A comparison of two distance metrics and two clustering algorithms for population identification using genetic data

Daniel John Lawson*and Daniel Falush†

January 17, 2012

Abstract

We compare algorithms for clustering genetics data into their underlying population structure in a simulation setting. Firstly, we compare two summaries of the data, the Identity-By-State (IBS) matrix and the Coancestry matrix produced by the ChromoPainter software of Lawson et al. (2012), which additionally can allow either using or ignoring linkage information. Secondly, we compare the generic clustering algorithm MCLUST and the population-genetics tool fineSTRUCTURE. We find that clustering based on the Coancestry matrix is uniformly better than clustering based on the IBS matrix. There is also a significant advantage to using linkage information when it is present. We also find that MCLUST and fineSTRUCTURE have similar performance when the free parameters in MCLUST are chosen optimally, but are unable to find a general way to set these parameters. fineSTRUCTURE outperforms MCLUST when the MCLUST parameters are not optimally set. We discuss the benefits to each approach and recommend the use of both tools in different contexts.

1 Introduction

We generate linked simulated data as described in Lawson, Hellenthal, Myers, and Falush (2012) using the population demographic simulation program SFS_CODE (Hernandez 2008), under a scenario roughly equivalent to European population history. We consider a hierarchical splitting scenario where a single ancestral population (subject to a bottleneck) splits into three populations and undergoes exponential growth. Two of these populations subsequently split at later times,

*Department of Mathematics, University of Bristol, Bristol, BS8 1TW, UK

†Environmental Research Institute, University College Cork, Ireland and Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig Germany

leaving 5 populations. We sample 20 individuals for each population and consider how well we can infer their structure using varying lengths of genome. The genome is split into n regions each of length 5Mb.

For IBS calculations, we use PLINK (Purcell, Neale, Todd-Brown, Thomas, Ferreira, Bender, Maller, Sklar, de Bakker, Daly, and Sham 2007) to first thin the data to exclude rare SNPs (which is necessary for computational reasons, and we have checked on our smaller datasets that it does not change inference significantly). We then use PLINK again to compute the IBS matrix, and to compute the Principal Components Analysis (PCA) Eigenvector matrix (using its multidimensional scaling method, which is formally identical to PCA). We then keep the top d components ranked by their Eigenvalue, and pass this to MCLUST.

For Coancestry based calculations, we compute the linked and unlinked coancestry matrices using ChromoPainter as described in Lawson, Hellenthal, Myers, and Falush (2012), with further details of all commands are at `http://www.paintmychromosomes.com`, in the "Complex Example" section. We then compute the PCA Eigenvector matrix of the coancestry matrix using R (R Development Core Team 2009) as described in the supplementary material. As above, the top d components are ranked by their Eigenvalue when passed to MCLUST.

The fineSTRUCTURE inference procedure is described in Lawson, Hellenthal, Myers, and Falush (2012). The correlation with the truth for MCLUST output was performed similarly to the STRUCTURE and ADMIXTURE analysis described in that paper. The clustering is represented as a vector of membership for each of the K inferred populations with a 1 in the appropriate column, which is the same format as STRUCTURE. Specifically, for all observed populations we compute the correlation with all true populations and we report the mean (weighted by population size) of the correlation with the best matching true population. 'Correlation' is defined on the sets of elements found, which is preferable to correlating the binary N-vectors as the result is not dependent on sample size. Let l_i be the number of individuals in population i and n_{ij} be the number of shared individuals between populations i and j. The correlation is $(n_{ij}/(l_i + l_j - n_{ij}))^2$.

2 Results

MCLUST has a free parameter, the number d of Eigenvectors that are passed to the clustering algorithm. Too few, and there is not enough information to infer the correct clustering. Too many, and the algorithm is attempting to fit to noise. We wish to compare best possible results of MCLUST and therefore choose d to maximise the correlation with the truth. We consider data containing $r = (5, 10, 25, 50, 75, 100, 150, 200)$ independent genetic regions of length 5Mb. Figure 1 shows how the average correlation with the truth depends on d for

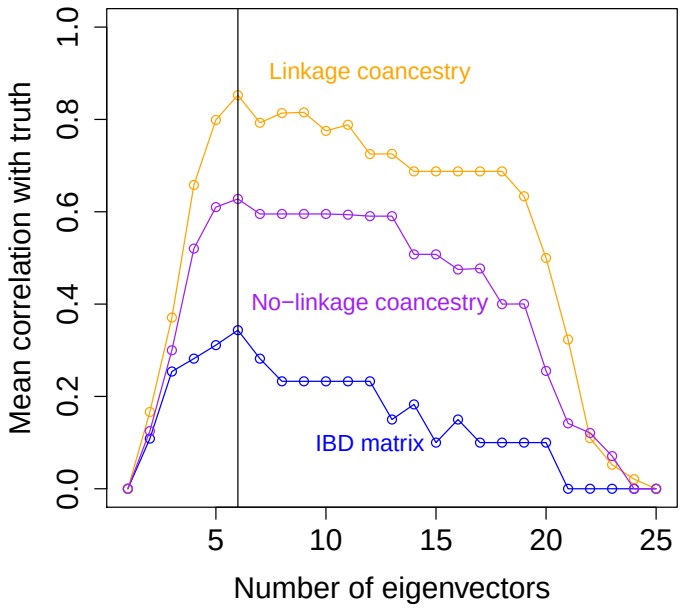

Figure 1: Correlation with truth as a function of the number of Eigenvectors retained for MCLUST analysis. $d = 6$ is shown as the optimal choice for all three matrices (linkage and no-linkage coancestry matrices produced by ChromoPainter, as well as the IBS matrix).

the three potential matrices (IBS, no-linkage coancestry and linkage coancestry). In all cases, the optimum value of d is at 6 components. The range of good values does depend strongly on the data matrix considered. MCLUST performs significantly better using linkage information than not, and significantly better using the coancestry matrix than the IBS matrix.

Continuing with the optimal choice of $d = 6$, we compare MCLUST to fineSTRUCTURE as a function of the number of data regions in Figure 2. On the linked coancestry matrix, fineSTRUCTURE and MCLUST perform almost identically. On the unlinked coancestry matrix, MCLUST performs similarly to fineSTRUCTURE although finds the optimum answer with 150 rather than 200 regions. MCLUST based on the IBS matrix performs somewhat similarly to ADMIXTURE, failing to find the correct solution with any amount of data.

If we are not able to make an optimal choice, we might have guessed d. Under such circumstances inference with MCLUST is significantly impacted; Figure 3 shows the range of correlations when (dense lines) $d \in [5, 8]$ and (sparse lines)

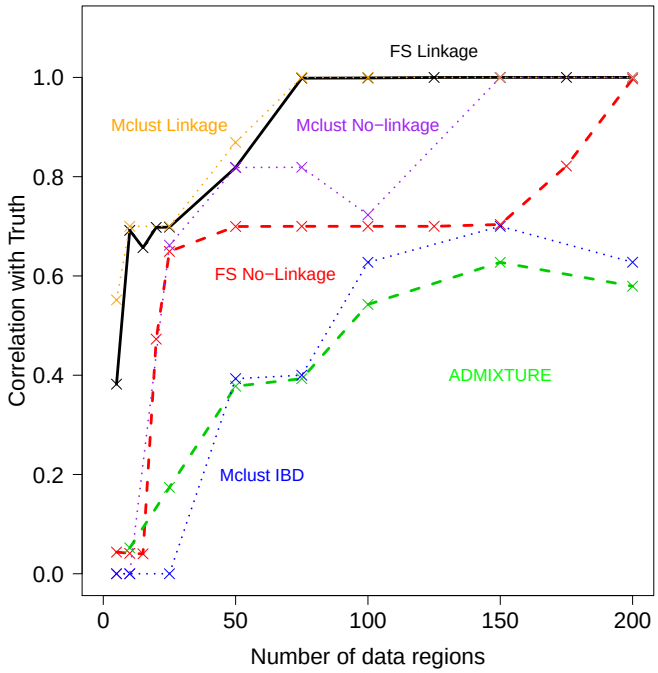

Figure 2: Correlation with truth as a function of the number of 5Mb simulated data regions when MCLUST performance is tuned to be optimal (using $d = 6$ from Figure 1). Also shown is the fineSTRUCTURE performance on the linkage and no-linkage coancestry matrices and the ADMIXTURE performance, repeated from Lawson, Hellenthal, Myers, and Falush (2012).

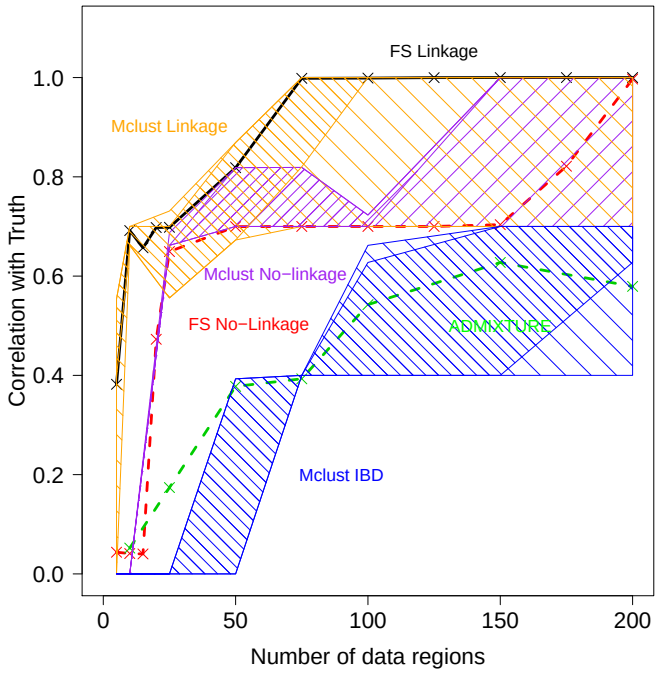

Figure 3: Correlation with truth as a function of the number of 5Mb simulated data regions when MCLUST performance is taken from a reasonable range of d (wide dashes: $d \in [5, 8]$, narrow dashes $d \in [4, 12]$). Other details are repeated from Figure 2.

$d \in [4, 12]$. We can potentially lose all value from the linkage information and can fail to infer the correct distribution at all. In the narrowest range, for the linked coancestry matrix fineSTRUCTURE has a small advantage and for the unlinked coancestry matrix the MCLUST advantage is significantly reduced. The non-monotonic curve of of MCLUST on the unlinked case implies that there may be some 'lucky guessing' for fewer than 100 regions, that increases the correlation but does not find very well resolved additional populations (although $K = 5$ is correctly inferred here, the final population split is not correct). fineSTRUCTURE, on the other hand, does not call the split until it is sure (and is able to correctly report any uncertainty).

We would wish to infer d from the data. A recent review (Shriner 2011) found that many methods over-fit (i.e. allow too many PCA components) when assumptions (such as Hardy-Weinberg equilibrium) are violated. They recommend a "minimum average partial" (MAP) test by Velicer (1976), but the generality and interpretation of such tests are always difficult to assess. We consider the MAP

test for the linkage and no-linkage coancestry maps for which it is easy to obtain data in a form suitable for use with pre-existing methods (it is implemented in package 'psych' in R). Unfortunately, the criterion does not always appear to have a well defined minima and is not consistent with the optimal choice (Figure 4). We note anecdotally that other methods implemented in this package also find that there is little ability to distinguish between a range of choices of d due to degeneracy, and no method tried consistently finds a good value. These estimates are of slightly higher variance that the larger limit of the range used in Figure 3, and are similarly bad at recovering the clusters (Figure 5).

3 Discussion

3.1 Comparison of ChromoPainter's coancestry matrix with the IBS matrix

The coancestry matrix of Lawson, Hellenthal, Myers, and Falush (2012) is uniformly better than the IBS matrix for the purpose of inferring population structure. Recall (from Proposition 1 therein) that the (unlinked) Coancestry matrix is created by weighting each shared SNP by the number of other individuals who share that SNP. This makes it suitable for the population structure problem since the relative importance of each shared SNP for determining recent ancestry is appropriately weighted. The IBS matrix is otherwise very similar in construction, but SNPs are not correctly normalised.

Propositions 1-4 of Lawson, Hellenthal, Myers, and Falush (2012) prove that the coancestry matrix is a sufficient statistic for the population assignment problem. The IBS matrix will not be, and therefore the coancestry matrix will be more suitable for many purposes. An additional benefit is that it generalises to the case of linked loci.

3.2 Advantage of utilising linkage information

The linkage information uniformly provides a significant improvement in the correct assignment of individuals to populations. The magnitude of this increase is not large, but it will always be significant when trying to disentangle very fine scale structure given a finite genome size. Although for practical reasons linkage information may be hard to exploit, where possible it should always be used.

3.3 MCLUST versus fineSTRUCTURE

Under optimum conditions (i.e. correct choice of the number of Eigenvectors to consider), MCLUST can match fineSTRUCTURE in its assignment, and does not overfit the number of populations. Since the likelihood of fineSTRUCTURE is

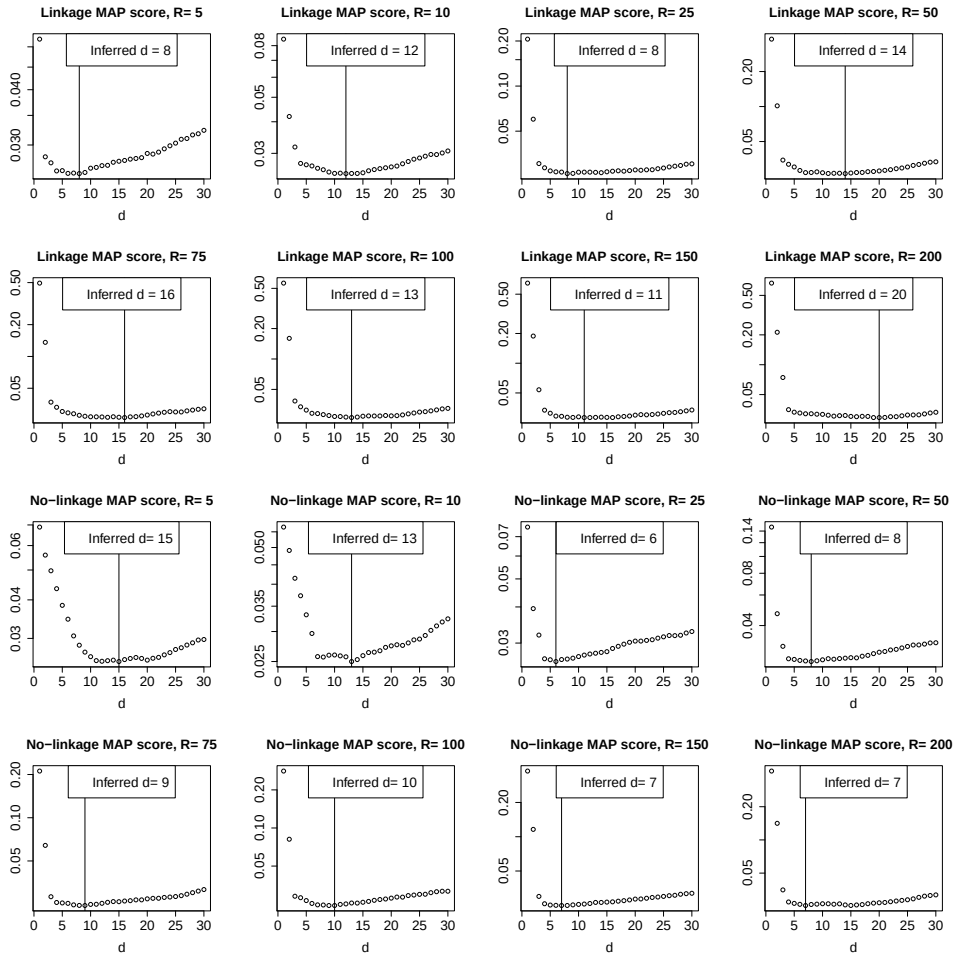

Figure 4: MCLUST inference of the optimal d using the MAP test of Velicer (1976), for a range of data quantities and for both the linkage and no-linkage coancestry matrices.

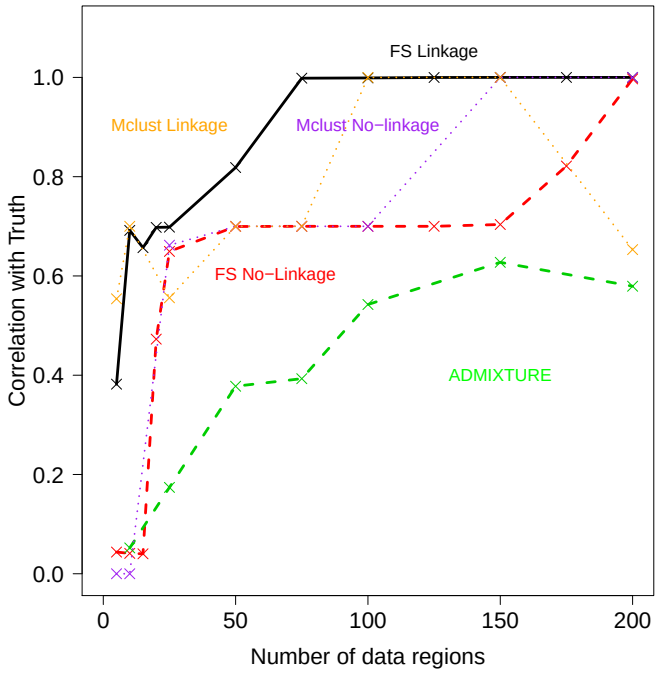

Figure 5: Correlation with truth as a function of the number of 5Mb simulated data regions when MCLUST performance is estimated according to Figure 4. Also shown is the fineSTRUCTURE performance on the linkage and no-linkage coancestry matrices and the ADMIXTURE performance, repeated from Lawson, Hellenthal, Myers, and Falush (2012).

asymptotically normal (Proposition 4 of Lawson, Hellenthal, Myers, and Falush 2012) the two models are very similar. Differences can arise by the use of a prior in fineSTRUCTURE, and from comparing a single maximum aposteriori state of MCLUST with an expectation of the MCMC over the posterior from fineSTRUCTURE. fineSTRUCTURE uses a conservative prior that favours merging two populations if they share correlated drift and are not strongly segregated (discussed in Lawson, Hellenthal, Myers, and Falush 2012).

Similarly, under optimum conditions, we cannot easily attribute a significant performance difference between ADMIXTURE and MCLUST performed on IBS data. There does not appear to be an underlying relationship between these two methods and the performance in both cases is probably dominated by the low signal-to-noise ratio.

The theory provided in Supplementary Section 4 of Lawson, Hellenthal, Myers, and Falush (2012)indicates that the distribution of the coancestry matrix should be multivariate normal under reasonable assumptions. The MCLUST likelihood is therefore a good choice as it is asymptotically correct. fineSTRUCTURE uses the fact that the multinomial distribution converges to the normal distribution and fixes the covariance matrix to simplify the inference problem, but apriori it is not clear which likelihood is to be preferred, and we observe in practise that the difference is not likely to be important. Therefore the choice of inference technique will depend not on the likelihood but on a) potential benefits of a well chosen prior, b) the search algorithm to explore clusterings, and c) computational efficiency.

MCLUST has several advantages. It is very fast, easy to use, and (conditional on choosing the correct number of retained components) it does not overfit. It also does not make any prior assumptions about the structure of the prior (which can be considered as both an advantage and a disadvantage). However, choosing the number of components to retain is a non-trivial problem that has been well studied but not generally solved (reviewed by Peres-Neto, Jackson, and Somers 2005). Out-of-the-box approaches such as attempted here do not appear to work well. Although better results can no doubt be achieved, for a dataset without a known truth it is hard to establish whether the value has been correctly chosen.

Conversely, fineSTRUCTURE accounts for uncertainty, allows for complex prior models, and has a very clear interpretation based on its biological model, which does not depend on any additional parameters. Potentially, selection of K in complex cases may be more accurate, as fineSTRUCTURE fully evaluates the posterior probability rather than e.g. using approximations such as BIC (although these simulations do not indicate any problems with BIC). Similarly, by using MCMC it can explore a much larger state space than a greedy algorithm, which may lead to better convergence for extremely difficult problems. This may be particularly important in the presence of admixture which may make greedy

algorithms converge to a sub-optimal solution. For computationally demanding problems, fineSTRUCTURE can be run in a greedy Maximum-aposteriori mode which should be fast enough for almost all datasets, though it is still not as fast as MCLUST. Finally, fineSTRUCTURE has some guarantee that it will not overfit; in all cases we have considered so far, it provides a close to optimal but conservative estimate of the number of clusters, and there are theoretical reasons (such as its conservative prior, and the extensive testing described in Lawson, Hellenthal, Myers, and Falush 2012) to back this up.

There are also downstream advantages to fineSTRUCTURE; we have spent a lot of effort on tools to interpret the clusters by building trees that make biological sense. These include merging individuals who are outliers in well understood ways, such as having close relatives in the sample, which breaks the assumption of normality in the likelihood. Such methods could be developed around a normally distributed likelihood model such as MCLUST but this would require further development.

3.4 Take home message

Overall, we find that ChromoPainter's coancestry matrices are significantly better than any competing "correlation based" summary of the data. Although MCLUST performs as well as fineSTRUCTURE under optimal conditions, the lack of a good method to estimate the number of Eigenvectors to use (which is itself an uncertain quantity) means that it cannot be used as a primary clustering tool. The additional computational burden of fineSTRUCTURE (relative to the very important ChromoPainter step for constructing the coancestry matrix) is not great, provides advantages of uncertainty assessment, conservative estimation, and a consistent modelling framework. However, since MCLUST potentially has the power to find the same features as fineSTRUCTURE there is significant value to exploring how MCLUST clusterings relate to fineSTRUCTURE's. If both methods find a particular population split this provides additional reassurance that it reflects a real feature of the data.

References

Hernandez, R., 2008 A flexible forward simulator for populations subject to selection and demography. Bioinformatics **24:** 2786–2787.

Lawson, D. J., G. Hellenthal, S. Myers, and D. Falush, 2012 Inference of population structure using dense haplotype data. PLoS Genetics**:** In Press.

Peres-Neto, P., D. Jackson, and K. Somers, 2005 How many principal components? Stopping rules for determining the number of non-trivial axes revisited. Comput Stat Data Anal. **49:** 974–997.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. Ferreira, D. Bender, J. Maller, P. Sklar, P. de Bakker, M. Daly, and P. Sham, 2007 PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics **81:** 559–75.

R Development Core Team, 2009 R: A Language and Environment for Statistical Computing. ISBN 3-900051-07-0.

Shriner, D., 2011 Investigating population stratification and admixture using eigenanalysis of dense genotypes. Heredity **107:** 413–420.

Velicer, W., 1976 Determining the number of components from the matrix of partial correlations. Psychometrika **41:** 321–327.