Comparison study in statistical estimation of gene effects based on a real data set

Xiaona Sheng1, Yuqiu Ma1,4, Jiabin Zhou2, Jingjing Zhou3

1School of Information Engineering, Harbin University, Harbin 150086, China
2School of Mathematical Sciences, Heilongjiang University, Harbin 150080, China
3Shenyang Number 88 middle school, Shenyang 110023, China
4Corresponding author’s e-mail: 41514018@qq.com

Abstract. Gene mapping study is very important task in current biostatistics and real life science. Most biological traits are mainly controlled by the genetic factors, i.e., the function of genes in the genome, and also affected by some environment factors. Reasonable statistical theory and methods should be used in the gene mapping study, so that the gene effects can be inferred precisely and reliable evidence can be further provided to practical domains of genetics and medicine. In this paper, we compare two statistical methods (conventional LSE method and the Dantzig Selector method) on estimating gene effects in the framework of linear model, and make some comparison on them. The two methods are illustrated by applying to a real data set. The analyzing results show that the Dantzig Selector method has some advantage although the two methods give similar parameter estimates.

1. Introduction

Life science has made rapid process in recent years, with the development of statistics and computer science. Now it is possible for researchers to consider the genetic mechanism of biology from more microcosmic aspect, so that more biological traits including all kinds of complex diseases. Scientists have found lots of gene loci that affect some human diseases. The main effects of these genes explain large proportion of heritability of the traits, which means that the statistical inference of main effects is key step in the gene mapping study. Of course, there is still additional heritability that needs to be detected [1, 2].

Many statistical methods were used and newly developed in the course of detecting causal genes or genetic variants. Using RFLP linkage maps, Lander and Botstein [3] considered the mapping problem of Mendelian factors underlying quantitative traits, which is a famous early work of this aspect. Haley [4] applied simple regression method to map quantitative trait in line crosses using information of flanking markers. Kwak et al. [5] mapped quantitative trait loci underlying function-valued phenotypes via a regression-based statistical method. To perform high resolution mapping of quantitative traits, Lou et al. [6] proposed the idea of unifying interval and linkage disequilibrium mapping. In additional, aiming at other kinds of traits, generalized linear mixed models can be considered in statistical analysis [7].

In this paper, we consider the estimating problem of gene effects in the framework of linear model, and use two statistical methods (conventional LSE method and the Dantzig Selector method [8, 9]) to infer the gene effects which mainly control the quantitative trait. The analysis of a real data set of
biology shows that the Dantzig Selector (DS) method has certain advantage over the LSE method on estimating gene effects.

2. Theory and methods
A quantitative trait is considered here, which is potentially controlled by multiple gene loci and one environment variable.

2.1. Statistical model
The statistical model that measures the relationship between the trait value and the genotype values is as follows

\[ t_i = a_0 + \sum_{m=1}^{M} a_m g_{im} + \sum_{j<k} b_{jk} g_{ij} g_{ik} + \varepsilon_i, \quad i = 1, 2, \ldots, n. \]  

(1)

where \( t_i \) denotes the trait value of the \( i \)th observed individual, \( g_{im} \) denotes the genotype value on \( m \)th locus of the \( i \)th observed individual, \( \varepsilon_i \) is an environment variable that follows a normal distribution with mean 0 and variance \( \sigma^2 \), \( a_m \) is the \( m \)th main effect and \( b_{jk} \) is the interaction effect between loci \( j \) and \( k \).

Since we focus on the inference of main effects in this paper, we set all \( b_{jk} = 0 \) (\( j < k \)) in model (1).

2.2. Methods
The trait values and the genotype values can be observed, therefore model (1) is a linear model. When the number of loci \( M \) is moderate and smaller than the sample size \( n \), the conventional LSE method can be directly used to estimate the effect parameters; but if \( M \) is larger than the sample size \( n \), the LSE method cannot be used, and other statistical shrinkage methods (e.g., the DS method) should be considered to deal with this problem.

The DS method has optimal \( l_2 \) rate properties under a sparsity scenario, that is to say, when the number of non-zero components of the true parameter vector is small, and it is also computationally feasible when the number of parameters is much larger than sample size. The DS method can be executed through the Matlab program given by its authors [8].

When the number of loci \( M \) is smaller than the sample size, both of these two methods can be used to estimate parameters, but different inference results may be given since the goal of the DS method is to find the best subset of variables via linear program, but the LSE method can provide the closed-form solution for the parameters. Hence reasonable explanations should be sufficiently considered in practical data analysis.

3. Real data analysis
To compare the performance of the LSE method and the DS method on estimating gene effects, we used barley data set from the North American Genome Mapping Project (Tinker et al. [10]) was used for illustration. The barley data set contained 145 individuals and 7 traits and 127 mapped markers were genotyped. In this section, we mainly analyze the trait of kernel weight.

Firstly, we standardized the trait data and fitted the distribution of the kernel weight trait (see Figures and 1 and 2). No matter the histogram of kernel weight trait or the fitted density all support that the trait follows a normal distribution. So model (1) can be used here to infer genetic parameters.
Secondly, the above-mentioned methods (DS and LSE) were used to estimate the 127 gene effects via our computer program. The total estimate values are presented in Figures 3 and 4.

| Locus | DS Method | Effect  | Locus | LSE Method | Effect |
|-------|-----------|---------|-------|------------|--------|
| 1     | 3.3386    |         | 1     | -4.8597    |        |
| 2     | 3.1898    |         | 2     | 9.1177     |        |
| 4     | 3.1879    |         | 13    | 5.5703     |        |
| 11    | 3.9179    |         | 14    | -4.8883    |        |
| 12    | 4.4646    |         | 19    | -7.2457    |        |
| 13    | 4.5101    |         | 31    | 4.6768     |        |
| 14    | 3.4721    |         | 37    | -5.2873    |        |
| 15    | 3.3683    |         | 49    | -6.3483    |        |
| 16    | 3.0140    |         | 53    | -4.7880    |        |
| 37    | -3.2050   |         | 54    | 5.5684     |        |
| 100   | 3.3736    |         | 61    | -5.4042    |        |
| 101   | 6.1979    |         | 62    | 7.0626     |        |
| 102   | 6.8254    |         | 79    | 5.9739     |        |
| 103   | 4.4160    |         | 90    | -4.6706    |        |
| 104   | 3.7554    |         | 102   | 11.2091    |        |
The results in Figures 3 and 4 for the two methods are certainly different, i.e., the results from the DS method is more centralized, and larger effects show more obviously, which presents better shrinkage property of the $l_1$ regularization problem; In contrast, the results from the LSE method is more uniform and only the largest estimated effect lies out of the group of all effects. In order to further show the difference between the two groups of estimate results, in Table 1 we listed the top 15 gene effects by methods DS and LSE, respectively. From this table, 6 main loci were common detected the two methods (see the bold type number in Table 1), and locus 102 showed the most significant main effect among all the gene loci considered.

4. Conclusion and discussion

The genotype main effects provide major contribution to the genetic variance, therefore effective statistical methods should be considered in the real data analysis. In fact, estimating the main effects is the first step in gene loci mapping, and the interaction effect analysis [11] can be done on the basis on the first step, although the main effects and interaction effects have no direct causal relationship. If the interaction effect analysis is also considered in our real data analysis part, then only the DS method can be used to do this, as the number of parameters is much larger than the sample size.

The DS method estimates parameters from the viewpoint of model selection, and it builds optimal mathematical properties under a sparsity scenario, which is significantly different from the conventional LSE method. But if the number of parameters is smaller than the sample size and the sparsity condition about parameters does not satisfy, the two methods may give similar results.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant No. 12071114) and the Natural Science Foundation of Heilongjiang Province of China (Grant No. LH2019A020).

References

[1] Manolio T, Collins F, Cox N, Goldstein D, Hindorff L, Hunter D, McCarthy M, Ramos E, Cardon L, Chakravarti A, Cho J, Guttmacher A, Kong A, Kruglyak L, Mardis E, Rotimi C, Slatkin M, Valle D, Whittemore A, Boehnke M, Clark A, Eichler E, Gibson G, Haines J, Mackay T, McCarroll S, and Visscher P 2009 Finding the missing heritability of complex diseases. Nature 461 747–53
[2] Robinson M R, Wray N R and Visscher P M 2014 Explaining additional genetic variation in complex traits. Trends in Genetics 30 124–132
[3] Lander E S and Botstein D 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121 185–99
[4] Haley C S, Knott S A 1992 A simple regression method for mapping quantitative trait in line crosses using flanking markers. Heredity 69 315–24
[5] Kwak I Y, Moore C R, Spalding E P and Broman K W 2014 A simple regression-based method to map quantitative trait loci underlying function-valued phenotypes. Genetics 197 1409–16
[6] Lou X Y, George C, Rory J T, Mark C K Y and Wu R 2005 A general statistical framework for unifying interval and linkage disequilibrium mapping: toward high resolution mapping of quantitative traits. J. Am. Stat. Assoc. 100 158–71
[7] Che X and Xu S 2012 Generalized linear mixed models for mapping multiple quantitative trait loci. Heredity 109 41–9
[8] Candes E and Tao T 2007 The Dantzig selector: statistical estimation when $p$ is much larger than $n$. Ann. Stat. 35 2313–51
[9] Ma W, Zhou Y and Zhang SL 2011 A two-step method for estimating QTL effects and positions in multi-marker analysis Genet. Res. 93 115–124
[10] Tinker N A, Mather D E, Rossnagel B G, Kasha K J and Kleinhofs A 1996 Regions of the genome that affect agronomic performance in two-row barley. Crop. Sci. 36 1053–62
[11] Lin X, Lee S, Wu M, Wang C, Chen H, Li Z and Lin X 2016 Test for rare variants by environment interactions in sequencing association studies. *Biometrics* 72 156–64