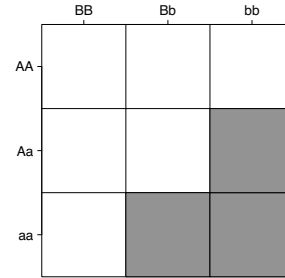


Biological and Statistical Interactions

Logic Regression as a Tool to Assess Interactions in SNP Association Studies

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health



$$(SNP_{Aa} \wedge SNP_{BB^c}) \vee (SNP_{bb} \wedge SNP_{AA^c})$$

Logic Regression

- X_1, \dots, 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. $L_j = (X_1 \vee X_2) \wedge X_3^c$.

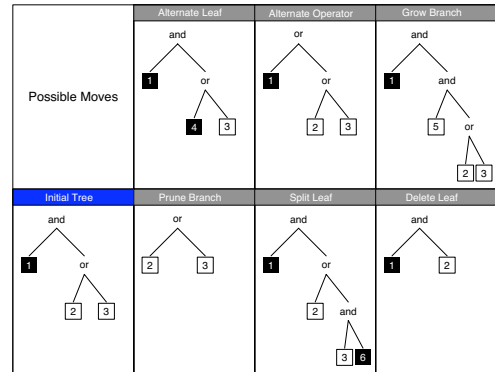
- Determine the logic terms L_j and estimate the b_j simultaneously.

SNP X	X.R	X.D
AA	0	0
AT	0	1
TT	1	1

- SNPs are coded as dominant and recessive:

Reference: Ruczinski I, Kooperberg C, LeBlanc M (2003). Logic Regression. *Journal of Computational and Graphical Statistics*, 12(3): 475-511.

The Move Set for Logic Regression

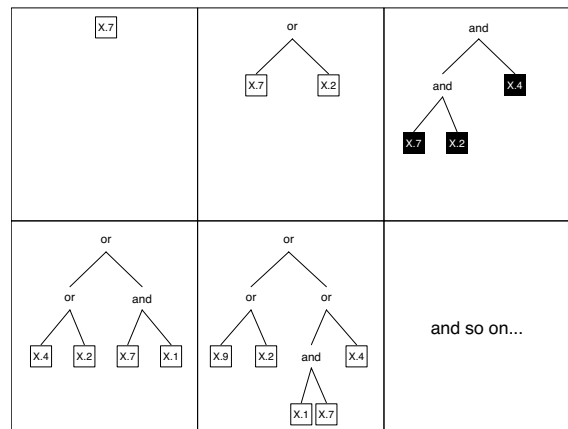


Simulated Annealing for Logic Regression

We try to fit the model $g(E(Y)) = b_0 + \sum_{j=1}^t b_j \cdot L_j$.

- 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.

Growing Logic Models

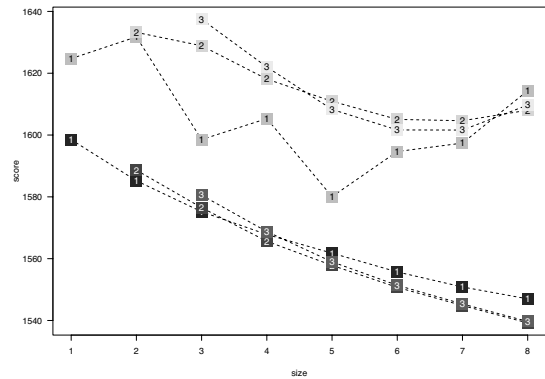


Model Selection

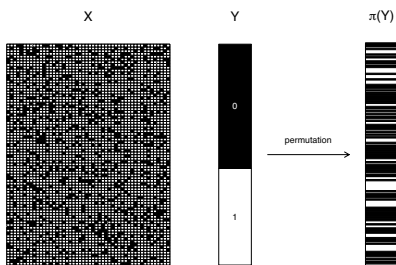
We implemented two flavors for the required model selection. Both approaches require a definition of [model size](#).

- **Cross-validation:**
This is most applicable when prediction is the main objective, i. e. not in SNP association studies.
- **Permutation tests:**
This is a test for association, i. e. the preferred test in SNP association studies. The model size is chosen via a sequence of hypothesis tests.

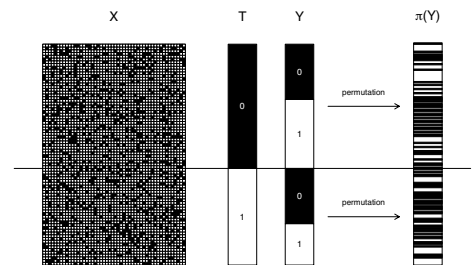
Model Selection 1 : Cross Validation



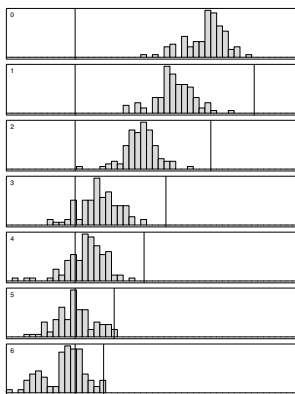
Model Selection 2 : Permutation Tests



Model Selection 2 : Permutation Tests



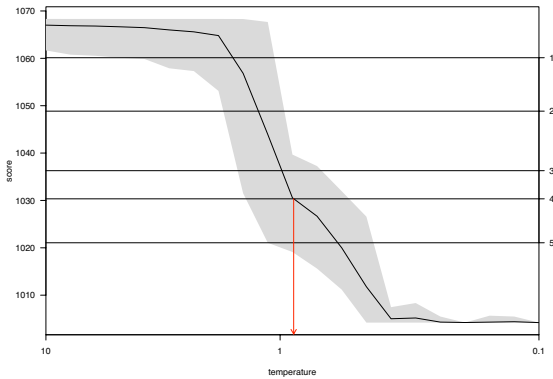
Model Selection 2 : Permutation Tests



Multiple Models 1 : Monte Carlo LR

- Goal: identify all models and combinations of covariates that are potentially associated with the outcome.
- Use reversible jumps to implement an MCMC algorithm with priors on models and model size.
- The prior on model size does influence the total number of SNPs selected.
- The prior on model size has virtually no influence on the relative ordering of the SNPs or combinations thereof.

Multiple Models 2 : Metropolis-Hastings



Multiple Models 2 : Metropolis-Hastings

Let γ_S be the score of a certain state S .

- We use the acceptance function

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

- 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 $\pi_t(S)$.
- If we know the model size where the signal ends and the noise starts, we can read off the corresponding temperature from the diagnostic plot!

Multiple Models 2 : Metropolis-Hastings

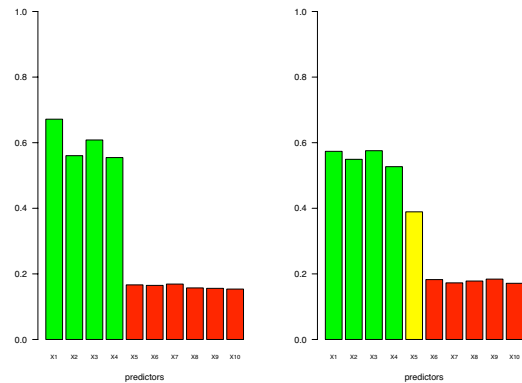
Example: Simulate 10 binary predictors X_1, \dots, X_{10} .

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

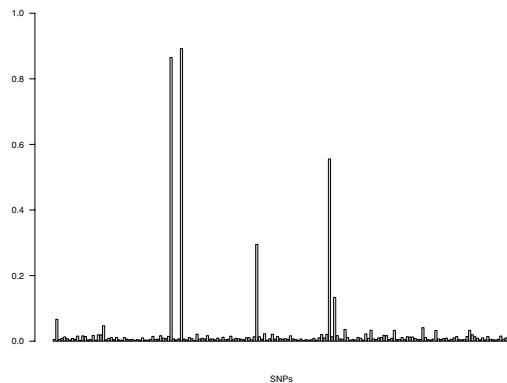
Run a homogeneous Markov chain during “crunch time” for two separate cases:

- 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 2 : Metropolis-Hastings



Multiple Models 2 : Metropolis-Hastings



<http://biostat.jhsph.edu/~iruczins>