Modelling haplotype effects based on phylogeny

Maria L. Selle* (maria.selle@ntnu.no), Ingelin Steinsland*, Finn Lindgren† and Gregor Gorjanc‡

*Norwegian University of Science and Technology, †The University of Edinburgh, ‡The Roslin Institute

Conclusion

- Including the haplotype phylogeny when modelling haplotype effects improves estimates compared to assuming independent haplotypes, especially when few observations for specific haplotypes
- The proposed approach performs similarly to modeling haplotype effects using the mutation model

Background and aim

- Accurate estimation of haplotypes with low frequency is challenging
- Most mutations have no causal effect
- Leveraging similarities between haplotypes could improve estimation

1. Propose sparse latent hierarchical model for haplotype effects by leveraging phylogeny between haplotypes
2. Compare the proposed model with a model assuming independent haplotypes and the mutation model

The haplotype network model

Assume conditional independence between haplotypes

$$h_j | h_{p(j)} \sim \mathcal{N}(\rho h_{p(j)}, \sigma_h^2),$$

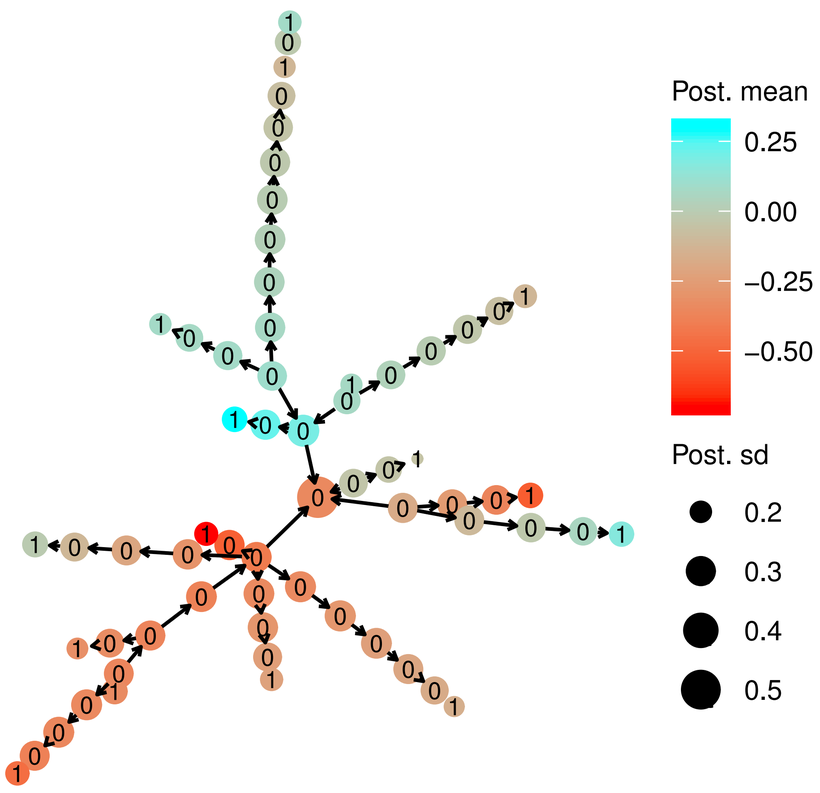
h_j haplotype effect one mutation away from parent haplotype $h_{p(j)}$. The common ancestral haplotype effect distributed as $h_{anc} \sim \mathcal{N}(0, \sigma_0^2)$,
$\sigma_h^2 = \sigma_0^2(1-\rho^2)$.

Joint density of $\boldsymbol{h} = (h_1, ..., h_n)^T$ Gaussian, $\boldsymbol{h}|\rho, \sigma_h^2 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}(\rho, \sigma_h^2)^{-1})$
Precision matrix $\boldsymbol{Q}$ sparse, and derived from the phylogeny

The dependency parameter, ρ Determines similarity between haplotypes
Prior distribution close to 1

Real data application

Posterior haplotype effects



Model

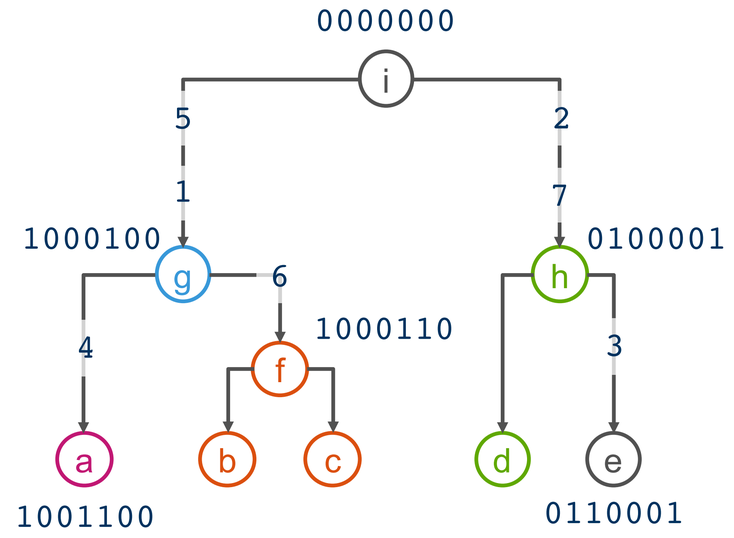
$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\gamma} + \boldsymbol{a} + \boldsymbol{Zh} + \boldsymbol{\varepsilon}$$

Data 381 cattle, milk yield as phenotype, information about age at calving, county, herd, year and season of calving Mitogenome haplotypes with phylogeny consisting of 63 unique haplotypes, where 16 of the haplotypes were observed in the cows

Inference INLA

Result Sharing of information between observed (1) and non-observed haplotypes (0)

Example phylogeny



Mutations uniquely identify haplotypes on which they appeared, which creates "network" known as genealogy or phylogeny

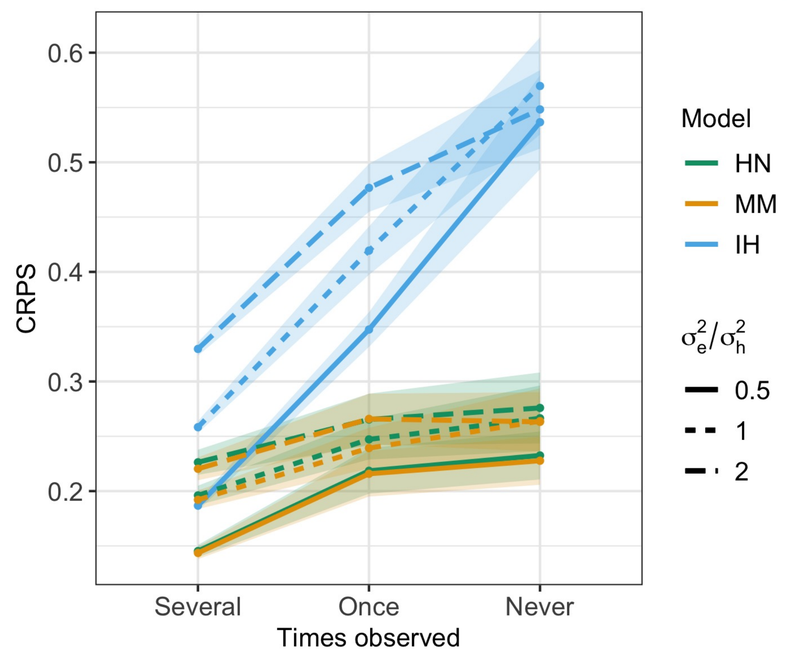
Simulation study

Compare models:

- Haplotype network (HN)
- Mutation model (MM), $\boldsymbol{h} = \boldsymbol{Uv}$,
  $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \sigma_v^2 \boldsymbol{I})$
- IID haplotype effects (IH),
  $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, \sigma_h^2 \boldsymbol{I})$

Results

- HN and MM similar in CRPS, and both better than IH
- Improvement largest when haplotypes observed only once or not at all



Simulated data from a mutation model with 10% causal variants, and varied the proportion of residual variance and haplotype variance

Extensions

- Extend to multiple phylogenies for different regions due to recombination
- Time as distance rather than mutation (Ornstein–Uhlenbeck process)
- Allow ρ to vary

Limitations

- Sparsity disappears if have polyploid individuals, or if much recombination
- Only focused on biallelic SNPs