Supplement: Bayesian Pedigree Analysis using Measure Factorization

Alexandre Bouchard-Côté
Statistics Department
University of British Columbia
bouchard@stat.ubc.ca

Bonnie Kirkpatrick
Computer Science Department
University of British Columbia
bbkirk@cs.ubc.ca

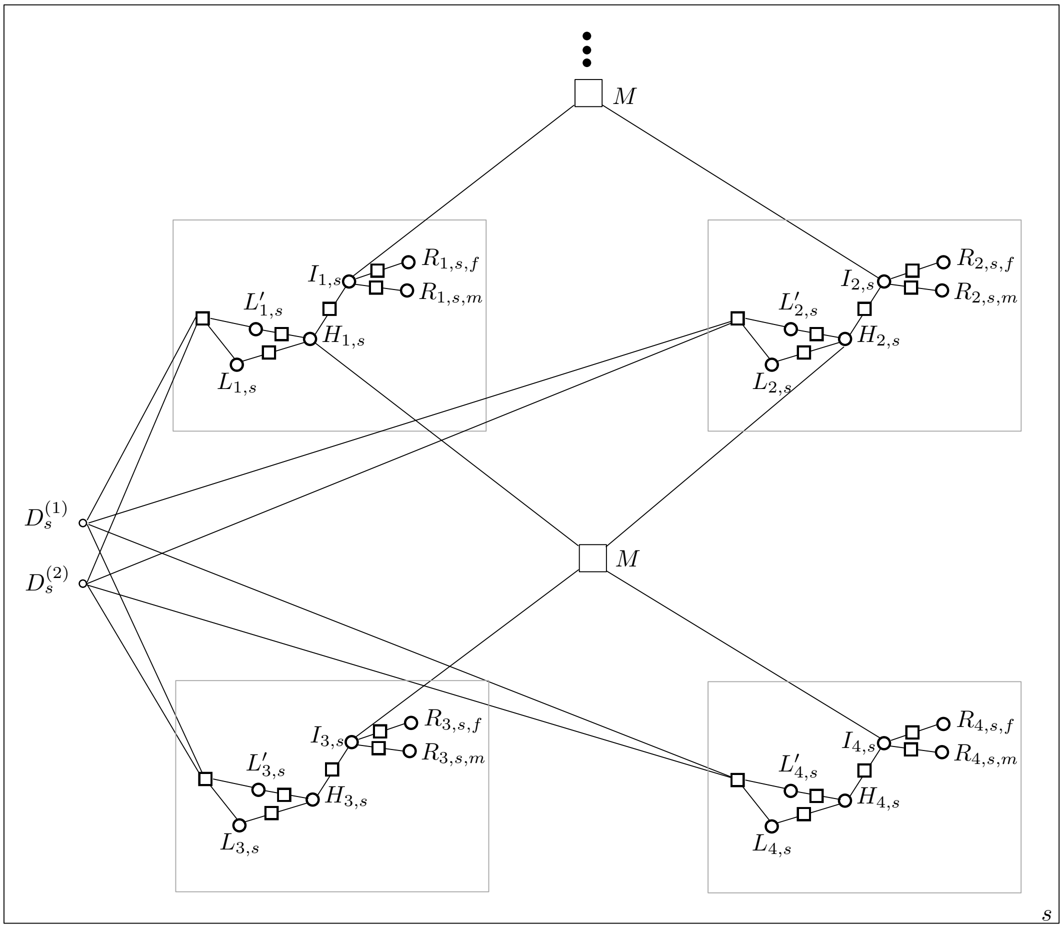

Figure 1: This figure illustrates the graphical model for a pedigree with independent sites s. The light-gray boxes are individuals. The nodes are labeled as follows: $R_{\cdot,f}$ and $R_{\cdot,m}$ for the two recombination indicators, I for inheritance, H for haplotype, L and L' for the two alleles, M for the marriage node, $D^{(1)}$ for the disease site indicator, and $D^{(2)}$ for the disease allele value. For dependent sites, there needs to be additional factors connecting to each $R_{i,s}$ and $R'_{i,s}$.

The pedigree graphical model is shown in Figure 1. The corresponding Figure 2 shows the pedigree graph and the data associated with it. In Figure 2(a), the pedigree corresponds to the four light-gray boxes of Figure 1 and contains a mother, father, and the two sons that are brothers. For all four individuals the figure shows the genotypes and haplotypes (vertical) inside the box or circle for each

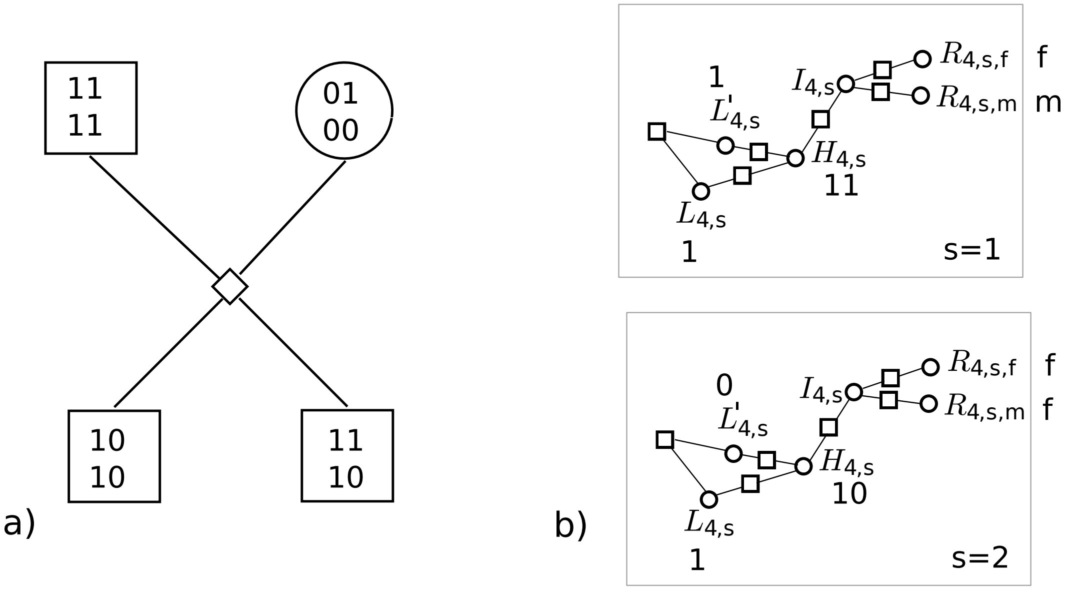

Figure 2: This figure illustrates both the pedigree and the data for the graphical model. (a) This is the pedigree fragment shown in detail in Figure 1. The genotypes are given horizontally and the haplotypes are given vertically in the node for each individual. The haplotypes are determined by Mendelian inheritance. (b) This shows the data.

individual. In this case the haplotypes of the children are determined by the rules of Mendelian inheritance. In Figure 2(b), for individual 4, the second son, we show how the data is translated into sampled assignments for the variables in the graphical model. The two loci are shown in the light-gray box as they appear in the graphical model, but without the connections to other individuals.

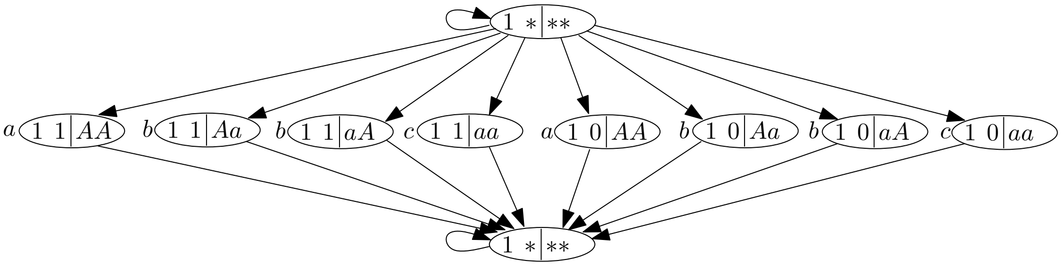

Figure 3: This figure shows the composition of the two transducers for D and R. Each oval contains four symbols, the two on the left are the disease state and the disease allele value, while the two on the right are the ordered genotype. This is the state space for a Markov chain, where the starting state is any state in the upper two layers and the ending state is any state in the bottom two layers. The weights a, b, c at the nodes correspond to a weight on the emission probability. For a sick individual, they are set as follows $a = f_2$, $b = f_1$, and $c = f_0$. For a healthy individual, the values are $a = 1 - f_2$, $b = 1 - f_1$, and $c = 1 - f_0$.

The composition of the transducers is shown in Figure 3

1 Experiments

Haplotype Reconstruction. The founder haplotypes were drawn i.i.d. at each site with a founder allele frequency of $1/2$. We simulated 150 centiMorgans with 1000 equally-spaced sites. Each experiment was replicated 10 times where for each replicate the founder haplotypes were sampled with a different random seed. This seed changes both the founder haplotype distribution and the recombination breakpoints for inheritance.

We compute the following metric. For each individual with missing genotype data that is not a founder or a child of a founder, denoted by set N, look at the chromosome that is inferred to have been inherited from the father (mother). Compare the held-out haplotype information, $S(i,s,x)$ for individual i, site s and chromosome x, to the inferred haplotype. Let m be the number of sites, and for each site in the haplotypes, sum the indicators of whether the two haplotypes are different at that site

$$\phi = \frac{\sum_{i \in N} \sum_s \sum_x \mathbb{I}\{H(i,s,x) \neq S(i,s,x)\}}{2|N|m}.$$

Disease Prediction. We took the haplotype distribution from the combined HapMap [1] populations JPT+CHB where each phased haplotype from the 2009-02 Phase III release contributed equally to the empirical distribution on full haplotypes. We simulated the first $10,000$ SNPs of Chromosome 1, sampling every 10th SNP as a site for our model. The recombination probabilities were taken from the physical distance between the sampled SNPs, Δx, using the conversion $1-\exp(k\Delta x)$ to genetic distance where $k = -9 \times 10^{-9}$.

Each experiment was replicated 10 times where for each replicate the founder haplotypes were sampled with a different random seed. Since the founder haplotype distribution is fixed, only the haplotypes and recombination breakpoints are resampled.

We compute the following metric. Both our Bayesian method and Merlin produce a list of numbers scoring each site. We sort each list and take as a metric $\psi = c_d + (c'_d - 1)/2$ where c_d is the number of sites scoring higher than the true disease site in the list and c'_d is the number of sites scoring equal to the true disease site.

References

[1] The International HapMap Consortium. The international HapMap project. *Nature*, 426:789–796, 2003.