Some models of genomic selection

Lev Utkin

Munich, December 2013

What is the talk about? Barley!

Steptoe x Morex barley mapping population

Steptoe x Morex barley mapping population genotyping from Close at al., 2009 and phenotyping from cite

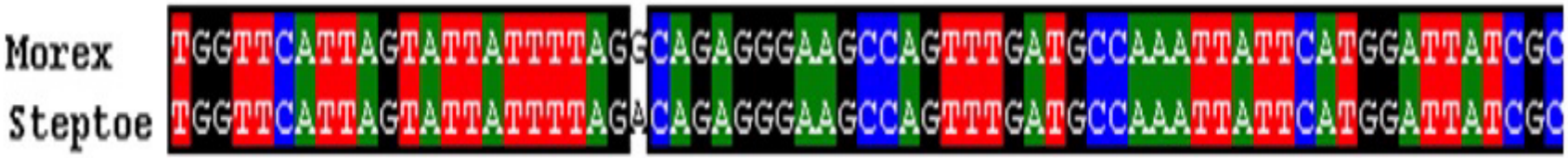
http://wheat.pw.usda.gov/ggpages/SxM/

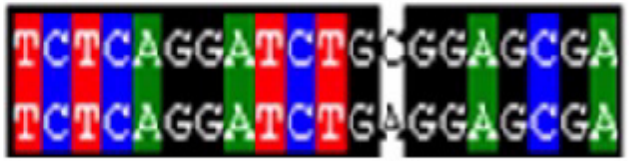
		Parents		Progeny lines									
	Data	Steptoe	Morex	SM001	SM002	SM003	SM004	SM005	SM006	SM010	SM011	SM012	SM013
The heading date early flowering of barley	*hd1_USA	149.5	150	156.5	149.5	150	157	149.5	161	163.5	161.5	155.5	155.5
	*hd2_USA	176	176.5	178	176.5	175	182	170.5	182	183.5	182	172.5	185.5
	*hd3_USA	198	200	199	197	201	199	204	204	202	202	197	204
	*hd4_USA	217	215.5	159	155	156	159	154.5	160.5	164	160	155	156
	*hd5_USA	193.5	197	200	193	197	200.5	191	203.5	203	202.5	194.5	198
	*hd6_USA	187	188	191	185	189	194	185	198	197	188	187	188
	*hd7_USA	198	198	191	186	186	193	187	191	195	191	186	189
	*hd8_USA	192	196	199	195	195	200	192	202	204	202	192	197
	*hd9_USA	186	187	191	187	187	192	180	195	198	193	187	188
	*hd10_USA	178	176	174	172.5	173	175.5	170	176.5	177.5	177	171	173
	*hd11_USA	165	163.5	167	163	162.5	167.5	162	171	172.5	172	163.5	166.5
	*hd12_USA	179	180	183	178.5	180	183.5	178	185.5	187	184	178	179
	*hd13_USA	191	189	189	189	186	189	189	191	196	193	186	189
	*hd14_USA	181	182	181	178	177	179	176	183.5	189	183	177	178
	*hd15_USA	181	183	183	179	179	184	179	188	189	189	182	183
	*hd16_USA	181	184	184	183	183	185	180	185	186	185	181	182
	*hd2012_Pus	175	175	178	175	176	182	175	188	178	182	183	183
SNPs	chromosome	Steptoe	Morex	SM001	SM002	SM003	SM004	SM005	SM006	SM010	SM011	SM012	SM013
1_0410	1H	S	M	S	M	S	M	M	S	M	M	S	M
1_0549	1H	S	M	S	M	S	M	M	S	M	M	S	M
3_0588	1H	S	M	S	M	S	M	S	S	M	M	S	M
3_0955	1H	S	M	S	M	S	M	S	S	M	M	S	M
2_0318	1H	S	M	M	M	S	M	S	S	M	M	S	M
1_0030	1H	S	M	M	M	S	M	S	S	M	M	S	M
2_0712	1H	S	M	M	S	S	S	S	S	M	M	S	M
1_0744	1H	S	M	M	S	S	S	S	S	M	M	S	M
2_0617	1H	S	M	M	S	S	S	S	S	M	M	S	M
1_0238	1H	S	M	M	S	S	S	S	S	M	M	S	M
1_1498	1H	S	M	M	S	S	S	S	S	M	M	S	M
3_0336	1H	S	M	M	S	S	S	S	S	M	M	S	M
1_0314	1H	S	M	M	S	S	S	S	S	M	M	S	M
1_0235	1H	S	M	M	S	S	S	S	S	M	M	S	M
1_0159	1H	S	M	M	S	S	S	S	S	M	M	M	M
1_0833	1H	S	M	M	S	S	S	S	S	M	M	M	M

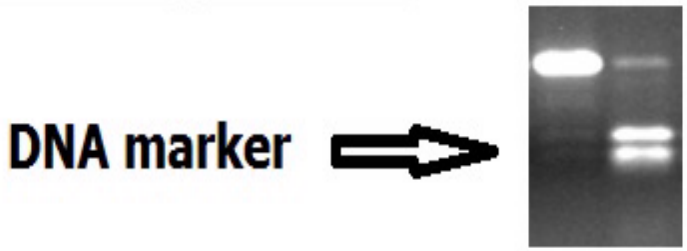
Some definition from biology

- **A phenotype** is any observable characteristic or trait of an organism: such as its morphology, development, biochemical or physiological properties, or behavior. We consider the heading date of early flowering of barley.
- The genetic contribution to the phenotype is called **the genotype** which is an individual's collection of genes. Some traits are largely determined by the genotype, while other traits are largely determined by environmental factors.
- **SNPs (Single nucleotide polymorphisms)** are the most common type of genetic variation among organisms. Each SNP represents a difference in a single DNA building block, called a nucleotide.
- **The linkage disequilibrium** is a nonrandom association of two genes on the same chromosome.

SNP explanation







SNP map

It is assumed the presence of $n = 83$ homozygous progeny, phenotypic observations in $l = 17$ environments, and a reasonably large set of mapped genetic markers $m = 395$.

		Parents		Progeny lines									
	Data	Steptoe	Morex	SM001	SM002	SM003	SM004	SM005	SM006	SM010	SM011	SM012	SM013
The heading date early flowering of barley	*hd1_USA	149.5	150	156.5	149.5	150	157	149.5	161	163.5	161.5	155.5	155.5
	*hd2_USA	176	176.5	178	176.5	175	182	170.5	182	183.5	182	172.5	185.5
	*hd3_USA	198	200	199	197	201	199	204	204	202	202	197	204
	*hd4_USA	217	215.5	159	155	156	159	154.5	160.5	164	160	155	156
	*hd5_USA	193.5	197	200	193	197	200.5	191	203.5	203	202.5	194.5	198
	*hd6_USA	187	188	191	185	189	194	185	198	197	188	187	188
	*hd7_USA	198	198	191	186	186	193	187	191	195	191	186	189
	*hd8_USA	192	196	199	195	195	200	192	202	204	202	192	197
	*hd9_USA	186	187	191	187	187	192	180	195	198	193	187	188
	*hd10_USA	178	176	174	172.5	173	175.5	170	176.5	177.5	177	171	173
	*hd11_USA	165	163.5	167	163	162.5	167.5	162	171	172.5	172	163.5	166.5
	*hd12_USA	179	180	183	178.5	180	183.5	178	185.5	187	184	178	179
	*hd13_USA	191	189	189	189	186	189	189	191	196	193	186	189
	*hd14_USA	181	182	181	178	177	179	176	183.5	189	183	177	178
	*hd15_USA	181	183	183	179	179	184	179	188	189	189	182	183
	*hd16_USA	181	184	184	183	183	185	180	185	186	185	181	182
	*hd2012_Pus	175	175	178	175	176	182	175	188	178	182	183	183
SNPs	chromosome	Steptoe	Morex	SM001	SM002	SM003	SM004	SM005	SM006	SM010	SM011	SM012	SM013
1_0410	1H	S	M	S	M	S	M	M	S	M	M	S	M
1_0549	1H	S	M	S	M	S	M	M	S	M	M	S	M
3_0588	1H	S	M	S	M	S	M	S	S	M	M	S	M
3_0955	1H	S	M	S	M	S	M	S	S	M	M	S	M
2_0318	1H	S	M	M	M	S	M	S	S	M	M	S	M
1_0030	1H	S	M	M	M	S	M	S	S	M	M	S	M
2_0712	1H	S	M	M	S	S	S	S	S	M	M	S	M
1_0744	1H	S	M	M	S	S	S	S	S	M	M	S	M
2_0617	1H	S	M	M	S	S	S	S	S	M	M	S	M
1_0238	1H	S	M	M	S	S	S	S	S	M	M	S	M
1_1498	1H	S	M	M	S	S	S	S	S	M	M	S	M
3_0336	1H	S	M	M	S	S	S	S	S	M	M	S	M
1_0314	1H	S	M	M	S	S	S	S	S	M	M	S	M
1_0235	1H	S	M	M	S	S	S	S	S	M	M	S	M
1_0159	1H	S	M	M	S	S	S	S	S	M	M	M	M
1_0833	1H	S	M	M	S	S	S	S	S	M	M	M	M

Biological problem

Biologists well know that the 57th gene or marked SNP is the Ppd-H1 gene (photoperiod response gene), that is, it is the most important.

Steptor and Morex have 2 alternative alleles of this gene.

But *biologists face the problem that this gene alone can not explain the variation that we observe. For example, parents have alternative alleles, and we are expecting one of them corresponds to early flowering and another one corresponds to late flowering. However, we have quite different dates for progeny. This implies that other genes are masked by the Ppd-H1 gene effect. How to find the SNP subset in order to take into account the heading date in any unknown barley varieties?*

The formal problem statement

1. m SNPs and n individuals
2. $Y = (y_1, ..., y_n)^{\mathrm{T}}$, $y_i \in \mathbb{R}$, is the response vector (quantitative trait phenotypes for n progeny in a given environment)
3. $\mathbf{X} = [\mathbf{x}_1^{\mathrm{T}}, ..., \mathbf{x}_n^{\mathrm{T}}]$, $\mathbf{x}_i^{\mathrm{T}} = (x_{i1}, ..., x_{im})$, $\mathbf{x}_i \in \{0, 1\}^m$, is the predictor matrix (matrix of genotypes for m background markers in n progeny)
4. The genotypes in $\mathbf{X}$ are coded as binary based on the parental origin of alleles.

Our goals

We solve two problems:

1. Selection of SNPs which impact on the difference between individuals with the heading date of early flowering and with the heading date late flowering (classification problem).
2. Selection of SNPs which impact on the heading date of early flowering (regression problem).

Filter methods

- Filter methods use statistical properties of the features to filter out poorly informative ones.
- The well-known and popular measures: t-statistics and Fisher criterion score.
- Other more or less popular measures and methods: Information Gain (IG), chi-square score, Relief-F method, Mann–Whitney–Wilcoxon U-test, mutual information, Pearson correlation coefficients, principal component analysis.
- An excellent review of filter methods is provided by Altidor et al. 2011.

Wrapper methods

- Feature selection is “wrapped” in a learning algorithm.
- One of the well-known wrapper methods is the Recursive Feature Elimination (RFE) (Guyon et al., 2001): removing a redundant feature leads to small changes of the risk measure.
- The wrapper methods are often used in combination with the filter methods.

Embedded methods

- Embedded methods perform feature selection in the process of model building.
- They mainly change the penalty terms in optimization problems:
  - l_1-norm support vector machine (Tibshirani, 1996);
  - l_0-SVM or Concave Feature Selection (FSV), based on the minimization of the "zero norm" (Bradley and Mangasarian,1998).
- Examples: Least Angle Regression, LASSO (Tibshirani, 1996) (least absolute shrinkage and selection operator).

Other interesting methods

- The Potential Support Vector Machine [Hochreiter and Obermayer, 2004]
- Feature vector machine [Li et al., 2005]

Main ideas (1)

1. We use step-wise **adding** of t "best" features (SNP) to an initially empty or non-empty feature set, and step-wise **removing of** r "worst" SNPs from the initial set (sequential bidirectional selection). (Dash and Liu 1997; Somol and Pudil 2000; Gheyas and Smith 2010). This strategy allows us to take into account some cases when the output strictly depends on a pair of SNPs, but weakly depends on every SNP from the pair separately.
2. For every SNP j, we estimate the probability p_j of M or S.
3. For every subset S of SNPs, we estimate the joint probability $p_i(S)$ of the i-th genotype and then the expected value of the phenotype.

Two important questions

1. How to select the "best" SNPs for adding? What is the objective function for the optimal selection?
2. How to select the "worst" SNPs for removing? What is the objective function for the optimal selection?

Discriminant analysis of binary data

The idea to use the joint probability mass function of genotype subsets for selecting "best" SNP stems from the Bayesian approach in classification (see, for example, the work of Lee and Jun, 2011: "Discriminant analysis of binary data following multivariate Bernoulli distribution".)
We use the joint probability mass function (or the likelihood function) of binary random variables $(x_{i1}, ..., x_{im}|\text{class } k)$ for solving the classification problem.
The classification rule reduces to

$$\text{Classify } (x_{i1}, ..., x_{im}) \text{ as class 1 if } p(\mathbf{x}|\text{class 1}) > p(\mathbf{x}|\text{class 2})$$

and as class 2, otherwise.

The objective function (1)

We select the "best" SNP and add it to the prior set S of the "best" SNPs in order to minimize the mean phenotype value. By having phenotype values, we can find mean phenotype value instead of the probability mass function:

$$R(S) = \sum_{k=1}^{n} y_k P_k(S) \to \min, \;\; P_k(S) = \frac{p_k(S)}{\sum_{j=1}^{n} p_j(S)}.$$

We would like to reduce the heading date of early flowering of barley.

The objective function (2)

In classification problems, we separate ordered values of the phenotype into two subsets with y_k^* and y_k^{**}:

$$|R^*(S) - R^{**}(S)| = \left| \sum_{k \in J^*} y_k^* P_k(S) - \sum_{k \in J^{**}} y_k^{**} P_k(S) \right| \to \max .$$

We would like to separate two samples.

The SNP selection algorithm

Require: s (number of important SNPs), T (training set), t, r.
Ensure: S (set of important SNPs)

repeat
  $l \leftarrow 1$
  repeat
    $S^* \leftarrow S \cup k$, $k_{\text{opt}} \leftarrow \arg\min_{k \in M \setminus S} R(S^*)$.
    $l \leftarrow l + 1$; $S \leftarrow S \cup k_{\text{opt}}$
  until $l > t$
  $l \leftarrow 1$
  repeat
    $S^* \leftarrow S \setminus j$, $j \in S$. $j_{\text{opt}} \leftarrow \arg\min_{j \in M \setminus S} R(S^*)$
    $l \leftarrow l + 1$; $S \leftarrow S \setminus j_{\text{opt}}$
  until $l > r$
until $card(S) > s$

An open question of using the algorithm

How to find probabilities of subsets SNPs S taking into account the **linkage disequilibrium** *that is the correlation between SNPs?*

Bahadur representation

We can use the Bahadur representation of the Bernoulli probability distribution: (Asparoukhov and Krzanowski, 2001, Lee and Jun, 2011):

$$p(S) = \left(\prod_{j=1}^{m} p_j^{x_i} q_j^{1-x_i} \right) \times \left(1 + \sum_{i<j} \rho_{ij} u_i u_j + \sum_{i<j<k} \rho_{ijk} u_i u_j u_k + \ldots + \rho_{1,2,\ldots,m} u_1 \cdots u_m \right).$$

$$U_j = \frac{X_j - p_j}{\sqrt{p_j q_j}}, \; u_j = \frac{x_j - p_j}{\sqrt{p_j q_j}}, \; \rho_{j_1 j_2 \ldots j_k} = \mathbb{E}\left[U_{j_1} U_{j_2} \ldots U_{j_k} \right].$$

Bahadur representation

The parameters can be estimated as follows:

$$\widehat{p}_j = \sum_l x_{jl}/n = 1 - \widehat{q}_j,\ \ j = 1, \ldots, m,$$

$$\widehat{\rho}_{j_1 j_2 \ldots j_k} = \sum_l \widehat{u}_{j_1 l} \cdots \widehat{u}_{j_k l}/n,$$

$$\widehat{u}_{jl} = \frac{(x_{jl} - \widehat{p}_j)}{\sqrt{\widehat{p}_j \widehat{q}_j}}$$

Bahadur representation: the pros and cons

1. It is the most precise representation for $p(S)$ taking into account the correlation between SNPs. $(+)$
2. It is computationally very hard. (-)
3. If we cut the sum and to use the so-called second-, three-, or higher-order Bahadur models, then we have a chance to get negative probabilities. (-)

Frechet inequalities

Let us use Frechet inequalities for $p(S)$:

$$\max\left\{0, \sum_{j\in S} p(x_j) - (m_S - 1)\right\} \leq p(S) \leq \min_{j\in S} p(x_j),$$

where $p(x_j)$ are marginal probabilities of the unit values for the j-th SNP.

Frechet inequalities

We can find only lower $\underline{R}(S)$ and upper $\overline{R}(S)$ bounds by solving the optimization problems:

$$\underline{R}(S) = \min_{p_k(S)} \frac{\sum_{k=1}^{n} y_k p_k(S)}{\sum_{j=1}^{n} p_j(S)}, \quad \overline{R}(S) = \max_{p_k(S)} \frac{\sum_{k=1}^{n} y_k p_k(S)}{\sum_{j=1}^{n} p_j(S)},$$

subject to Frechet inequalities for $p(S)$.

- The problems can be reduced to linear problems by using the Charnes-Cooper transformation (Charnes and Cooper, 1962).

The objective function in the imprecise case (1)

We exploit the minimax strategy for regression problems:

$$R(S) = \min_S \max_P \sum_{k=1}^{n} y_k P_k(S) = \min_S \overline{R}(S).$$

This implies that we have to find only upper bound for R.

The objective function in the imprecise case (2)

In classification problems, we use maximin strategy: we are looking for the largest difference $|R^*(S) - R^{**}(S)|$ by worst conditions, i.e., by minimizing the difference over the probabilities $p^{(i)}(S)$:

$$|R^*(S) - R^{**}(S)| = \max_S \min_P \left| \sum_{k \in J^*} y_k^* P_k(S) - \sum_{k \in J^{**}} y_k^{**} P_k(S) \right|$$

$$\max_S \left| \overline{R}^*(S) - \underline{R}^{**}(S) \right|$$

We assume here that Y is ordered.

Frechet inequalities

The idea of using the Frechet inequalities is interesting, but for many sets S the upper bound $\min_{j \in S} p(x_j)$ does not change and the lower bound is 0. So, two sets S might be undistinguished.

How to improve it?

Another idea with Frechet inequalities

We find $m(m-1)/2$ probabilities of all pairs of SNPs in S. The probability of pair (t, i) is

$$\Pr(X_t = x_t, X_i = x_i) = p(x_t, x_i) = \left(p_t^{x_t} q_t^{1-x_t} p_i^{x_i} q_i^{1-x_i}\right)(1 + \rho_{tj} u_t u_t).$$

It is simply to prove by using the dual form of the natural extension theorem

$$p(S) \geq \max\left\{0, \sum_{t=1}^{m} \sum_{k>t} p(x_t, x_k) - \frac{m(m-1)}{2} + 1\right\},$$

$$p(S) \leq \min_{t=1,\ldots,m} \min_{k>t} p(x_t, x_k).$$

In the same way, we can consider triples (quadruples, etc.) of SNPs in S.

The simplest way for computing probabilities of subsets

A strange result due to Scheinok words (Scheinok, 1972; Norusis, 1973):

> *"The superposition of Bayes's Theorem over Bahadur's Distribution led to posterior probabilities, equal to the original frequencies of occurrence of the diagnoses for each individual patient profile."*

The estimates of $p(S)$ are identical with the simple actuarial estimates:

$$p(S) = \frac{\text{the number of individuals with the genotype } S}{\text{total number of individuals } (n)}$$

The simplest way for computing probabilities of subsets (Example)

Example: S consists of $m = 2$ SNPs:
Let $T_1, ..., T_{2^m}$ be all possible genotypes consisting of m elements.
$T_1 = \{0, 0\}$, $T_2 = \{0, 1\}$, $T_3 = \{1, 0\}$, $T_4 = \{1, 1\}$.
There are 6 values of the phenotype: 3 individuals have genotype $T_2 = \{0, 1\}$, 2 individuals have genotype $T_3 = \{1, 0\}$, one has genotype $T_4 = \{1, 1\}$. Then

$$R(S) = \frac{(y_1 + y_2 + y_3)\,3/6 + (y_4 + y_5)\,2/6 + y_6/6}{9/6 + 4/6 + 1/6}.$$

The main problem: when S is large, genotypes have very small probabilities.

Walley's imprecise Dirichlet model

We change Scheinok's estimate by using the IDM as

$$\underline{p}(S) = \frac{\text{the number of individuals with the genotype } S}{n+t},$$

$$\overline{p}(S) = \frac{\text{the number of individuals with the genotype } S \ +t}{n+t},$$

here t is the hyperparameter.

Walley's imprecise Dirichlet model (Example)

Example: S consists of $m = 2$ SNPs: $T_1 = \{0, 0\}$, $T_2 = \{0, 1\}$, $T_3 = \{1, 0\}$, $T_4 = \{1, 1\}$.

$$\underline{R}(S) = \min_{p_k(S)} \frac{\sum_{k=1}^{n} y_k p_k(S)}{\sum_{j=1}^{n} p_j(S)}, \quad \overline{R}(S) = \max_{p_k(S)} \frac{\sum_{k=1}^{n} y_k p_k(S)}{\sum_{j=1}^{n} p_j(S)},$$

subject to

$$0 \le p_k(0,0) \le \frac{1+t}{6+t}, \quad \frac{3}{6+t} \le p_k(0,1) \le \frac{3+t}{6+t},$$
$$\frac{2}{6+t} \le p_k(1,0) \le \frac{2+t}{6+t}, \quad \frac{1}{6+t} \le p_k(1,1) \le \frac{1+t}{6+t}$$

Lasso penalized regression models

$$\beta = \arg\min_{\beta \in \mathbb{R}^p} \|Y - \mathbf{X}\beta\|^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| ,$$

Weights are assigned in accordance with the following sources of prior knowledge:

1. Genotyping errors such that the unreliable variants should be penalized more (Zhou et al. 2011).
2. The allele frequencies can be used (Madsen and Browning, 2009) with weights $w = 2\sqrt{\pi(1-\pi)}$ where π is the population frequency by arguing that smaller penalties are assigned to rarer variants.

Adaptive Lasso models

$$\beta = \arg\min_{\beta \in \mathbb{R}^p} \|Y - \mathbf{X}\beta\|^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| ,$$

If $w_j = 1/|\beta_{init,j}|$, where $\beta_{init,j}$ is a prior estimator of β_j, for example, the least square estimator (Zou, 2006), then the corresponding Lasso problem is referred as the adaptive Lasso.

- It has many nice properties improving the performance of the Lasso.
- It can be a basis for constructing the boosting Lasso (Buhlmann and van de Geer, 2011).

Correlations and penalty terms (Tutz and Ulbricht 2009)

The correlation based penalty is given by

$$\lambda \sum_{i=1}^{p-1} \sum_{j>i} \left\{ \frac{(\beta_i - \beta_j)^2}{1-\rho_{ij}} + \frac{(\beta_i + \beta_j)^2}{1+\rho_{ij}} \right\}.$$

ρ_{ij} is the empirical correlation between the i-th and the j-th predictors.

- If $\rho_{ij} \longrightarrow 1$, then the first term in the sum becomes dominant. When $\rho_{ij} \longrightarrow -1$, then the second term becomes dominant. Both these cases lead to the approximate equality $\beta_i \approx \beta_j$.
- If $\rho_{ij} \longrightarrow 0$, the corresponding model is reduced to the ridge regression.

Some peculiarities of the applied problem

1. Our aim is not to find the "best" regression model, but to select SNPs which impact on the smallest values of the phenotype, i.e., on the heading date of early flowering of barley.
2. The genotype values corresponding to every SNP make up a binary vector.

The ideas from the above

1. We mention ideas provided by Madsen and Browning, 2009: "the weights depend on the allele frequencies" and by Tutz and Ulbricht, 2009: "the empirical correlations between pairs of predictors impact on penalty terms".
2. The allele frequencies and correlations indirectly impact on the smallest values of the phenotype!

New ideas

1. The contribution of the k-th SNP to the mean phenotype value denoted as $\widetilde{R}_k$ can be estimated by the average contributions of pairs (the k-th and the i-th) SNPs to the mean phenotype value:

$$\widetilde{R}_k = \frac{1}{m-1} \sum_{i=1, i \neq k}^{m} R_{ki}.$$

2. Every pair of SNPs is characterized by the mean phenotype value R_{ki} corresponding to the k-th and the i-th SNPs

$$R_{ki} = \sum_{j=1}^{n} \pi(x_{jk}, x_{ji}) y_j.$$

3. The joint distribution $\pi(x_{jk}, x_{ji})$ is

$$\pi(x_k, x_i) = p_k^{x_k} q_k^{1-x_k} \cdot p_i^{x_i} q_i^{1-x_i} \cdot (1 + \rho_{ki} u_k u_i).$$

Weights

The smaller values of the average mean expected value $\widetilde{R}_k$ give us more important SNP and exert less penalty w_k:

$$w_k = \frac{\widetilde{R}_k - \min_{k=1,\dots,m} \widetilde{R}_k}{\max_{k=1,\dots,m} \widetilde{R}_k - \min_{k=1,\dots,m} \widetilde{R}_k}.$$

Advantages

The obtained weights take into account:

1. the correlation between predictors (SNPs).
2. the allele frequencies.
3. binary data.
4. the fact that the smallest values of the phenotype are more important in comparison with other values because we are looking for the SNPs which impact on the heading date of early flowering of barley.

The SNP selection algorithm

Require: $Y = (y_1, ..., y_n)^{\mathrm{T}}$, $\mathbf{X} = [X_1, ..., X_m]$
Ensure: $\beta = (\beta_1, ..., \beta_m)$
repeat
$k \leftarrow 1$
Compute $\pi(x_{jk}, x_{ji})$, $i = 1, ..., m$, $i \neq k$, for all $j = 1, ..., n$.
Compute $R_{ki} = \sum_{j=1}^{n} \pi(x_{jk}, x_{ji}) y_j$, for all $i = 1, ..., m$, $i \neq k$.
Compute $\widetilde{R}_k = \frac{1}{m-1} \sum_{i=1, i \neq k}^{m} R_{ki}$.
Normalize $\widetilde{R}_k = \frac{\widetilde{R}_k - \min_{k=1,...,m} \widetilde{R}_k}{\max_{k=1,...,m} \widetilde{R}_k - \min_{k=1,...,m} \widetilde{R}_k}$
Compute new variables $\widetilde{x}_{ik} = x_{ik} / \widetilde{R}_k$, $\widetilde{\beta}_k = \beta_k \widetilde{R}_k$.
until $k > m$
Compute $\widetilde{\beta}^{\mathrm{opt}}$ by using the standard Lasso with $\widetilde{\beta}$ and $\widetilde{\mathbf{X}}$ instead of $\widetilde{\beta}$ and $\mathbf{X}$.
Compute $\beta_k = \widetilde{\beta}_k / \widetilde{R}_k$, $k = 1, ..., m$.

Results of Lasso SNP selection algorithms

1. Modified Lasso method (SNP): 8;64;306;57;73;58;354;101
2. Wrapper method with the Bahadur representation: 64;8;57;247;254;263
3. Standard Lasso method (SNP): 57;73;58;354;56;163;74;215
4. t-statistics: 57;58;56;59;55;54;53;60
5. F-criterion: 57;58;59;56;60;53;55;54

What to do when genotypes are not binary?

Goodman and Johnson, 2005: "Multivariate dependence and the Sarmanov-Lancaster expansion"
Clear examples in the paper:

- Distributions with Gaussian marginals
- Distributions with uniform marginals
- Distributions on the integers: the Bahadur expansion is the Sarmanov-Lancaster Expansion for integer-valued random variables taking two values.

Thank you for your attention

Questions

?