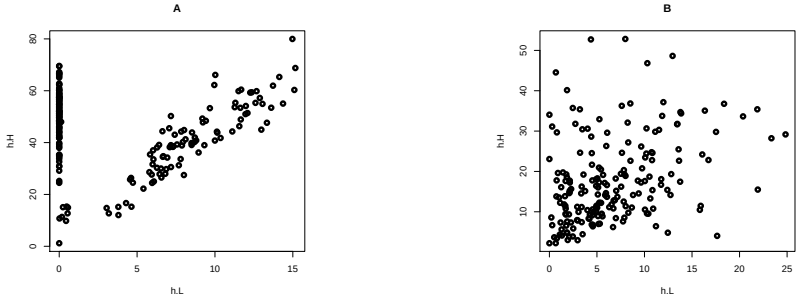
Finite Mixture Model Clustering of SNP Data

Norma Coffey[1,⋆], **John Hinde**[1], **Augusto Franco Garcia**[2]

1.National University of Ireland, Galway
2. Department of Genetics, ESALQ/USP, Piracicaba, Brazil
⋆Contact author: norma.coffey@nuigalway.ie

Keywords: Clustering, finite mixture models, orthogonal regression, SNPs.

Sugarcane is polypoid, i.e. has 8 to 14 copies of every chromosome, with individual alleles in varying numbers. It is therefore important to develop methods that identify the many different alleles and associated genotypes. One way of doing this is through the analysis of single nucleotide polymorphisms (SNPs). The frequency of a SNP base at a gene locus will vary depending on the number of alleles of the gene containing the SNP locus. Capturing this information accurately across several SNPs can give an indication as to the number of allele haplotypes present for a gene. Such information could have implications for sugarcane breeding since high yield potential may be due to the presence of and/or different number of copies of, a specific allele(s) present at a gene locus. The figures below display the data collected for two SNPs of the sugarcane plant. Each point represents the intensity of two SNP bases; h.L is the intensity of the A base, h.H is the intensity of the T base. The data in Figure A can clearly be clustered into two groups - the group along the y-axis and the group along the line with a particular (unknown) angle. These groups correspond to two genotypes and thus clustering is essential for genotyping. In Figure B it is not clear how many clusters (genotypes) are present and therefore it is necessary to develop a technique that can determine the number of clusters present, determine the angles between the clusters to identify different genotypes, and provide a probabilistic clustering to identify points that have high probability of belonging to a particular cluster (have a particular genotype) and those that are regarded as an unclear genotype.



The above criteria indicate that model-based cluster analysis techniques could be useful for analysing these data. However standard model-based cluster analysis techniques such as those implemented in the *R* package **mclust** (Fraley and Raftery, 2002) attempt to fit sperical/ellipsoildal components thus failing to cluster these data in an appropriate way and do not provide estimates of the angles between the clusters. To determine these angles it is necessary to fit a regression line to the data in each cluster. Using finite mixtures of linear regression lines is also inappropriate since it is not clear for these data which is the response variable and which is the explanatory variable. Problems are also encountered when attempting to fit a regression line to the group parallel to the y-axis in Figure A since this line has infinite slope in the usual least squares setting. As a result we propose to use finite mixtures of *orthogonal* regression lines to cluster the data, which ensures that using either variable as the response variable yields the same clustering results and that a regression line can be fitted to the group parallel to the y-axis. We implement this technique in *R* and show its usefulness in clustering these data.

References

Fraley, C. and A. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*, 611–631.