Genetic Epidemiology 21(Suppl 1): S272–S277 (2001)

Marker Selection by Akaike Information Criterion and Bayesian Information Criterion

Wentian Li and Dale R. Nyholt

Laboratory of Statistical Genetics, The Rockefeller University, New York, New York

We carried out a discriminant analysis with identity by descent (IBD) at each marker as inputs, and the sib pair type (affected-affected versus affected-unaffected) as the output. Using simple logistic regression for this discriminant analysis, we illustrate the importance of comparing models with different number of parameters. Such model comparisons are best carried out using either the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). When AIC (or BIC) stepwise variable selection was applied to the German Asthma data set, a group of markers were selected which provide the best fit to the data (assuming an additive effect). Interestingly, these 25-26 markers were not identical to those with the highest (in magnitude) single-locus lod scores.

Key words: Akaike information criterion, Bayesian information criterion, discriminant analysis, identity by descent, logistic regression, model selection

INTRODUCTION

There are two frameworks behind the analysis carried out in this contribution. The first is discriminant analysis. This can be best viewed as a process to distinguish data points with two different labels in a high-dimensional space. Here we want to distinguish affected-affected (AA) and affected-unaffected (AU) sib pairs. The purpose of a discriminant analysis is to learn how the location of the data point in this high dimensional space (i.e., the coordinates) contributes to the classification (label) of the point. In our case, a data point is a sib pair, and the coordinates of a data point are the average identity-by-descent (IBD) allele sharing at marker loci. Through a discriminant

Address reprint requests to Dr. Wentian Li, Laboratory of Statistical Genetics, The Rockefeller University, Box 192, 1230 York Avenue, New York, NY 10021.

analysis, we would like to find out how the IBD at different markers jointly contributes to the different types of sib pair, i.e., AA versus AU.

The second framework, which has rarely been applied to genetic linkage data, is the model selection process. In particular, we adopt the Akaike information criterion (AIC) and Bayesian information criterion (BIC) model selection procedures. The purpose of model selection is to identify a model that best fits the available data set, *with model complexity being corrected/penalized.* Fisher's likelihood provides a measure of the goodness-of-fit of the data (if that of the full model is subtracted), but it does not correct for the model complexity. The common practice in Fisher's likelihood framework is the likelihood ratio test: if the two models differ by ΔK degrees of freedom (*df*), the null distribution of twice the log likelihood ratio is χ^2 (df = ΔK). The main limitation of this framework is that the simpler model (null) has to be a subset of the more complicated model (alternative). Therefore, one cannot compare any two models in this framework.

This issue was addressed by Akaike and others in the 1970s, leading to the concepts of AIC [Akaike, 1974; Burnham and Anderson, 1998] and BIC [Schwarz, 1978] being proposed. AIC and BIC are defined as:

$$\text{AIC} = -2\log(\hat{L}) + 2\,\text{K} + \ldots \qquad \text{BIC} = -2\log(\hat{L}) + \log(\text{N})\,\text{K} + \ldots ,$$

where $\hat{L}$ is the maximum likelihood, K the number of parameters to be estimated in the model, and N the sample size. High-order terms are ignored here. A model is better than another model if it has a smaller AIC (or BIC) value. Both AIC and BIC have solid theoretical foundations: Kullback-Leibler distance in information theory (for AIC), and integrated likelihood in Bayesian theory (for BIC). If the complexity of the true model does not increase with the size of the data set, BIC is the preferred criterion, otherwise AIC is preferred [Burnham and Anderson, 1998].

We restrict ourselves to a subset of the model selection problem: the variable selection. When variables are combined linearly or nonlinearly, each variable has its coefficient. In this context, adding a new variable (a new dimension or coordinate in the high-dimensional space, i.e., a new marker) also adds a new parameter. Since selecting the number of parameters is a model selection problem, so is the selection of the number of variables (markers). Using too many parameters/variables can fit the data perfectly, but it can be an overfitting. Using too few parameters/variables may not fit the data set at all, thus underfitting. Variable selection by AIC/BIC will provide an answer to this problem. The idea of model selection and AIC/BIC has also been applied recently to epidemiology [Li et al., 2000], microarray data analysis [Li and Yang, 2001], and DNA sequence analysis [Li, 2001].

METHODS

Equation. A label of binary values is assigned to each sib pair: $y = 0$ for AA and $y = 1$ for AU. The mean (average) IBD allele sharing at each marker (i) is x_i = Prob(IBD = 1) + 2 × Prob(IBD = 2) ($i = 1, 2, \ldots, p$; where p is the number of markers). For simplicity, we use logistic regression (LR) for the discriminant analysis:

$$\text{Prob}(y = 1) = 1/(1 + \exp(-a_0 - \Sigma_i\, a_i\, x_i)) .$$

The likelihood, $L = \prod P(y = 1)^y\,(1-P(y = 1))^{1-y}$, is the product of these probabilities for all data points. Specifically, we would like to find a LR with p' variables ($p' < p$) that

coefficient with the lod score obtained from the ALLEGRO program (multipoint, using the nonparametric exponential score-pairs function), the two results were mostly consistent (result not shown). We represent top-ranked markers by symbol T, whereas those top-ranked markers with positive ALLEGRO lod are represented by symbol PA (e.g., T25, PA25). Another obvious limitation on the initial pool of markers for stepwise variable selection is that the number of markers/parameters has to be smaller than the number of parameters (334). Clearly, it is impossible to carry out a LR with all markers. Actually, the number of makers included should not exceed 34 for this data set.

Table I shows the result of several stepwise variable selection starting from PA100, T50, PA50, using either AIC or BIC. It lists -2log maximum likelihood, AIC, BIC (absolute and relative to the constant prediction null), and the within-sample prediction rate. Perfect prediction of the sib pair types (152 out of 152) has been achieved with either 25 or 26 markers. Notice that BIC and AIC may lead to a different final set of markers. Also, as expected, different initial pools of markers may lead to different final sets of markers. Figures 1 and 2 show the location of markers selected from T50 and PA100 (AIC and BIC lead to similar results), as well as the lod score obtained from ALLEGRO (Figures 1 and 2) and that from single-variable LR (the null–1 model).

DISCUSSION

Model selection is not hypothesis testing. It does not draw conclusions as to whether a null model is wrong; instead, it explores and ranks various alternative models. For

TABLE I. AIC and BIC of Variable Selection Results

Model	K	−2log(L)	AIC	ΔAIC	BIC	ΔBIC	P_{within}
PA100.AIC.25[a]	26	0.68[d]	52.68	−130.40	131.30	−54.81	152/152
T50.AIC.26[b]	27	0.99[d]	54.99	−128.09	136.64	−49.47	152/152
T50.BIC.25[c]	26	3.91[d]	55.91	−127.17	134.53	−51.58	152/152
PA50.AIC.46	47	6.53[d]	100.53	−82.55	242.65	+56.54	152/152
T9	10	134.70	154.70	−28.39	184.94	−1.17	118/152
T25	26	104.62	156.62	−26.46	235.24	+49.14	129/152
$m_{121}+m_7$	3	161.48	167.48	−15.60	176.56	−9.55	112/152
$m_{121}+m_{178}$	3	162.18	168.18	−14.90	177.26	−8.85	109/152
$m_{121} * m_7$	2	164.64	168.64	−14.44	174.69	−11.42	109/152
$m_{121}+m_7+ m_{121}\ m_7$	4	161.46	169.46	−13.62	181.55	−4.56	112/152
m_{121} (best)	2	169.90	173.90	−9.18	179.95	−6.16	109/152
m_7 (2^{nd} best)	2	172.94	176.94	−6.14	182.99	−3.12	109/152
m_{293} (worst)	2	181.08	185.08	+2	191.13	+5.02	109/152
null-1	1	181.08	183.08	0	186.11	0	109/152
null-0	0	210.72	210.72	+27.64	210.72	+24.61	76/152

[a]PA100.AIC.25 is the set of markers (31, 36, 37, 45, 59, 86, 89, 91, 98, 121, 123, 125, 146, 149, 173, 175, 178, 192, 213, 222, 232, 291, 317, 320, 325) that are obtained by stepwise variable selection (max iteration = 10) from the top 100 markers with the positive ALLEGRO lod scores using AIC (BIC leads to the same result).

[b]T50.AIC.26 is the set of markers (7, 19, 31, 60, 72, 97, 119, 121, 123, 124, 138, 164, 165, 178, 192, 201, 213, 222, 223, 229, 237, 240, 275, 285, 291, 321) that are obtained by stepwise variable selection (max iteration = 10) from the top 50 markers using AIC.

[c]T50.BIC.25 differs from T50.AIC.26 by removing marker 19.

[d]Due to perfect fitting, these minus twice the log-likelihood values will gradually approach 0 as the max iteration allowed is increased.

complex human diseases, one should not expect that statistical methods alone will provide the final answer to the location of susceptibility genes. Exploratory approaches such as model selection offer a better framework where other information, such as biological knowledge, may be more flexibly included.

The conclusion that there are perhaps 20 or so markers that best explain the German Asthma data set may not stand true for other data sets, or may be altered if any one of the following conditions are changed: (i) if the independent variable is not the mean IBD (using mean IBD implies that the maternal and paternal IBD is treated equally and additively); (ii) if the mean IBDs from different markers are not combined additively (additivity is equivalent to a linear hyperplane used in the discrimination, instead of a nonlinear surface such as those used in artificial neural networks [Li et al., 1999]); and (iii) if genetic heterogeneity is assumed (tree-based discriminations, and conditional approaches that separate main and minor effects, may provide a solution to this issue). Clearly, for such a small data set, a best model can be obtained by chance. Stepwise variable selection also does not guarantee an exhaustive search of all possibilities.

Our stepwise variable selection starts with a set of markers that are the best single-variable predictors. Due to the limitation of sample size (152), we cannot start a stepwise variable selection from more than 151 markers. This may exclude some markers that themselves are weak predictors, but may nevertheless be more important in an interaction term or in a conditional context. This may perhaps explain the fact that some markers considered to be involved in interactions [Colilla et al., 2001] are not present in our selected markers. Furthermore, even for the selected marker sets using different criteria, T25, T50.AIC.26, and PA100.AIC.25, the number of overlapping markers was disappointedly small: markers 31, 121, 178, 213, and 291. Consistently selected markers should be those that have major affect, and will be selected no matter what approach is used.

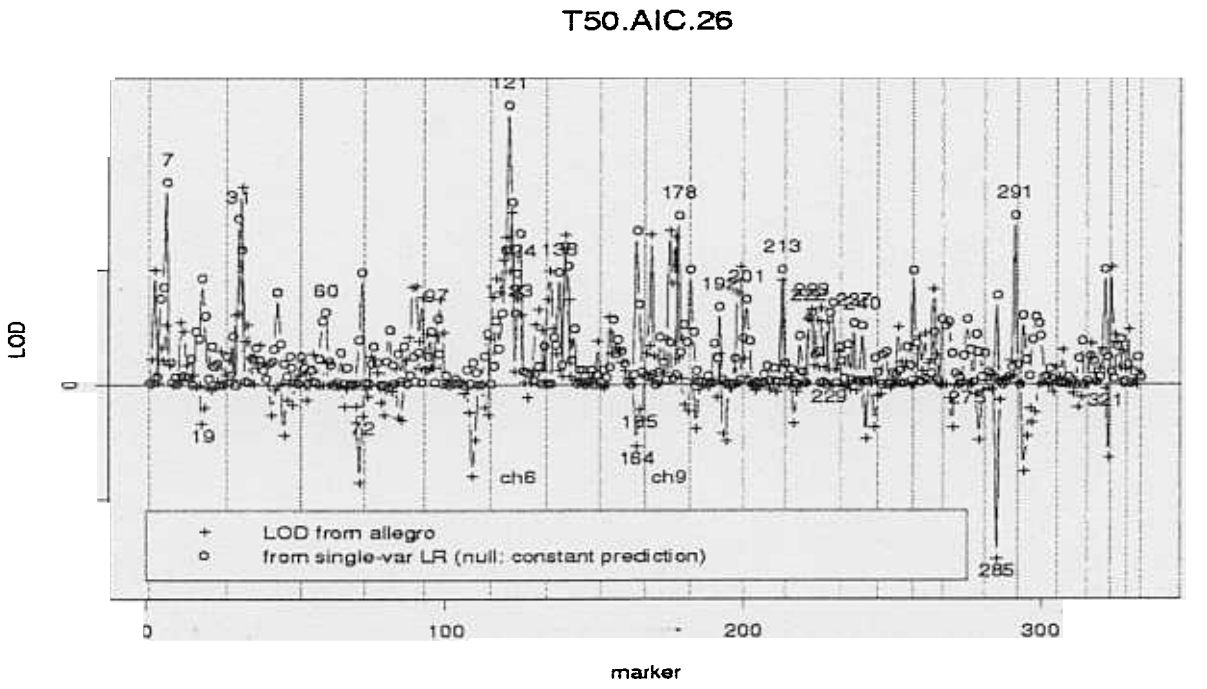

Fig. 1. T50.AIC.26 is the set of 26 markers that are selected in a stepwise fashion from the top 50 markers with the best single-variable LR likelihoods (max iteration = 10).

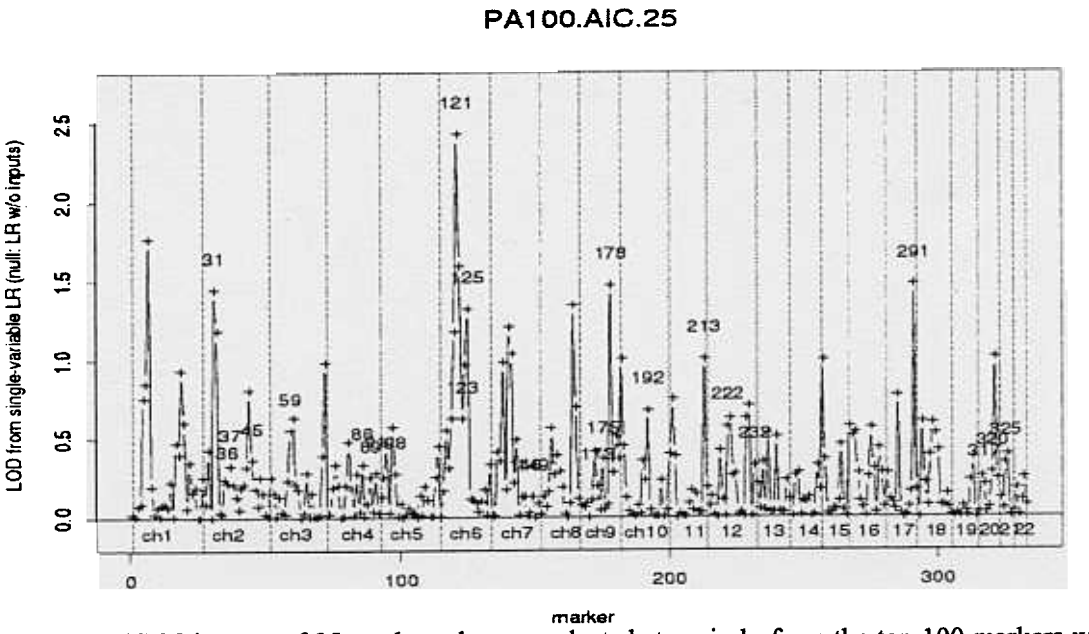

Fig. 2. PA100.AIC.25 is a set of 25 markers that are selected stepwisely from the top 100 markers with the best single-variable LR likelihoods plus a positive ALLEGRO lod score (max iteration = 10).

Topics not discussed here but worth further investigations include: model uncertainty and model averaging (e.g., via bootstrapping); cross-validation using test data sets; reversal of input and output (with IBD = 1,0 as the output and sib pair type as dummy inputs [Rice et al., 1999]); and logistic regression without the AU pairs (as related to LR of matched case-control pairs where the controls are simulated by the null [Langefeld and Boehnke, 1999]).

ACKNOWLEDGMENTS

This work is supported by NIH grants K01HG00024 and MH44292.

REFERENCES

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans Automatic Control 19:716-3.

Burnham KP, Anderson DR. 1998. Model Selection and Inference. Berlin: Springer.

Colilla et al. 2001. Genome-wide approaches for identifying interacting susceptibility regions for asthma. Genet Epidemiol. 20: this volume.

Langefeld CD, Boehnke M. 1999. Multiple trait nonparametric linkage regression [abstract]. Am J Hum Genet 65(Suppl.):A258.

Li W, Haghighi F, Falk CT. 1999. Design of artificial neural networks and its applications to the analysis of alcoholism data. Genet Epidemiol 17(Suppl.):S223-8.

Li W, Sherriff A, Liu X. 2000. Assessing risk factors of human complex diseases by Akaike and Bayesian information criteria (abstract). Am J Hum Genet 67(Suppl.):S222.

Li W. 2001. New stopping criteria for segmenting DNA sequences. Phys Rev Lett 85(25):5815-8.

Rice JP, Rochberg N, Neuman RJ, et al. 1999. Covariates in linkage analysis. Genet Epidemiol 17(Suppl. 1):S691-5.

Schwarz G. 1978. Estimating the dimension of a model. Ann Stat 6:461-4.

Wjst M, Fischer G, Immervoll T, et al. 1999. A genome-wide search for linkage to asthma. Genomics 58:1-8.