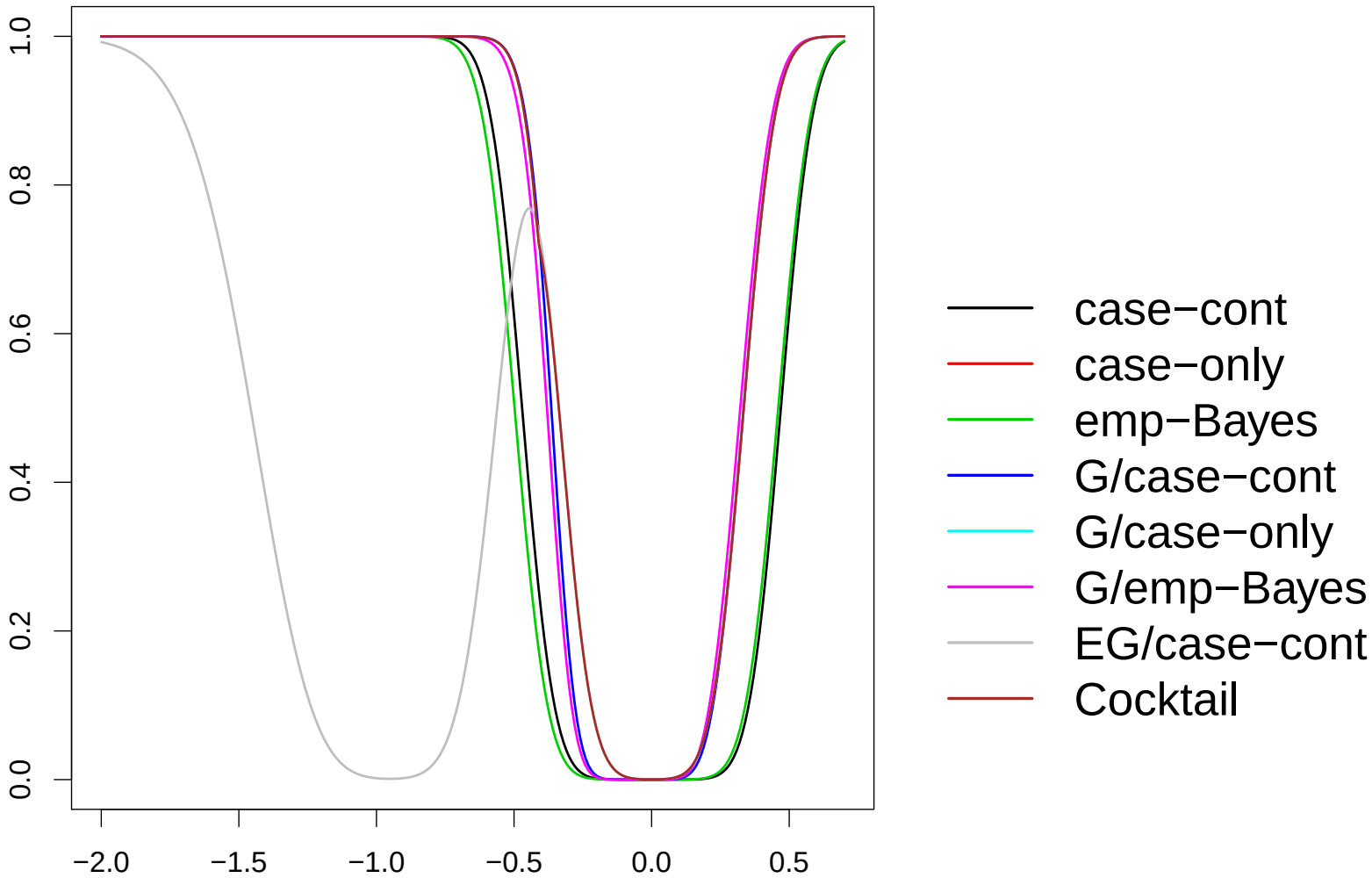
Last lecture revisited

October 20, 2015

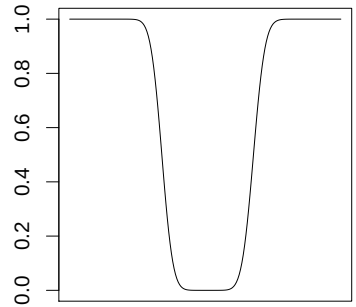
Does the grey curve go up again...

$N = 5000/5000$, $P(G = 1) = 0.3$, $P(E = 1) = 0.5$, 250,000 SNPs.
$\mathrm{OR}(G, E) = 1.5$

$$\mathrm{logit}(Y = 1|G, E) = \beta_0 + 0G + 0.5E + \beta_3 GE$$



More in "general"

- The power of (most of) the two stage procedures is like the product of two typical power curves 

, with their minimum maybe located at different spots - so you can have two minima in the product, but otherwise the curve should be smooth.
- The one (I think) exception is "Cocktail", since it also involves taking the minimum of two P-values. This is close to taking the maximum of two power curves (with a little smoothing aorund the point where the curves cross). After that this maximum curve is still multiplied by another power curve.

Higher order interactions

October 20, 2015

Why?

I ended the last lecture about GG and GE interactions with

A sobering note

There likely have been more papers written about methods to identify *GxE* and *GxG* interactions, than the number of interactions that have successfully been identified.

So why would we be interested in higher order interactions???

Targeted regions

For many reasons

- power,
- computational, and
- interpretation,

we should only be interested in higher order interactions when we focus attention on a few targeted regions (e.g. genes), selected because of

- studies (carried out on other data sets),
- biology,
- ...

It is not a surprise that...

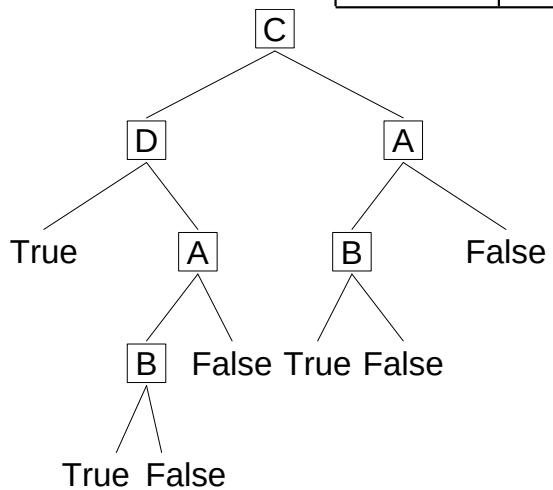
- The power is small.
- As such we may want to see these methods as "hypothesis generating" - i.e. we may identify a limited number of interactions that we can follow up on in new studies.

Models

- SNPs as 3 level categorical variables:

low	low	low
high	low	high
low	high	high

- Decision tree models.



- Boolean rules like:

  You are at increased risk if you have at least one mutant for SNP1 or two mutants for SNP2.

- Classical interaction model

$$g[E(Y|\mathbf{G})] = \beta_0 + \beta_1 G_1 + \beta_2 G_2 + \beta_3 G_3 + \beta_4 G_1 G_2 + \beta_5 G_1 G_3 + \beta_6 G_2 G_3 + \beta_7 G_1 G_2 G_3,$$

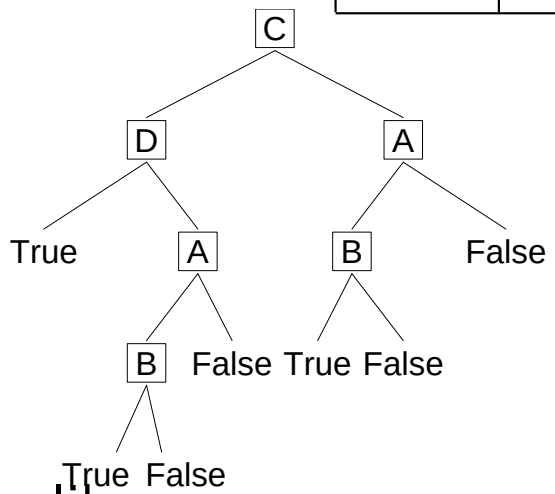
Issues: interpretation, computation, power.....

Models

MDR SNPs as 3 level categorical variables:

low	low	low
high	low	high
low	high	high

CART Decision tree models.



Logic Regression Boolean rules like:

You are at increased risk if you have at least one mutant for SNP1 or two mutants for SNP2.

- Classical interaction model

$$g[E(Y|\mathbf{G})] = \beta_0 + \beta_1 G_1 + \beta_2 G_2 + \beta_3 G_3 + \beta_4 G_1 G_2 + \beta_5 G_1 G_3 + \beta_6 G_2 G_3 + \beta_7 G_1 G_2 G_3,$$

Issues: interpretation, computation, power.....

SNPs

A typical dataset may contain:

- Response
- Environmental and demographic variables
- Sequence (SNP) data

Regression models

We want to find a regression model:

$$g(\mathrm{E}[Y|\mathbf{G},\mathbf{E}]) = f(\mathbf{G},\mathbf{E}),$$

- Y is the response, and $g()$ some link function,
- $\mathbf{E}$ are environmental/demographic variables,
- $\mathbf{G}$ are the SNPs, and
- $\mathbf{f}$ is a regression function.

Adaptive model selection

Consider

$$g(\mathrm{E}[Y|\mathbf{E},\mathbf{G}]) = \sum \beta_j B_j(\mathbf{E},\mathbf{G}).$$

$B_i()$: *basis functions* that may depend on the environmental variables **E** and/or the SNPs **G**.

Attempt to select basis functions B_i and estimate coefficients β_i.

Examples:

- CART - Classification And Regression Trees - Breiman, Friedman, Olshen and Stone (1984).
- MARS - Multivariate Adaptive Regression Splines - Friedman (1991).

Methods for other responses (e.g. survival, logistic) exist. The CS literature contains many proposals of methods.

Tree based methods

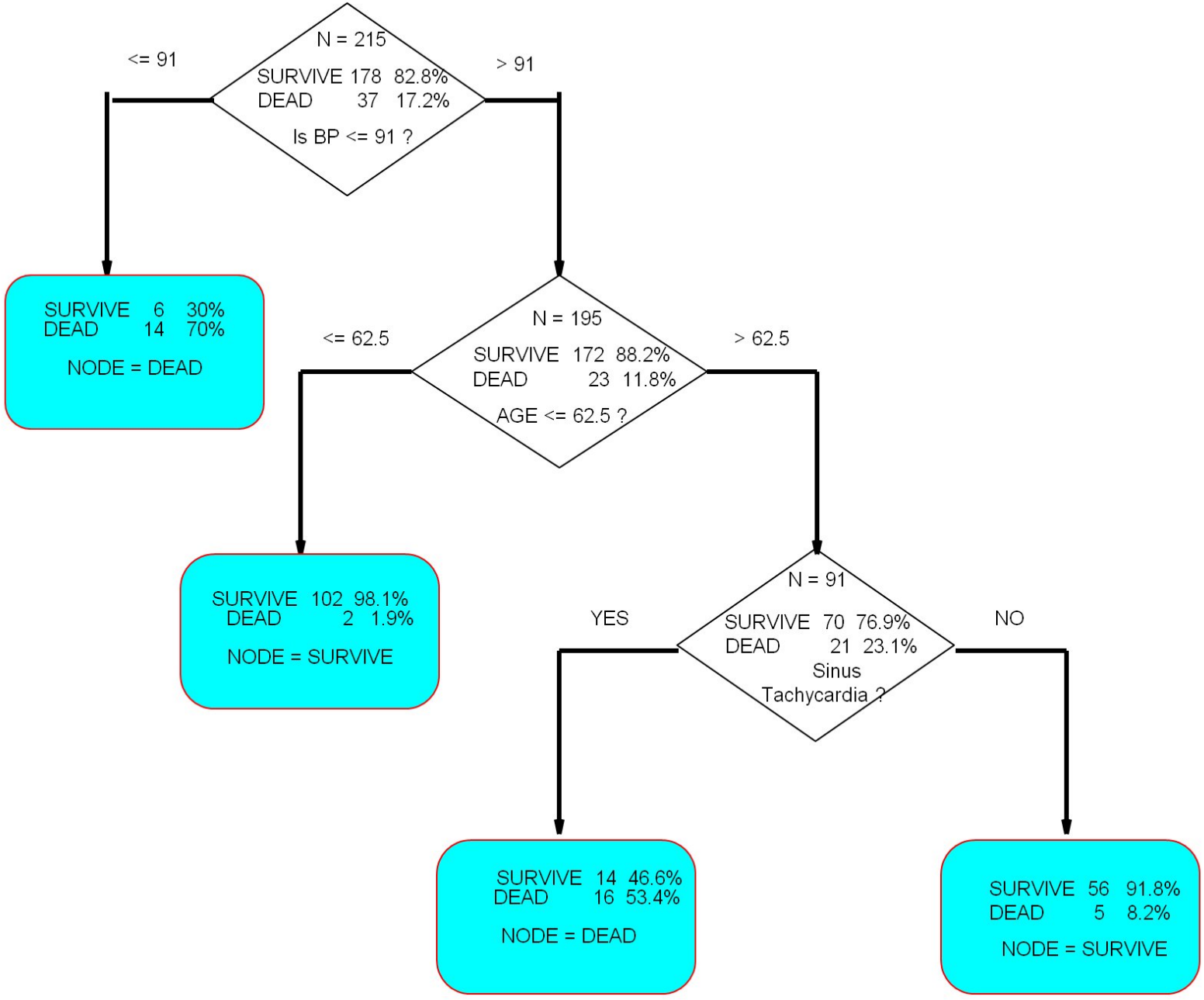
AKA "recursive partitioning", CART

[Breiman, Friedman, Olshen, Stone (1984) *Classification and Regression Trees* Wadsworth]

- Feature space recursively partitioned into rectangular areas such that observations with similar response are grouped.
- When you stop, you provide a common prediction Y for subjects in the same group.

Example

- UCSD Heart Disease study:
- Given the diagnosis of a heart attack based on Chest pain, Indicative EKGs, Elevation of enzymes typically released by damaged heart muscle
- Predict who is at risk of a 2nd heart attack and early death within 30 days Prediction will determine treatment program (intensive care or not)
- For each patient about 100 variables were available, including demographics, medical history, lab results


N = 215
SURVIVE 178 82.8%
DEAD 37 17.2%
Is BP <= 91 ?
<= 91
> 91
SURVIVE 6 30%
DEAD 14 70%
NODE = DEAD
N = 195
SURVIVE 172 88.2%
DEAD 23 11.8%
AGE <= 62.5 ?
<= 62.5
> 62.5
SURVIVE 102 98.1%
DEAD 2 1.9%
NODE = SURVIVE
N = 91
SURVIVE 70 76.9%
DEAD 21 23.1%
Sinus Tachycardia ?
YES
NO
SURVIVE 14 46.6%
DEAD 16 53.4%
NODE = DEAD
SURVIVE 56 91.8%
DEAD 5 8.2%
NODE = SURVIVE

Difference from linear (logistic) regression

- Not linear in predictors.
- Can have multiple splits of the same predictor.
- Non-linear and even non-monotone associations are identified in data-adaptive way.
- Modeling involves interactions, but the focus is identification of association/variable importance.
- Entire tree represents a complete analysis or model.
- Every data point goes from the *root* node, through (possibly multiple) splits and ends in a *terminal* node.

Note:

- These models can be written in the basis function set up. Basis functions are products of indicator functions.

These things are nice

- Universally applicable to both classification and regression problems with no assumptions on the data structure.
- Good properties:
  - Variable selection.
  - Deals well with missing data.
  - Deals well with outliers.
  - Deals well with multiple types of predictors.
  - No need to transform predictors.
  - Deals well with large dimensionality (though maybe not GWAS).
- A simple and easy to comprehend model:
  - Has the form of a decision tree.
  - Picture of the tree gives valuable insights into which variables are important and where.
  - Terminal nodes suggest natural clustering of data into homogeneous groups.

Elements of tree construction

- Tree growing
  - This is like stepwise addition in regression models.
  - Typically we want splits that split a less homogeneous node into two more homogeneous daughter nodes.
  - Usually done in a greedy way.
  - Continue growing until the tree is "too large".

Elements of tree construction

- Finding the right size of the tree.
  - Innovation in CART: cost-complexity pruning.
  - Pruning at a node means making that node terminal by deleting all its decendants.
  - For each α we can find the best tree that minimizes

$$R_\alpha(T) = R(T) + \alpha|T|$$

    where R is a cost measure, and $|T|$ the size of tree T.
  - For different αs the best trees are nested.
  - Cross-validation or an external data set allow us to pick the best α.

A few more plusses and minuses

+ Easy to interpret.
+ Natural way to decide which variables are (not) important, and when (i.e. age is not relevant if BP is low).
– Modest accuracy.
– Instability: changing the data a little can change the tree (sometimes) a lot.

Ensemble versions

- Bagging (Breiman 1996): Fit many trees to bootstrap resampled versions of the training data, and classify by majority vote.
- Boosting (Freund & Schapire 1996): Fit many trees to reweighted versions of the training data. Classify by weighted majority vote.
- Random Forest (Breiman 2001): Bootstrap both cases and randomly select predictors. Use the not-selected cases to estimate the accuracy.

Ensemble versions have much better prediction and stability, but loose interpretation.

Logic Regression

- $X_1, \ldots, X_k$ are 0/1 (False/True) predictors.
- Y is a response variable.
- Fit a model

$$g(\mathrm{E}[Y|\mathbf{E}, \mathbf{X}]) = \beta_0 + \sum_{j=1}^{t} \beta_j L_j + \sum_k \gamma_k E_k,$$

  where L_j is a Boolean combination (logic term) 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 Regression for SNP data

Think of a SNP as a variable G which takes values 0, 1 or 2.

A "dominant" SNP would have effect when $G \geq 1$, a "recessive" SNP when $X = 2$. Thus it makes some sense to recode:

$$X_1 = 1 \quad \text{if } X \geq 1,$$

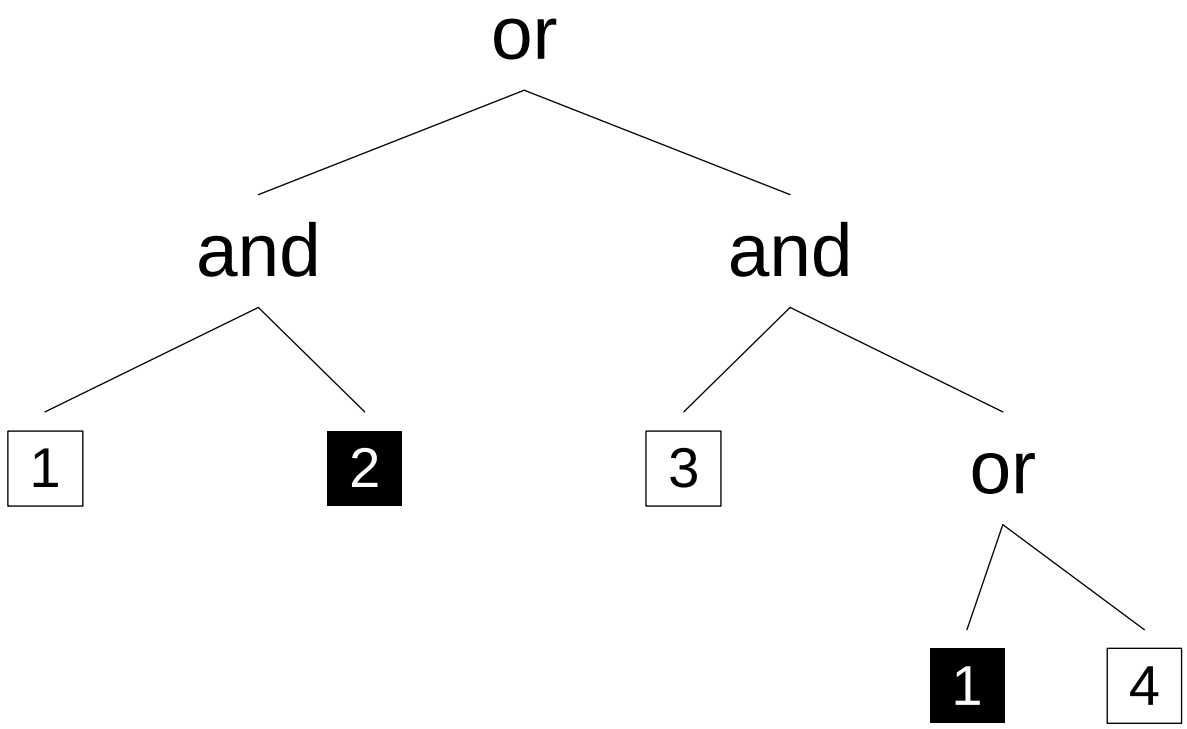
and

$$X_2 = 1 \quad \text{if } X = 2.$$

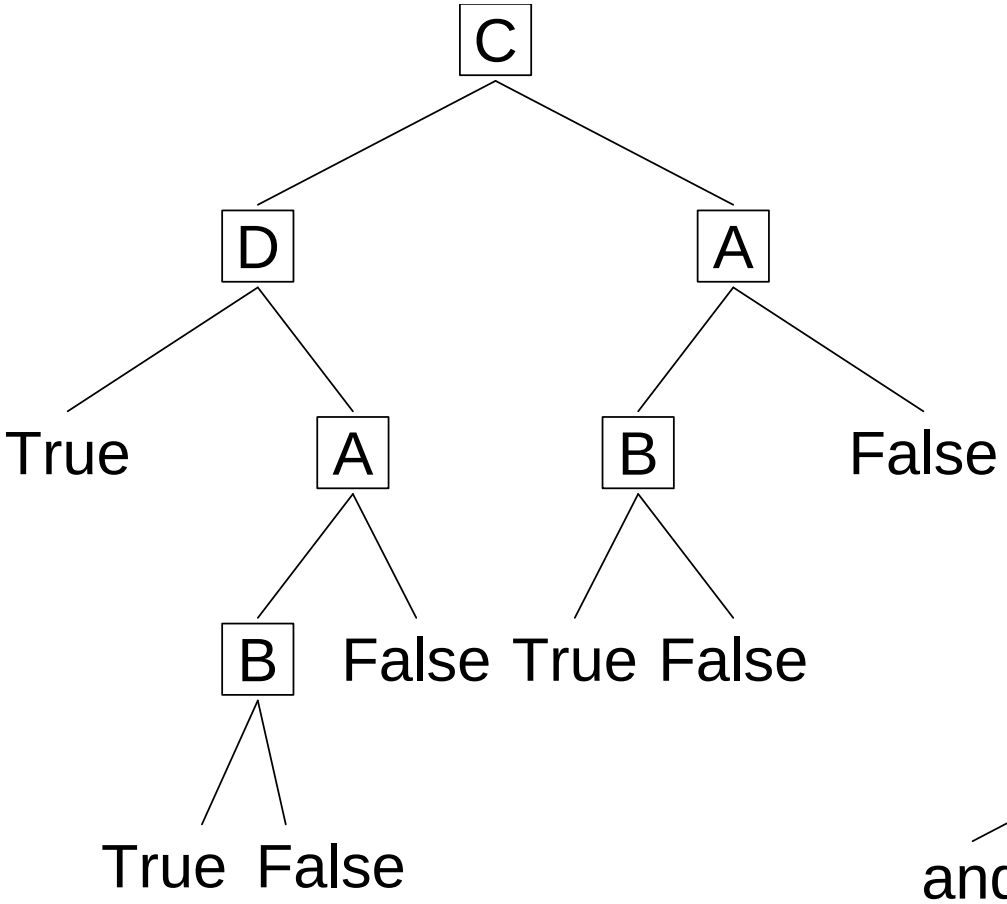
Logic Trees

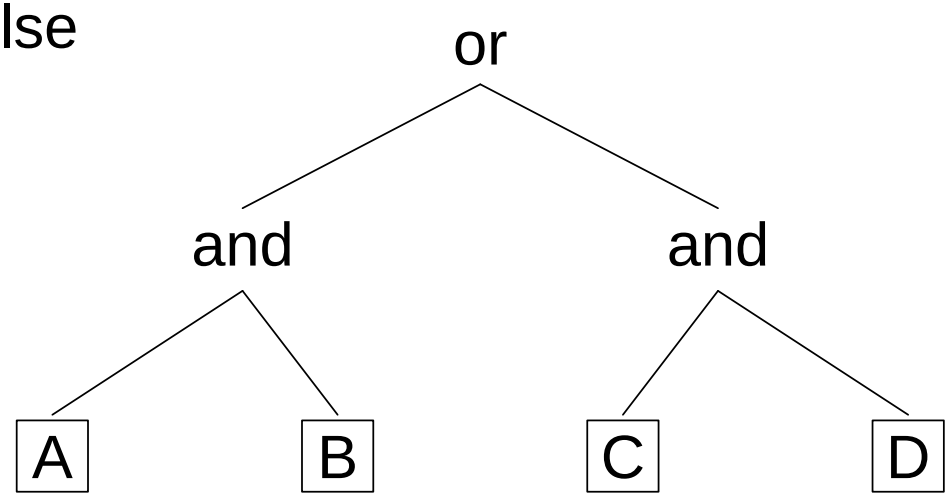
The Logic Tree representation of the logic term

$$(X_1 \wedge X_2^c) \vee (X_3 \wedge (X_1^c \vee X_4))$$



A decision tree (CART) is something different!





Greedy search

Typical way to select basis functions for adaptive regression models

$$g(\mathrm{E}[Y|\mathbf{V},\mathbf{X}]) = \sum \beta_i B_i(\mathbf{V},\mathbf{X}),$$

is *stepwise*:

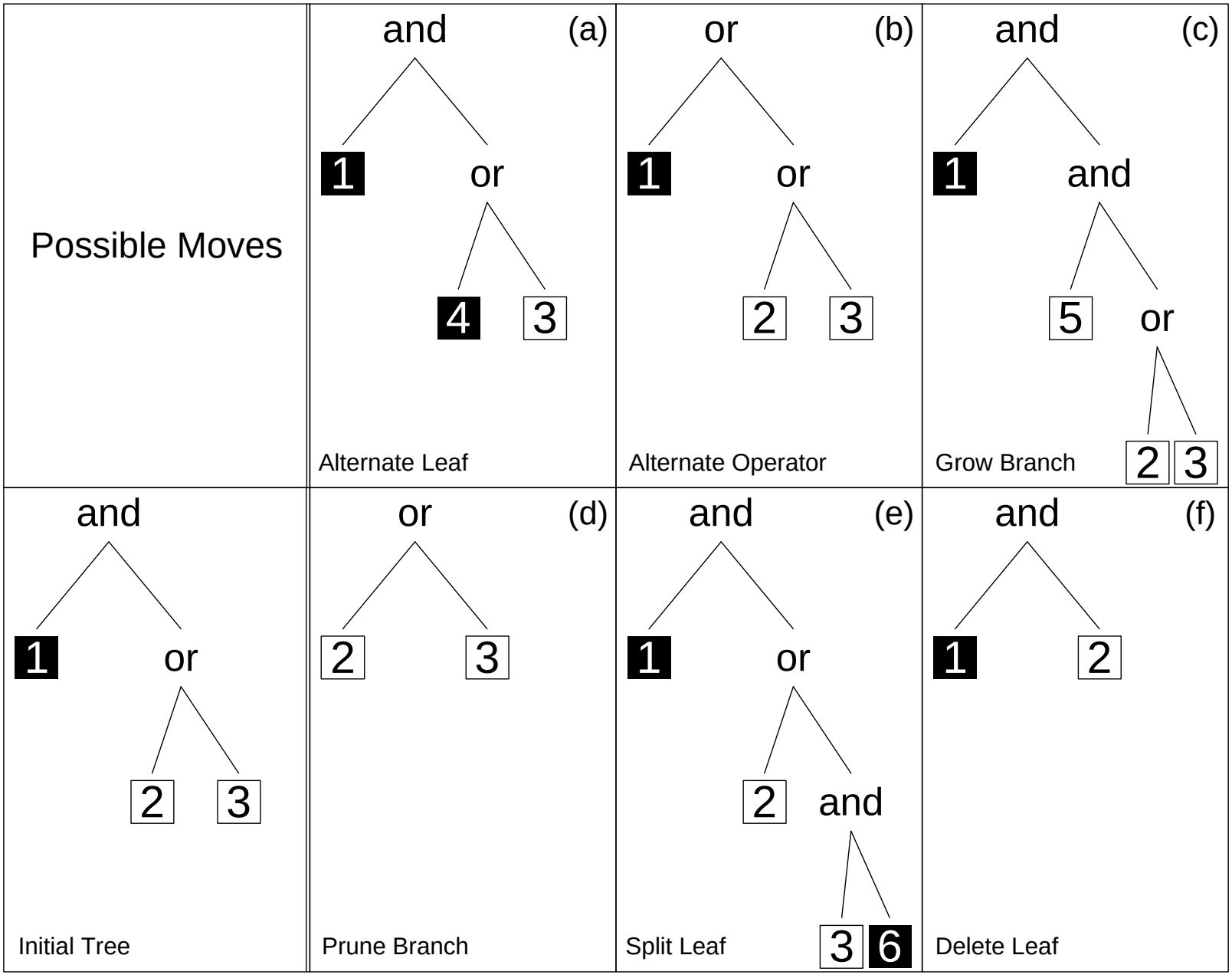
- Find the single best basis function to include in the model.
- Given the basis function already in the model, find the next best basis function to add.
- Continue until a largest model size with *stepwise addition*.
- Now remove basis functions one at a time, each time removing the least significant one.
- Select one model out of all models considered.

CART and MARS algorithms can be rephrased in this format.

Greedy search for logic regression

Problems:

- We want to keep the number of basis functions small, but rather find potentially quite complicated ones.
- Changes should thus not be in adding basis functions, but in making them more complicated. [See next slide...]
- The search space is quit messy, with many local optima, suggesting that greedy algorithms may not be very useful.


Possible Moves
and
(a)
1
or
4
3
Alternate Leaf
or
(b)
1
or
2
3
Alternate Operator
and
(c)
1
and
5
or
2
3
Grow Branch
and
1
or
2
3
Initial Tree
or
(d)
2
3
Prune Branch
and
(e)
1
or
2
and
3
6
Split Leaf
and
(f)
1
2
Delete Leaf

Simulated annealing for Logic Regression

We try to fit the model

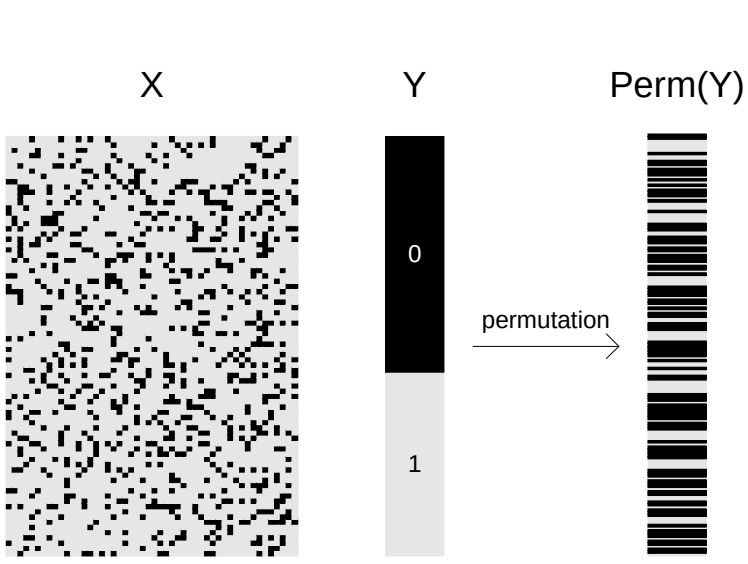
$$g(\mathrm{E}[Y|\mathbf{E},\mathbf{X}]) = \beta_0 + \sum_{j=1}^{t} \beta_j L_j.$$

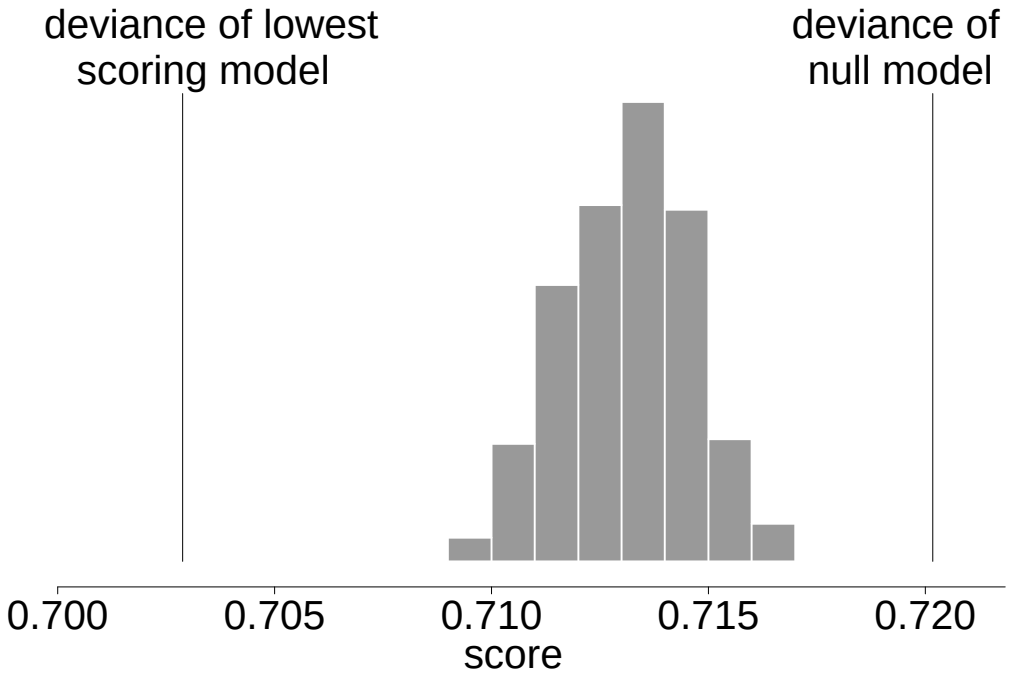
- Select a scoring function (RSS, log-likelihood, ...).
- Pick the maximum number of Logic Trees.
- Pick the maximum number of leafs in a tree.
- Carry out a Simulated Annealing algorithm:
  - Propose a move.
  - Accept or reject the move, depending on scores and temperature: $\alpha(s_{\text{old}}, s_{\text{new}}, t) = \min\{1, \exp([s_{\text{old}} - s_{\text{new}}]/t\}$.
  - Verrrrrrrrrry slowly reduce t.

Cardiovascular Health Study (CHS) MRI data

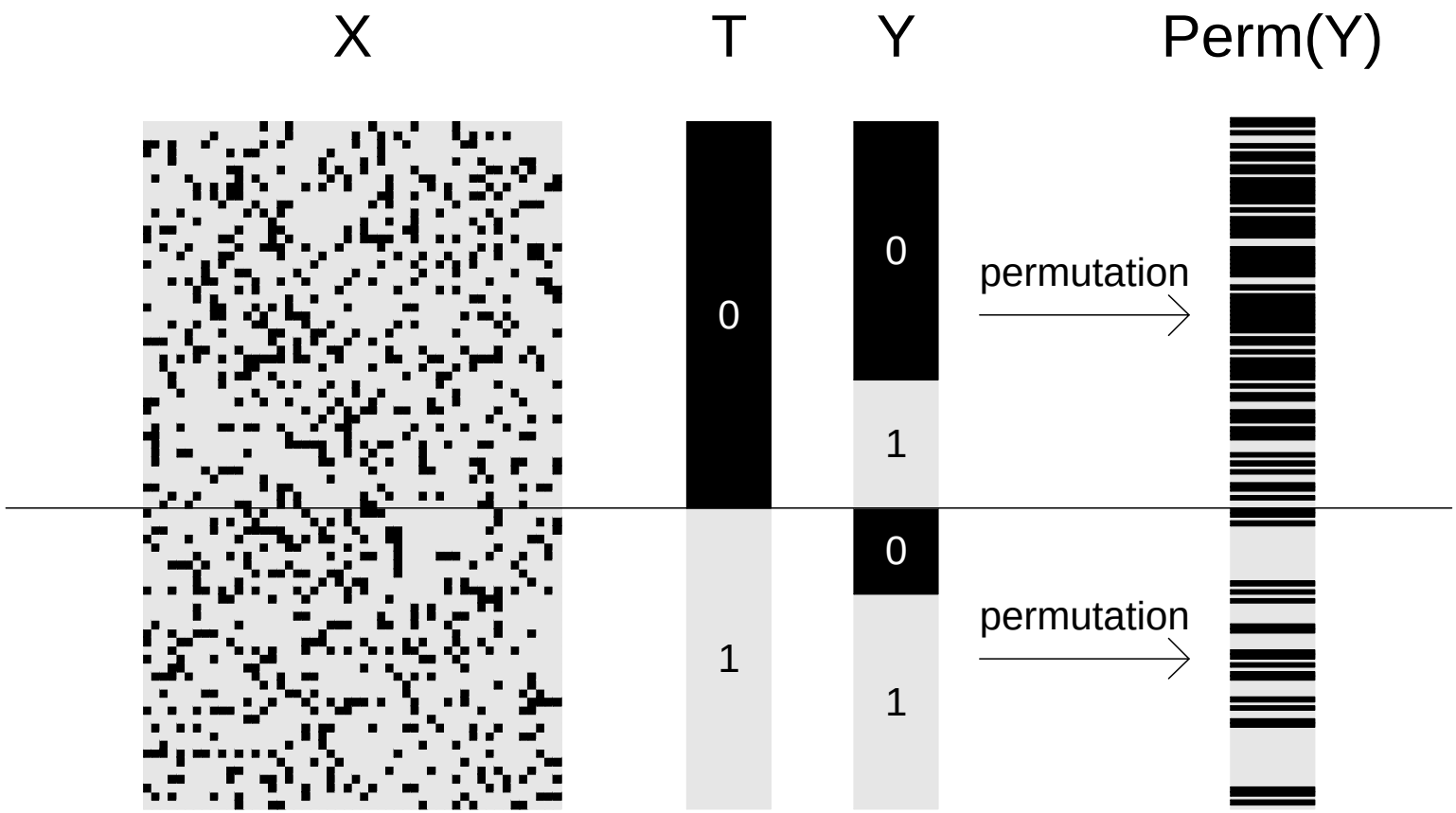
- CHS is a study of coronary heart disease and stroke in elderly people. Between 1989 and 1993, 5888 subjects over the age of 65 were recruited in four communities in the US.
- During 1992–94, a subset of these patients had an MRI scan.
- For 3647 CHS participants, MRI detected strokes (infarcts bigger than 3mm that led to deficits in functioning) were recorded as entries into a 23 region atlas of the brain.
- The mini-mental state examination is a screening test for dementia. The response Y is a variable derived by transforming the mini-mental score.
- We investigated models of the form
  $Y = \beta_0 + \beta_1 \times L_1 + \cdots + \beta_p \times L_p + \epsilon$

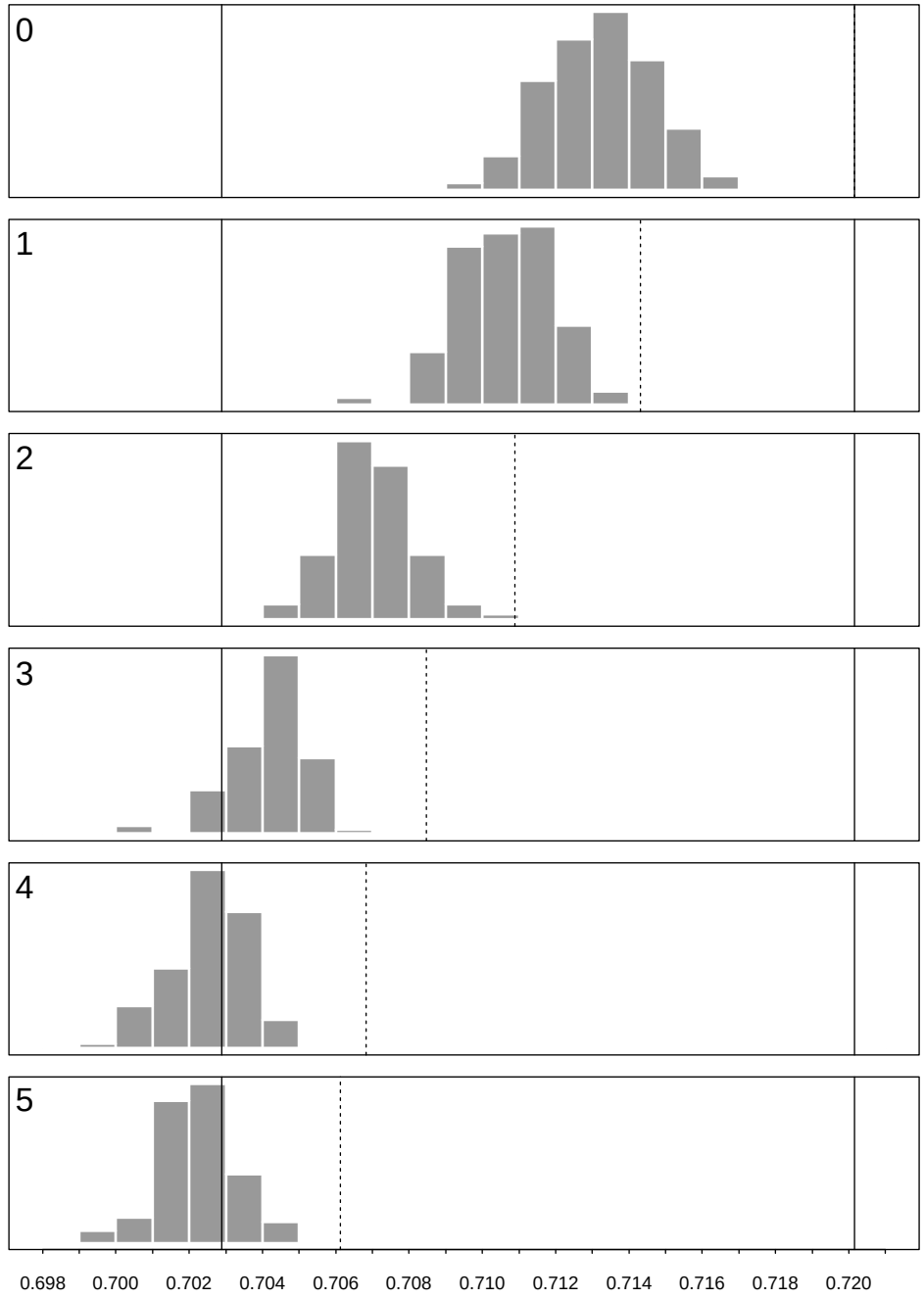
Null model (permutation) test



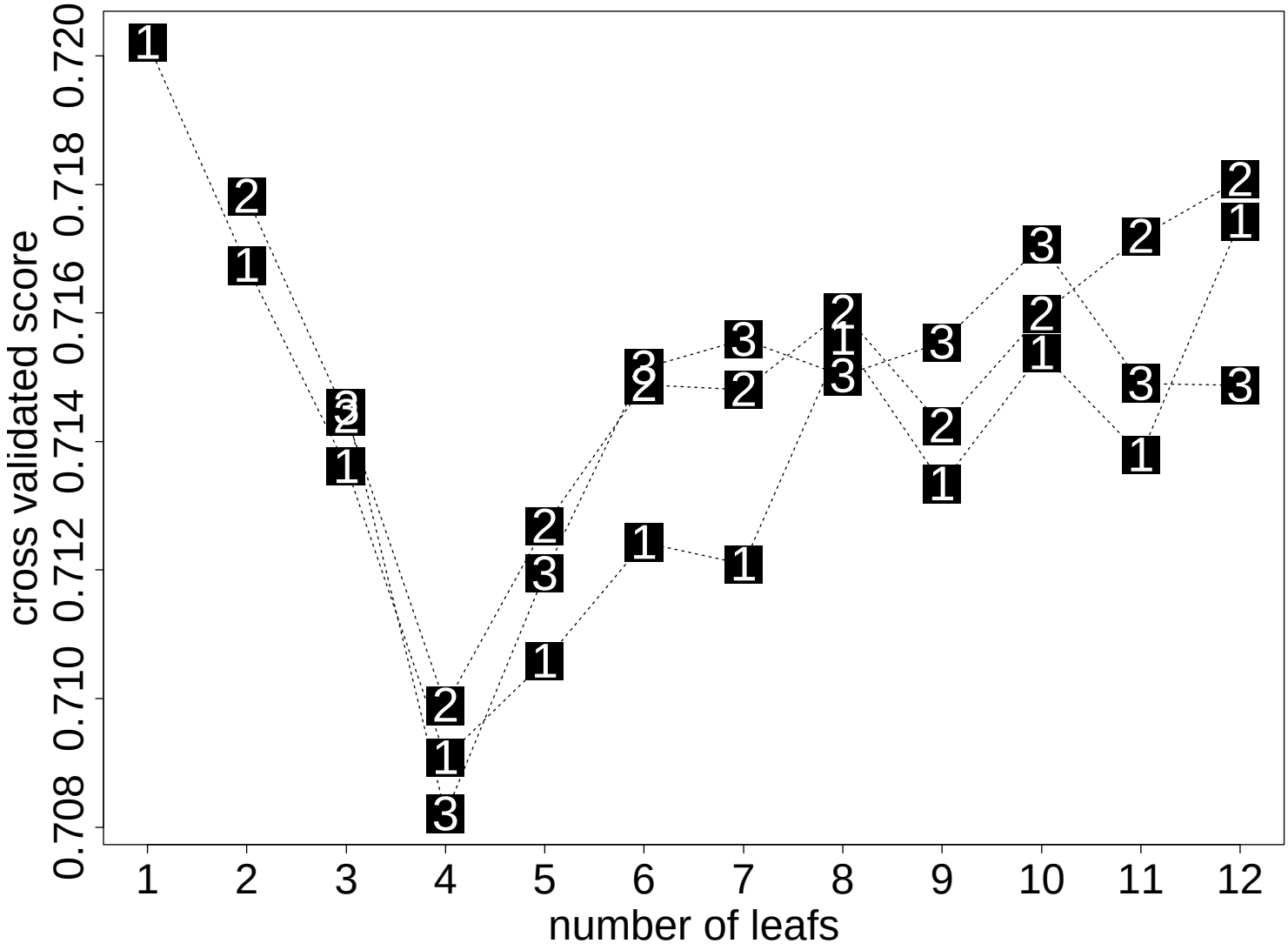


A sequential permutation test for model size.

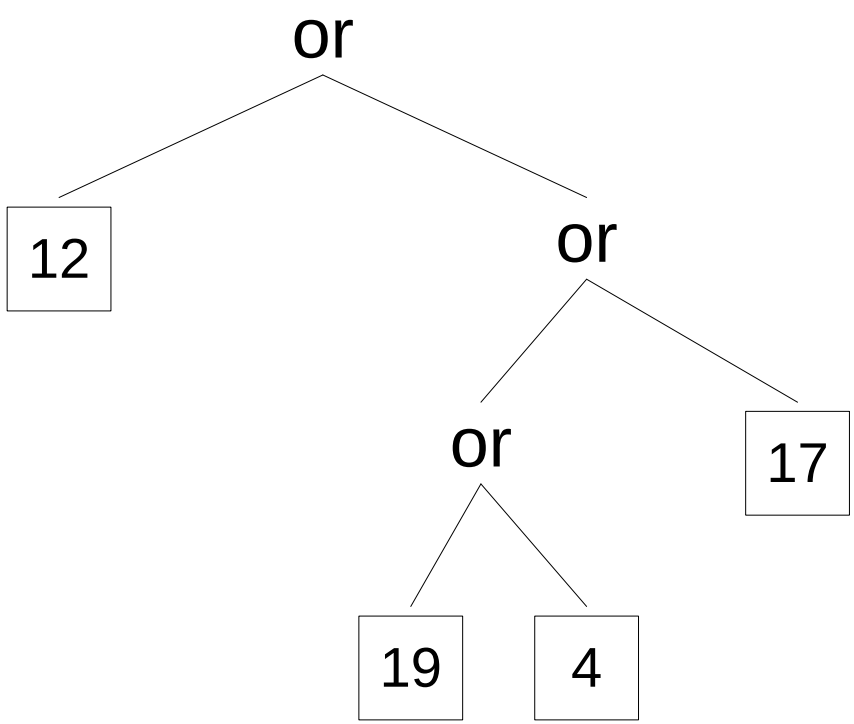



0
1
2
3
4
5
0.698
0.700
0.702
0.704
0.706
0.708
0.710
0.712
0.714
0.716
0.718
0.720

Cross validation



The selected model



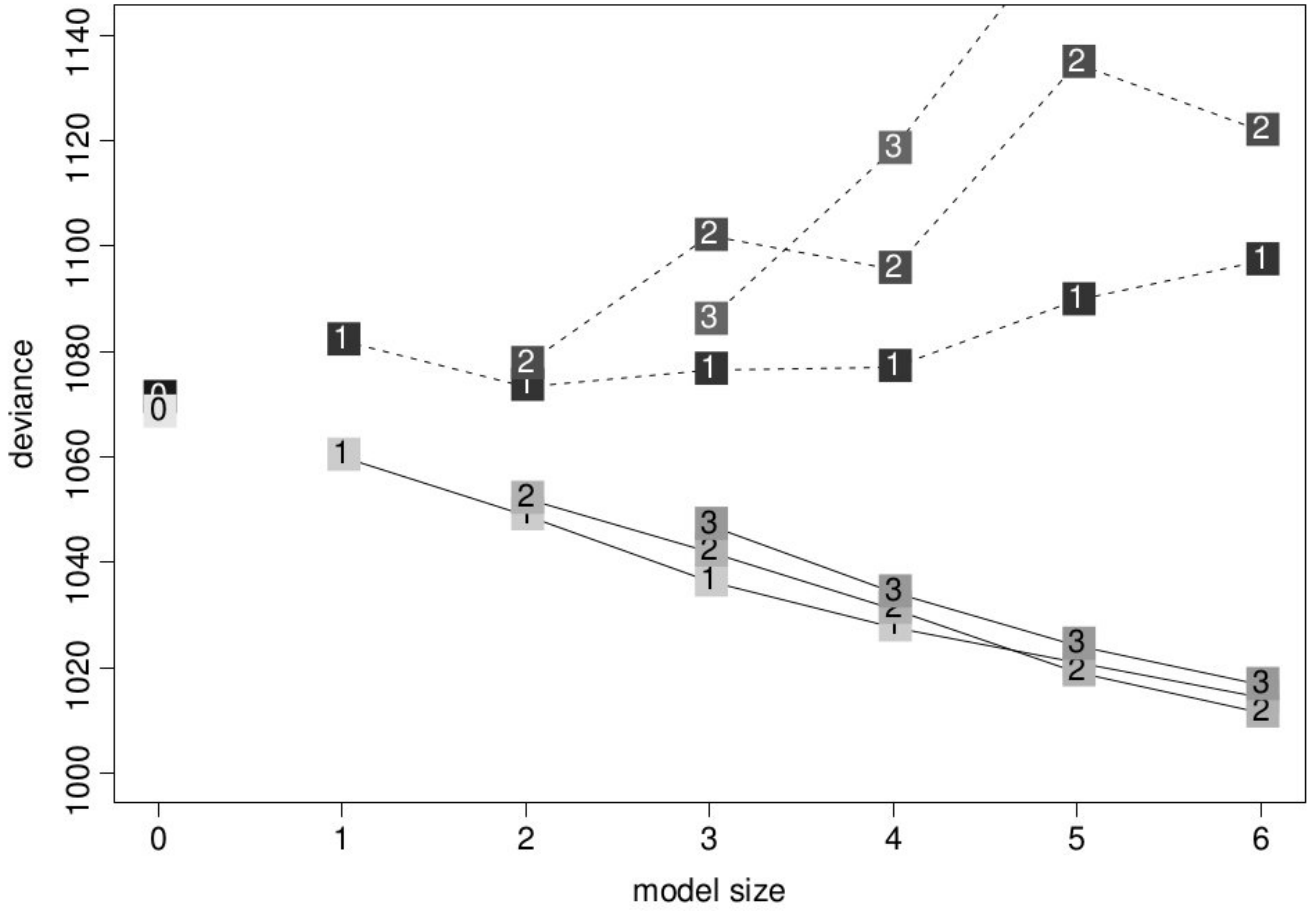
The model we found is $Y = 1.96 + 0.36 \times L$,
with $L = X_4$ or X_{12} or X_{17} or X_{19}.

Heart disease data

(Jurg Ott, Rockefeller University)

- 779 heart disease patients, all undergone angioplasty
- 342 experienced restenosis, 437 did not
- 63 candidate genes identified, 1-2 SNPs per gene; total 89 SNPs; recode in 178 binary predictors.
- no other variables

Cross validation



Potential problems with model selection

- The best size may not be clear cut.
- There may be several models of the same size that are (almost) as good.
- Sample sizes may currently be limited, reducing power to find interactions.

Bayesian Logic Regression

Reversible jump McMC (Green, 1995) requires

1. Prior on size of the model.
   What is size?
   Most natural to take a geometric prior relative to model size: this is equivalent to an AIC type penalty on size. This requires us to count the number of possible models of a given size (nontrivial!).
2. Prior on models, given the size.
   Also: probably want it uniform on model given the size. It's not totally obvious what this means. E.g. there are more ways to write $(X_1 \vee X_2 \vee X_3)$ than $(X_1 \vee (X_2 \wedge X_3))$. But ignore that right now.

Prior on size

$P(\text{size} = i) \propto a^i$.

Average posterior model size

a	mean of the prior	number of fitted logic trees 1	2	3	4
$1/\sqrt{2}$	2.41	3.00	5.07	6.34	7.02
$1/2$	1.00	1.76	2.24	2.48	2.55
$1/3$	0.50	1.04	1.13	1.22	1.25

Median over 25 permutations (Null model test)

a	mean of the prior	number of fitted logic trees 1	2	3	4
$1/\sqrt{2}$	2.41	2.66	4.19	4.92	5.58
$1/2$	1.00	1.54	1.86	2.01	2.14
$1/3$	0.50	0.92	1.02	1.08	1.13

Fraction of times in model

Top 15 SNPs	3 trees $a = 1/\sqrt{2}$	2 trees $a = 1/2$
TP53(P72R)$_d$	.379	.200
CD14$_d$	.353	.201
MDM2$_d$	.135	.054
CBS(I278T)$_r$	.132	.054
TNFR1$_d$	.119	.055
CBS(68bp ins)$_r$	.112	.046
IL4RA(I50V)$_r$	.110	.048
TNFR1$_r$	.105	.042
APOC3(T3206G)$_d$	.096	.038
LTA$_r$	.076	.032
GNB$_d$	.073	.026
ADRB3$_r$	.064	.025
NOS3$_r$	.063	.026
LPA(G21A)$_d$	.055	.024
ITGB3$_r$	.053	.023

Top seven two-SNP interactions.

		3 trees, $a = 1/\sqrt{2}$			2 trees, $a = 1/2$		
SNP 1	SNP 2	obs	exp	ratio	obs	exp	ratio
TP53(P72R)$_d$	CD14$_d$	.182	.072	2.52	.084	.037	2.23
TP53(P72R)$_d$	CBS(I278T)$_r$	.077	.027	2.85	.028	.010	2.77
APOC3(T3206G)$_d$	TNFR1$_r$	.074	.006	13.42	.030	.001	20.16
CD14$_d$	CBS(I278T)$_r$	.061	.025	2.43	.023	.010	2.34
TP53(P72R)$_d$	CBS(68bp ins)$_r$	.061	.023	2.67	.022	.008	2.55
CD14$_d$	CBS(68bp ins)$_r$	.047	.021	1.60	.018	.009	1.26
TP53(P72R)$_d$	MDM2$_d$	.044	.028	1.60	.013	.010	1.26

Top three three-way interactions.

SNP 1	SNP 2	SNP 3	3 trees $a = 1/\sqrt{2}$ obs	2 trees $a = 1/2$ obs
TP53(P72R)$_d$	CD14$_d$	CBS(I278T)$_r$	.0581	.0223
TP53(P72R)$_d$	CD14$_d$	CBS(68bp ins)$_r$	.0439	.0167
TP53(P72R)$_d$	CD14$_d$	APOC3(T3206G)$_d$	.0204	.0073

Logic Regression references

- Ruczinski, I., Kooperberg, C., and LeBlanc, M. L. (2003). Logic regression. *Journal of Computational and Graphical Statistics*, **12**, 475–511.
- Kooperberg, C. and Ruczinski, I. (2005). Identifying interacting SNPs using Monte Carlo Logic Regression. *Genetic Epidemiology*, **28**, 157–170.
- Kooperberg, C., Bis, J. C., Marciante, K. D., Heckbert, S. R., Lumley, T., and Psaty, B. M. (2007). Logic Regression for the analysis of the association between genetic variation in the renin-angiotensin system and myocardial infarction or stroke. *American Journal of Epidemiology*, **165**, 334–343.
- CRAN package `LogicReg`

Multifactor Dimensionality Reduction

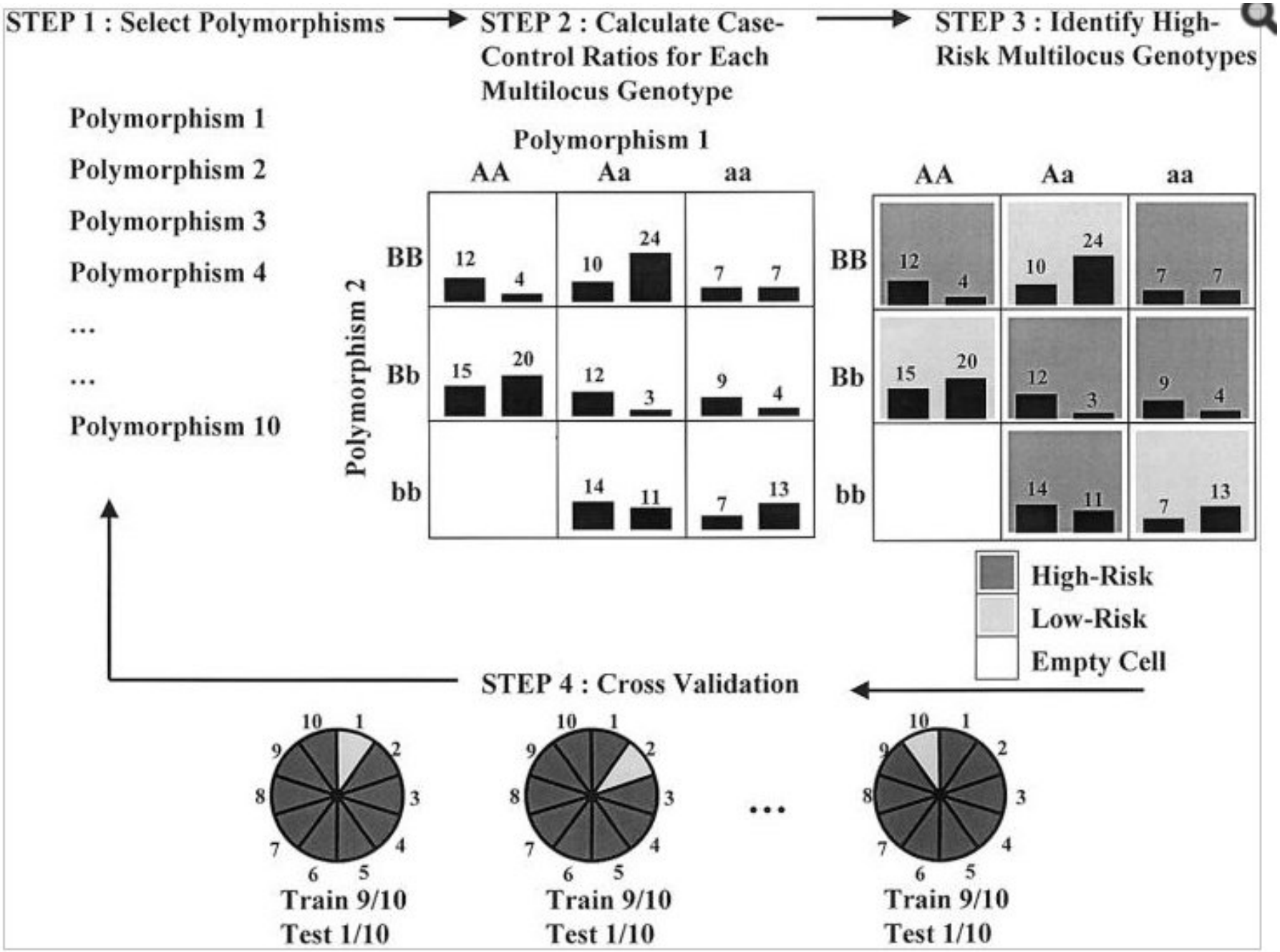
[Ritchie et al. (2001) *Am J Hum Gen* **69**:138–47]
[Hahn et al. (2003) *Bioinformatics* **19**:376–82]
modification of
[Nelson et al. (2001) *Genome Res* **11**:458–70]

- Complex interactions are hard to detect because of sparse data via standard parametric models
- Inaccurate parameter estimates and large standard errors with relatively small sample sizes.
- Reduce the dimensionality and identify SNP combinations that lead to high risk of disease.

Hunting for:

low	low	low
high	low	high
low	high	high

MDR


STEP 1 : Select Polymorphisms
STEP 2 : Calculate Case-Control Ratios for Each Multilocus Genotype
STEP 3 : Identify High-Risk Multilocus Genotypes
Polymorphism 1
Polymorphism 2
Polymorphism 3
Polymorphism 4
...
...
Polymorphism 10
Polymorphism 1
AA
Aa
aa
Polymorphism 2
BB
Bb
bb
12
4
10
24
7
7
15
20
12
3
9
4
14
11
7
13
High-Risk
Low-Risk
Empty Cell
STEP 4 : Cross Validation
Train 9/10
Test 1/10
Train 9/10
Test 1/10
Train 9/10
Test 1/10

MDR

For a particular model with M SNPs (or environmental factors):

- 10-fold Cross-validation
  1. Consider each "cell" (if factors are SNPs, there are 3^M).
  2. On 9/10th of the data decide whether a cell is "high" or "low" risk (for a case-control study the typical cut-off in each cell would be the case/control ratio in the study).
  3. Evaluate the prediction on the remaining 1/10th of the data.
  4. Check how many of the MDR models are the same. Not entirely clear how this is done - if each cell should be consistent, this would work against models that have (m)any cells that are close to 50/50.
- Repeat this a number of times - to achieve stability of the cross-validation. If you have enough computing power, always a good idea.
- Select the model with the lowest prediction error, provided the consistency is better than by chance.

Sporadic breast cancer

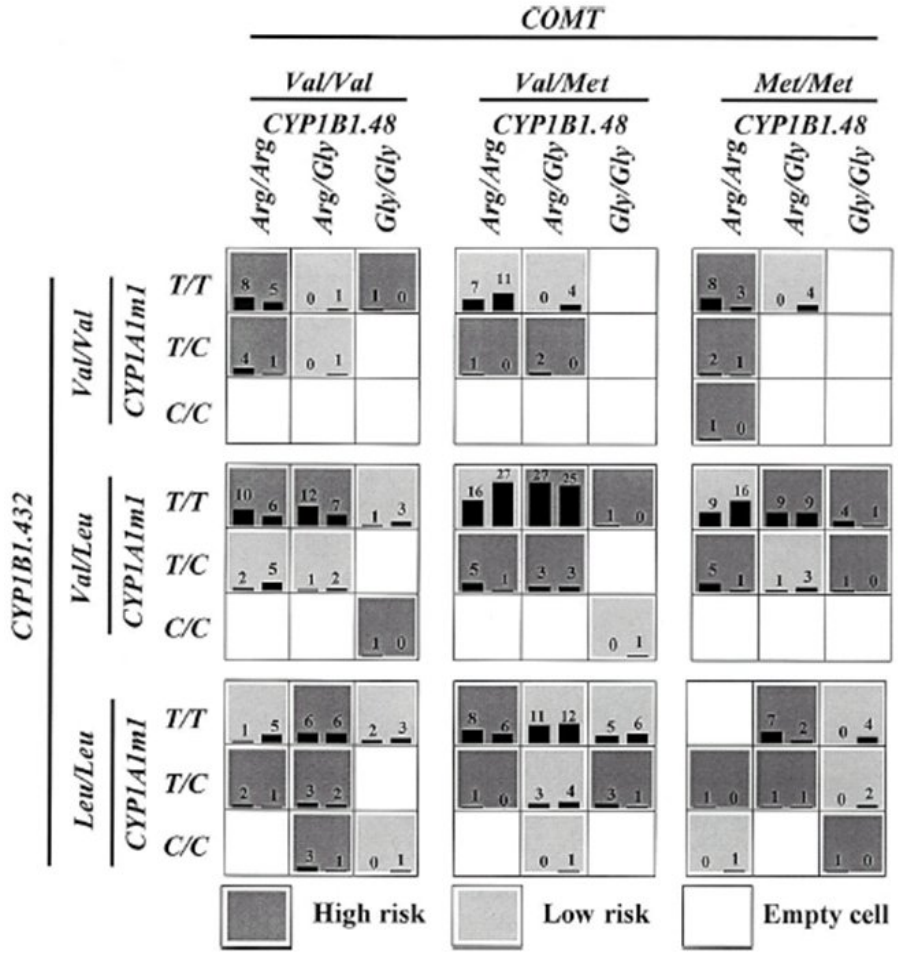
200 women with sporadic primary invasive breast cancer with age-matched hospital based controls, 10 estrogen metabolism SNPs

Summary of Results for Breast Cancer

No. of Loci	Cross-Validation Consistency	Prediction Error
2	7.00	51.06
3	4.17	51.35
4	***9.80****	***46.73***
5	4.71	50.26
6	5.00	48.61
7	8.60	47.15
8	8.20	52.55
9	7.10	53.40

NOTE.—The multilocus model with maximum cross-validation consistency and minimum prediction error is indicated in boldface italic type.

* $P < .001$.



Issues

- While making things binary helps, computation can explode if the number of SNPs in the study is substantial.
- The selected models do not adhere to the usual parsimony that we like in statistics: if a model with, say, 4 factors is ϵ better than a model with 3 factors, MDR will pick 4 factors. Usually we would prefer 3. Conceivably this could be changed fairly easy. The MDR implementation of cross-validation makes this worse, however (next slide).
- The models are very hard to interpret.
- To me, it would make more sense to identify a smaller number of cells with "extreme high" or "extreme low" risk.

Bias in their implementation of Cross Validation

- Consider the number of models with M SNPs out of a total T.

	0	1	2	3	4	5	6	7	8	$\cdots$
10	1	10	45	120	210	252	210	120	45	$\cdots$
25	1	30	435	4060	27405	142506	593775	2035800	5852925	$\cdots$

- Imagine what happens if there is no signal, and every model is equally likely, which size would we most likely end up with...
- The consistency reduces this problem a little, but not by much. Think about the situation where there is one SNP with a strong effect...

Take home message well beyond MDR

When using cross-validation for model selection, if the number of models of size M is different for different M, you can use cross-validation to find the best model of each size, but you cannot use it to find the best size. You need another test dataset for that!