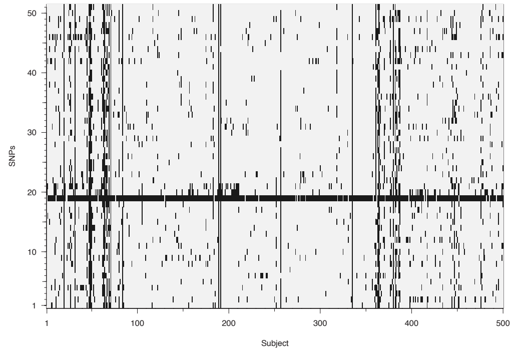
October 20, 2006 @ Georgetown University

On Missing Data and Interactions in SNP Association Studies

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Missing Data



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, Finkle WD (1995). A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses. *American Journal of Epidemiology*, 142 (12): 1255-64.

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.

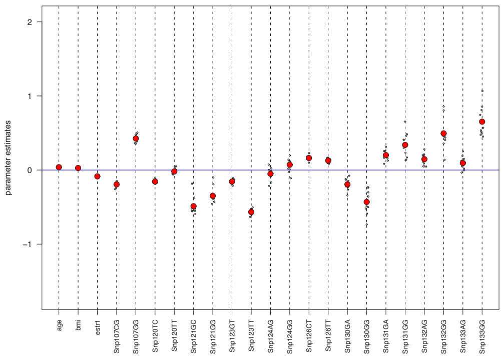
References:
- Little RJ, Rubin DB (1987). Statistical Analysis with Missing Data. *John Wiley Sons*, New York.
- Schafer JL (1997). Analysis of Incomplete Multivariate Data. *Chapman & Hall.*

Not Missing at Random

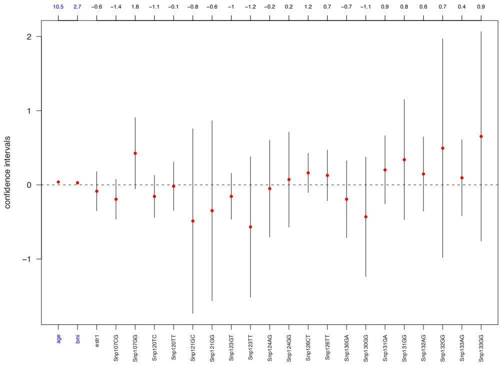
Method	Confidence Threshold	Overall Call Rate	Hom Call Rate	Het Call Rate
DM	0.26	**94.16%**	97.24%	86.32%
DM	0.33	**95.96%**	98.24%	90.16%
BRLMM	0.3	**97.40%**	97.40%	97.75%
BRLMM	0.4	**98.27%**	98.30%	98.48%
BRLMM	0.5	**98.79%**	98.82%	98.93%
BRLMM	0.6	**99.15%**	99.18%	99.25%

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

Multiple Imputation



Multiple Imputation



Example 1

- **The CLUE II Population:** A nested case-control study was conducted using the population-based CLUE II cohort, established in 1989. Individuals residing in Washington County, Maryland, and surrounding regions were invited to donate blood for cancer and heart disease research.

The CLUE II cohort consists of 32,892 individuals, including approximately 30% of county residents. Cancers that developed among cohort participants were ascertained through linkage to the Washington County and, since 1992, the Maryland State Cancer Registries. In 1996, and about every two years afterwards, participants were asked to complete a follow-up questionnaire asking about health events, medication use, and cancer risk factors.

Example 1

- **The Study Population:** For this study, 321 cases of breast cancer that occurred after blood donation to CLUE II in 1989 were identified. Cases were excluded if they had a diagnosis of any other cancer except for non-melanoma skin cancer and cervical cancer in situ. Controls were individually matched to cases (1:1) by age at blood donation and menopausal status at blood donation. Selected controls were cancer-free.

- **Family History:** Information on breast cancer risk factors was obtained from several sources, including a questionnaire that was sent in 1995 to a portion of the CLUE II breast cancer cases and controls who were part of a case-control study on organochlorine compounds. In addition, a 1996 follow-up questionnaire that was sent to all CLUE II participants. The questionnaires contained detailed information on family history of breast cancer, reproductive history, medication history, and selective dietary intake.

Reference:
Brewster AM, Jorgensen TJ, Ruczinski I, Huang HY, Hoffman S, Thuita L, Newschaffer C, Lunn RM, Bell D, Helzlsouer KJ (2006). Polymorphisms of the DNA Repair Genes XPD (Lys751Gln) and XRCC1 (Arg399Gln and Arg194Trp): Relationship to Breast Cancer Risk and Familial Predisposition to Breast Cancer. *Breast Cancer Research and Treatment*, 95(1): 73-80.

Example 1 - Statistical Inference

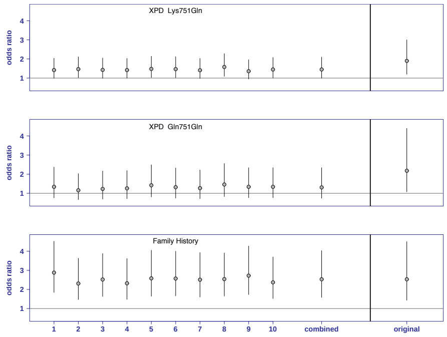
	Number of Pairs	Odds Ratio	Confidence Interval
		XPD Lys751Gln	
original data set	202	1.90	(1.20 – 3.00)
multiple imputations	321	1.45	(1.00 – 2.10)
		XPD Gln751Gln	
original data set	202	2.18	(1.08 – 4.40)
multiple imputations	321	1.31	(0.74 – 2.34)
		Positive Family History	
original data set	202	2.53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

Example 1 - Missing Data Patterns

The polymorphisms of the DNA repair gene XPD (751) for case/control pairs and family history reporting status, in the breast cancer study. This highlights the fact that missing data can not simply be ignored.

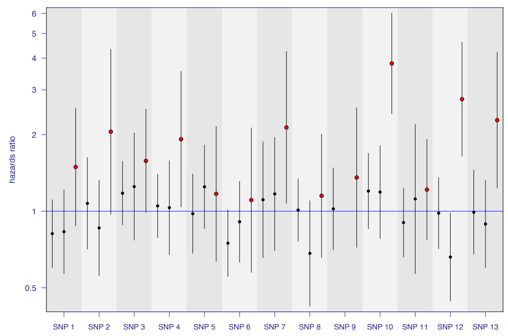
	Family History not complete				Family History complete			
	AA	AC	CC	na	AA	AC	CC	na
	raw numbers							
case	43	54	5	5	61	121	25	7
control	35	57	12	3	90	102	22	0
	percentages							
case	40.2	50.5	4.7	4.7	28.5	56.5	11.7	3.3
control	32.7	53.3	11.2	2.8	42.1	47.7	10.3	0.0

Example 1 - Imputation



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

Example 2



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, Ruczinski I, LeBlanc M, Kooperberg C (2006). A Comparison of Haplotype-based and Tree-based SNP Imputation in Association Studies. *Genetic Epidemiology*, (in press).

Tree-based Imputation

- For each individual i, let $\mathbf{M}_i = (M_{i1}, M_{i2}, \ldots, M_{ip})$ be the vector of p variables consisting of the covariates $\mathbf{X}_i = (x_{i1}, \ldots, x_{ir})$ and the unphased SNP data $\mathbf{G}_i = (g_{i1}, \ldots, g_{ik})$ which have missing entries ($1 \le p \le r + k$).
- Let $\mathbf{C}_i$ be the vector of the remaining covariates and unphased SNP data for which all data are available. We assume that the outcome $\mathbf{D}_i$ is always observed.
- The joint probability distribution of the missing data for individual i given the observed data, Pr$(M_{i1}, M_{i2}, \ldots, M_{ip}|\mathbf{C}_i, \mathbf{D}_i)$, is difficult to get. An obvious problem is that the sets of missing data $\mathbf{M}_i$ and complete data $\mathbf{C}_i$, respectively, are different for each individual i.
- Instead of modeling the joint distribution, we use the Gibbs sampler, a Markov chain Monte Carlo technique that uses conditional (low-dimensional) distributions to draw samples from a high-dimensional distribution.

Tree-based Imputation

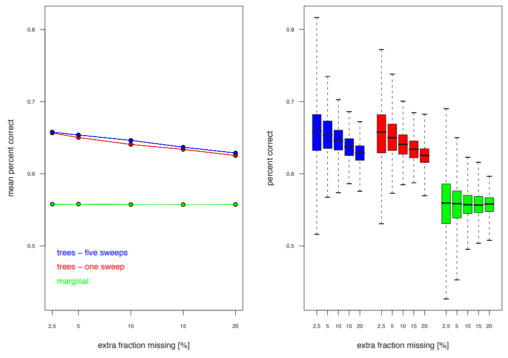
Specifically, we consider iteratively sampling from the following sequence of the full conditional distributions in the $(n+1)^{\text{th}}$ iteration:

$$\begin{aligned}
M_1^{(n+1)} &\sim \text{Pr}(M_1|M_2^{(n)}, M_3^{(n)}, \ldots, M_p^{(n)}, \mathbf{C}, \mathbf{D}) \\
M_2^{(n+1)} &\sim \text{Pr}(M_2|M_1^{(n+1)}, M_3^{(n)}, \ldots, M_p^{(n)}, \mathbf{C}, \mathbf{D}) \\
&\vdots \\
M_p^{(n+1)} &\sim \text{Pr}(M_p|M_1^{(n+1)}, M_2^{(n+1)}, \ldots, M_{p-1}^{(n+1)}, \mathbf{C}, \mathbf{D}).
\end{aligned}$$

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 1



Simulation 2

Haplotype frequencies (see *Kraft et al, 2005*) used in this simulation study. Haplotype **1000** is associated with the disease outcome.

Haplotype	Frequency
0000	0.3265
0001	0.1327
0100	0.0306
0101	0.0408
1000	**0.1633**
1010	0.0408
1100	0.0204
1110	0.2449

We added a signal through a logistic penetrance function:

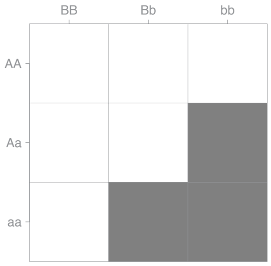
$$\text{logit}(\text{Pr}(\mathbf{D} = 1|\mathbf{H})) = -3 + \beta \cdot (\text{number of copies of } h_{1000})$$

Simulation 2 - Results

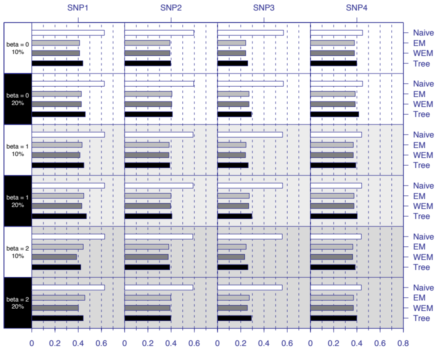
Mean imputation errors in the simulated data of four SNPs on the PGR gene for four imputation approaches:

		10% missing data				20% missing data			
	Approach	SNP1	SNP2	SNP3	SNP4	SNP1	SNP2	SNP3	SNP4
$\beta = 0$									
	Naive	0.625	0.596	0.568	0.449	0.625	0.595	0.567	0.449
	EM	0.412	0.390	0.243	0.379	0.427	0.407	0.271	0.385
	WEM	0.412	0.390	0.243	0.379	0.427	0.406	0.271	0.385
	Tree	0.440	0.397	0.260	0.399	0.461	0.411	0.292	0.415
$\beta = 1$									
	Naive	0.627	0.589	0.560	0.441	0.627	0.589	0.560	0.441
	EM	0.433	0.383	0.245	0.369	0.448	0.399	0.273	0.375
	WEM	0.415	0.381	0.241	0.369	0.431	0.396	0.269	0.375
	Tree	0.449	0.389	0.263	0.389	0.471	0.407	0.296	0.403
$\beta = 2$									
	Naive	0.628	0.587	0.557	0.438	0.627	0.588	0.557	0.438
	EM	0.443	0.380	0.246	0.365	0.457	0.397	0.273	0.371
	WEM	0.386	0.375	0.233	0.363	0.402	0.391	0.257	0.370
	Tree	0.422	0.388	0.262	0.385	0.443	0.398	0.292	0.399

Double Penetrance Model



Simulation 2 - Results



Logic Regression

- $X_1, \ldots, X_k$ 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.
- SNPs are coded as dominant and recessive:

SNP X	X.R	X.D
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.

Motivation

"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."

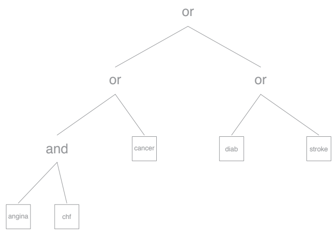
". . . patterns of interactions between several loci, for example, disease phenotype caused by locus A and locus B, or A but not B, or A and (B or C), clearly make identification of the involved loci more difficult. While the simultaneous analysis of every single two-way pair of markers can be feasible, 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.

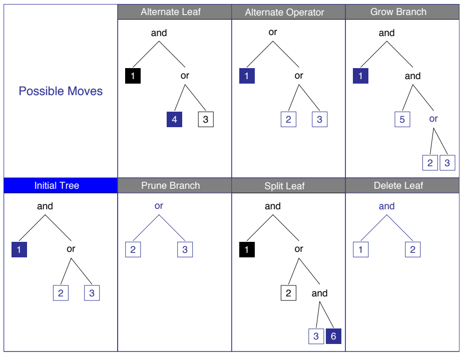
Example: The WHAS

$$p = \Pr(\text{death in round j} \mid \text{survival to round j-1, X, age})$$

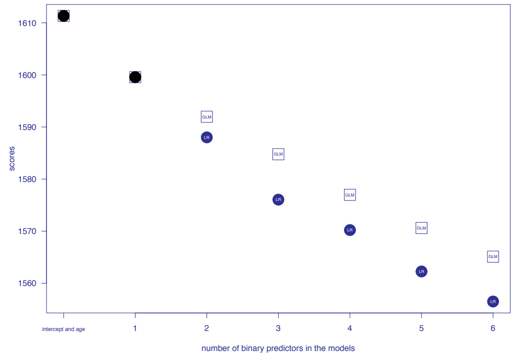
$$\text{logit}(p) = -9.01 + 0.06 \cdot \text{age} + 1.07 \cdot L(X)$$



The Move Set for Logic Regression



Example: The WHAS



Simulated Annealing for Logic Regression

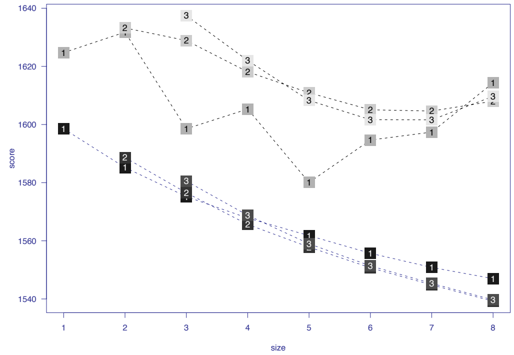
We try to fit the model $g(E(Y)) = b_0 + \sum_{j=1}^{t} b_j \cdot L_j$.

- 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:

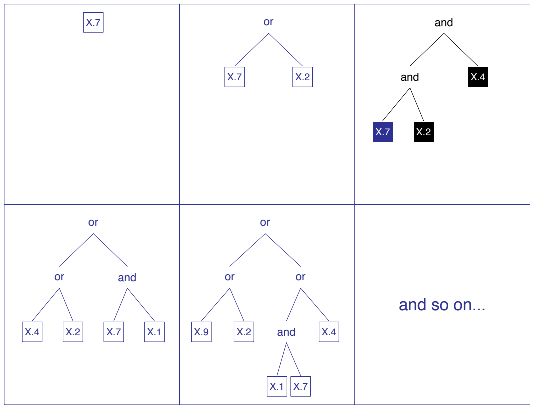
  → Propose a move.

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

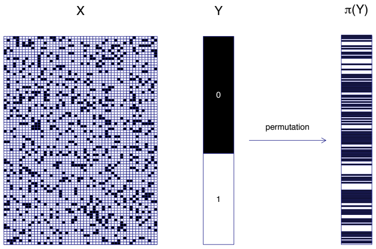
Model Selection 1 : Cross Validation



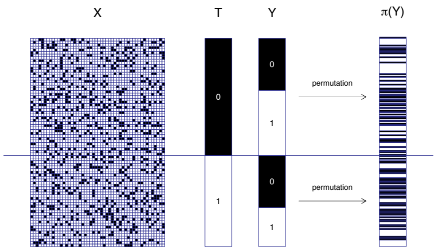
Growing Logic Models



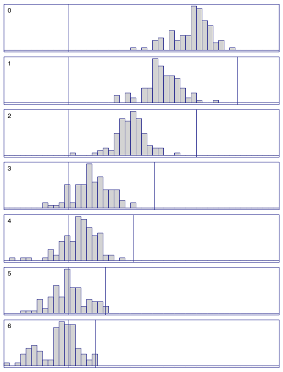
Model Selection 2 : Permutation Tests



Model Selection 2 : Permutation Tests



Model Selection 2 : Permutation Tests

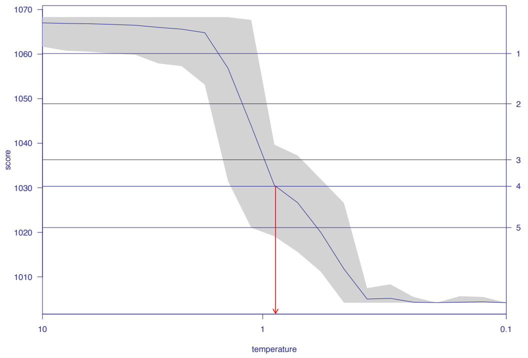


Multiple Models 1 : Monte Carlo LR

- 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.

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

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

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

- 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 $\pi_t(S)$.
- 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

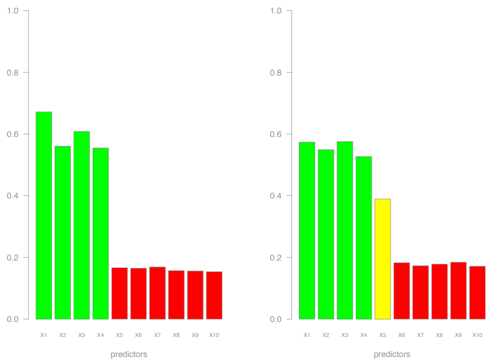
Example: Simulate 10 binary predictors $X_1, \ldots, X_{10}$.

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

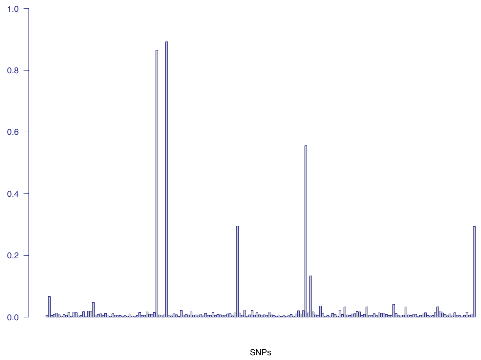
Run a homogeneous Markov chain during “crunch time” for two separate cases:

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