Finding Interactions and Assessing Variable Importance in SNP Association Studies

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins University

Email: ingo@jhu.edu

The slides and software used for this presentation are available at http://biostat.jhsph.edu/∼iruczins

An Example

[With Kathy Helzlsouer and Han-Yao Huang]

The odyssey cohort study consists of 8,394 participants who donated blood samples in 1974 and 1989 in Washington County, Maryland. The cohort has been followed until 2001, and environmental factors such as smoking and dietary intake are available. The goals of the study include finding associations between polymorphisms in candidate genes and disease (including cancer and cardiovascular disease). Particularly, gene-environment and gene-gene interactions associated with disease are of interest. Currently, SNP data from 51 sites are available for some 1600 subjects.

Motivation

[Lucek and Ott]

"Current methods for analyzing complex traits include analyzing and localizing disease loci one at a time. However, complex traits can be caused by the interaction of many loci, each with varying effect."

"... patterns of interactions between several loci, for example, disease phenotype caused by locus A and locus B, or A but not B, or A and (B or C), clearly make identification of the involved loci more difficult. While the simultaneous analysis of every single two-way pair of markers can be feasible, it becomes overwhelmingly computationally burdensome to analyze all 3-way, 4-way to N-way 'and' patterns, 'or' patterns, and combinations of loci."

Logic Regression

[With Charles Kooperberg and Michael LeBlanc]

$X_1, \ldots, X_k$ are 0/1 (False/True) predictors.

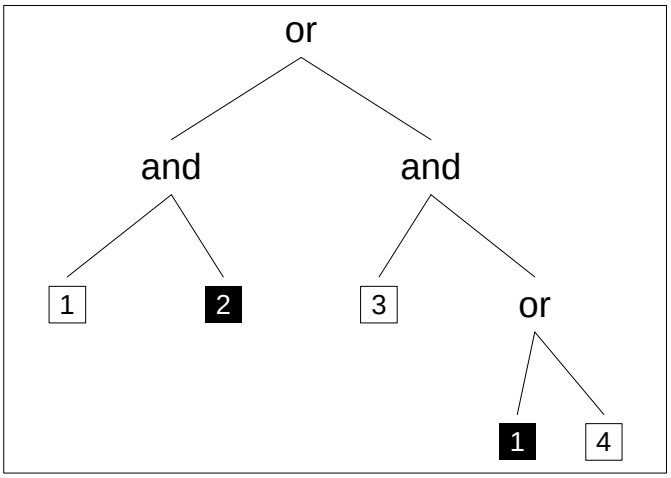
Y is a response variable.

Fit a model $g(E(Y)) = b_0 + \sum_{j=1}^{t} b_j \cdot L_j$, where L_j is a Boolean combination of the covariates, e.g. $L_j = (X_1 \vee X_2) \wedge X_4^c$.

Determine the logic terms L_j and estimate the b_j simultaneously.

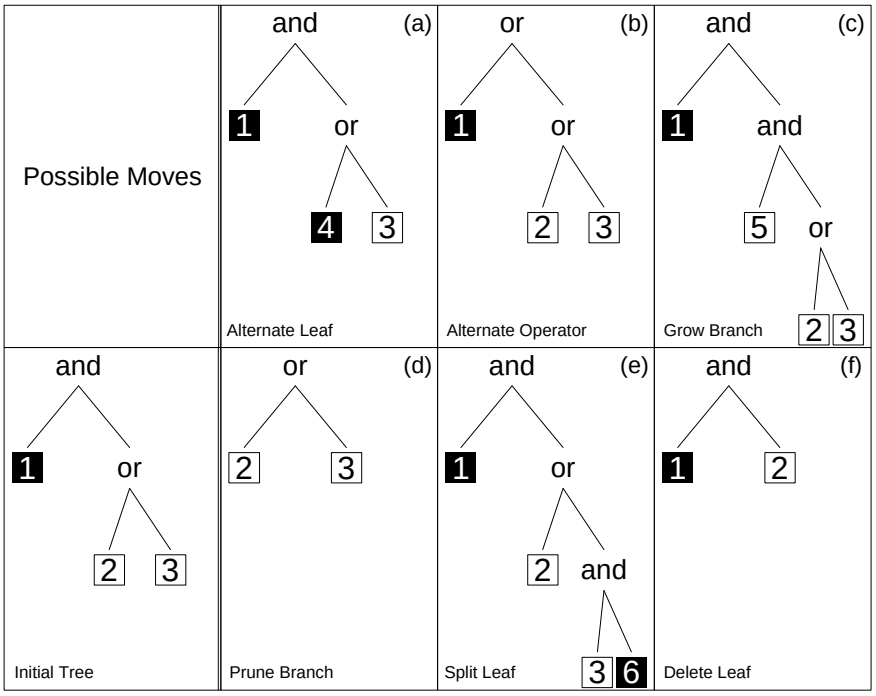
Logic Trees

An equivalent representation of $(X_1 \wedge X_2^c) \vee (X_3 \wedge (X_1^c \vee X_4))$ is the following:



This is a Logic Tree!

The Move Set

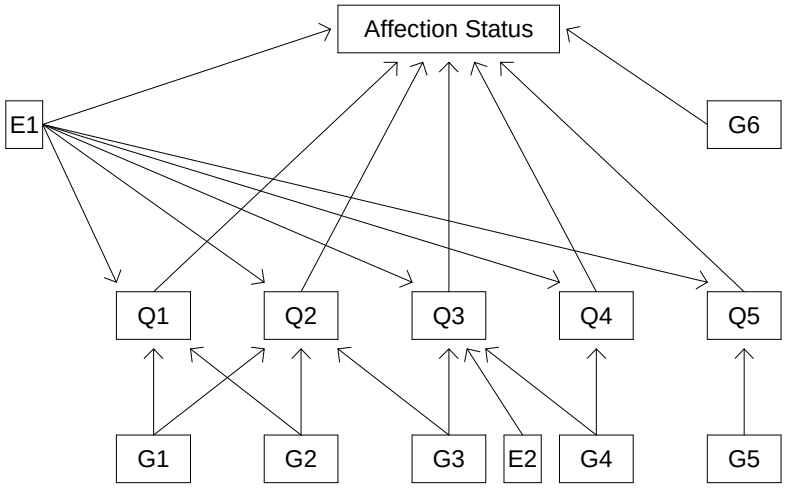


Simulated Annealing for Logic Regression

We try to fit the model $g(E(Y)) = \beta_0 + \sum_{j=1}^{t} \beta_j \times L_j$.

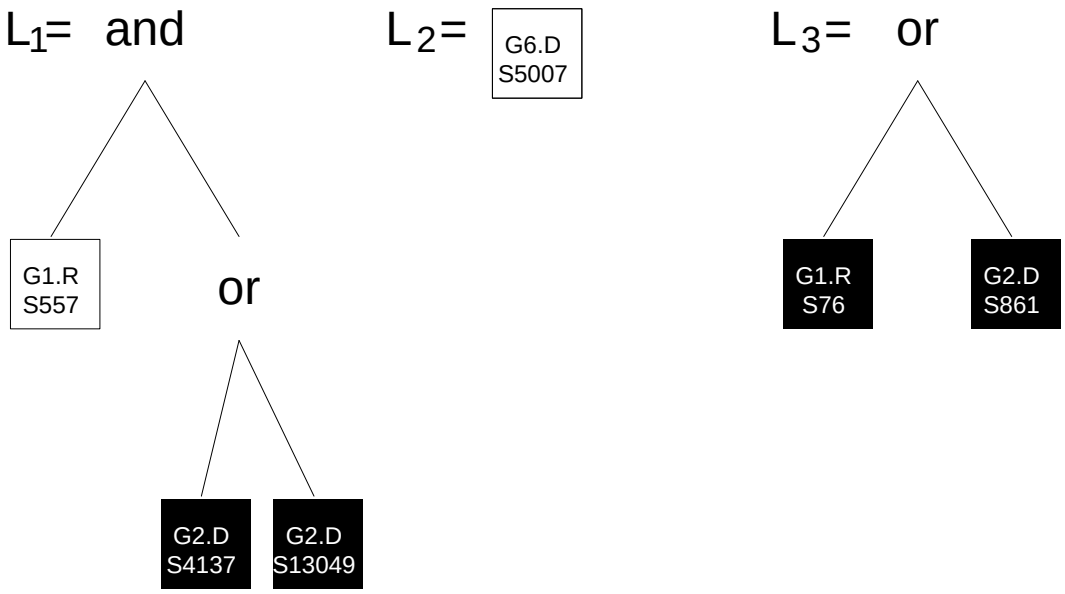
- Select a scoring function (RSS, log-likelihood, . . .).
- Pick the maximum number of Logic Trees.
- Pick the maximum number of leaves in a tree.
- Initialize the model with $L_j = 0$ for all j.
- Carry out the Simulated Annealing Algorithm:
  - Propose a move.
  - Accept or reject the move, depending on the scores and the temperature.

Genetic Analysis Workshop GAW 12

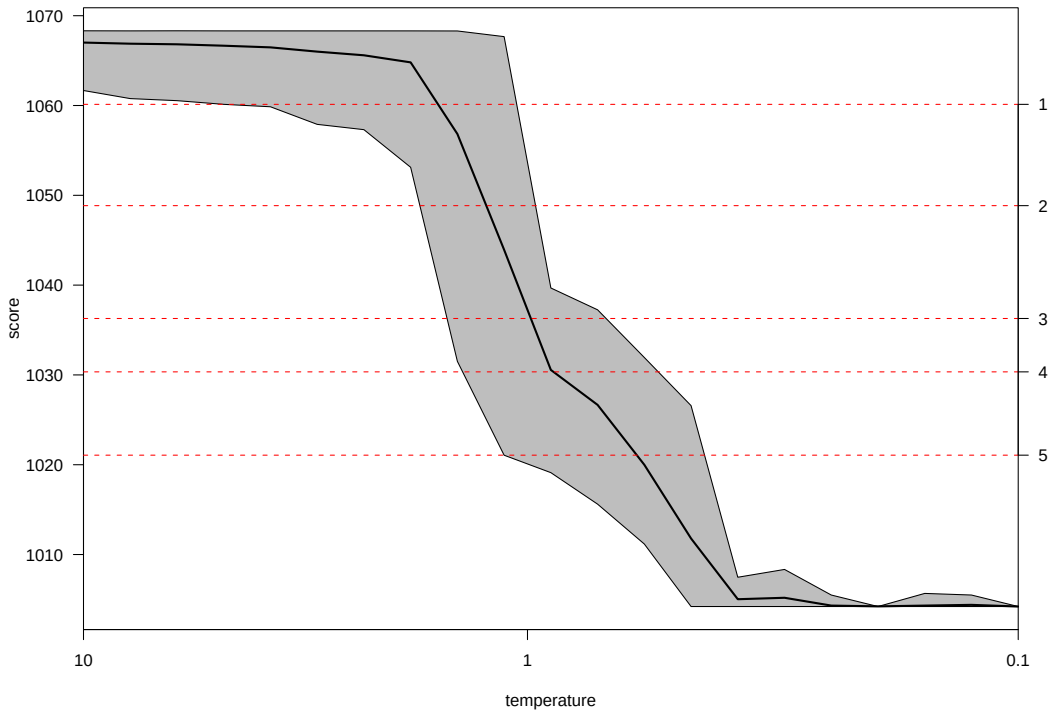


Genetic Analysis Workshop GAW 12

$$\text{logit(affected)} = \beta_0 + \beta_1 \times \text{ENV}_1 + \beta_2 \times \text{ENV}_2 + \beta_3 \times \text{GENDER} + \sum_{i=1}^{K} \beta_{i+3} \times L_i$$



Multiple Models



Multiple Models

Let γ_s be the score of a certain state S.

- We use the acceptance function

$$\alpha(\gamma_{\text{old}}, \gamma_{\text{new}}, t) = \min\{1, \exp([\gamma_{\text{old}} - \gamma_{\text{new}}]/t)\}$$

- If we keep the temperature constant, this defines a homogeneous Markov chain.

- We constructed the move set to be irreducible and aperiodic, therefore each homogeneous Markov chain has a limiting distribution $\pi_t(S)$.

Multiple Models

Simulate 10 binary predictors $X_1, \ldots, X_{10}$.

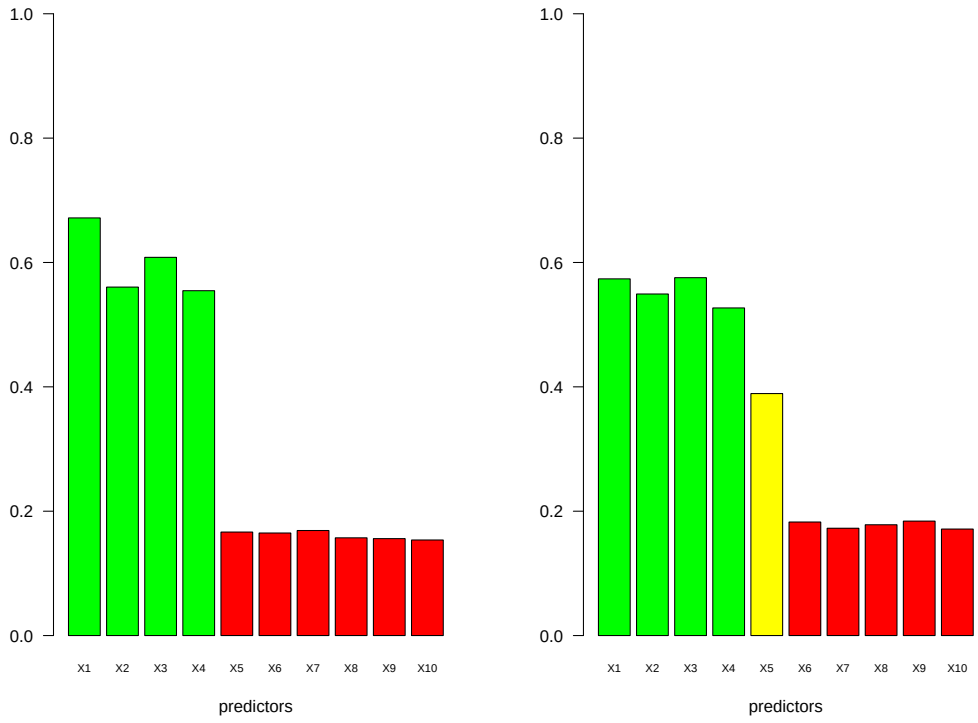
Let $Y = 5 + 1 \times L(X_1, X_2, X_3, X_4) + \epsilon, \quad \epsilon \sim \text{N}(0,1)$.

Run a homogeneous Markov chain during "crunch time" for two separate cases:

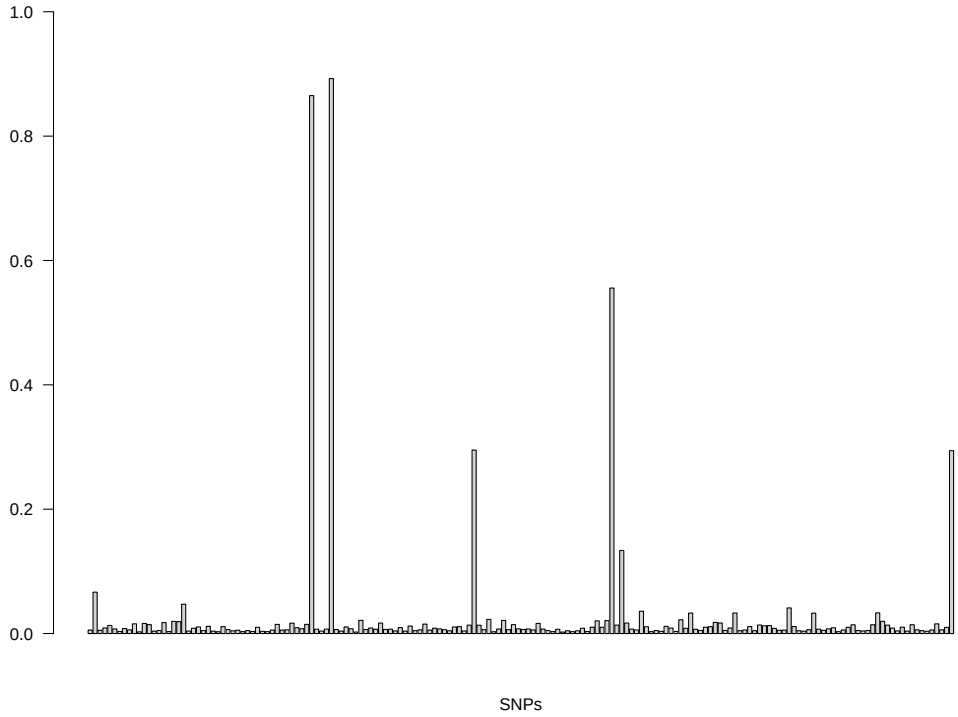
Case 1: All X are independent.

Case 2: All X are independent, except X_4 (in the signal) and X_5 (not in the signal), which are heavily correlated.

Multiple Models

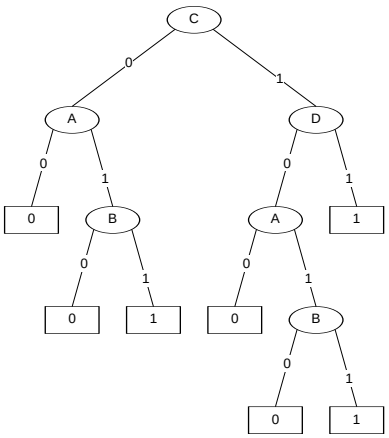


Multiple Models

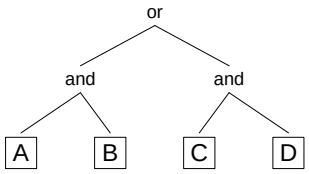


Comparison to Decision Trees

Decision Tree

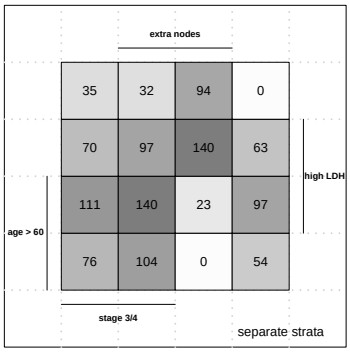


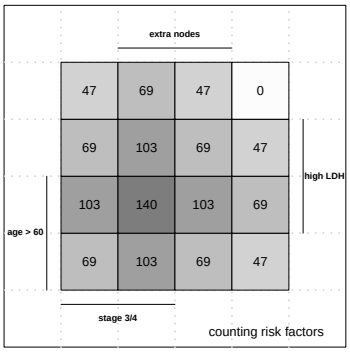
Logic Tree

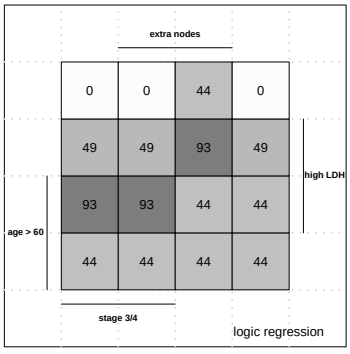


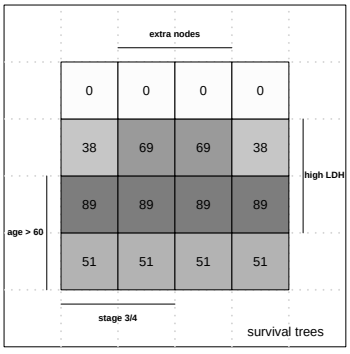
A Decision Tree (CART) is something different!

Another Useful Way to Look at it. . .









References

- Kooperberg, C., Ruczinski, I., LeBlanc, M., and Hsu, L. (2001).
  Sequence Analysis using Logic Regression.
  Genetic Epidemiology, 21 (S1), 626-631.
- Ruczinski, I., Kooperberg, C., and LeBlanc, M. (2003).
  Logic Regression.
  Journal of Computational and Graphical Statistics, 12 (3), 475-511.
- Ruczinski, I., Kooperberg, C., and LeBlanc, M. (2004).
  Exploring Interactions in High Dimensional Genomic Data: An Overview of Logic Regression, With Applications.
  Journal of Multivariate Analysis, in press.

Software and manuscripts available at: http://biostat.jhsph.edu/∼iruczins/