Bayesian Inference of Interactions and Associations

Jun Liu
Department of Statistics
Harvard University

http://www.fas.harvard.edu/~junliu

Based on collaborations with Yu Zhang, Jing Zhang, Yuan Yuan, Ke Deng, Zhi Geng

A segment of Chromosome 7 of two random individuals compared

2200 base pairs
Introduction

• Single Nucleotide Polymorphism (SNP)
  …ACAA…AGTCT….TAGACG…
  …ACCA…AGACT….TAACCG…

– Mostly SNPs are biallelic
– About 10 million “common” SNPs with minor allele frequencies > 1%
– Cover the entire human genomes and Commonly used markers in genetics.

Fine Mapping of Disease Genes

• Genetic Disease
  – Genetic variants affect one’s susceptibility to certain disease
• Map genes related with disease
  – Association method using unrelated individuals is very powerful!

Two disease mutations. They may interact to increase disease susceptibility
Problem

• Given genotypes at multiple loci for both cases and controls, find most likely positions where a disease-related mutation may have occurred
  – Complex Disease:
    • Multiple mutations, low risks (1.2~1.3)
    • Espistasis, environmental exposure, individual parameters
  – Epistasis (multi-locus interaction):
    • Alleles at one locus “affect” the behavior of alleles at other loci
    • Examples: breast cancer (Ritchie et al. 2001)
      post-PTCA stenosis (Zee et al. 2002)
      essential hypertension (Williams et al. 2004)
      atrial fibrillation (Tsai et al. 2004)
      type 2 diabetes (Cho et al. 2004)

Detecting Interactions among unlinked markers

• Generalizing/simplifying existing models to handle genome-wide association study (with unlinked markers)
  – Diseased individuals may form distinct “haplotype patterns” among the disease-related markers.
  – Our Bayesian model attempts to infer such patterns by contrasting with the control individuals.
• Simulation Studies (haplotype types):
  – (a) 1000 cases, 1000 controls; 1000 candidate markers; 3 interacting markers; ~40% phenocopies
  – (b) 200 cases, 200 controls; 100 candidate markers; 6 interacting markers; ~60% phenocopies. A total of $2^6=64$ haplotype patterns

Assigned risk=5 to six patterns and risk=7.5 to one pattern
Methods for Detecting Epistasis

- Parametric modeling:
  - too many parameters, no sufficient information
- Non-parametric modeling:
  - Machine learning: complicated, work for small datasets
  - CART: Classification and Regression Trees (Breiman et al. 1984)
  - MARS: Multivariate Adaptive Regression Splines (Friedman 1991)
  - CPM: Combinatorial Partitioning Method (Nelson et al. 2001)
  - RPM: Restricted Partitioning Method (Culverhouse et al. 2004)
  - MDR: Multifactor Dimension Reduction (Ritchie et al. 2001)
  - Monte Carlo Logic Regression (Kooperberg and Ruczinski, 2005)
  - BGTA: Backward Genotype-Trait Association (Lo et al. 2005)
  - and More...

- computationally very expensive!
- over-fitting, sensitive to test data and new data
- multiple testing issue: False Discovery Rate (Benjamini, Hochberg 1995; Storey, 2002)

General: Regression and Classification

| Ind 1 | x_{11}, x_{12}, \ldots, x_{1p} | Y_1 |
| Ind 2 | x_{21}, x_{22}, \ldots, x_{2p} | Y_2 |
| \vdots | \vdots | \vdots |
| Ind N | x_{N1}, x_{N2}, \ldots, x_{NP} | Y_N |

\[ P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} \]

How to model this?
A digression: Naïve Bayes Classifier

\[ P(Y \mid X_1, \ldots, X_p) = \frac{P(X_1, \ldots, X_p \mid Y)P(Y)}{P(X_1, \ldots, X_p)} \propto P(Y)\prod_{j=1}^{p} P(X_j \mid Y) \]

Our approach: one step beyond NB
Our Approach: Beyond NB

- **Partition $L$ markers into three groups**
  - **Group 1:** $l_1$ markers have marginal effects only
    - Genotype frequencies are different between cases and controls
  - **Group 2:** $l_2$ markers have epistasis effect
    - Genotypes of markers are correlated, consider a vector of genotypes with unknown frequencies $\{\rho\}_{1..p}$
    - Different from multiplication of single marker frequencies $\rho = f_1 \times \cdots \times f_i$
  - **Group 0:** $L - l_1 - l_2$ markers have no association
    - Genotype frequencies are the same between cases and controls
    - 1st-order Markov chain to account for Linkage Disequilibrium

Another graphical Illustration
Generalization

• Modeling the covariates

For cases:
\[ P(X \mid Y = 1) = \int P(X \mid I_G) P(I_G \mid Y = 1) dI_G \]

\( G \) is a vector of indicators, taking values in \( \{0, 1, 2\} \)

For controls:
\[ P(X \mid Y = 0) = \prod_{j=1}^{p} P(X_j \mid Y = 0) \]

Probabilities of a grouping \( I \)

• Group 1:
\[ P(G_1^j \mid \Theta_1^j) P(\Theta_1^j) = \prod_{j=1}^{p} \prod_{k=1}^{N} P(g_{jk}^j \mid \Theta_1^j) P(\Theta_1^j) = \prod_{j=1}^{p} (\Theta_1^j)^{\alpha_j} P(\Theta_1^j) \]

Integrate out \( \Theta_1^j \):
\[ P(G_1^j) = \prod_{j=1}^{p} \left\{ \frac{\Gamma(N + |\beta_j|)}{\Gamma(N + |\beta_j|) + \beta_j} \right\} \]

• Group 2:
\[ P(G_2^j) = \prod_{j=1}^{p} \left\{ \frac{\Gamma(N + |\beta_j|)}{\Gamma(N + |\beta_j|) + \beta_j} \right\} \]

• Group 0:
\[ P(G_0^j, G^*) = \prod_{j=1}^{p} \left\{ \frac{\Gamma(N + |\beta_j|)}{\Gamma(N + |\beta_j|) + \beta_j} \right\} \]

Integrate out \( \Theta_1^j \):

\( n_k^j \): number of genotype \( k \) at marker \( j \) in group 1

\( n_k \): number of genotype vector \( k \) at markers in group 2

\( m_k \): number of genotype \( k \) at marker \( j \) in group 0 and controls
Markov Chain Monte Carlo Sampling

- Joint Likelihood: $P(G^d, G^*, I) = P(G^d_1 | I)P(G^d_2 | I)P(G^d_3 | G^* | I)P(I)$
  - $P(I)$: multinomial prior for the number of markers in each group

- Randomly assign markers to group 0, 1 or 2

- Update the marker membership and accept the change according to the Metropolis-Hastings Ratio
  - A quarter million iterations takes 3 minutes on P4-1.6GHz PC

- The output is a sample of markers from the posterior distribution
  - assess the significance of disease association based on the posterior density of markers in group 1 and 2

Simulation

- Model 1: two markers, marginal effects
- Model 2: a pair of interacting markers
- Model 3: threshold model
- Model 4: 3-interacting loci
- Model 5: two pairs of interacting loci
- Model 6: a 6-way interaction
Results

- Compare to and step-wise logistic regression and Chi-square
  B: The full Bayesian model;  S: Step-wise B-stat
  L: Step-wise logistic  C: Chi-square test

- 1,000 markers in N cases and N controls, where N = 1,000 (black bar) or 2,000 (grey bar)
- Power is averaged over 50 tests
- Type I error rate is at 0.1 with multiple correction
Compare to Other Methods

- MDR (Ritchie et al. 2001)
- Logic Regression (Kooperberg and Ruczinski 2005)
- BGTA (Zheng and Lo 2006)
- Chi-square (a single-marker approach)

Impact of MAF discrepancy

Figure 2: Impact of MAF discrepancy and LD on the powers of BEAM (B), the stepwise B-stat (S), the stepwise logistic regression (L) and the 2-d.f. $\chi^2$ test (C). The comparison is based on model 2, where the allele frequencies of the second disease locus are unmatched by that of the associated marker. The marginal effect size per disease locus is 0.5. Under each setting, the power is calculated from 50 data sets containing 1,000 markers genotyped from 1,000 cases and 1,000 controls. The power is the proportion of 50 data sets in which all associated markers are identified at a significance threshold of 0.1 after Bonferroni correction.
Analyzing the whole-genome AMD data

- From J. Hoh’s group (Klein et al. 2005)
- 116,204 SNP markers typed for 96 cases and 50 controls
- After filtering, 96,932 SNPs left for analysis
- We found the two markers reported (marginally significant), but no interactions
- We did further simulations using this data set

Posterior probabilities prior=10^{-3}
MCMC Convergence

AMD Data, 100K SNPs
Posterior for a simulated case

Simulating AMD-like data

- 500 cases and 500 controls, 100K SNPs
- With genotype frequencies and LD structures similar to the AMD data
- Insert interactions based on Models 2 & 4
- Both BEAM and Logistic regression runs about 5 hrs

Augmented Naïve Bayes Classifier

Basic setting for the classification problem:
- Y: class label (1,2,...,K)
- X: covariates (1,2,...,m)
  - Discrete valued

Difficulties
- Large number of covariates
  - Redundancy and colinearity would affect most classifiers
  - Variable selection is necessary
- Different classes have different associated covariates.
  - Methods that select one group of variables for all classes would work poorly

Naïve Bayes model

\[ P(Y|X_1, \ldots, X_m) = \frac{P(Y) \prod_{j=1}^{m} P(X_j|Y)}{P(X_1, \ldots, X_m)}. \]
Tree-Augmented Naïve Bayes

(Pearl 1988; Friedman 1997)

Augmented Naïve Bayes

Group 0

Group 1

Group 2
Classification

- Different classes are indicated by \( I \)'s
- Sample \( I \) from posterior distribution
- Calculate posterior probability ratio for each class \( k \) and put sample into the class with the highest ratio.

\[
\frac{P(Y_{test} = k | X_{test}, X, y)}{P(Y_{test} \neq k | X_{test}, X, y)} = \frac{P(Y_{test} = k | X, y) P(X_{test} | Y_{test} = k, X, y)}{P(Y_{test} \neq k | X, y) P(X_{test} | Y_{test} \neq k, X, y)}
\]

Simulation study

- 5 classes, \( p=(0.1, 0.1, 0.3, 0.2, 0.3) \)
- \( N=2000 \) samples
- \( m=200 \) covariates
- Each class is associated with 5 covariates (1 overlap, total 24)
  - Multinomial distributed with parameters randomly sampled from a Dirichlet distribution
Simulation study

- Naive Bayes: 49.1% (5-fold CV)
- Random forest: 47.95% (5-fold CV)
- CART: 55.2% (no CV)
- TAN: 55.65% (5-fold CV)

- Our method: 72% (5-fold CV)

Simulation study

- 22 of the 24 truly associated covariates have the highest posterior probability of selection in Group 1

- Some covariates are selected in Group 2 too. However, all of them are associated with other classes.
Some other real data

<table>
<thead>
<tr>
<th></th>
<th>NB</th>
<th>TAN</th>
<th>C4.5</th>
<th>ANB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast</td>
<td>97.4%</td>
<td>96.9%</td>
<td>94.7%</td>
<td>97.4%</td>
</tr>
<tr>
<td>Cleveland</td>
<td>82.8%</td>
<td>81.8%</td>
<td>73.3%</td>
<td>83.5%</td>
</tr>
<tr>
<td>Iris</td>
<td>93.3%</td>
<td>94%</td>
<td>94%</td>
<td>94%</td>
</tr>
<tr>
<td>Heart</td>
<td>81.5%</td>
<td>83.3%</td>
<td>81.1%</td>
<td>84.1%</td>
</tr>
<tr>
<td>Soybean</td>
<td>91.2%</td>
<td>92.2%</td>
<td>92%</td>
<td>91.6%</td>
</tr>
</tbody>
</table>

Another Example: HIV-1Drug

- Protease Inhibitors (PIs) target HIV-1 protease enzyme which is responsible for the posttranslational processing of the viral gag- and gag-pol-encoded poly proteins to yield the structural proteins and enzymes of the virus.
How to detect drug resistance Mutations

• Protease sequences from treated patients (949 cases)
  VVTIRGGQL#EALLDTGAD
  IVTIRGGQL#EALLDTGAD
  RVTRGGQL#EALLDTGAD

• Sequences from untreated patients (4146 controls)
  LVTIRGGQL#EALLDTGAD
  IVTIRGGQL#EALLDTGAD
  LVTIRGGQL#EALLDTGAD

Which ones contributes to drug resistance?
Drug resistance mutations

• The IAS-USA Drug Resistance Mutations list in HIV-1 updated in Fall 2006
• For IDV, mutations on the list are 10, 20, 24, 32, 36, 46, 54, 71, 73, 77, 82, 84, 90
• The ones we detect 10, 24, 32, 46, 54, 71, 73, 82, 90

Posterior plots
Interactions

• What is known:
The occurrence of changes at L10, L24, M46, I54, A71, V82, I84, L90 was highly significantly correlated with phenotypic resistance.
Minor mutations influence drug resistance only in combination with other mutations. 73 + 90, 32+47, 84+90, 48+54+82, 88+90,
Our results are consistent with above.

Part II

Themes discovery with generalized dictionary model
Example: market basket

- Analyze tables of transactions

<table>
<thead>
<tr>
<th>Customer</th>
<th>Basket</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>Chips, Salsa, Cookies, Crackers, Coke, Beer</td>
</tr>
<tr>
<td>C2</td>
<td>Lettuce, Spinach, Oranges, Celery, Apples, Grapes</td>
</tr>
<tr>
<td>C3</td>
<td>Chips, Salsa, Frozen Pizza, Frozen Cake</td>
</tr>
<tr>
<td>C4</td>
<td>Lettuce, Spinach, Milk, Butter</td>
</tr>
</tbody>
</table>

- Which items are frequently purchased together by customers?

Generalized dictionary model:
from sequences to combinations

- A set of basic “elements” \( \mathcal{E} = \{\omega_1, \ldots, \omega_K\} \)
- A theme dictionary \( \mathcal{D} = \{\alpha_1, \ldots, \alpha_n\} \), where each \( \alpha_i \subset \mathcal{E} \).
- A sequence is generated by drawing themes independently with theme-specific probabilities

Example:

\[
P = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 \\
AB & CD & A & B & C & D \\
\end{bmatrix}
\]

\[
\text{Probability } p_{AB} p_{CD} p_{A} p_{B} p_{C} p_{D}
\]

Under this model, the likelihood function of sentence \( S = \{\alpha_{s_1}, \ldots, \alpha_{s_k}\} \) is

\[
P(S|p) = \prod_{\omega \in S} p_\omega \times \prod_{\omega \in D / S} (1 - p_\omega) = \prod_{\omega \in S} \frac{p_\omega}{1 - p_\omega} \times \prod_{\omega \in D / S} (1 - p_\omega).
\]
Application

text mining in *The Stone Story*

- 108,296 sentences and 4,502 Chinese characters are involved
- Mean length of sentences is 6.72

Application II (cont.)

(4937 themes found)

text mining in *The Stone Story*

<table>
<thead>
<tr>
<th>Group I</th>
<th>Group II</th>
<th>Group III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relationships among characters</td>
<td>Important places</td>
<td>Important characters</td>
</tr>
<tr>
<td>玉王争霸</td>
<td>贾贾母政</td>
<td>玉王争霸</td>
</tr>
<tr>
<td>贾贾母政</td>
<td>贾贾母政</td>
<td>贾贾母政</td>
</tr>
<tr>
<td>贾贾母政</td>
<td>贾贾母政</td>
<td>贾贾母政</td>
</tr>
<tr>
<td>贾贾母政</td>
<td>贾贾母政</td>
<td>贾贾母政</td>
</tr>
<tr>
<td>贾贾母政</td>
<td>贾贾母政</td>
<td>贾贾母政</td>
</tr>
<tr>
<td>贾贾母政</td>
<td>贾贾母政</td>
<td>贾贾母政</td>
</tr>
<tr>
<td>贾贾母政</td>
<td>贾贾母政</td>
<td>贾贾母政</td>
</tr>
<tr>
<td>贾贾母政</td>
<td>贾贾母政</td>
<td>贾贾母政</td>
</tr>
<tr>
<td>贾贾母政</td>
<td>贾贾母政</td>
<td>贾贾母政</td>
</tr>
<tr>
<td>贾贾母政</td>
<td>贾贾母政</td>
<td>贾贾母政</td>
</tr>
</tbody>
</table>
Acknowledgement

- Yu Zhang (Penn State U, Statistics Dept)
- Wei Zhang (Harvard Statistics)
- Jing (Maria) Zhang (Harvard Statistics)
- Yuan Yuan
- Ke Deng
- Zhi Geng
- Josephine Hoh