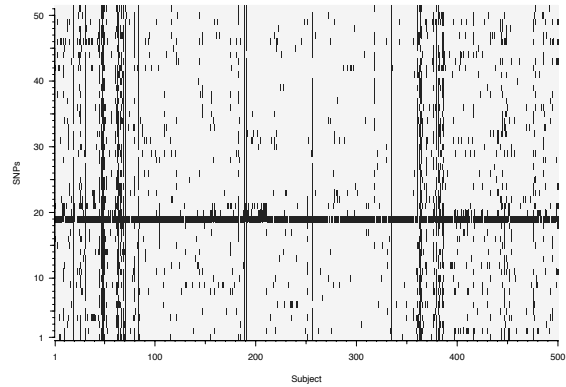


On Missing Data and Interactions in SNP Association Studies

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health



Missing Data - Approaches

- The most common approach for dealing with missing data is to omit the observations that have missing records in the model's covariates. This approach can have several shortcomings, including:

- Loss of power.
- Bias in the parameter estimates.

A good reference on this topic is [Greenland and Finkle \(1995\)](#).

- Some other used approaches are:
 - To impute a value from the marginal distribution of the covariate.
 - To create an extra level indicating *missingness*, if the covariate is a factor.

These choices tend to be not so great either.

Reference: Greenland S and Finkle WD (1995). A Critical Look at Methods for Handling Missing Covariates ... *Am J of Epidemiol*, 142(12): 1255-64.

Not Missing at Random

Method	Confidence Threshold	Overall Call Rate	Hom Call Rate	Het Call Rate
DM	0.26	94.16%	97.24%	86.32%
DM	0.33	95.96%	98.24%	90.16%
BRLMM	0.3	97.40%	97.40%	97.75%
BRLMM	0.4	98.27%	98.30%	98.48%
BRLMM	0.5	98.79%	98.82%	98.93%
BRLMM	0.6	99.15%	99.18%	99.25%

From the "white paper", http://www.affymetrix.com/support/technical/product_updates/brlmm_algorithm.affx

Missing Data - Approaches

- Multiple imputation can be used to draw valid statistical inference from data with missing values when the data are missing at random ([Little and Rubin 1987](#), [Schafer 1997](#)).

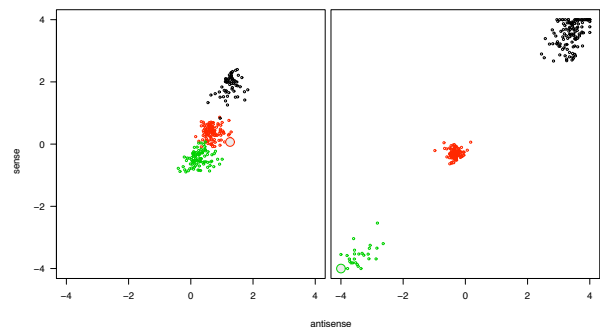
- In essence, multiple imputation acknowledges the uncertainty due to missing data, instead of simply ignoring it: several complete data sets are generated, and the uncertainty in the model parameter estimates incorporates the standard errors of the parameter estimates as well as the variability between the parameter estimates from the replicate data sets.

- While the hypothesis of missing at random cannot formally be tested, it is a lot less stringent than the requirement of missing completely at random, which is the underlying assumption made when observations are omitted.

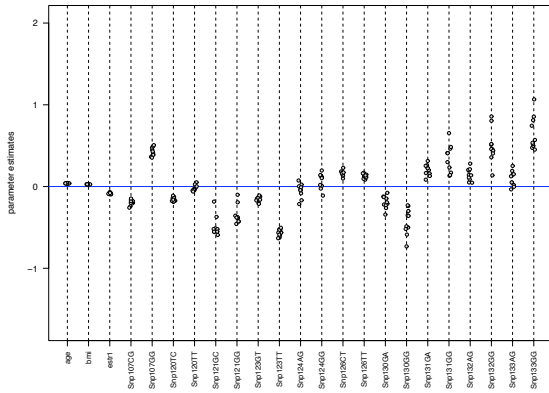
References:

- Little RJ, Rubin DB (1987). *Statistical Analysis with Missing Data*. John Wiley Sons, New York.
- Schafer JL (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

Not Missing at Random



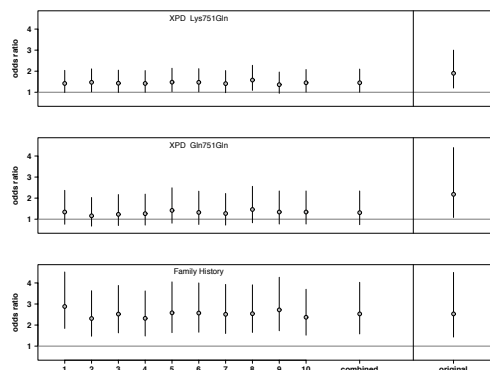
Multiple Imputation



Example 1

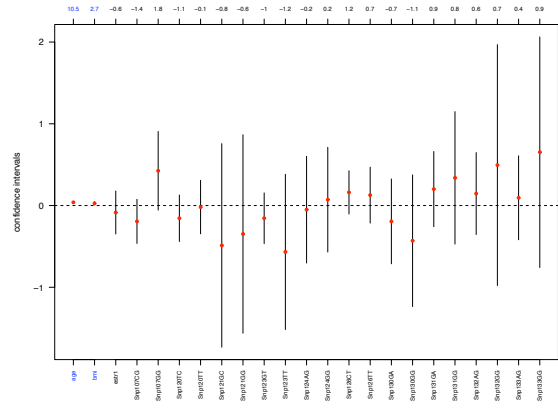
	Number of Pairs	Odds Ratio	Confidence Interval
XPD Lys751Gln			
original data set	202	1.90	(1.20 – 3.00)
multiple imputations	321	1.45	(1.00 – 2.10)
XPD Gln751Gln			
original data set	202	2.18	(1.08 – 4.40)
multiple imputations	321	1.31	(0.74 – 2.34)
Positive Family History			
original data set	202	2.53	(1.43 – 4.50)
multiple imputations	321	2.53	(1.58 – 4.03)

Example 1



The missing data were imputed using decision trees. In a minute ...

Multiple Imputation

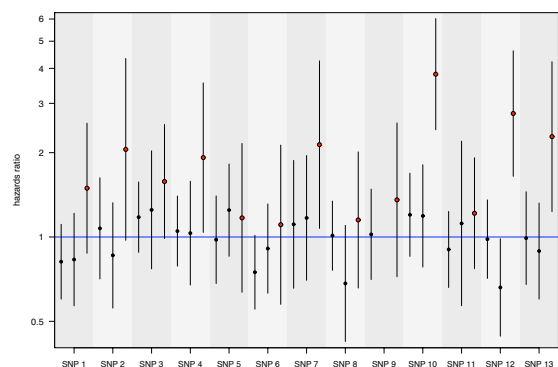


Example 1

	Family History not complete				Family History complete			
	AA	AC	CC	na	AA	AC	CC	na
raw numbers								
case	43	54	5	5	61	121	25	7
control	35	57	12	3	90	102	22	0
percentages								
case	40.2	50.5	4.7	4.7	28.5	56.5	11.7	3.3
control	32.7	53.3	11.2	2.8	42.1	47.7	10.3	0.0

Reference:
Brewster AM et al (2006). Polymorphisms of the DNA Repair Genes XPD (Lys751Gln) ... *Breast Cancer Res Treat*, 95(1): 73-80.

Example 2



Unpublished data.

Multiple Imputation

We looked into two approaches:

1. Haplotype-based imputation

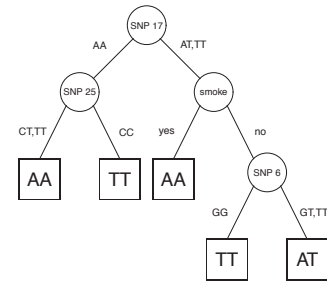
→ The idea here is to reconstruct the haplotypes (for example via the EM algorithm), and impute the missing values from the estimated haplotype frequencies.

2. Tree-based imputation

→ The idea here is to use decision trees to impute the genotype data, borrowing information from neighboring SNPs and other variables.

Reference: Dai J et al (2006), A Comparison of Haplotype-based and Tree-based SNP Imputation ... *Genet Epidemiol*, 30(8): 690-702.

Tree-based Imputation



→ Classification trees are great for categorical data!

Tree-based Imputation

Specifically, we consider iteratively sampling from the following sequence of the full conditional distributions in the $(n + 1)^{\text{th}}$ iteration:

$$M_1^{(n+1)} \sim \Pr(M_1 | M_2^{(n)}, M_3^{(n)}, \dots, M_p^{(n)}, \mathbf{C}, \mathbf{D})$$

$$M_2^{(n+1)} \sim \Pr(M_2 | M_1^{(n+1)}, M_3^{(n)}, \dots, M_p^{(n)}, \mathbf{C}, \mathbf{D})$$

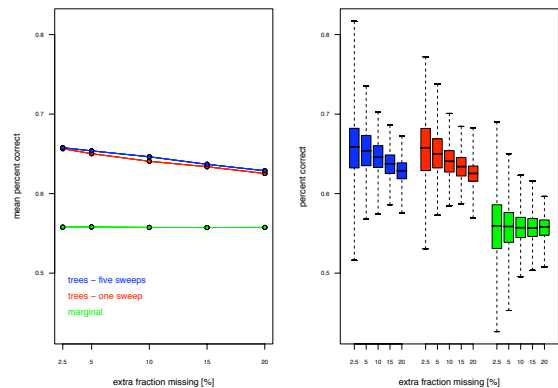
$$\vdots$$

$$M_p^{(n+1)} \sim \Pr(M_p | M_1^{(n+1)}, M_2^{(n+1)}, \dots, M_{p-1}^{(n+1)}, \mathbf{C}, \mathbf{D}).$$

where each full conditional distribution is modeled by CART.

→ A convenient property of surrogate splits in CART is that we do not have to guess the initial values of the missing data in \mathbf{M} . As a result only a very short burn-in of the above sampler is required.

Simulation (an example)



Take Home Message

- Do something about the missing data!
 - They can introduce bias, and reduce the power in the analysis.
- Only using the marginal distributions for imputation doesn't count!
- Haplotype based imputation slightly beats the tree based imputation if the genotype / phenotype relationship is best described by haplotypes, and you have few or no environmental variables.
- Tree based imputation is fast, accepts many environmental variables, and does remarkably well given its simplicity.
- Always take the response variable into account!

Detecting SNP-SNP Interactions

Overview

Many statistical approaches have been used in the literature for the analysis of SNP-SNP or SNP-environment interactions:

- Logistic Regression with Higher Order Interactions
- Multifactor Dimensionality Reduction (MDR)
- Classification and Regression Trees (CART)
- Random Forests
- Boosting
- Multivariate Adaptive Regression Splines (MARS)
- Neural Networks
- Logic (Boolean) Regression and Monte Carlo Logic Regression

References:

- Heidema AG et al (2006). The Challenge for Genetic Epidemiologists: How to Analyze Large Numbers... *BMC Genetics* 7: 23.
- McKinney BA et al (2006). Machine Learning for Detecting Gene-Gene Interactions: A Review. *Applied Bioinformatics* 5(2): 77-88.
- Musani SK et al (2007). Detection of Gene x Gene Interactions in Genome-Wide Association Studies of... *Hum Hered* 63(2): 67-84.

Interactions

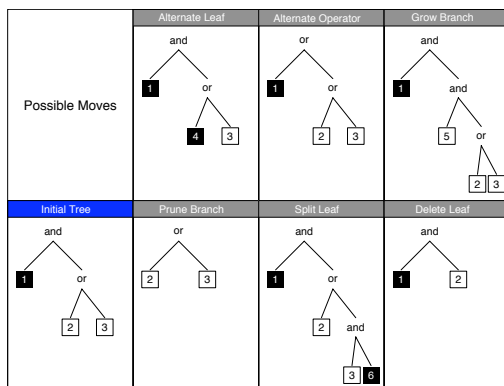
[Lucek and Ott]

“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.”

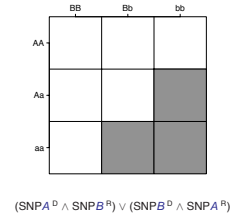
“... patterns of interactions between several loci, for example, disease phenotype caused by locus A and locus B, or A but not B, or A and (B or C), clearly make identification of the involved loci more difficult. While the simultaneous analysis of every single two-way pair of markers can be feasible,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

Reference: Lucek PR, Ott J (1997). Neural Network Analysis of Complex Traits. *Genetic Epidemiology* 14(6):1101-6.

The Move Set for Logic Regression



Biological and Statistical Interactions



→ Statistical interaction: Deviation from additivity in a linear statistical model.

→ Epistasis: Masking of phenotype expressed by one gene by the effects of another gene.

Reference: Moore JH (2005). A Global View of Epistasis. *Nature Genetics*, 37: 13-4.

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
• SNPs are coded as dominant and recessive:	AA	0	0
	AT	0	1
	TT	1	1

Reference:

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

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.

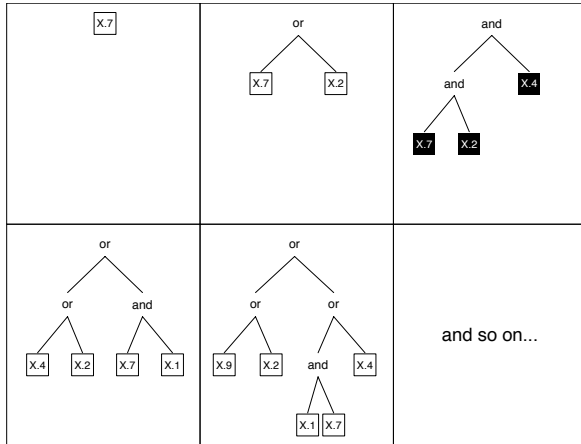
• Initialize the model with $L_j = 0$ for all j .

• Carry out the Simulated Annealing Algorithm:

→ Propose a move.

→ Accept or reject the move, depending on the scores and the temperature.

Growing Logic Models



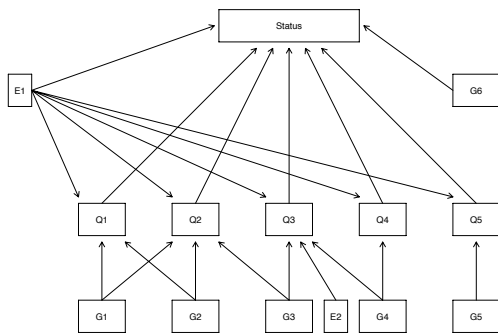
Model Selection

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 necessarily the best option 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.

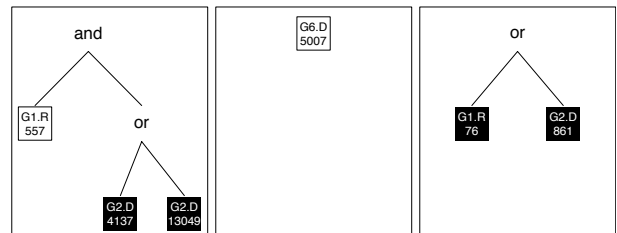
Reference:
Ruczinski I, Kooperberg C, LeBlanc M (2003). Logic Regression. *Journal of Computational and Graphical Statistics*, 12(3): 475-511.

Genetic Analysis Workshop GAW 12



Genetic Analysis Workshop GAW 12

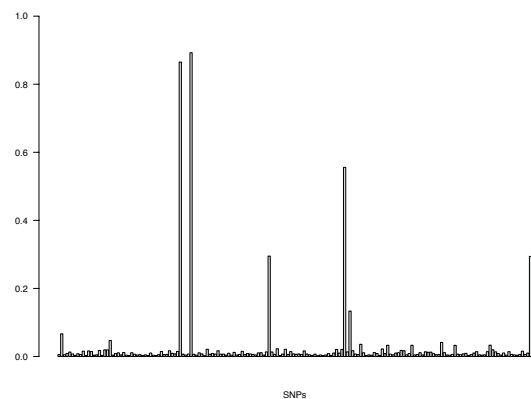
$$\text{logit}(\text{affected}) = \beta_0 + \beta_1 \times \text{ENV}_1 + \beta_2 \times \text{ENV}_2 + \beta_3 \times \text{GENDER} + \sum_{i=1}^K \beta_{i+3} \times L_i$$



Monte Carlo Logic Regression

- 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.

Monte Carlo Logic Regression



Reference:
Kooperberg C, Ruczinski I (2005). Identifying Interacting SNPs using Monte Carlo Logic Regression. *Genetic Epidemiol.*, 28(2): 157-70.