Pasteur, Paris 75015, France.
Etienne Patin
Unit of Human Evolutionary Genetics, Institut Pasteur, Paris 75015, France.
Centre National de la Recherche Scientifique, URA3012, Paris 75015, France.
The Milieu Interieur Consortium

Milieu Interieur

- The Milieu Interieur consortium have collected data of several different modalities with the aim of looking for environmental and genetic associations.
- Several hundred variables related to the life and environment of donors have also been collected.
- We want to use the database of environmental variables to find controls and to decrease the residual variance in GWAS models in an automatized way.

Support Estimation for a Linear Model

- Environment variables typically enter as linear predictors in the GWAS model. Therefore we want to estimate the support of a linear model

$$y = X\beta + \varepsilon,\ \varepsilon \in \mathcal{N}(0, \sigma^2 I),\ X \in \mathbb{R}^{n\times p},\ \beta \in \mathcal{R}^p. \tag{1}$$

- Denote $X_{i\in I} : I \subset \{1, \ldots, p\}$ as the matrix created by $\forall i \in I$ collecting the ith column of X. Then we are looking for a set

$$I \subset \{1, \ldots p\} : y = X_{i\in I}\beta_{i\in I} + \varepsilon.$$

- We use complementary pairs stability selection [1, 2] with an elastic net support estimator, detailed below.

Complementary Pairs Stability Selection

1. Subsample data in M pairs.
2. $\forall$ subsets B_i estimate support $\hat{S}(B_i)$ using for instance the elastic net.
3. $\forall$ variables x^k estimate the selection probability $\mathbb{P}_k = \frac{1}{2B}\sum_{j=1}^{2B} \mathbb{1}_{k\in\hat{S}(A_j)}$.
4. Compute a threshold for $\mathbb{P}_k$ that controls the false discovery rate on a chosen level.

Support Estimation for CD 4 T_{EMRA} cells

- The inferred selection probabilities for the 50 environmental variables of highest probability of being included in the support for levels of CD 4 positive T_{EMRA} cells are given in figure 1.
- The first 6 variables are either batch variables or related to the study design so they are included a priori. For this example variables related to age and infection of the cytomegalovirus were included in the support. The threshold was 0.60.

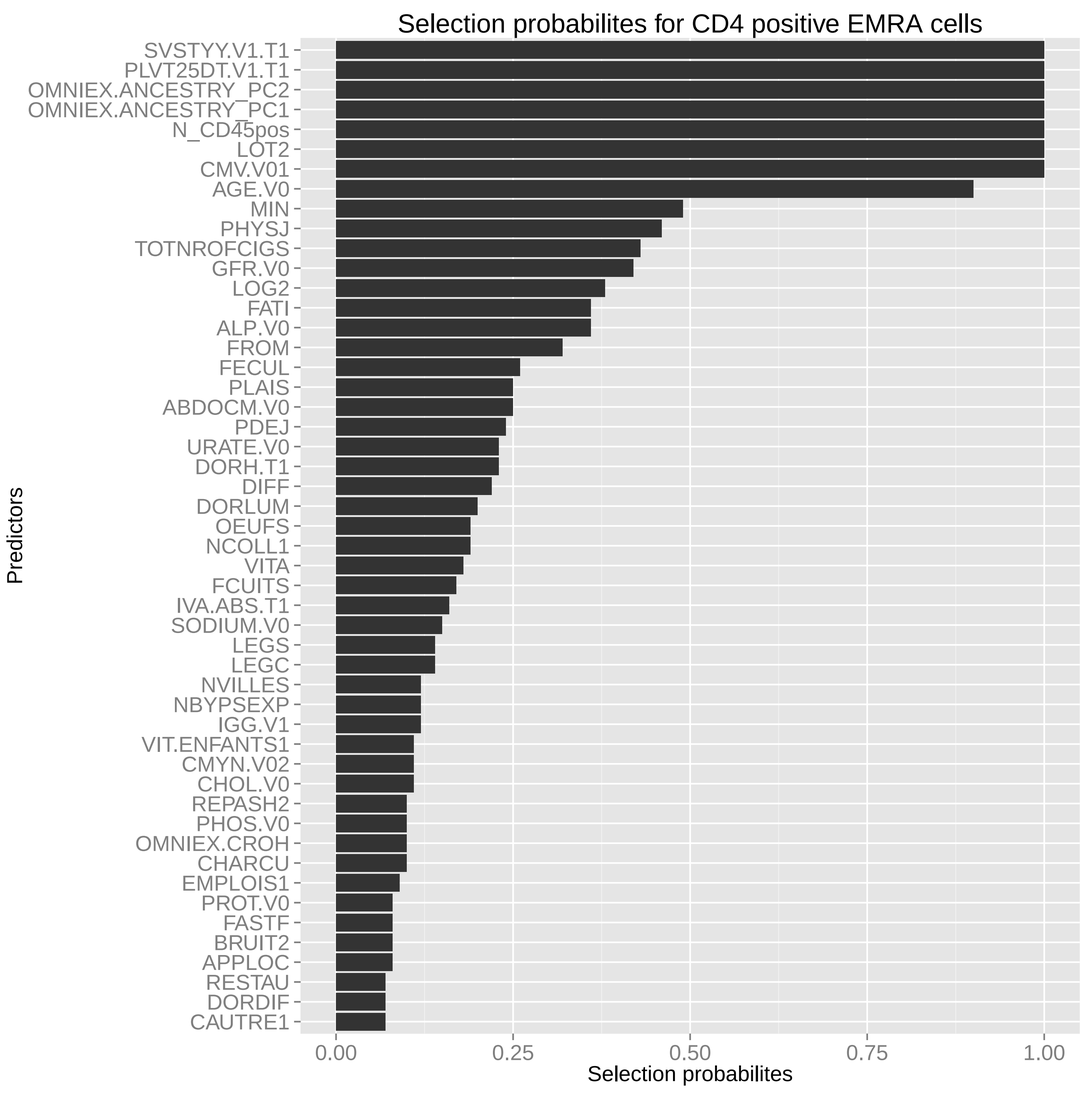

Figure 1: Selection probabilites for support estimation of levels of CD 4 EMRA T cells in the blood.

A Variable Importance Measure

- By counting the number of times a particular variable was selected across all phenotypes of a certain type we get a multivariate variable importance measure. A measure like that for 76 immunophenotypes is shown in figure 2. It is clear that age, cytomegalovirus infection and tobacco smoking is related to levels of immune cells in the blood.

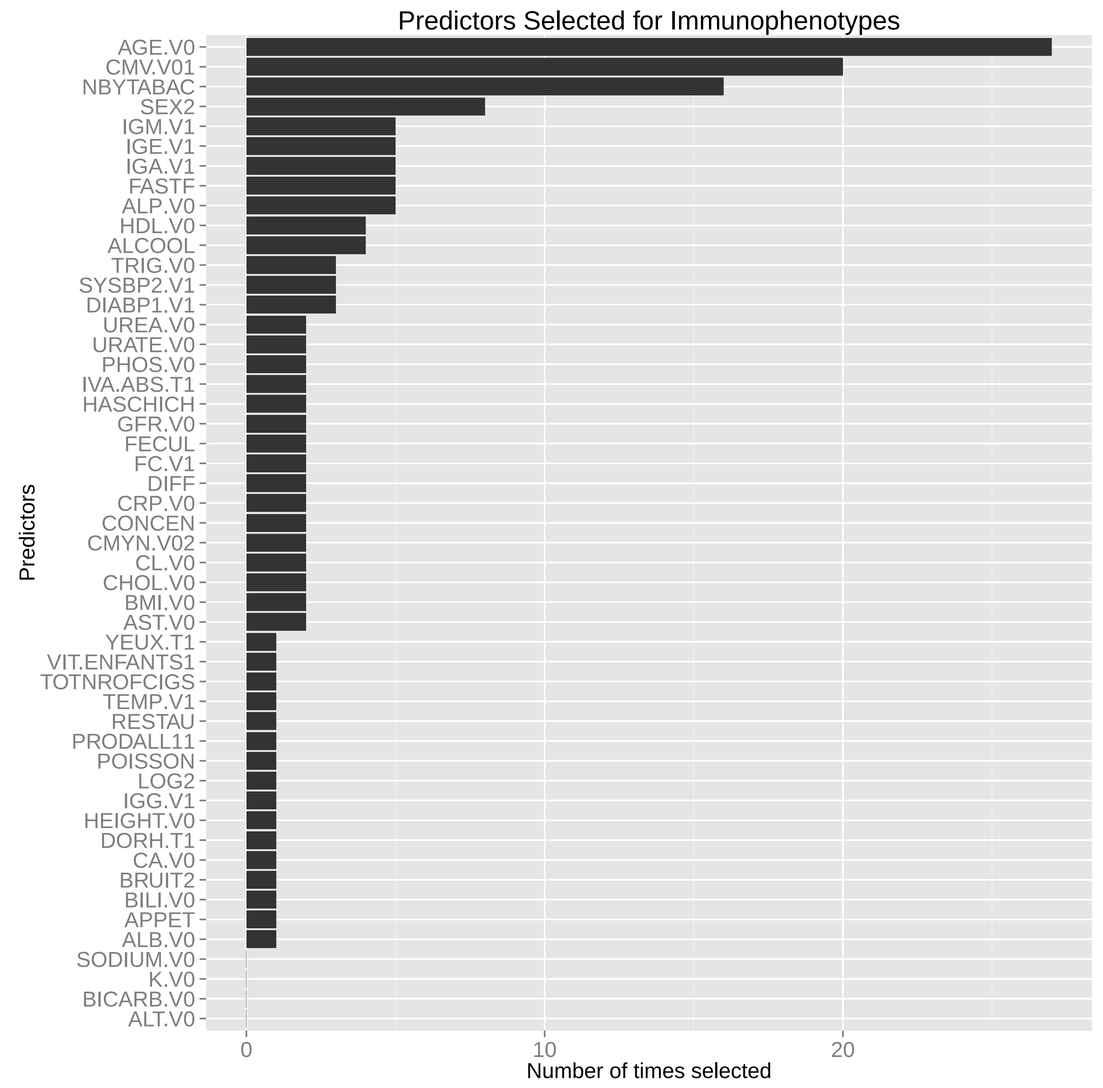

Figure 2: Variable importance measures for 76 immunophenotypes.

Impact on GWAS Results

- Significant signals of two different genome wide association studies on the same phenotypes are compared in table 1.

Phenotype	min log p without selection	min log p with selection
MFI of CD69 in CD16 hi	1.2E-21	1.5E-21
MFI of CD69 in CD8a+ CD16 hi	4.8E-17	5E-16
Count of CD69+ hi CD16	3.6E-19	7.6E-24
MFI of HLADR in CD8a+ 16 hi	1.3E-25	8.3E-33
MFI of CD16 in CD16 hi	1.9E-39	9E-45
MFI of CD16 in CD56 hi	2.8E-22	1.7E-14

Table 1: Comparison of studies with and without covariate selection

Discussion

- The stability selection scheme with the elastic net as support estimator gives stable support estimation and can be used to select for controls in GWAS.
- Gives a variable importance measure across phenotypes.
- Can directly be used to find importance measures and selection probabilites for SNPs since the elastic net works in the $p \gg N$ setting.
- The support estimation algorithm can leverage biological information by using the group lasso as support estimator.
- Several other generalizations are possible.

References

[1] Meinshausen, Bühlmann, "Stability Selection", Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2010.

[2] Shah, Samworth, "Variable selection with error control: another look at stability selection", Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2013.