A PENALIZED LIKELIHOOD FRAMEWORK FOR HAPLOTYPE PROBABILITY ESTIMATION

Paul H.C. Eilers[†], Hae-Won Uh

Leiden University Medical Center, Leiden, The Netherlands

[†] E-mail: *p.eilers@lumc.nl*

Haplotypes, the configuration of alleles of SNPs on each chromosome of a pair, can be coded as binary vectors, representing presence or absence of the reference allele. We cannot observe haplotypes directly, but only the sum of the two vectors. Multiple haplotypes can give identical genotypes, so we are in a missing data situation. The EM algorithm and Bayesian methods have been proposed as solutions. Here we present an alternative, combining the composite link model of Thompson and Baker (1981) and penalized likelihood.

If we consider a diplotype, an ordered pair of haplotypes, having probabilities p_k and p_l, then, under the assumption of "random mating", the probability of the pair is $p_k p_l$. With s SNPs we have $n = 2^s$ possible haplotypes and $4^s = n^2$ possible diplotypes. If we code the diplotypes from 1 to n^2 and set $\beta = \log p$, we can write the diplotype probabilities as the vector $\gamma = \exp(X\beta)$. X is a matrix with $n - 2$ zeros and 2 ones each row, indicating the corresponding haplotypes (or $n - 1$ zeros and one 2, for the pairs of identical haplotypes). An indicator matrix C with 3^s rows and 4^s columns connects each diplotype to a genotype. We then can write:

$$\mu = E(y) = tC\gamma = tC\exp(X\beta),$$

where y gives the genotype frequencies, and $t = \sum_i y_i$. This is a variant of the CLM, with a Poisson response.

Thompson and Baker presented a weighted regression algorithm, very similar to the usual one for generalized linear models. In the haplotype problem it can become unstable, because some elements of p can be very small, leading to problems because some elements of β may become large negative numbers. To avoid this, we add a ridge penalty: $Pen = \kappa \sum_k (\beta_k - \alpha_k)^2$, where α is the logarithm of a (uniform or structured) prior distribution. We optimize κ using AIC.

The model is flexible and effective. It also offers a uniform approach to estimation of linkage disequilibrium and deviations from Hardy-Weinberg equilibrium. Because of the regression framework, covariates, case-control situations and the computation of standard errors and various sensitivity measures can be handled straightforwardly. The model can be extended to handle missing of uncertain data.