ssdnm
środowiskowe
studia doktoranckie
z nauk matematycznych

Damian Brzyski

Uniwersytet Jagielloński

THE DANTZIG SELECTOR IN LOCALIZING INFLUENTIAL GENES

Praca semestralna nr 2

(semestr zimowy 2012/13)

Opiekun pracy: Krzysztof Topolski

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

THE DANTZIG SELECTOR IN LOCALIZING INFLUENTIAL GENES

DAMIAN BRZYSKI

Abstract. Finding regions of the genome that influence quantitative traits (QTLs) is very important in biological research. This issue is however very hard in general due to huge disproportion between number of DNA sequence which can impact on trait and number of individuals participating in the study. In this paper we compare results obtained for three model selection methods for genome wide association studies, mBIC, mBIC2 and EBIC (extensions of BIC method), with basing on different approach - the Dantzig selector method (DS). All simulations are conducted with datasets modeled on real-world data using in experiments. We show that DS algorithm can be successfully used under GWAS motivated design and that it is able to achieve not worse results than modiffications of commonly used BIC method.

1. Introduction

Genome wide association study (GWA study, or GWAS) often focues on identification of Single Nucleotide Polymorphisms (SNPs) and other variants in DNA which are associated with traits that are quantitative in nature (for example height or survival time). Finding such locations, called quantitative trait loci (QTLs), is issue of fundamental biological importance.

There are two main statistical problems occurring at this point. First, there are more than 180 million SNPs in humans and each reasonable statistical model taking into account all of them would be too computationally burdensome. Despite the fact that actual experiments include only SNPs which, according to researchers, are most likely to be associated with trait (usually less than million SNPs), number of variables is still a huge and thus in each proposed algorithm the complexity is one of the most important issues. Secondly, size of a sample consisting of genotypes of genetic markers which can be found in different variants (alleles) is relatively very small (i.e. there are much more variables than observations). In this context the fact that in most cases only a small number of SNPs affect the variation of given biological trait is invaluable for further analysis.

Situation presented above leads directly to the statistical issue of "best variables" selection. There exist many various approaches to this problem which has drawn increasing attention recently. Most of them assume that there is the linear dependence between observations and explanatory variables. The aim is then to choose the solution with small number of nonzero coefficients (finding a so-called "sparse solution"). A common procedure is to use a penalized likelihood criterion (with a penalty for the number of variables included in the model) like *the Akaike*

Date: January 16, 2012.

Key words and phrases. Variables selection, Linear model, GWAS, BIC, mBIC, mBIC2, EBIC, the Dantzig selector, localizing influential genes, SNPs, QTL.

Information Criterion (AIC) [3] or *the Bayesian Information Criterion* (BIC) [4]. More recently an alternative class of penalization methods represented by *the Least Absolute Shrinkage and Selection Operator* (LASSO)[1] has been popular. The motivation for LASSO came from Breiman's *non-negative garotte* [5]. The idea is to minimize the residual sum of squares subject to the absolute sum of coefficients of explanatory variables is below some tuning parameter. This means that in contrast to the AIC and BIC when using LASSO we try to minimize the L_1 norm of coefficients vector instead of number of variables taken into account. It is known that this method can discover the correct sparse representation of the model. Some extensions of LASSO have been suggested among them *The Adaptive Lasso* [7], *Relaxed Lasso* [8] and *Forward-Lasso adaptive shrinkage* [9]. The alternative approaches to variables selection problem have been proposed by Efron et al. *the Least Angle Regression* (LARS) [6] and by Zou and Hastie *the Elastic Net* [2].

In the literature it is possible to find a lot of extensions of mentioned above methods. In the context of the problems raised in this paper the modification of BIC method is most important. Broman and Speed [14] observed that in certain situations the original BIC used in GWAS has a tendency to overestimate the QTL number. This phenomenon has been also discussed for example by Siegmund [15] and Bogdan et al. [11]. The problem is connected with the fact that BIC was based on the assumption that there is a given number of available regressors and that sample size tends to infinity which is totally different from the situation taking place in actual experiments in which sample size is relatively very small. The same problem is observed for AIC. Several new modifications of BIC have been recently proposed to deal with this limitation. In this paper we cinsider three of them: mBIC [11], mBIC2 [18], [12], [16] and EBIC [18]. In [12] other possible modifications can be found.

Introduced by Bogdan et al. mBIC algorithm was derived in a Bayesian setting with the assumption that for each considered models prior probability was assigned. There was used a binomial prior constructed in a way that for large group of avaliable regresors and small number of variables suspected to be trait influencing the sparse solutions would be strongly preferred. The same approach was used in extended BIC (EBIC) (where noninformative, uniform prior was applied) while in mBIC2, one of the modiffications of mBIC presented in [12], the additional penalty for model dimension was added. There was shown that all three modiffications of BIC substantially outperform classical BIC. Comparison of these methods ([16], [13]) shows in particular that mBIC2 seems to have slightly better characteristics than EBIC ([16], [13]).

In the article we will focus on the linear model $y = X\beta + z$ with n dimensional vector y, p predictor variables $X_1, \ldots, X_p$ and normally distributed error term z. We will show results of simulations in which matrix X was generated in such way that it imitated datasets used in real-world GWAS experiments. The main purpose is to compare the results presented by P. Szulc and M. Bogdan in [13] given by three model selection methods (mBIC, mBIC2 and EBIC) with results obtained by the LASSSO related algorithm - *the Dantzig selector* [19] (DS).

DS is a new approach which minimizes the L_1 norm subject to an L_∞ constraint on the correlation of the residuals with the predictors. The method was introduced by Candes and Tao in 2007 [19] and it shed new light on highdimensional model selection. Authors used known properties of L_1 norm to find a sparse representation.

This property of L_1 norm was reported and intensively studied in [20] and by Donoho in [21]. Fascinating is the fact that considered by Candes and Tao problem:

$$\min_{\beta \in R^p} \|\beta\|_1 \text{ subject to } \|X^\star r\|_\infty \leq \lambda,$$

where r is residual vector $y - X\beta$, can be simply recast as a linear program with affine constraints and solved, usually faster than the existing methods, by using one of many available algorithms for such problems. Moreover it turned out that the Dantzig selector is very accurate - under certain assumptions about the matrix X, with large probability DS mimics the risk of the oracle estimator up to a logarithmic factor of p.

2. BIC related methods

In the paper we consider used in QTL searching problem the homoscedastic linear regression model of the following form:

$$y = X\beta + z, \tag{2.1}$$

where y is the n dimensional vector of the observations (the traits data) and X is predictor matrix (containing informations about genetic markers from each of the patients) with n rows and p columns. It means that p predictive variables $X_1, \ldots, X_p$ are initially taken into account. The vector z consists of stochastic errors of measurment and we assume that $z \sim N(0, \sigma^2 I_n)$.

We will choose models from the set $\{M_i : i \in I\}$ for $I = \{0,1\}^p$ in the sense that taking into account the M_i model for $i = (i_1, \ldots, i_p)$ rely on rejection from consideration the variables subset $\{X_{i_j} : i_j = 0\}$. This could be also expressed by saying that M_i has form:

$$y = X\beta_i + z, \tag{2.2}$$

where β_i is some p length vector with fixed zeros indexed by zeros of vector i. It should be noted that number of possible choices grows rapidly with p since number of all available models is equal to 2^p.

We will use the notation p_i for number of variables taken ino account in model M_i (number of vector's i coefficients with are equal to one). Furthermore we will say that selected model M_i is *S-sparse* if we have $p_i \leq S$ and we will generally use the term *sparse* with regard to the case when p_i is "very small" in relation to p.

In [13] author conducted simulations and compared the results for three model selection methods: mBIC, mBIC2 and EBIC. All of these methods were constructed under assumption that the number of variables which should be found by algorithm is small in respect to number of all initially considered variables making its appropriate tools in GWAS research (in the context of assumption that number of influential genes is small). We will briefly present mentioned methods after saying a few words about the *Bayesian Information Criterion*.

BIC was originally developed under the assumptions that observations come from exponential family of distributions. The idea is to choose the model with the lowest value of combination of expression containing posterior probability and some penalty for the model dimension. This value could occur in various forms since basing on this selection procedure the additive or multiplicative, greater than zero constant does not affect the model choice.

If we use the notation L_i for the maximal value of the likelihood function for model M_i than in general case in BIC method we calculate the value

$$-2\ln(L_i) + p_i \ln(n), \tag{2.3}$$

for this model.

It is worth noting that for $n \geq 8$ BIC penalizes model dimension more strongly than based on a similar principles AIC method.

If we denote by $RSS_i(\beta_i)$ the residual sum of squares from model M_i, i.e.

$$RSS_i(\beta_i) = (y - X\beta_i)^T(y - X\beta_i),$$

and assume that $z \sim N(0, \sigma^2 I_n)$ then likelihood function $L_i(y|\beta_i, \sigma)$ associated with this model can be expressed as

$$L_i(y|\beta_i, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\Big(- \frac{RSS_i(\beta_i)}{2\sigma^2}\Big).$$

Lets $\widehat{\beta_i}$ be the vector which minimizes $RSS_i(\beta_i)$ (i.e. least squares regression estimator). In the case when λ is unknown we can use the estimator

$$\widehat{\sigma^2} = \Big(\frac{RSS_i(\widehat{\beta_i})}{n}\Big)$$

and consider function $L_i(y|\beta_i) = L_i(y|\beta_i, \widehat{\sigma})$ which achieves maximum at $\widehat{\beta_i}$. Hence in this situation the BIC value under the expression 2.3 could be defined as:

$$BIC(i) = n\ln(RSS_i) + p_i \ln n,$$

where $RSS_i = RSS_i(\widehat{\beta_i})$.

BIC usually selects the true model with large probability but it was observed that this method has tendency to overestimate number of influential factors under circumstances occurring in GWAS research which is connected with the fact that BIC allows the the same prior probability to all considered models [11]. Criterion must be modified to make it useful for QTL mapping.

Introduced in [11] mBIC algorithm is one of such attempts. It developed for a wider family of models than (2.1), incuding first order interaction between variables. For such model M_i involving p_i main effects from p available and q_i interactions from m available the prior distribution $\pi(M_i)$ was imposed as

$$\pi(M_i) = \alpha^{p_i}\nu^{q_i}(1-\alpha)^{p-p_i}(1-\nu)^{m-q_i},$$

where α and ν are respectively probabilities that the ith main effect and the jth interaction term appears in the model.

Adding this prior distribution into BIC resulting in algorithm relying on choosing model with the lowest value

$$n\ln(RSS_i) + (p_i + q_i)\ln n + 2p_i \ln\left(\frac{1}{\alpha} - 1\right) + 2q_i \ln\left(\frac{1}{\nu} - 1\right).$$

For the situation in which we assume that the true model of the form (2.2) is S-sparse and we choose $\alpha = \frac{c}{p}$ where c is equal to the expected value of S we obtain

$$mBIC(i) = n\ln(RSS_i) + p_i \ln n + 2p_i \ln\left(\frac{p}{c} - 1\right).$$

The second component of the penalty part for model dimension (dependent on p) reduces the tendency of overestimation in case when p is greater than n.

If estimaing of expected valu(2.2)e of n is impossible then authors of [17] propose to take $c = 4$ in case when model (2.2) is considered or $c = 2.2$ when interactions are taken into account.

In article [17] relationship between mBIC and the well-known Bonferroni correction was also discussed. Described in [12] mBIC2 is related with another popularly used multiple testing rule, the Benjamini–Hochberg procedure [23], what is reflected in different penalty for the model dimension with respect to mBIC. Modified criterion can be expressed in the form:

$$mBIC2(i) = n\ln(RSS_i) + p_i \ln n + 2p_i \ln\left(\frac{p}{n}\right) - 2\ln(p_i!).$$

It was shown that in genome-wide association study good results could be achieved by using mBIC2 [10].

Comparing penalties for model dimension, P_{mBIC} and P_{mBIC2}, coming from mBIC and mBIC2 models, under assumption that p is much larger then c and c is sufficiently lower than n, we have

$$\frac{1}{2}(P_{mBIC} - P_{mBIC2}) = \ln\left(\frac{n}{c} \cdot \frac{p-c}{p}\right) + \ln(p_i!) > 0,$$

what means that mBIC2 will potentially select models containing more variables than mBIC. Some asymptotic optimality properties of mBIC2 were analyzed (in the case when matrix X is orthogonal) in [12] and it was shown than these properties hold for wide range of sparsity levels (much larger than for the original mBIC).

The last of the BIC modifications described in this paper, EBIC [18], was constructed by taking into account some prior distribution π on the model size (as in mBIC). This results in considering in the general case the following value calculating for model M_i instead of (2.3):

$$-2\ln(\pi(p_i)L_i(y|\theta_i)) + p_i \ln n, \tag{2.4}$$

where θ_i is vector of parameters associated with the error distribution.

There was assumed that probabilty that true model is k dimensional is proportional for some $\kappa \in [0,1]$ to $\binom{n}{k}^{\kappa}$ and hence prior probability for each model consisting of k variables is equal to

$$\pi(p_i) = \delta\binom{n}{k}^{\kappa}\binom{n}{k}^{-1} = \delta\binom{n}{k}^{\kappa-1} \tag{2.5}$$

for some $\delta > 0$. Formulas (2.4) and (2.5), under assumption of normally distributed error and up to the additive constant, lead directly to definition

$$EBIC_\kappa(i) = n\ln(RSS_i) + p_i \ln n + 2(1-\kappa)\ln\binom{n}{p_i}.$$

In [18] results obtained for different values of κ was discussed but there is no suggestion which specific parameter should be chosen. It is worth to note however that taking $\kappa = 0$ wll result in the uniform distribution on the model dimension while taking $\kappa = 1$ corresponds to the original BIC.

In the article [16] was shown that if some conditions about maximal allowable model are satisfied then all three mentioned modiffications of BIC are asymptotically equivalent and that each of them signifficantly improves original BIC in situation when we consider that true model is sparse. Results of an extensive simulation study and a real data analysis was also present in this article and it turned out

that mBIC2 and EBIC have a higher power and false discovery rate (FDR) than mBIC. Moreover simulations reported in [18] show that EBIC could be very good tool under sparse designs.

3. The Dantzig selector

Dantzig selector (DS) was introduced by E. Candes and T. Tao in [19] to deal with variables selection problem under assumption that model (2.1) is considered. DS is a solution of the optimization problem

$$\min \|\widehat{\beta}\|_1 \text{ subject to } \|X^T(y - X\widehat{\beta})\|_\infty \leq \lambda_p \cdot \sigma, \tag{3.1}$$

where λ_p is tuning parameter. The procedure suggested by authors of [19] is to set parameter λ_p on the level which guarantees that with large probability the true solution of (3.1) is feasible (this idea let also simply extend the method to situations when non-gaussian error distribution is assumed). Two following facts show why DS algorithm is potentially good in practical applications. First, under some assumptions about X matrix, L_1 norm minimization in consider situation has tendency to sparse solution recovery and secondly, problem (3.1) can be recast as a linear program

$$\begin{aligned} \min \textstyle\sum_{i=1}^p u_i \quad \text{subject to} \qquad & -u_i \leq \widehat{\beta}_i \leq u_i && , i = 1, \ldots, p \\ & -\lambda_p \sigma \leq X^T(y - X\widehat{\beta})_i \leq \lambda_p \sigma && , i = 1, \ldots, p \end{aligned}$$

what ensures that estimation procedure is computationally tractable (One can use for example a primal dual interior point algorithm [24] to deal with such optimization problems).

Following [19] we will denote by X_T the $n \times |T|$ submatrix formed by extracting the columns of X corresponding to the indices in T for some $T \subset \{1, \ldots, p\}$. Definices introduced below are used in theoretical results obtained in the cited manuscript.

Definition 1. *We say that δ_S is the S-restricted isometry constant of X if δ_S is the smallest quantity such that*

$$(1 - \delta_S)\|c\|_2^2 \leq \|X_T c\|_2^2 \leq (1 + \delta_S)\|c\|_2^2$$

for all $T \subset \{1, \ldots, p\}$ with $|T| \leq S$ and coefficient sequences $(c_j)_{j \in T}$.

Definition 2. *Let S, S' be such numbers that $S + S' \leq p$. We say that $\theta_{S,S'}$ is S, S'-restricted orthogonality constants if $\theta_{S,S'}$ is the smallest quantities such that*

$$| \langle X_T c, X_{T'} c' \rangle | \leq \theta_{S,S'} \|c\|_2 \|c'\|_2$$

for all disjoint sets $T, T' \subseteq \{1, \ldots, p\}$ such that $|T| \leq S$, $|T'| \leq S'$ and coefficient sequences $(c_j)_{j \in T}$, $(c'_j)_{j \in T'}$.

Small value of δ_S can be interpreted as property that every set of not more than S columns approximately behaves like an orthonormal system while small value of $\theta_{S,S'}$ indicates that disjoint subsets of covariates span nearly orthogonal subspaces.

All theorems presented in [19], in particular concrete values of λ_p, was derived under assumption that the columns of X have unit Euclidean norm with the proviso that these results can be extended to other cases. For this purpose one only need to consider $\widetilde{\lambda_p} = M \cdot \lambda_p$ since all the columns of X have norm less than M. We present one of results which shows impressive accuracy of the Dantzig selector.

Theorem 1. *Suppose that $\beta \in \mathbb{R}^p$ is any S-sparse vector of parameters obeying $\delta_{2S} + \theta_{S,2S} < 1$ with $\lambda_p = \sqrt{2 \ln p}$ in (3.1). Then with large probability $\widehat{\beta}$ obeys*

$$\|\widehat{\beta} - \beta\|_2^2 \leq C_1^2 \cdot (2 \ln p) \cdot S \cdot \sigma^2,$$

for $C_1 = 4/(1 - \delta_S - \theta_{S,2S})$.

The above theorem is especially interesting in the context of the following observation. Suppose that we know the each nonzero entries of parameter vector. Then we can estimate solution by using least-squares method under restriction to known support and determine the others coefficients as zero. Denoting this ideal estimator by β^* one can show that

$$E\|\beta^* - \beta\|_2^2 \geq \frac{1}{1 + \delta_S} \cdot S \cdot \sigma^2.$$

Hence it turns out that if we use the Dantzig selector to estimate β (and Theorem's 1 assumptions are satisfied) the estimation error we achieved is expected not to be much larger than in situation when the true support is a priori known (up to constant times a logarithmic factor).

Naturally we want to provide that condition

$$\delta_{2S} + \theta_{S,2S} < 1 \tag{3.2}$$

is hold for as large S as it is possible. There are some very important results useful in this field. For instance if X is a random matrix with i.i.d. Gaussian entries or matrix with i.i.d. entries taking values from the set $\{1, -1\}$ with probability $1/2$ then there is an overwhelming chance that the condition (3.2) is satisfied for $S = O(n/\ln(p/n))$ [20]. Similar results were also obtained for other matrices [20] but these issues are beyond the scope of the paper.

4. Simulation results for BIC modifications (from existing article)

In paper [13] simulations for five different scenarios which vary in number of all avaliable variables (p), number of sample (n), number of trait influencing variables (s) and the range of effect sizes (RES) for mBIC, mBIC2 and EBIC were conducted. There are:

- Scenaro 1: p=50, n=50, s=3, RES=[1.2, 2],
- Scenaro 2: p=100, n=75, s=5, RES=[1.2, 2],
- Scenaro 3: p=150, n=110, s=7, RES=[1.2, 2],
- Scenaro 4: p=200, n=150, s=9, RES=[1.2, 2],
- Scenaro 4b: p=200, n=150, s=9, RES=[0.8, 1].

For each initial setup the traits were simulated 1000 times according to the model (2.2) with $\sigma = 1$ and X matrix generated in a way that it imitates database containing genotypes of genetic markers (this procedure was described in detail in next section).

Matrix X was generated again for each step and s coefficients representing influencing variables were randomly selected. Results for all quantities reported by author are collected in Table 1. These are:

- probability of choosing the true model (T),
- average proportion of trait influencing variables detected by method (P),
- average proportion of false discoveries among all discoveries (FDR).

TABLE 1. Results obtained in [13]

	mBIC			**mBIC2**			**EBIC**		
	T	P	FDR	T	P	FDR	T	P	FDR
Sc 1	0.355	0.736	0.134	0.279	0.804	0.229	0.272	0.630	0.109
Sc 2	0.714	0.953	0.048	0.467	0.979	0.130	0.675	0.934	0.056
Sc 3	0.750	0.967	0.038	0.491	0.983	0.097	0.714	0.973	0.046
Sc 4	0.868	0.994	0.014	0.470	0.998	0.076	0.761	0.996	0.027
Sc 4b	0.245	0.762	0.081	0.335	0.915	0.107	0.299	0.781	0.081

As it has been reported in [13] these simulations show that mBIC outperforms EBIC in terms of probabilities of finding the true model, average numbers of false positives and false discovery rates while in scenarios 1–4 mBIC2 performs worse than the other criteria. However in scenario 4b the power of mBIC2 is significantly higher with respect to other modifications.

5. SIMULATION

The purpose of the performed simulations was to see how DS method works under GWAS inspirated design in comparison with other commonly used methods for which results was reported in [13]. The most important therefore was to restore settings used in simulations conducted in the cited paper. In our simulations we considered the same scenarios, method of establishing sparse vector and kind of random disturbances. We used also the same method of X matrix constructing , i.e. we generate basic p length vector of genetic markers (randomizing independently on each coordinate we choose equal probability value from the set $\{1, -1\}$) and randomize for each "patient" locations for varying alleles with respect to this basic vector (coordinates which change their sign). This procedure guarantees that individual genotypes moderately different from each other as it occurs in real-world situation.

One of the main practical problem when we apply DS algorithm is the appropriate choice of λ. The larger value of λ is the larger the feasible region for the linear program is. If λ is large enough then the vector with all coordinates equal to zero will be indicated as solutions of (3.1). Furthermore, it is known that with the growth of parameter λ the coefficients of estimated solution are pulled to zero. There are several methods to deal with this phenomenon [22], [19]. For the problems raised in the paper the issue of estimating the support of the true solution is more important then the recovering of the true effect sizes. For this reason we use the method mentioned in [19]. We estimated support I of true solution β by taking for some $\alpha > 0$ the set $\widehat{I} = \{i : |\widehat{B_i}| > \alpha \cdot \sigma\}$ where $\widehat{B_i}$ is given by DS algorithm with fixed λ. This means that we considered two tuning parameters: shrinkage parameter λ and the "cutting parameter" α. This is rather obvious that if these parameters are both "small" then sensitivity of support's elements indication increases (there is more likely that $I \subset \widehat{I}$), while taking "large" values will lead to increase in specificity which means that we have greater chance that $\widehat{I} \subset I$. These two tuning parameters provide therefore potentially a wide range of possible

TABLE 2. Results of simulations

	SETUP 1					SETUP 2				
	λ	T	P	FDR	Supp	λ	T	P	FDR	Supp
Sc 1	24	0.540	0.962	0.151	4.020	2.4	0.847	0.994	0.042	3.367
Sc 2	31	0.349	0.970	0.166	5.977	3.1	0.946	0.999	0.009	5.048
Sc 3	38	0.277	0.984	0.147	8.218	3.8	0.981	1.000	0.003	7.019
Sc 4	46	0.252	0.992	0.138	10.513	4.6	0.993	1.000	0.001	9.007
Sc 4b	46	0.242	0.945	0.135	9.920	4.6	0.800	0.976	0.000	8.788

properties of DS estimators. In table 2 we present results obtained for two various settings of λ and α. For each scenarios to propose a reasonable level of shrinkage parameter in first setup (SETUP 1) we conducted 1000 simulations in which we observed the value $\|A^T z\|_\infty$ for $z \sim N(0, I_n)$ and we established λ such that with the large probability (greater than 98%) we have $\|A^T z\|_\infty < \lambda$ (i.e. with the large probability the searched solution is feasible when normal error distribution is assumed). Besides we determined $\alpha = 0.1$ which is rather small with respect to deviation of measurement errors. To study different kind of settings we tested another setup (SETUP 2) in which we greatly reduced the shrinkage level (tenfold relative to previous value) and fixed higher α parameter equal to 0.5. We repeated all simulations 1000 times. For each scenarios results for probability of finding the true model (T), proportion of trait influencing variables detected by method (P), false discovery rate (FDR) and average size of model chosen by DS was reported.

It can be seen that results obtained by DS for considered settings are more homogeneous than presented results given by mBIC, mBIC2 and EBIC for which differences for scenarios 1 and 4b compared to others scenarios are generally significantly different. In each scenarios of SETUP 1 nearly the whole support of true solution was recovered with FDR in range (0.13–0.17). It means that in this case DS algorithm was rather sensitive and has tendency to choose greater number of variables than it was determined in true model (what is also visible in an average support size). Intuitively small increase of cutting parameter α should therefore improve performance and we actually observed such effect (results not showed) but despite checking many densely scattered values of α we were not able to achieve such a high estimation accuracy as in SETUP 2 in which the probability of finding the true model is very large.

Even after reducing effect sizes in SETUP 1 (scenario 4b) there are still far from the value of α parameter using in this variant and differences in results in relation to scenario 4 are negligible. Scenario 4b in SETUP 2 leads however to situation when cutting parameter is close enough to the nonzero coefficients of the true solution to cause that DS algorithm is too specific (for considered $\lambda = 4.6$). These simulations confirm the expected α parameter impact on results. If it is set "too hight" then the lowest nonzero coefficients of true solution can be rejected from the estimated support.

References

[1] Tibshirani R., *Regression Shrinkage and Selection via the Lasso*, Journal of the Royal Statistical Society: Series B, Vol. 58, No. 1, 267-288, 1996

[2] Zou H., Hastie T., *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society: Series B, Vol. 67, No. 2, 301320, 2005

[3] Akaike H., *A New Look at the Statistical Model Identification*, IEEE Transactions on Automatic Control, Vol. AC-19, KO. 6, 1974

[4] Schwarz G., *Estimating the dimension of a model*, The Annals of Statistics, Vol.6, No.2, 461-464, 1978

[5] Breiman L., *Better Subset Regression Using the Nonnegative Garrote*, Technometrics, Vo.37, No.4, 373-384, 1995

[6] Efron B., Hastie T., Johnston I., Tibshirani R., *Least angle regression (with discussion)*, The Annals of Statistics, Vol. 32, No. 2, 407-499, 2004

[7] , Zou H., *The Adaptive Lasso and Its Oracle Properties*, Journal of the American Statistical Association, Vol. 101, No. 476, 1418-1429, 2006

[8] Meinshausen N., *Relaxed Lasso*, Computational Statistics and Data Analysis, Vol.52, No.1, 374-393, 2007

[9] Radchenko P., James G. M., *Improved variable selection with Forward-Lasso adaptive shrinkage*, The Annals of Applied Statistics, Vol.5, No.1, 427448, 2011

[10] Frommlet F., Ruhaltinger F., Twarg P., Bogdan M., *Modified versions of Bayesian Information Criterion for genome-wide association studies*, Computational Statistics and Data Analysis, Vol.56, No.5 ,10381051, 2011

[11] Bogdan M., Ghosh J. K., Doerge R. W., *Modifying the Schwarz Bayesian Information Criterion to Locate Multiple Interacting Quantitative Trait Loci*, Genetics Society of America, Vol. 167, No. 2, 989-999, 2004

[12] Frommlet F., Chakrabarti A., Murawska M., Bogdan M., *Asymptotic Bayes optimality under sparsity for generally distributed effect sizes under the alternative*, Technical report, arXiv:1005.4753, 2011

[13] Szulc P., Bogdan M., *Localizing Influential Genes with Modified Versions of Bayesian Information Criterion*, Mathematica Applicanda, Vol. 40, No. 1, 3-14, 2012

[14] Broman K. W., Speed T. P., *A model selection approach for the identification of quantitative trait loci in experimental crosses*, Journal of the Royal Statistical Society: Series B, Vol.64, No.4, 641656, 2002

[15] Siegmund D. *Model selection in irregular problems: Applications to mapping quantitative trait loci*, Biometrika, Vol.91, No.4, 785-800, 2004

[16] , Żak-Szatkowska M., Bogdan M., textitModiffed versions of Bayesian Information Criterion for sparse Generalized Linear Models, Vol. 55, No. 11, 29082924, 2011

[17] Bogdan M., Jayanta K. G., Żak-Szatkowska M., *Selecting Explanatory Variables with the Modified Version of the Bayesian Information Criterion*, Quality and Reliability Engineering International Vol.24, No.6, 627641, 2008

[18] Chen J., Chen Z., *Extended Bayesian information criteria for model selection with large model spaces*, Biometrika, Vol. 95, No. 3, 759771, 2008

[19] Candes E., Tao T., *The Dantzig selector: statistical estimation when p is much larger than n*, The Annals of Statistics, Vol. 35, No. 6, 23132351, 2007

[20] Candes E., Tao T., *Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?*, Information Theory, Vol. 52, No. 12, 5406 - 5425, 2006

[21] Donoho, D.L., *For most large underdetermined systems of linear equations the minimal L1-norm solution is also the sparsest solution*, Communications on Pure and Applied Mathematics, Vol. 59, No. 6, 797-829, 2006

[22] James, G. M., Radchenko, P., *A generalized Dantzig selector with shrinkage tuning*, Biometrika, Vol. 96 No. 2, 323337, 2009

[23] Benjamini Y., Hochberg Y., *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society: Series B, Vol.57, No.1, 289300, 1995

[24] Boyd S., Vandenberghe L., *Convex Optimization*, Cambridge University Press, The Edinburgh Building, Cambridge, CB2 8RU, UK, 2004

E-mail address: damian.brzyski@uj.edu.pl