14. Exploring Interactions in Genomic Data Using Logic Regression

Ingo Ruczinski, [1] **Charles Kooperberg,** [2] **Michael LeBlanc,** [3]

Keywords: adaptive model selection, boolean logic, genomic data, interactions, simulated annealing, SNPs.

1 Abstract

Exploring higher order interactions is a statistical challenge frequently occurring in the analysis of genomic data. For example, interactions may be present between single nucleotide polymorphisms or chromosomal deletions when considering associations with disease or disease stages, respectively. Logic regression is an adaptive regression methodology that can be used to address the problem of detecting such interactions, by constructing predictors as Boolean combinations of binary covariates. The logic regression software is freely available with the manual at `http://bear.fhcrc.org/~ingor/logic/`.

2 Introduction

Interactions play a key role in the analysis of genomic data. For example, the occurrences of diseases such as cancer are often related to the interaction of multiple SNPs rather than to single variation sites. This creates novel challenges in the statistical analysis how SNPs relate to disease, since the number of possible interactions between nucleotides (and therefore the respective model search space) is immense. Statistical techniques such as neural networks (Cheng and Titterington [2]) have been developed to deal with high-dimensional search spaces. However, the interpretation of the results and especially the rules, which are of key interest, is almost impossible. Adaptive spline and tree-based methods such as MARS (Friedman [3]) and CART (Breiman et. al. [1]) generate rules that are much easier to interpret. MARS however is efficient on data that has interactions in at most a few variables, and CART for example only generates rules in disjunctive normal form (DNF). There are many simple Boolean expressions that have a rather complicated DNF, and hence CART would likely not discover those rules. Both CART and MARS are greedy type algorithms, and therefore do not necessarily find the globally best rules. Further, it is often necessary to implement and use very specific objective functions (for example to account for dependencies in the data), which is not an option in CART and MARS. Logic regression is an adaptive regression methodology that addresses the above described problem by constructing predictors as Boolean combinations of binary covariates, using simulated annealing.

[1]Department of Biostatistics, Johns Hopkins University, 615 N Wolfe St, Baltimore, MD 21205, USA. E-mail: `ingo@jhu.edu`

[2]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave, MP 1002, Seattle, WA 98109, USA. E-mail: `clk@fhcrc.org`

[3]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave, MP 557, Seattle, WA 98109, USA. E-mail: `mikel@crab.org`

3 Methodology

Logic regression was introduced by Ruczinski, Kooperberg and LeBlanc [5] to address problems arising when data of mostly binary covariates are analyzed, and the interactions between those predictors is of main interest. Given a set of binary predictors X (such as indicator variables in SNP data whether a variation at a particular site is present), we try to create new, better predictors for the response by considering combinations of those binary predictors. For example, if the response is binary as well (which is not required in general), we attempt to find decision rules such as "if X_1, X_2, X_3 and X_4 are true", or "X_5 or X_6 but not X_7 are true", then "the response is more likely to be in class 0". In other words, we try to find Boolean statements involving the binary predictors that enhance the prediction for the response. In more specific terms: Let $X_1, ..., X_k$ be binary predictors, and let Y be a response variable. We try to fit regression models of the form $g(E[Y]) = b_0 + b_1 L_1 + ... + b_n L_n$, where L_j is a Boolean expression of the predictors X, such as $L_j = [(X2 \wedge X4^c) \vee X7]$. The above framework includes many forms of regression, such as linear regression ($g(E[Y]) = E[Y]$) and logistic regression ($g(E[Y]) = \log(E[Y]/(1 - E[Y]))$). For every model type, we define a score function that reflects the "quality" of the model under consideration (for example, the residual sum of squares for linear regression and for the deviance logistic regression). We try to find the Boolean expressions in the regression model that minimize the scoring function associated with this model type, estimating the parameters b_j simultaneously with the Boolean terms L_j. In general, any type of model can be considered, as long as a scoring function can be defined. For example, we also implemented the Cox proportional hazards model, using the partial likelihood as the score. A detailed introduction to logic regression is available in Ruczinski et. al. [5]; for an application to SNP data see Kooperberg et. al. [4].

4 Software

The logic regression software is currently a stand-alone program written in Fortran 90 that can be downloaded at http://bear.fhcrc.org/~ingor/logic/. A beta version as an R/Splus package is also available. The results generated by the executable can be directly fed into standard statistical software packages to generate graphical representations of the output. A feature of the logic regression methodology is that it is easy to include and use ones own scoring function if that is desired. Online help for all these issues can be found by clicking through the menu of the logic regression webpage (see also Ruczinski et. al. [6]).

References

[1] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C.J. 1984. *Classification and Regression Trees.* Belmont, CA: Wadsworth.

[2] Cheng, B. and Titterington, D.M. 1994. Neural Networks: A Review from a Statistical Perspective. *Statistical Science*, 9:2-75 (with discussion).

[3] Friedman, J. H. 1991. Multivariate Adaptive Regression Splines (with discussion). *Annals of Statistics*, 19:1-141.

[4] Kooperberg, C., Ruczinski, I., LeBlanc, M., and Hsu, L. 2001. Sequence Analysis using Logic Regression. *Genetic Epidemiology* 21 (S1), 626-631.

[5] Ruczinski, I., Kooperberg, C., and LeBlanc, M. 2002. Logic Regression. *Journal of Computational and Graphical Statistics*, to appear.

[6] Ruczinski, I., Kooperberg, C., and LeBlanc, M. 2002. Logic Regression - Methods and Software. In: *Proceedings of the MSRI workshop on Nonlinear Estimation and Classification (Eds: D. Denison, C. Holmes, M. Hansen, B. Mallick, B. Yu),* Springer. pp. 333-344.