

Logic Regression and Interactions in High Dimensional Genomic Data

Ingo Ruczinski

Department of Biostatistics
Johns Hopkins University

Email: ingo@jhu.edu

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

With Charles Kooperberg and Michael LeBlanc, FHCRC

Motivation

[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 genes are available.

Table 1. Amino Acid Substitution Variants Identified in DNA Repair and Repair-Related Genes (Source: Mohrenweiser et al 2002, and Goode et al 2002)

Gene Name	Exon	Codon	Common Residue	Variant Residue	Allele Frequency
Base Excision Repair					
ADPRT	17	761	Val	Ala	0.18
APE1	5	148	Asp	Glu	0.33
OGG1	7	326	Ser	Cys	0.15-0.45
OGG1		Nucleotide 7143*	A	G	0.15
OGG1		Nucleotide 11657*	A	G	0.15
POLD1	1	19	Arg	His	0.12
POLD1	3	119	Arg	His	0.15
POLD1	4	173	Ser	Asn	0.05
XRCC1	6	194	Arg	Trp	0.13
XRCC1	10	399	Arg	Gln	0.24
Nucleotide Excision Repair					
ERCC2	10	312	Asp	Asn	0.4
ERCC2	23	751	Lys	Gln	0.32
ERCC4	8	415	Arg	Gln	0.06
ERCC5	15	1104	Asp	His	0.18
RAD23B	7	249	Ala	Val	0.10
XPC	8	499	Ala	Val	0.24
XPC	15	939	Lys	Gln	0.38
Double Strand Break/Recombination Repair					
NBS1	5	185	Gln	Glu	0.34
XRCC2	3	188	Arg	His	0.05
XRCC3	7	241	Thr	Met	0.43
XRCC4	5	247	Ala	Ser	0.08
Damage Recognition, Repair and Cell Cycle Check point					
CDKN2A	2	148	Ala	Thr	0.05
RAD52	8	287	Ser	Asn	0.05
MLH1	8	219	Ile	Val	0.12
Mismatch Repair					
MSH3	10	514	Glu	Lys	0.05
MSH3	21	940	Arg	Gln	0.1
MSH3	23	1036	Thr	Ala	0.3
MSH6	1	39	Gly	Glu	0.24

* Amino acid substitution variants of these SNPs have not been published. However, these nucleotide substitutions occur in a gene of particular interest (see section B.2.c.), and have been found to associate strongly with risk of prostate cancer. (Goode et al 2002)

Motivation

Lucek and Ott (1997):

“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

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

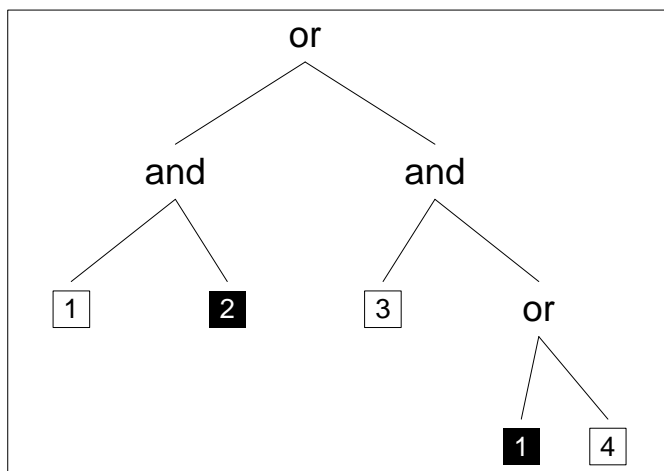
Y is a response variable.

Fit a model $g(E(Y)) = \beta_0 + \sum_{j=1}^t \beta_j \times L_j$, 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 β_j simultaneously.

Logic Trees

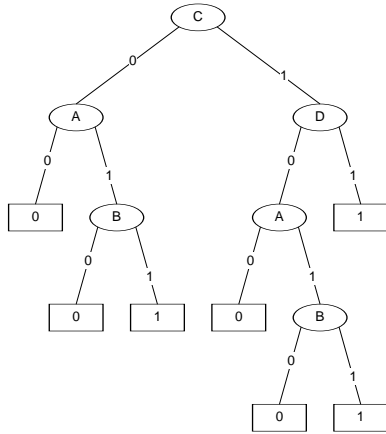
An equivalent representation of $(X_1 \wedge X_2^c) \vee (X_3 \wedge (X_1^c \vee X_4))$ is the following:



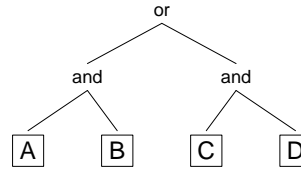
This is a Logic Tree!

Comparison to Decision Trees

Decision Tree

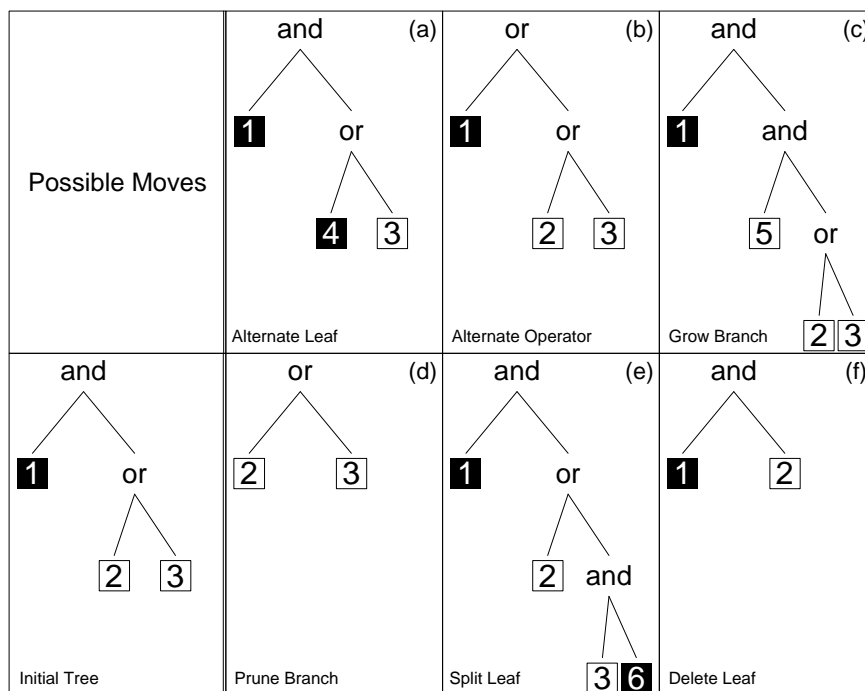


Logic Tree



A Decision Tree (CART) is something different!

The Move Set

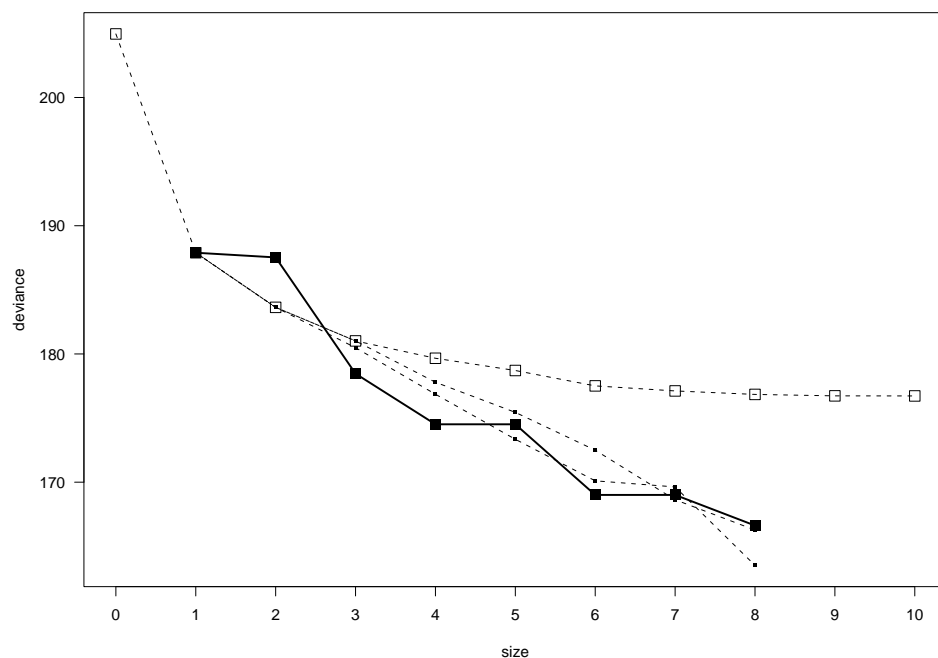


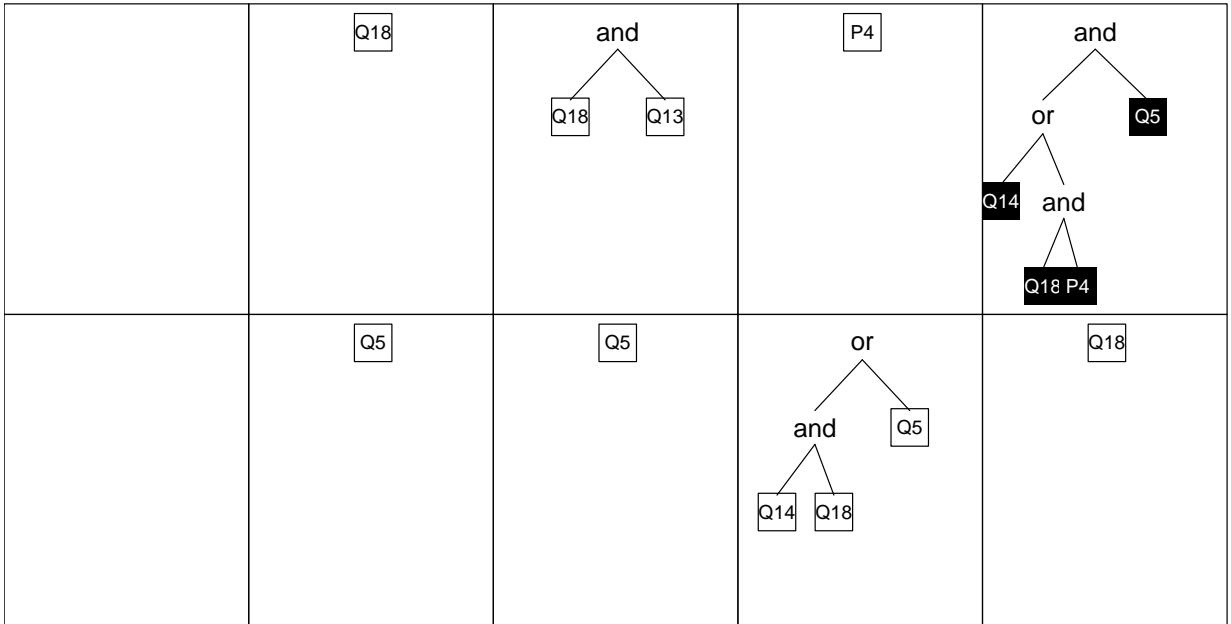
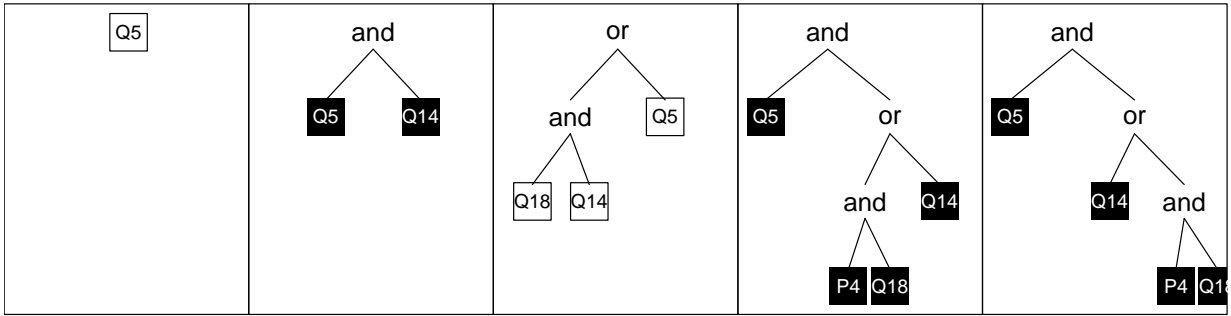
Simulated Annealing for Logic Regression

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.

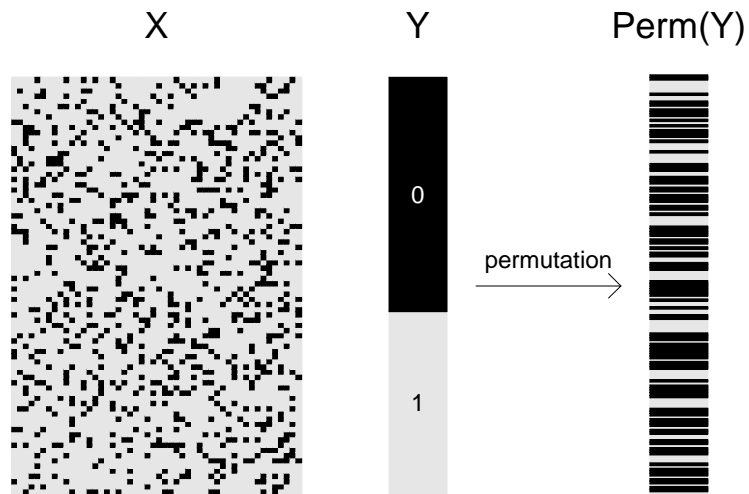
Bladder Cancer Example



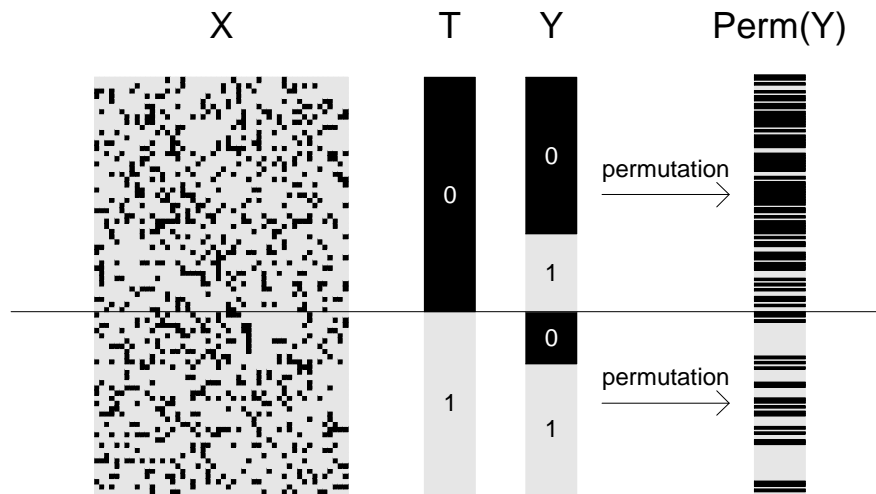


		Q18	and <div style="display: flex; justify-content: space-around; align-items: center;"> <div style="border: 1px solid black; padding: 2px;">P9</div> <div style="border: 1px solid black; padding: 2px;">P11</div> </div>	Q18
		Q5	Q18	Q5
		P4	Q5	and <div style="display: flex; justify-content: space-around; align-items: center;"> <div style="display: flex; flex-direction: column; align-items: center;"> <div style="margin-bottom: 5px;">and</div> <div style="display: flex; justify-content: space-around; align-items: center;"> <div style="border: 1px solid black; padding: 2px;">Q14</div> <div style="border: 1px solid black; padding: 2px;">Q13</div> </div> </div> <div style="border: 1px solid black; padding: 2px;">Q9</div> </div>

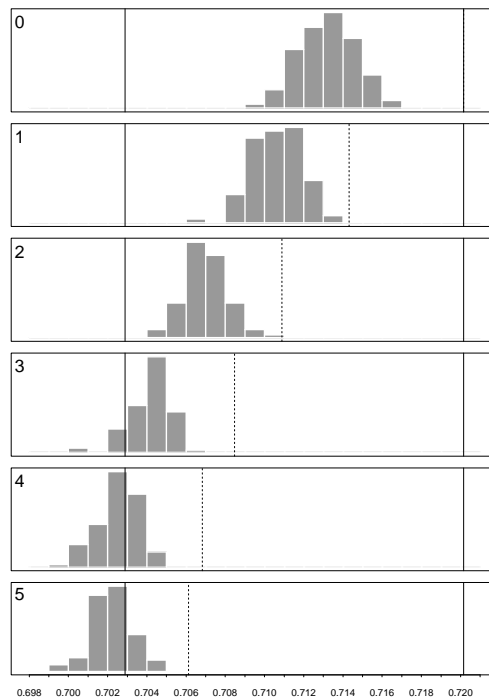
A Global Randomization Test of Association



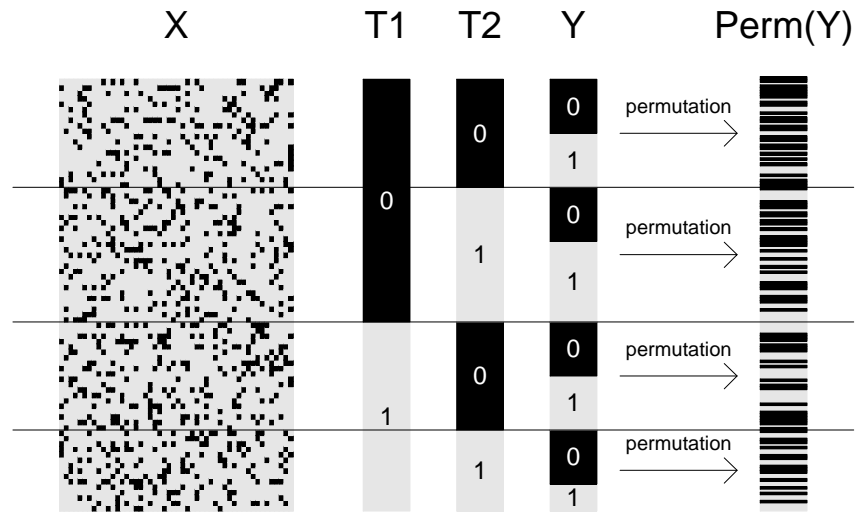
A Sequential Randomization Test for Model Size



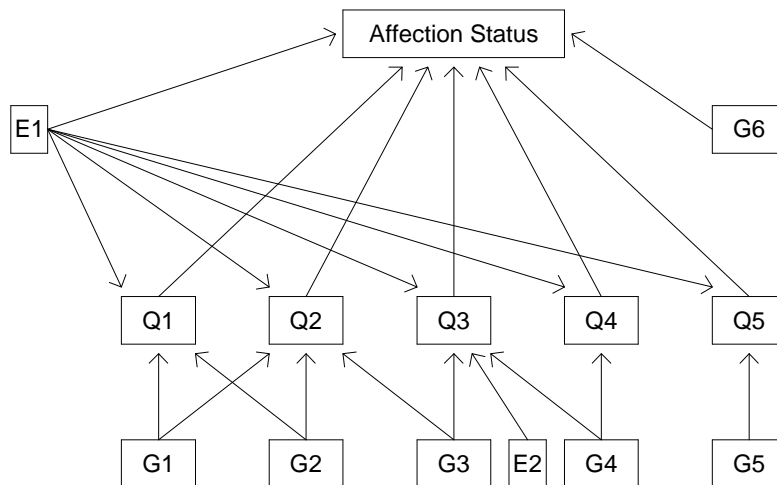
A Sequential Randomization Test for Model Size



Sequential Randomization Test for 2 Trees:

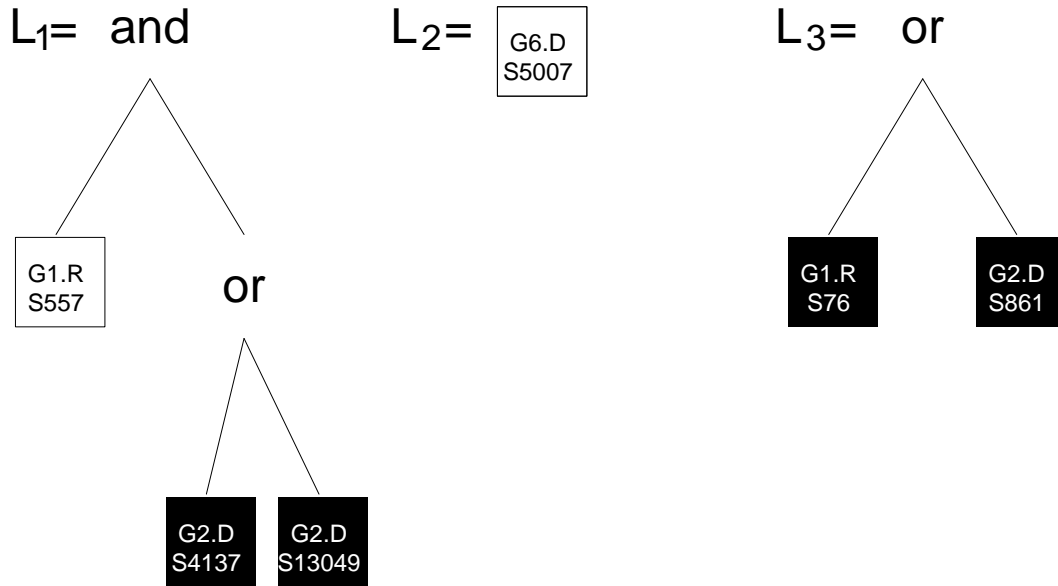


Genetic Analysis Workshop GAW 12

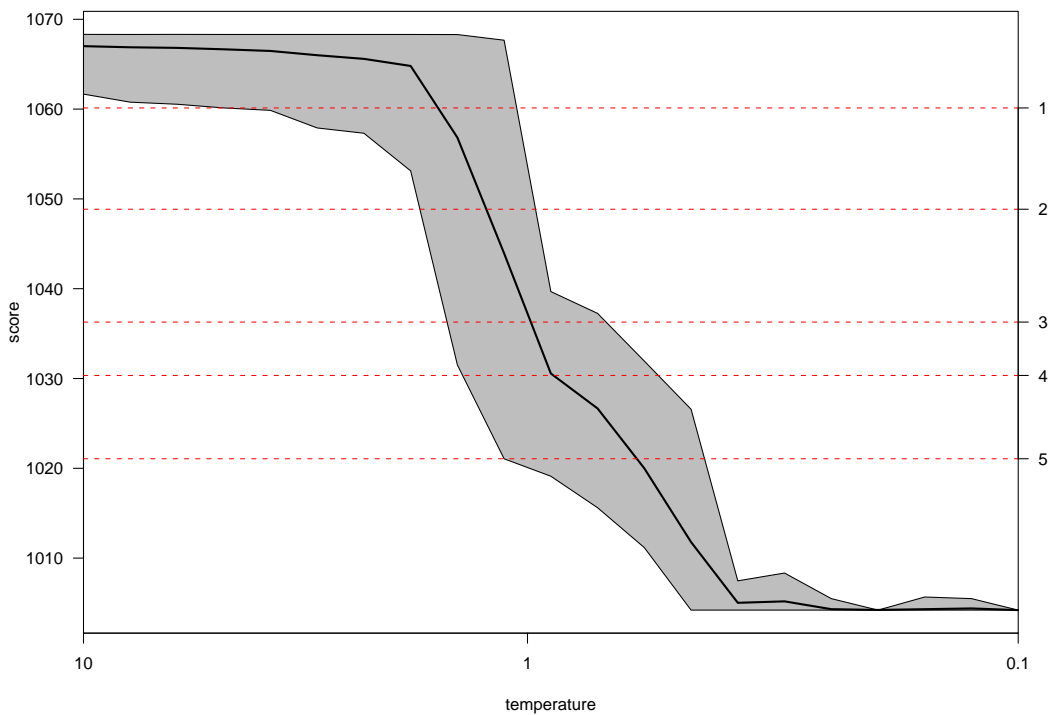


Genetic Analysis Workshop GAW 12

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



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, \dots, X_{10} .

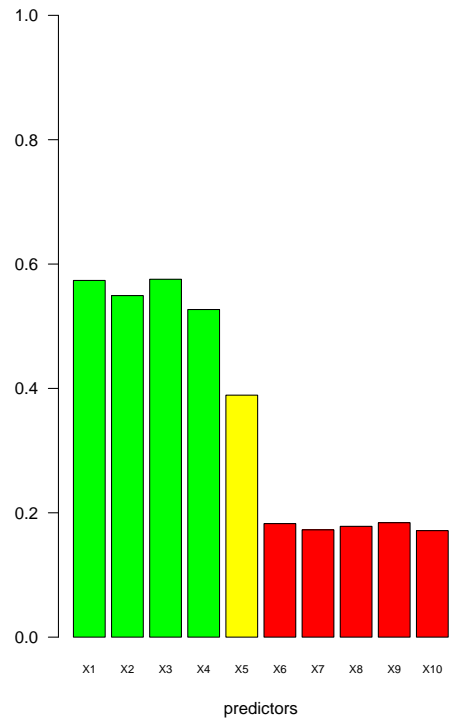
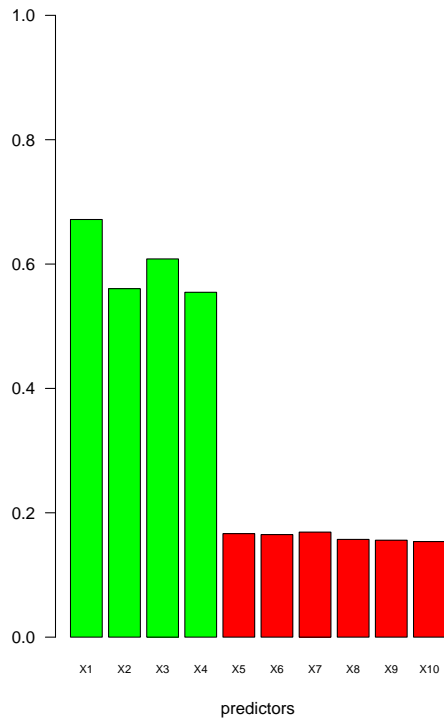
Let $Y = 5 + 1 \times L(X_1, X_2, X_3, X_4) + \epsilon$, $\epsilon \sim N(0,1)$.

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

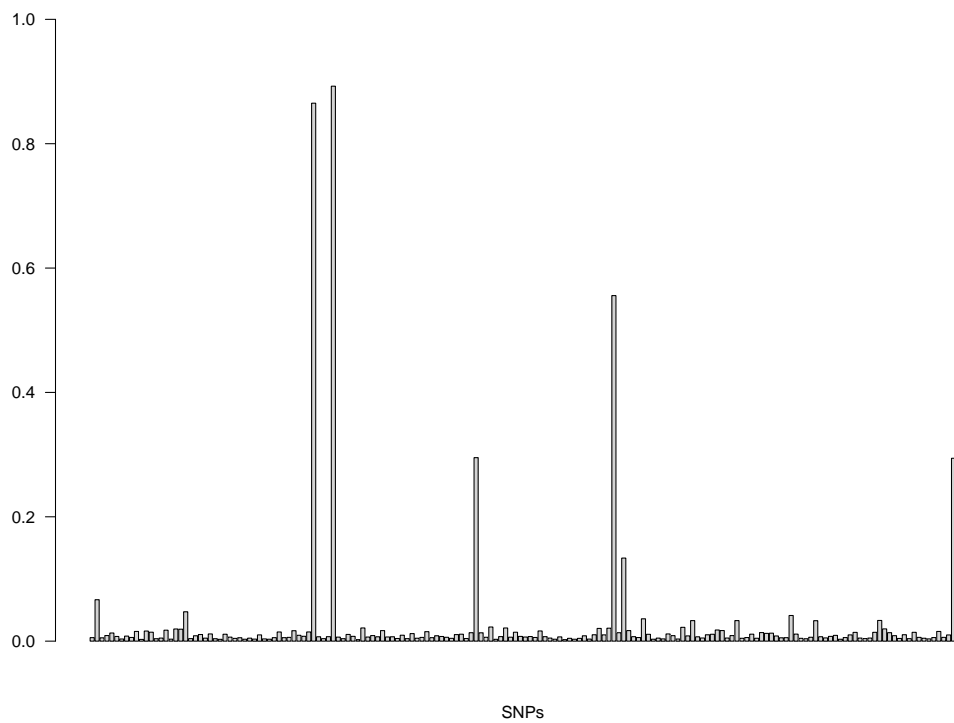
Case 1: All X are independent.

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

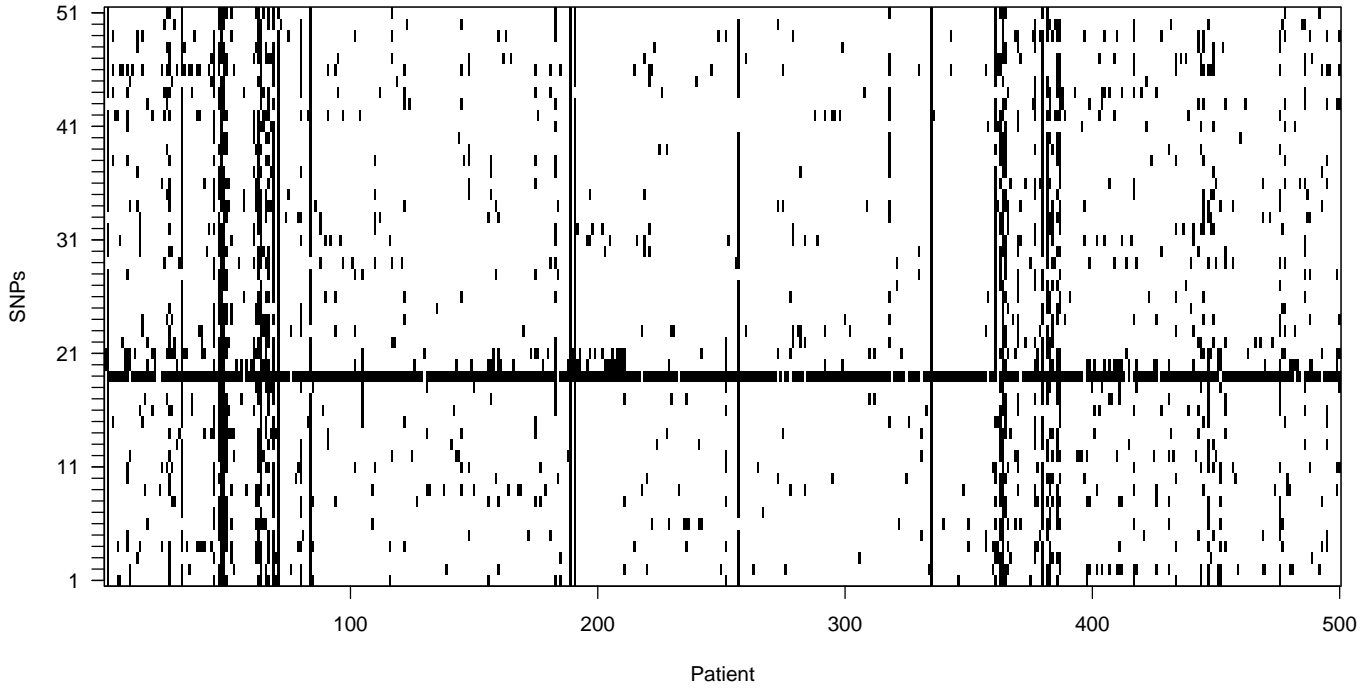
Multiple Models



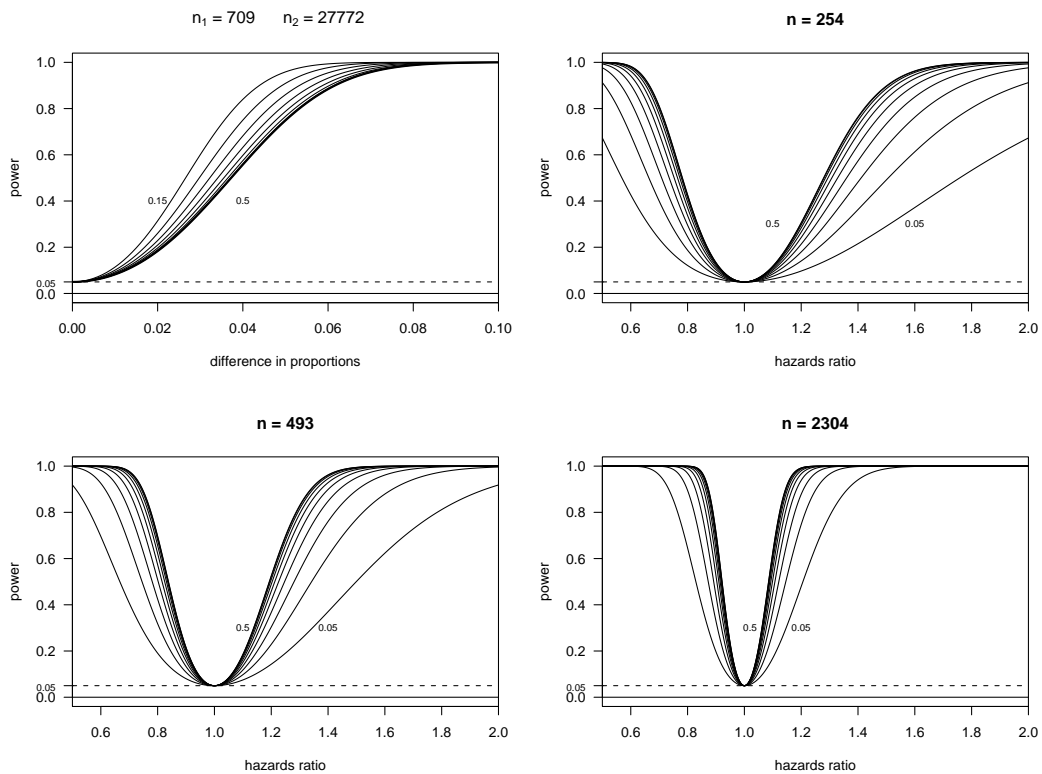
Multiple Models



Statistical Issues: Missing Values



Statistical Issues: Power



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. (2002), *Logic Regression - Methods and Software*, Proceedings of the MSRI workshop on Nonlinear Estimation and Classification (Eds: D. Denison, C. Holmes, M. Hansen, B. Mallick, B. Yu), Springer.
- Ruczinski, I., Kooperberg, C., and LeBlanc, M. (2003), *Logic Regression* Journal of Computational and Graphical Statistics (to appear). Available at: <http://biostat.jhsph.edu/~iruczins/>

The Bibliography of this paper contains all the references in this presentation.