Model-Based Clustering using multi-allelic loci data with loci selection

Wilson Toussile, Elisabeth Gassiat

Model-Based Clustering using multi-allelic loci data with loci selection

Wilson TOUSSILE[*,a,b,c], Elisabeth GASSIAT[c]

[a] *UR016, Institut de Recherche pour le Développement (IRD)*
[b] *Ecole Nationale Supérieure Polytechnique (ENSP), Université de Yaoundé 1, Cameroun*
[c] *Laboratoire de Mathématiques d'Orsay (LMO), UMR 8628, Université Paris-Sud 11*

Abstract

A long standing issue in population genetics is the identification of genetically homogeneous populations. The most widely used measures of population structure are Wright's F statistics (Wright 1931). But the fundamental prerequisite of any inference based on these statistics is the definition of populations and this definition is typically subjective (based on linguistic, cultural or physical characters, geographical location). The population structure may be difficult to detect using visible characters.

We propose a Model-Based Clustering (MBC) method combined with loci selection using multi-allelic loci data. The loci selection problem is regarded as a model selection problem and models in competition are compared with the Bayesian Information Criterion (BIC). The resulting procedure selects the subset $\widehat{S}_n$ of clustering variables, the number $\widehat{K}_n$ of clusters, estimates the proportion of each population and the allelic frequencies within each cluster. We prove that the selected model $\left(\widehat{K}_n, \widehat{S}_n\right)$ converges in probability to the true model (K_0, S_0) under a single realistic assumption as the size n of the sample tends to infinity. The proposed algorithm named **MixMoGenD** ('Mixture Model for Genetic Data') has been implemented using $C++$ and C programming languages. An interface with **R** was created. Numerical

[*]Wilson Toussile, Laboratoire de Mathématiques d'Orsay, Bât 425, 91405 Orsay cedex, Tel : +33 (0) 1 69 15 76 18, +33 (0) 6 12 98 44 43

Email addresses: Wilson.Toussile@math.u-psud.fr (Wilson TOUSSILE), Elisabeth.Gassiat@math.u-psud.fr (Elisabeth GASSIAT)

URL: http://www.math.u-psud.fr/~toussile (Wilson TOUSSILE), http://www.math.u-psud.fr/~gassiat (Elisabeth GASSIAT)

experiments on simulated data sets was conducted to highlight the interest of the proposed loci selection procedure.

Key words: Model-Based Clustering, Model Selection, Variable Selection, BIC, Population Genetics

1. Introduction

Evolutionary processes have produced an immense array of biological diversity on our planet, with species displaying complex adaptations to their environments. Understanding this diversity and complexity, its origins, and its implications is a daunting challenge. Population genetics provides tools to meet this challenge. It is concerned with origin, amount, and distribution of genetic variation present in populations of organisms and fate this variation through space and time. A long standing issue in population genetics is the identification of genetically homogeneous populations. The most widely used measures of population structure are Wright's F statistics (Wright 1931). But the fundamental prerequisite of any inference based on these statistics is the definition of populations and this definition is typically subjective, based on linguistic, cultural or physical characters, or geographic location. The population structure may be difficult to detect using visible characters.

This article is concerned with population structure that is difficult to detect using visible characters (such as linguistic, cultural, physical characters, or geographic location), but may be significant in genetic terms. For example, the problem of cryptic population arises in the context of DNA fingerprinting for forensics, where it is important to assess the degree of population structure to estimate the probability of false matches (*D. J. Balding and Nichols RA* 1994 et 1995 [3], [2], [4]; *Foreman et al.* (1997) [9]).

Several Model Based-Clustering (MBC) for multi-locus genetic data have been developped in recent years: **STRUCTURE** by *J. K. Pritchard et al.* (2000) [15], **BAPS** by *J. Corander. et al.* (2004) [7], **FASTRUCT** by *Olivier Franois et al.* (2006) [10]. These methods attempt to group samples into clusters of random mating individuals so that the *Hardy-Weinberg* (HW) and linkage disequilibria (LD) are minimized across the sample. Although Hardy-Weinberg and linkage equilibria models are based on several simplifying assumptions that can be unrealistic, they have still proven to be useful in describing many population genetics attributes and will serve as a

useful base model in the development of more realistic models of microevolution. **STRUCTURE** and **BAPS** are bayesian methods that use MCMC algorithms and thus require much longer computations than frequentist likelihood methods using Expectation-Maximization (EM) algorithm [8].

Multi-locus data sets are becoming increasingly large due to the explosion of genomic projects. But, the structure of interest may be contained in only a subset of available loci, the others being useless or even harmful to detect a reasonable clustering structure. It then becomes necessary to select the optimum subset S_0 of loci which cluster in the best way the population. None of the methods cited above include a loci selection procedure. Defining the optimum set S_0 of loci and optimum number K_0 of subpopulations requires a suitable variable selection procedure.

In this article, we propose a new clustering method and an associated algorithm named Mixture Model for Genetic Data (**MixMoGenD**) that has three benefits. First, it is model-based. Second, it is based on EM algorithm, so it is relatively fast compared to its counterparts based on MCMC [10]. The main benefit of our proposed method is that it is coupled with a loci selection procedure based on Bayesian Information Criterion (BIC) and on a backward stepwise method. Recall that in clustering, classification is not observed and there is no a priori knowledge of the structure being looked for in the analysis, and of the subset of available loci that are relevant for discrimination. So there is no simple pre-analysis screening technic available to use. Thus it makes sense to include loci selection procedure as a part of the clustering algorithm as recommended by *C. Maugis et al.* (2007) [14] in a gaussian framework. The resulting procedure selects the subset S of clustering variables, the number K of clusters, estimates proportion and allelic frequencies within each cluster.

We recast loci selection problem and the estimation of the number of clusters as a model selection problem. Bayes factors, the ratio of integrated likelihood for models, are used to compare models, so that the models to be compared can be non-nested. Since integrated likelihood is usually difficult to compute, the Bayesian Information Criterion (BIC) for the competing models is used to approximate the log-likelihood. *Raftery et al.* (2006) [16] showed that compared to clustering methods based on all variables, variable selection method based on the BIC consistently yielded more accurate estimates of the number of clusters in a gaussian context. The proposed method can be applied to various types of markers (e.g. microsatellites, restriction frag-

ment length polymorphisms (RFLPs), or single nucleotide polymorphisms (SNPs)).

The model and methods are presented in section 2. The consistency of the estimators of the number of populations and the set of loci relevant for discrimination is proved in Section 3 under a single realistic assumption. The proposed algorithm has been implemented using $C++$ and C programming languages. In Section 4, Numerical experiments on simulated data sets was conducted to highlight the interest of the proposed loci selection procedure.

The program, sample project files and their simulation parameters, and documentation for linux OS are available free of charge at :
`http://www.math.u-psud/~toussile`.

2. Model and methods

In this section, we present the model and our clustering method using loci selection based on the Bayesian Information Criterion. The loci selection procedure is presented in sub-section 2.2, the identifiability of models in competition and of parameters are discussed in sub-section 2.4, and the EM equations are given in sub-section 2.3.

2.1. Notations and estimation method

The data set we shall deal with consists of genotypes of n diploid individuals labeled $1, \ldots, \ i, \ \ldots, \ n$ at L loci labeled $1, \ldots, \ l, \ldots, \ L$. The observations are written as $x = \left(x_i^l\right)_{i=1,\ldots,n;\ l=1,\ldots,L}$, where $x_i^l = \left\{x_{i,1}^l, \ x_{i,2}^l\right\}$ is the genotype of the i^{th} individual at the l^{th} locus. The data set x is assumed to be a realization of a random vector $X = \left(X_i^l\right)_{i=1,\ldots,n,\ l=1,\ldots,L}$, where $X_i^l = \left\{X_{i,1}^l, \ X_{i,2}^l\right\}$, with $X_{i,1}^l$ and $X_{i,2}^l$ taking values in the set $\{1, \ldots, \ l, \ldots, \ A_l\}$, and $1, \ldots, \ j, \ldots, \ A_l$ denote the labels of distinct alleles that are observed at locus l. We assume that the variables $X_i = \left(X_i^l\right)_{l=1,\ldots,\ L}$, $i = 1, \ldots, \ n$ are independent and identically distributed.

Let :

- z_i be the (unobserved) population of origin of individual i;
- $\pi_k := P\left(z_i = k\right)$ be the proportion of population k;
- $\alpha_{k,l,j} := P\left(X_{i,1}^l = j |\ z_i = k\right) = P\left(X_{i,2}^l = j |\ z_i = k\right)$ be the frequency of the j^{th} allele at locus l in population k;

- $\mathcal{X}$ be the set of possible genotypes from observed alleles;

and let $z = (z_1, \ldots,\ z_n)$, $\pi = (\pi_1, \ldots,\ \pi_K)$ and $\alpha = (\alpha_{k,l,j})_{k=1,\ldots,K;\ l=1,\ldots,L;\ j=1,\ldots,A_l}$. The π_k's are called the mixing proportions and represent the prior probability of an individual coming from each population k.

Our main modeling assumptions are

($\mathcal{H}_1$): Hardy-Weinberg Equilibrium (HWE) within populations and
($\mathcal{H}_2$): complete Linkage Equilibrium (LE) within populations.

Model-based methods proceed by assuming that observations from each cluster are drawn from some parametric model and the overall population is a finite mixture of these populations. Thus, without loci selection, observations $x = (x_1, \ldots,\ x_n)$ are supposed to be a sample from the probability distribution with the likelihood contribution of individual i given by the following equation

$$P_K\left(x_i|\ \theta\right) := P\left(x_i|\ K,\ \theta\right) = \sum_{k=1}^{K} \pi_k \left[\prod_{l=1}^{L} P\left(x_i^l|\ z_i = k,\ \alpha_{k,l,\cdot}\right)\right], \qquad (1)$$

where $\theta = (\pi,\ \alpha)$ is a parameter ranging in a certain space Θ_K, for a given number K of populations.

In this model of probability distributions, all the L loci are supposed to be relevant for clustering. Now, the structure of interest may be contained in only a subset S of available loci, the others being useless or even harmful to detect a reasonable clustering structure. Let S^c be the subset of loci that are irrelevant for clustering ($S \cup S^c = \{1, \ldots,\ L\}$). The natural third hypothesis is the following.

($\mathcal{H}_3$): the alleles of the loci of S^c are identically distributed in the overall population, i.e

$$\alpha_{1,l,j} = \alpha_{2,l,j} = \ldots = \alpha_{K,l,j} =: \beta_{l,j},\ \forall l \in S^c \text{ and } \forall j \in \{1, \ldots,\ A_l\}. \qquad (2)$$

The allele frequencies are given by the Hardy-Weinberg model:

$$P\left(x_i^l|\ z_i = k,\ \alpha_{k,l,\cdot}\right) = \left(2 - \mathbb{1}_{\left[x_{i,1}^l = x_{i,2}^l\right]}\right) \alpha_{k,l,x_{i,1}^l} \times \alpha_{k,l,x_{i,2}^l}. \qquad (3)$$

Although this model makes several simplifying assumptions that are unrealistic in some cases, it has still proven to be useful in describing many population genetics attributes and will serve as a first tool in the development of more realistic models of microevolution.

Under the three assumptions $(\mathcal{H}_1)$, $(\mathcal{H}_2)$ and $(\mathcal{H}_3)$, and given the number K of populations and the subset S of relevant loci, the observations are supposed to be realizations of a sample from a probability distribution of the form

$$\begin{aligned} P_{(K,\ S)}\left(x_i|\ \theta\right) &:= P\left(x_i|\ K,\ S,\ \theta\right) \\ &= \left[\sum_{k=1}^{K} \pi_k \prod_{l \in S} P\left(x_i^l|\ z_i = k,\ \alpha_{k,l,\ .}\right)\right] \times \prod_{l \in S^c} P\left(x_i^l|\ \beta_{l,\ .}\right) \end{aligned} \quad (4)$$

where $\theta := \left(\pi,\ (\alpha_{.,l,\ .})_{l \in S}\,,\ (\beta_{l,\ .})_{l \in S^c}\right)$ is a multidimensional parameter ranging over some space $\Theta_{(K,\ S)}$. Each individual is assumed to originate in one of the K (unknown) populations, each with its own allele frequencies. Thus these parameters verify the following properties :

$$\begin{cases} 0 < \pi_k \leq 1,\ k = 1, \ldots,\ K; \\ \sum_{k=1}^{K} \pi_k = 1. \end{cases} \quad (5)$$

$$\begin{cases} 0 \leq \alpha_{k,\ l,\ a} \leq 1,\ k = 1, \ldots,\ K,\ l \in S,\ a = 1, \ldots,\ A_l; \\ \sum_{a=1}^{A_l} \alpha_{k,\ l,\ a} = 1,\ k = 1, \ldots,\ K,\ l = 1, \ldots,\ L. \end{cases} \quad (6)$$

$$\begin{cases} 0 \leq \beta_{l,\ a} \leq 1,\ l \in S^c,\ a = 1, \ldots,\ A_l; \\ \sum_{a=1}^{A_l} \beta_{l,\ a} = 1,\ l \in S^c. \end{cases} \quad (7)$$

The number K of populations, the subset S of relevant loci for clustering, the proportions π of populations, the allele frequencies α and $\beta := (\beta_{l,j})_{l \in S^c;\ j=1,\ldots,\ A_l}$ are treated as the parameters of the model, which have to be infered. The variable z_i, the assignment of individual i to its population is not observed and has to be predicted.

Infering on K and S is regarded as a model selection problem. In fact, each value of $(K,\ S)$ defines a parametric model

$$\mathcal{M}_{(K,\ S)} = \left\{P_{(K,\ S)}\left(\ \cdot\ |\ \theta\right);\ \theta \in \Theta_{(K,\ S)}\right\}$$

of probability distributions. Let $K_{\max}$ be the maximum number of clusters. $K_{\max}$ has to be specified by the user for identifiability purposes discussed in the sub-section 2.4 hereafter. Let us consider the collection $\mathcal{C}$ of competing models:

$$\mathcal{C} = \left\{ \mathcal{M}_{(K,\ S)} \ :\ K \in \{1, \ldots,\ K_{\max}\} \text{ and } S \in \mathcal{P}^*(L) \right\}, \tag{8}$$

where $\mathcal{P}^*(L)$ is the set of non-empty subsets of the available loci $\{1, \ldots,\ L\}$.

In a Bayesian framework, the model $\mathcal{M}_{\left(\widetilde{K}_n,\ \widetilde{S}_n\right)}$ maximizing the posterior probability is to be chosen :

$$\left(\widetilde{K}_n,\ \widetilde{S}_n\right) = \arg\max_{(K,\ S)} P\left[(K,\ S) \mid x\right]. \tag{9}$$

By Bayes Theorem and assuming a non informative uniform prior distribution $P\left[(K,\ S)\right]$ on the competing models $\mathcal{M}_{(K,\ S)},\ K = 1, \ldots,\ K_{\max},\ S \in \mathcal{P}^*(L)$, one has

$$\left(\widetilde{K}_n,\ \widetilde{S}_n\right) = \arg\max_{(K,\ S)} P\left[x|\ (K,\ S)\right]. \tag{10}$$

The quantity $P\left[x|\ (K,\ S)\right]$ is the integrated likelihood of model $\mathcal{M}_{(K,\ S)}$, namely

$$P\left[x|\ (K,\ S)\right] = \int_{\theta \in \Theta_{(K,\ S)}} \left(\prod_{i=1}^{n} P_{(K,\ S)}\left(x_i|\ \theta\right) \right) P\left[\theta|\ (K,\ S)\right] d\theta, \tag{11}$$

where $d\theta$ is a measure on the parameter space $\Theta_{(K,\ S)}$ and $P\left[\theta|\ (K,\ S)\right]$, the prior distribution (*Kass and Raftery 1995* [12]). This integrated likelihood is analytically difficult to compute. An asymptotic approximation of $2 \ln P\left[x|\ (K,\ S)\right]$ is generally used; this approximation is the Bayesian Information Criterion (BIC) defined by

$$BIC\left(K,\ S\right) = 2 \sum_{i=1}^{n} \ln P_{(K,\ S)}\left(x_i|\ \widehat{\theta}_{ML,(K,S)}\right) - d_{(K,\ S)} \ln n, \tag{12}$$

where $d_{(K,\ S)}$ is the dimension of the parameter space $\Theta_{(K,\ S)}$ and $\widehat{\theta}_{ML,(K,S)}$, the maximum likelihood estimate of θ in $\Theta_{(K,\ S)}$. Thus, the selected model is given by

$$\left(\widehat{K}_n,\ \widehat{S}_n\right) = \arg\max_{(K,\ S)} BIC\left(K,\ S\right). \tag{13}$$

The maximum likelihood estimate $\widehat{\theta}_{ML,\left(\widehat{K}_n,\ \widehat{S}_n\right)}$ yields the Maximum a Posteriori (MAP) prediction rule defined by

$$\widehat{z}_i = \arg \max_{k \in \left\{1,\ldots,\ \widehat{K}\right\}} \widehat{\pi}_k P\left(x_i |\ z_i = k,\ \widehat{\theta}_{ML,\left(\widehat{K}_n,\ \widehat{S}_n\right)}\right). \tag{14}$$

One can notice that $\widehat{\theta}_{ML,\ (K,\ S)} = \left(\widehat{\gamma}_{ML,\ (K,\ S)},\ \widehat{\beta}_{ML,(K,S)}\right)$, where $\gamma = (\pi,\ \alpha)$. The maximum likelihood estimate $\widehat{\gamma}_{ML,(K,S)}$ is computed using the Expectation Maximization (EM) algorithm (*Dempster et al.* (1977) [8]), and the likelihood estimate $\widehat{\beta}_{ML,(K,S)}$ is given by the observed frequencies of the alleles of the loci of S^c.

Before giving the EM equations, let us present the loci selection procedure.

2.2. Combined Loci selection and clustering procedure

The space of competing models can be very large, consisting of all combinations of all $\left(2^L - 1\right)$ non-empty subsets of the available loci with each possible number of populations. Thus an exhaustive research of an optimum model is very painful in most situations. We adopt a two nested-step algorithm as proposed by *C. Maugis et al.* (2007) [14] :

Step 1. For all $K \in \{1, \ldots,\ K_{\max}\}$, we reseach

$$\widehat{S}_n(K) = \arg \max_{S \in \mathcal{P}^*(L)} BIC(K,\ S) \tag{15}$$

by a backward stepwise procedure detailed hereafter.

Step 2. We determine

$$\widehat{K}_n = \arg \max_{K \in \{1,\ldots,\ K_{\max}\}} BIC\left(K,\ \widehat{S}_n(K)\right). \tag{16}$$

We prefer a backward stepwise procedure rather than a forward stepwise as in *Kass and Raftery* (1995) [12] because starting the selection algorithm with all loci included allows the model to take loci interactions into account. At each stage, the algorithm searches for a locus to remove, and then assesses whether one of the current irrelevant loci can be selected. Thus the algorithm is making use of an exclusion and an inclusion procedures described hereafter. The decision of excluding or including a locus from the set of clustering loci is based on the BIC approximation of the Bayes factor.

Backward Stepwise selection procedure.

1 - Initialisation : $S = \{1, \ldots, L\}$, $S^c = \emptyset$.

2 - Exclusion step : The proposed locus for removal from the currently selected clustering loci S is chosen to be the one from this set without which the model is the best among the models with $\sharp S - 1$ loci.

$$c_{ex} = \arg\max_{l \in S} BIC\left(K, S \setminus \{l\}\right). \tag{17}$$

This candidate c_{ex} is excluded if the model $(K,\ S \setminus \{c_{ex}\})$ is better than $(K,\ S)$, i.e.

$$BIC\left(K,\ S\right) - BIC\left(K,\ S \setminus \{c_{ex}\}\right) \leq 0. \tag{18}$$

3 - Inclusion step : The proposed new clustering locus for inclusion in the currently selected clustering loci set S is chosen to be the one from the set S^c of currently non-selected loci which shows most evidence of multivariate clustering including the previous selected loci. This locus is accepted as relevant for clustering if its evidence for clustering is stronger than not clustering, namely

$$BIC\left(K,\ S \cup \{c_{in}\}\right) - BIC\left(K,\ S\right) > 0, \tag{19}$$

where

$$c_{in} = \arg\max_{l \in S^c} BIC\left(K,\ S \cup \{l\}\right). \tag{20}$$

The algorithm repeats **2** and **3** and stops when the proposed candidate for inclusion is the locus removed in the previous step or when S^c is empty.

2.3. EM equations

Here we describe the EM equations. To assign individual i to a cluster, we compute the posterior assignment probabilities $\tau_{ik} = P\left(z_i = k|\ x_i\right)$. Hereafter, we write $\gamma^{(r)} = \left(\pi^{(r)},\ \alpha^{(r)}\right)$ for the estimate of $\gamma = (\pi,\ \alpha)$ at iteration r of the EM algorithm. The $\tau_{ik}^{(r)}$ can be describe as

$$\tau_{ik}^{(r)} = \frac{\pi_k^{(r)} \prod_{l \in S} P\left(x_i^l|\ z_i = k,\ \alpha_{k,l,\ .}^{(r)}\right)}{\sum_{h=1}^{K} \pi_h^{(r)} \prod_{l \in S} P\left(x_i^l|\ z_i = k,\ \alpha_{h,l,\ .}^{(r)}\right)} \tag{21}$$

Then the update formulae for the parameters can be derived using the standard method of the EM algorithm

$$\pi_k^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n} \tau_{ik}^{(r)} \tag{22}$$

and

$$\alpha_{k,l,j}^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)} \left(\mathbb{1}_{\left[x_{i,1}^{l}=j\right]} + \mathbb{1}_{\left[x_{i,2}^{l}=j\right]} \right)}{2\sum_{i=1}^{n} \tau_{ik}^{(r)}}. \tag{23}$$

When applying the EM algorithm to data, we need to provide values for $\tau_{ik}^{(0)}$. There are mainly two types of initialization methods : random initialization methods and clustering-based initialization methods (McLachlan & Peel 2000). The random initialization methods assign individuals into clusters randomly, while the clustering-based initialization methods assign individuals into clusters according to some distance criteria.

2.4. Identifiability

Identifiability of models is necessary to have statistical consistency, which is a minimal requirement for an inference method. The parameters are of two types: $(K,\ S)$ on one hand, and $(\pi,\ \alpha,\ \beta)$ on the other hand for a given $(K,\ S)$. The identifiability of the parameter $(K,\ S)$ is discussed in sub-section 2.4.1. For a given $(K,\ S)$, the parameter β is always identifiable. The identifiability of the parameter $(\pi,\ \alpha)$ is discussed in sub-section 2.4.2 using the results of *Elizabeth Allan et al.* [1] which is given here in a multi-allelic multilocus genomic data framework.

2.4.1. Identifiability of the parameters $(K,\ S)$

Let $\mathcal{D} = \bigcup_{(K,\ S)} \mathcal{M}_{(K,\ S)}$ be the set of all probability distributions defined by the models $\mathcal{M}_{(K,\ S)}$ in competition. We assume that the true probability distribution P_0 of the observations that we are dealing with is an element of $\mathcal{D}$.

For a given $P \in \mathcal{D}$, let us define $K(P)$ and $S(P)$ as follow.

Definition 2.1. *For every P in $\mathcal{D}$,*

$$K(P) = \min_{K} \left\{ K \ :\ P \in \mathcal{M}_{(K,\ \cdot)} \right\}, \tag{24}$$

$$S(P) = \min_S \left\{ S \ : \ P \in \mathcal{M}_{(\ \cdot,\ S)} \right\}, \tag{25}$$

where $\mathcal{M}_{(K,\ \cdot)} = \bigcup_S \mathcal{M}_{(K,\ S)}$ *and* $\mathcal{M}_{(\ \cdot,\ S)} = \bigcup_K \mathcal{M}_{(K,\ S)}$.

This definition is justified by the following lemmas 2.1 and 2.2.

Lemma 2.1. *For every* $(K,\ S)$ *and* $(K',\ S')$, *if* $K \leq K'$ *and* $S \subseteq S'$, *then* $\mathcal{M}_{(K,\ S)} \subseteq \mathcal{M}_{(K',\ S')}$.

Proof of Lemma 2.1 : Let $P \in \mathcal{M}_{(K,\ S)}$ and let $\theta = (\pi,\ \alpha,\ \beta) \in \Theta_{(K,\ S)}$ be the parameter defining P. Let for instance define $\theta' = (\pi',\ \alpha',\ \beta') \in \Theta_{(K+1,\ S)}$ as follows

$$\begin{array}{rcl} \pi'_k & = & \pi_k,\ k = 1, \ldots,\ K-1 \\ \pi'_K > 0 & \text{and} & \pi'_{K+1} > 0 \text{ such that } \pi'_K + \pi'_{K+1} = \pi_K \\ \alpha'_{(k,\ \cdot,\ \cdot)} & = & \alpha_{(k,\ \cdot,\ \cdot)},\ k = 1, \ldots,\ K \\ \alpha'_{(K+1,\ \cdot,\ \cdot)} & = & \alpha_{(K,\ \cdot,\ \cdot)} \\ \beta' & = & \beta. \end{array}$$

Then we have $\theta' \in \Theta_{(K+1,\ S)}$ and $P = P_{(K+1,\ S)}(\ \cdot \mid \theta') \in \mathcal{M}_{(K+1,\ S)}$.

We have just showed that $\mathcal{M}_{(K,\ S)} \subseteq \mathcal{M}_{(K+1,\ S)}$ and there remains to show that $\mathcal{M}_{(K,\ S)} \subseteq \mathcal{M}_{(K,\ S')}$ for every S and S' such that $S \subseteq S'$. For such non empty subsets S and S' of available loci, the parameter space $\Theta_{(K,\ S)}$ can be regarded as a subset of $\Theta_{(K,\ S')}$ defined by the following equations :

$$\alpha_{1,l,\ .} = \ldots = \alpha_{K,l,\ .}\ \forall l \in S' \setminus S. \tag{26}$$

■

Lemma 2.2. *For every* $K_1,\ K_2 \in \{1, \ldots,\ K_{\max}\}$ *and* $S_1,\ S_2 \in \mathcal{P}^*(L)$, *we have* $\mathcal{M}_{(K_1,\ S_1)} \cap \mathcal{M}_{(K_2,\ S_2)} = \mathcal{M}_{(K_1 \wedge K_2,\ S_1 \cap S_2)}$, *where* $K_1 \wedge K_2 = \min\{K_1; K_2\}$.

Proof of Lemma 2.2 : Let P be a probability distribution in $\mathcal{M}_{(K_1,\ S_1)} \cap \mathcal{M}_{(K_2,\ S_2)}$. Then for every x in $\mathcal{X}$, $P(x)$ is given by the following two equations.

$$P(x) = \left[\sum_{k=1}^{K_1} \pi_k^1 \prod_{l \in S_1} P\left(x^l \mid \left(\alpha^1_{k,\ l,\ .}\right)\right) \right] \times \prod_{l \in S_1^c} P\left(x^l \mid \left(\beta^1_{l,\ .}\right)\right), \tag{27}$$

$$P(x) = \left[\sum_{k=1}^{K_2} \pi_k^2 \prod_{l \in S_2} P\left(x^l \mid \left(\alpha^2_{k,\ l,\ .}\right)\right) \right] \times \prod_{l \in S_2^c} P\left(x^l \mid \left(\beta^2_{l,\ .}\right)\right). \tag{28}$$

Assume without lost of generality that $K_1 \leq K_2$ and denote $A := S_1 \setminus (S_1 \cap S_2)$, $B := S_2 \setminus (S_1 \cap S_2)$ and $C = L \setminus S_1 \cup S_2$. Using equation (27), the marginal probability distribution of the sub-vector $x^{S_2} := \left(x^l\right)_{l \in S_2}$ is given by

$$P\left(x^{S_2}\right) = \left[\sum_{k=1}^{K_1} \pi_k^1 \prod_{l \in S_1 \cap S_2} P\left(x^l |\ \left(\alpha_{k,\ l,\ .}^1\right)\right)\right] \times \prod_{l \in B} P\left(x^l |\ \left(\beta_{l,\ .}^1\right)\right), \tag{29}$$

which using equation (28) becomes

$$\begin{aligned} P(x) &= \left[\sum_{k=1}^{K_1} \pi_k^1 \prod_{l \in S_1 \cap S_2} P\left(x^l |\ \left(\alpha_{k,\ l,\ .}^1\right)\right)\right] \times \prod_{l \in B} P\left(x^l |\ \left(\beta_{l,\ .}^1\right)\right) \\ &\quad \times \prod_{l \in A \cup C} P\left(x^l |\ \left(\beta_{l,\ .}^2\right)\right) \\ &= \left[\sum_{k=1}^{K_1} \pi_k^1 \prod_{l \in S_1 \cap S_2} P\left(x^l |\ \left(\alpha_{k,\ l,\ .}^1\right)\right)\right] \times \prod_{l \in A \cup B \cup C} P\left(x^l |\ \left(\beta_{l,\ .}^3\right)\right), \end{aligned}$$

which implies that $P \in \mathcal{M}_{(K_1 \wedge K_2,\ S_1 \cap S_2)}$ ■

Obviously, by definition 2.1, for every P_1 and P_2 in $\mathcal{D}$,

$$(P_1 = P_2) \Longrightarrow \left[K(P_1) = K(P_2) \text{ and } S(P_1) = S(P_2)\right]. \tag{30}$$

We will denote $(K_0,\ S_0) := (K(P_0),\ S(P_0))$, and this definition is well compatible with the use of the BIC criterion which aims at selecting the smallest model dimension in statistical adjustment.

2.4.2. Identifiability of parameter $\gamma = (\pi,\ \alpha)$ in the model $\mathcal{M}_{(K,\ S)}$

The classical definition of an identifiable model $\mathcal{M}_{(K,\ S)}$ of probability distributions requires that for any two different parameter values θ and θ' in parameter space $\Theta_{(K,\ S)}$, the corresponding probability distributions $P_{(K,\ S)}(\ \cdot \mid \theta)$ and $P_{(K,\ S)}(\ \cdot \mid \theta')$ be different. This is to require injectivity of the parameterization map Ψ for this model, which is defined by $\Psi(\theta) = P_{(K,\ S)}(\ \cdot \mid \theta)$.

In our context of finite mixtures, the above map will not strictly be injective because the latent classes can be freely relabeled without changing the

distribution underlining the observations. This is known as 'label swapping'. In such a case, the above map is always at least $K!$-to-one.

If the model is identifiable up to label swipping, then the number of independent parameters is at most equal to the number of distinct genotypes :

$$K - 1 + K\sum\nolimits_{l\in S}(A_l - 1) \leq \prod_{l\in S}\left(\begin{pmatrix} 2 \\ A_l \end{pmatrix} + A_l\right) - 1. \tag{31}$$

Despite that this condition is not sufficient, it gives an upper bound on $K_{\max} = \max_S K(S)$ of the number of populations where

$$K(S) := \frac{\prod_{l\in S}\frac{A_l(A_l+1)}{2}}{1 + \sum_{l\in S}(A_l - 1)}. \tag{32}$$

We use that upper bound to define the collection of models in competition given by equation (8).

Assume that the frequencies of the distinct observed genotypes are the parameters of interest. For a given K and S, we refer to the finite mixture model (1) as the K-class, $|S|$-feature model, with state space $\prod_{l\in S}\{1, \ldots, G_l\}$, as $\mathcal{M}\left(K\ ;\ (G_l)_{l\in S}\right)$, where $G_l = \frac{A_l(A_l+1)}{2}$ is the number of distinct observed genotypes at locus l and $|S|$ the cardinality of S.

Elizabeth S. Allman et al. (2008) [1] has proved that finite mixtures of multinomial distributions are *generically* identifiable. In the case of parametric setting, 'generic' means that the set of points for which identifiability does not hold has zero-measure. Here is the result of *Elizabeth S. et al.* relevant to our setting.

Theorem 2.1. *Consider the model $\mathcal{M}\left(K\ ;\ (G_l)_{l\in S}\right)$ where $|S| \geq 3$. Assume there exists a tripartition of the set S into three disjoint non-empty subsets S_1, S_2 and S_3, such that if $\mathcal{G}_i = \prod_{l\in S_i} G_l$, then*

$$\min(K,\ \mathcal{G}_1) + \min(K,\ \mathcal{G}_2) + \min(K,\ \mathcal{G}_3) \geq 2\cdot K + 2. \tag{33}$$

Then the model is generically idenfiable, up to label swapping. Moreover, the statement remains valid when the proportions of the groups $\{\pi_k\}_{k=1,\ldots,\ K}$ are held fixed and positive.

This result implies that one needs a minimum of genetic variability to guarantee the identifiability of the models in competition. For example, it will be difficult to detect 4 subpopulations with 3 biallelic loci such as Single Nucleotide Polymorphims (SNP).

3. Consistency

In this section, it is proved that the probability of selecting the true model $(K_0,\ S_0)$ by maximizing criterion (12) tends to 1 as $n \to \infty$ under the following single assumption:

$$(H)\ :\ \forall u \in \mathcal{X},\ P_0(u) > 0, \tag{34}$$

where $\mathcal{X}$ is the set of distinct genotypes defined by the observed alleles, and P_0 the true probability distribution of the observations. Assumption (H) is realistic because our method is proposed for experiments in which only observed alleles are considered.

Theorem 3.1. *Under assumption* (H)*,*

$$\lim_{n\to\infty} P_0\left[\left(\widehat{K}_n,\ \widehat{S}_n\right) = (K_0,\ S_0)\right] = 1.$$

Proof of Theorem 3.1 :

We need to prove that $\lim_{n\to\infty} P_0\left[\left(\widehat{K}_n,\ \widehat{S}_n\right) \neq (K_0,\ S_0)\right] = 0.$

$$P_0\left[\left(\widehat{K}_n,\ \widehat{S}_n\right) \neq (K_0,\ S_0)\right] \leq \sum_{(K,\ S)\neq(K_0,\ S_0)} P_0\left[\left(\widehat{K}_n,\ \widehat{S}_n\right) = (K,\ S)\right],$$

so that since the number of possible $(K,\ S)$ is finite, the theorem is proved if for every $(K,\ S) \neq (K_0,\ S_0)$, $\lim_{n\to\infty} P_0\left[\left(\widehat{K}_n,\ \widehat{S}_n\right) = (K,\ S)\right] = 0.$

Let $(K,\ S) \in \{1, \ldots,\ K_{\max}\} \times \mathcal{P}^*(L)$ such that $(K,\ S) \neq (K_0,\ S_0)$. We have

$$P\left[\left(\widehat{K}_n,\ \widehat{S}_n\right) = (K,\ S)\right] \leq P\left[2 \sup_{\theta\in\Theta_{(K,\ S)}} \ell_n\left(P_{(K,\ S)}(\ \cdot \mid \theta)\right)\right.$$
$$\left. - 2 \sup_{\theta\in\Theta_{(K_0,\ S_0)}} \ell_n\left(P_{(K_0,\ S_0)}(\ \cdot \mid \theta)\right) > \left(d_{(K,\ S)} - d_{(K_0,\ S_0)}\right) \ln n\right] \tag{35}$$

where $d_{(K,\ S)}$ is the number of independent parameters of model $\mathcal{M}_{(K,\ S)}$, and

$$\ell_n(P) = \sum_{u\in\mathcal{X}} n_u \ln P(u)$$

is the log-likelihood. Here n_u is the number of individuals in the sample with genotype u. Two cases are considered : $P_0 \in \mathcal{M}_{(K,\ S)}$ and $P_0 \notin \mathcal{M}_{(K,\ S)}$.

• **Case 1 :** $P_0 \in \mathcal{M}_{(K,\ S)}$.

Let $\mathcal{D}(\mathcal{X})$ denote the set of all probability distributions on the set $\mathcal{X}$ of distinct observed genotypes. Since $\mathcal{M}_{(K,\ S)} \subset \mathcal{D}(\mathcal{X})$,

$$\ell_n(P_0) \leq \sup_{\theta \in \Theta_{(K,\ S)}} \ell_n\Big(P_{(K,\ S)}(\ \cdot \mid \theta)\Big) \leq \sup_{P \in \mathcal{D}(\mathcal{X})} \ell_n(P),$$

so that

$$0 \leq \sup_{\theta \in \Theta_{(K,\ S)}} \ell_n\Big(P_{(K,\ S)}(\ \cdot \mid \theta)\Big) - \ell_n(P_0) \leq \sup_{P \in \mathcal{D}(\mathcal{X})} \ell_n(P) - \ell_n(P_0).$$

But it is well known that $2\sup_{P \in \mathcal{D}(\mathcal{X})} \ell_n(P) - 2\ell_n(P_0)$ converges in distribution to a chi-square variable with $|\mathcal{X}| - 1$ numbers of freedom, where $|\mathcal{X}|$ denote the cardinality of $\mathcal{X}$. Also, if $P_0 \in \mathcal{M}_{(K,\ S)}$ and $(K,\ S) \neq (K_0,\ S_0)$, $d_{(K,\ S)} - d_{(K_0,\ S_0)} > 0$. Thus in this case

$$\lim_{n \to \infty} P\Bigg[2 \sup_{\theta \in \Theta_{(K,\ S)}} \ell_n\Big(P_{(K,\ S)}(\ \cdot \mid \theta)\Big) - 2 \sup_{\theta \in \Theta_{(K_0,\ S_0)}} \ell_n\Big(P_{(K_0,\ S_0)}(\ \cdot \mid \theta)\Big) > \Big(d_{(K,\ S)} - d_{(K_0,\ S_0)}\Big) \ln n\Bigg] = 0.$$

• **Case 2 :** $P_0 \notin \mathcal{M}_{(K,\ S)}$

For every $\delta > 0$, let

$$\Theta^{\delta}_{(K,\ S)} = \left\{\theta \in \Theta_{(K,\ S)} \ : \ \forall x \in \mathcal{X},\ P_{(K,\ S)}(x \mid \theta) \geq \delta\right\}.$$

The key point is the following:

Proposition 3.1. *Under assumption* (H)*, there exists a real* $\delta > 0$ *such that for every* $(K,\ S)$*,*

$$\sup_{\theta \in \Theta_{(K,\ S)}} \frac{1}{n} \ell_n\Big(P_{(K,\ S)}(\ \cdot \mid \theta)\Big) = \sup_{\theta \in \Theta^{\delta}_{(K,\ S)}} \frac{1}{n} \ell_n\Big(P_{(K,\ S)}(\ \cdot \mid \theta)\Big) + o_{P_0}(1). \tag{36}$$

Proof of Proposition 3.1 :

$$\frac{1}{n}\ell_n\left(P_{(K,\ S)}\left(\ \cdot\mid\theta\right)\right)=\sum_{u\in\mathcal{X}}\frac{n_u}{n}\ln P_{(K,\ S)}\left(u\mid\ \theta\right)$$
$$=\sum_{u\in\mathcal{X}}\left[P_0\left(u\right)+o_{P_0}\left(1\right)\right]\times\ln P_{(K,\ S)}\left(u\mid\ \theta\right). \qquad (37)$$

The set of $\widetilde{\delta}>0$ such that $\Theta^{\widetilde{\delta}}_{(K,\ S)}\neq\emptyset$. Let $\widetilde{\delta}>0$ be such a real and $\widetilde{\theta}$ an element of $\Theta^{\widetilde{\delta}}_{(K,\ S)}$. Since for any u, $P_0(u)>0$, using (37),

$$\frac{1}{n}\ell_n\left(P_{(K,\ S)}\left(\ \cdot\mid\widetilde{\theta}\right)\right)\geq\sum_{x\in\mathcal{X}}P_0\left(x\right)\ln\widetilde{\delta}+o_{P_0}\left(1\right)=\ln\widetilde{\delta}+o_{P_0}\left(1\right). \qquad (38)$$

Let δ be a real such that $0<\delta<\widetilde{\delta}^{\frac{1}{\inf_{u\in\mathcal{X}}P_0(u)}}$ and $\delta\leq\inf_{u\in\mathcal{X}}P_0\left(u\right)$. Remark that $0<\inf_{u\in\mathcal{X}}P_0\left(u\right)\leq 1$ and $0<\widetilde{\delta}<1$ imply $\frac{1}{\inf_{\mathcal{X}}P_0(x)}\ln\widetilde{\delta}\leq\ln\widetilde{\delta}$, so that $0<\delta<\widetilde{\delta}^{\frac{1}{\inf_{\mathcal{X}}P_0(x)}}\leq\widetilde{\delta}$.

Then $\Theta^{\widetilde{\delta}}_{(K,\ S)}\subset\Theta^{\delta}_{(K,\ S)}$, and thus

$$\sup_{\theta\in\Theta^{\delta}_{(K,\ S)}}\frac{1}{n}\ell_n\left(P_{(K,\ S)}\left(\ \cdot\mid\theta\right)\right)\geq\sup_{\theta\in\Theta^{\widetilde{\delta}}_{(K,\ S)}}\frac{1}{n}\ell_n\left(P_{(K,\ S)}\left(\ \cdot\mid\theta\right)\right).$$

If now $\theta\in\Theta_{(K,\ S)}\setminus\Theta^{\delta}_{(K,\ S)}$, then there exists a genotype $u_\delta\in\mathcal{X}$ such that $P_{(K,\ S)}\left(u_\delta\mid\ \theta\right)<\delta$. In such a case

$$\frac{1}{n}\ell_n\left(P_{(K,\ S)}\left(\ \cdot\mid\theta\right)\right)\leq\inf_{\mathcal{X}}P_0\left(u\right)\ln\delta+o_{P_0}\left(1\right)$$
$$\leq\inf_{\mathcal{X}}P_0\left(u\right)\ln\widetilde{\delta}^{\frac{1}{\inf_{\mathcal{X}}P_0(x)}}+o_{P_0}\left(1\right)\ =\ln\widetilde{\delta}+o_{P_0}\left(1\right)$$
$$\leq\sup_{\theta\in\Theta^{\widetilde{\delta}}_{(K,\ S)}}\frac{1}{n}\ell_n\left(P_{(K,\ S)}\left(\ \cdot\mid\theta\right)\right)+o_{P_0}\left(1\right)$$
$$\leq\sup_{\theta\in\Theta^{\delta}_{(K,\ S)}}\frac{1}{n}\ell_n\left(P_{(K,\ S)}\left(\ \cdot\mid\theta\right)\right)+o_{P_0}\left(1\right)$$

Thus,

$$\sup_{\theta\in\Theta_{(K,\ S)}}\frac{1}{n}\ell_n\left(P_{(K,\ S)}\left(\ \cdot\mid\theta\right)\right)=\sup_{\theta\in\Theta^{\delta}_{(K,\ S)}}\frac{1}{n}\ell_n\left(P_{(K,\ S)}\left(\ \cdot\mid\theta\right)\right)+o_{P_0}\left(1\right)$$

which is the desired result. ■

Now, the set of functions $\{\ln P_{(K,\ S)}(\ \cdot\mid\theta),\theta\in\Theta^{\delta}_{(K,\ S)}\}$ is obviously Glivenko-Cantelli, so that

$$\sup_{\theta\in\Theta^{\delta}_{(K,\ S)}}\frac{1}{n}\ell_n\left(P_{(K,\ S)}(\ \cdot\mid\theta)\right)=\sup_{\theta\in\Theta^{\delta}_{(K,\ S)}}E_{P_0}\left[\ln P_{(K,\ S)}(U\mid\theta)\right]+o_{P_0}(1),$$

and Proposition 3.1 yields, for any (K,S),

$$\sup_{\theta\in\Theta_{(K,\ S)}}\frac{1}{n}\ell_n\left(P_{(K,\ S)}(\ \cdot\mid\theta)\right)=\sup_{\theta\in\Theta^{\delta}_{(K,\ S)}}E_{P_0}\left[\ln P_{(K,\ S)}(U\mid\theta)\right]+o_{P_0}(1).$$

Also,

$$\sup_{\theta\in\Theta^{\delta}_{(K_0,\ S_0)}}E_{P_0}\left[\ln P_{(K_0,\ S_0)}(U\mid\theta)\right]=E_{P_0}\ln P_0(U),$$

since $P_0\in\mathcal{M}_{(K_0,\ S_0)}$ and $P_0(u)\geq\delta\ \forall u\in\mathcal{X}$. Thus

$$\frac{1}{n}\sup_{\theta\in\Theta_{(K,\ S)}}\ell_n\left(P_{(K,\ S)}(\ \cdot\mid\theta)\right)-\frac{1}{n}\sup_{\theta\in\Theta_{(K_0,\ S_0)}}\ell_n\left(P_{(K_0,\ S_0)}(\ \cdot\mid\theta)\right)=$$
$$-\inf_{\theta\in\Theta^{\delta}_{(K,\ S)}}E_{P_0}\left[\ln P_0(U)-\ln P_{(K,\ S)}(U\mid\theta)\right]+o_{P_0}(1).$$

But on the compact set $\Theta^{\delta}_{(K,\ S)}$, the function $\theta\mapsto E_{P_0}\left[\ln P_0(U)-\ln P_{(K,\ S)}(U\mid\theta)\right]$ is continuous and attains its infimum at a point $\overline{\theta}$. But since $P_0\notin\mathcal{M}_{(K,\ S)}$, $P_0(\cdot)\neq P_{(K,\ S)}\left(\cdot\mid\overline{\theta}\right)$, and

$$E_{P_0}\left[\ln P_0(U)-\ln P_{(K,\ S)}\left(U\mid\overline{\theta}\right)\right]>0.$$

Noticing that $\lim_{n\to\infty}\frac{(d_{(K,S)}-d_{(K_0,S_0)})\ln n}{n}=0$, one gets

$$\lim_{n\to+\infty}P\left[2\sup_{\theta\in\Theta_{(K,\ S)}}\ell_n\left(P_{(K,\ S)}(\ \cdot\mid\theta)\right)-2\sup_{\theta\in\Theta_{(K_0,\ S_0)}}\ell_n\left(P_{(K_0,\ S_0)}(\ \cdot\mid\theta)\right)\right.$$
$$\left.>\left(d_{(K,\ S)}-d_{(K_0,\ S_0)}\right)\ln n\right]=0.$$

■

4. Simulation examples

MixMoGenD has been implemented using $C++$ and C programming languages. The main goal of the simulation examples was to confirm in practice the consistency of the loci selection procedure and to highliht its benifits. Before that, preliminary simulations were conducted to regulate certain known problems of the EM algorithm, in particular convergence towards the maximum likelihood and the low speed of convergence in certain cases. In fact, the EM algorithm converges almost always towards a local maximum under certain conditions of regularity. Thus it is not certain whether the algorithm converges towards a local or global maximum when there are several maxima. To reduce the dependence of the convergence point to the initial parameter of the algorithm, we opt for the strategy of at least 50 initial parameters, and the maximum likelihood estimate is the one maximasing the likelihood. For each initial parameter, we stop the EM algorithm when the difference between two consecutive likelihoods of the complete data is less than a certain positive real $\varepsilon > 0$ to be chosen by the user.

4.1. Consistency of the selection procedure

The goal here was to see how the increase of the size of the sample improves the capacity of our clustering method to select the true model $\mathcal{M}_{(K_0,\ S_0)}$. An interface between **MixMoGenD** and **R** was created for these simulations. In these experiments, we started with $n = 100$ individuals, and gradually increased this number to 400 by a step of 50. We assumed $K_0 = 2$ populations, $L = 4$ loci with 2 alleles by locus, S_0 with cardinality $|S_0| = 2$. For each value n of the sample size, 100 data sets were generated. The parameters of simulation are given in Table 1. The figure 1 shoes that **MixMoGenD** consistently identify the true model as $n \to \infty$. Other simulated data with $K_0 = 3$, $|S_0| = 4$ and $|S_0^c| = 2$ confirmed these results. These results confirm the theoretical result on the consistency that we showed in Section 3.

4.2. Benefits of the selection procedure

Two series of simulations were conducted to highlight the importance of the loci selection procedure. First, we independently generated 100 data sets, each of them contained 1 000 individuals. We assumed $K_0 = 3$ populations with the proportions given by $\pi = (0.20,\ 0.30,\ 0.50)$, $L = 6$ loci with the numbers of alleles given by $(3,\ 4,\ 3,\ 3,\ 3,\ 4)$, $S_0 = \{1,\ 2,\ 3,\ 4\}$ and allele

frequencies given in Table 2. Using all the 6 loci, the true model was selected 39 times against 61 for the model with $\widehat{K}_n = 2$. When including the selection procedure, **MixMoGenD** selected the true model $(K_0,\ S_0)$ 90 times against 10 for $(K,\ S) = (2,\ S_0)$. It appears that the number of populations is under estimated when considering all available loci as relevant for clustering.

To confirm this result, a second series of simulations with more variability was conducted. In these simulations, each of the data sets consisted of 1 000 individuals structured into 5 subpopulations of equal proportions. We assumed $L = 10$ loci each with 10 alleles, and two different cardinalities for S_0: 8 and 6. Instead of choosing manualy the allelic frequencies, we adopt the following strategy. For the loci in S_0, we first use the program **EASYPOP** [5] to simulate some data sets at some levels of F_{ST} between 0.03 and 0.04. Second, we estimated the allelic frequencies of the loci in S_0 by EM algorithm. And thirdly, we used these estimates and uniform probability distribution on loci in S_0^c to simulate the data sets with **R** program [17]. Sample project files and their simulation parameters are available on `http://www.math.u-psud/~toussile`.

Results

As expected, the results in the Table 3 show that the integrated loci selection procedure significantly improves the inference on the number K of subpopulations and the quality of the prediction. The benefit of the selection is more important with the increase of cardinality of the subset S_0^c of loci that are not relevant for clustering. The important result is that for these simulations, **MixMoGenD** perfectly selected the true subset S_0 of relevant loci for clustering. For each data set for which $\widehat{K}_n < K_0$, we calculated the square matrix of the pairwise F_{ST} between individuals sampled using the function *Fstat* of the package *Geneland* [11] of **R** program. We observed that there exists a threshold $F_{ST_{\max}}$ of pairwise F_{ST} for which two subpopulations with $F_{ST} < F_{ST_{\max}}$ are clustered together. This threshold are approximately 0.027 on the simulated data sets we used. The more striking example is the data set 5 in Table 3 (d). The square matrix of pairwise F_{ST} is given in Table 4. The F_{ST} between population 4 and the others are all < 0.026. On this data set, MixMoGenD produces 4 clusters and we observed that Pop4 was uniformly distributed in the 4 clusters.

5. Discussion

We believe that **MixMoGenD** will be useful for two main reasons. First, like **FASTRUCT**, **MixMoGenD** is based on the EM algorithm, so that both share certain qualities, particularly they are faster than their counterparts based on a bayesian approach [10].

The key point of our proposed method is that it is combined with a loci selection procedure. That is the main reason for which our method will be very useful, and it is our main contribution. In fact, the results obtained on simulated data show how the selection procedure improves significantly the inference on the number K of subpopulations and the prediction capacity. In addition, due to the explosion of genomic projects, data sets are becoming increasingly large. The space of the models in competion can then be very large. Then an exhaustive research of an optimum model is very painful in most situation and could not be achieved by methods based on MCMC algorithm as mentioned by *O. Francois et al.* (2006) [6]. Thus methods like frequentist likelihood methods using EM algorithm will then become useful because they require much shorter computations than the methods based on MCMC algorithm. For example, *E. K. Latch* (2005) [13] reported that a data set with 5 subpopulations, 100 individuals in each subpopulation, 10 loci and 10 alleles by locus take approximately 3 h to run without loci selection on **STRUCTURE** [15], and 30 h on **PARTITION** (all times provided are appropriate for a computer with a 2.2 GHz Celeron processor and 512 MB of RAM). For such data sets, **MixMoGenD** and its selection procedure take approximately 2 h 30 to run. This was made possible thanks to the Backward-Stepwise algorithm, which enabled us to avoid an exhaustive research of the optimum model among all the models in competition.

Acknowledgements

This work was supported by a doctoral fellowship from the "Institut de Recherche pour le Développement" (IRD), in the framework of the research unit UR06 of this insititut. We are gratefull to *Isabelle Morlais* for the explanations of the biological concepts we needed, and to *Henri GWET* for his constructive suggestions and his encouragement.

References

[1] E. S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of latent class models with many observed variables. arxiv.

[2] D. J. Balding. Estimating products in Forensic identification using dna profiles. *Journal of the American Statistical Association*, 90(431):839–844, SEP 1995.

[3] D. J. Balding and R. A. Nichols. DNA profile match probability calculation - How To allow for Population Stratification, relatedness, database selection and single bands. *Forensic Science Iinternational*, 64:125–140, FEB 1994.

[4] D. J. Balding and R. A. Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96:3–12, JUN 1995.

[5] F Balloux. EASYPOP (version 1.7): A computer program for Population Genetics simulations. *J Hered*, 92(3):301–2, 2001.

[6] C. Chen, F. Forbes, and O. Francois. fastruct: model-based clustering made faster. *Molecular Ecology Notes*, 6(4):980–983, 2006.

[7] J. Corander, P. Waldmann, P. Marttinen, and M.J. Sillanpaa. BAPS 2: enhanced possibilities for the analysis of genetic population structure, 2004.

[8] A. P. Dempster, N. M. Lairdsand, and D. B. Rubin. Maximum likelihood from in- complete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[9] L. A. Foreman, A. F. Smith, and I. W. Evett. A Bayesian approach to validating STR multiplex databases for use in forensic casework. *Int J Legal Med*, 110(5):244–250, 1997.

[10] O. François, S. Ancelet, and G. Guillot. Bayesian clustering using hidden markov random fields in spatial population genetics. *Genetics*, 174(2):805–16, oct 2006.

[11] G. Guillot, F. Mortier, and A. Estoup. Geneland: a computer package for landscape genetics. *Molecular Ecology Notes*, 5(3):712–715, 2005.

[12] R.E. Kass and A.E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[13] E. K. Latch, Guha Dharmarajan, Jeffrey C. Glaubitz, and Olin E. Rhodes Jr. Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation Genetics*, 7(2):295, 2006.

[14] C. Maugis, G. Celeux, and M.L. Martin-Magniette. Variable Selection for Clustering with Gaussian Mixture Models. Technical report, Technical Report 6211, INRIA, 2007.

[15] J K Pritchard, M Stephens, and P Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–59, June 2000.

[16] A.E. Raftery and N. Dean. Variable Selection for Model-Based Clustering. *Journal-American Statistical Association*, 101(473):168, 2006.

[17] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

Locus	Allele	Pop1	Pop2	Locus	Allele	Pop1	Pop2
1	1	0.70	0.25	3	1	0.85	0.85
	2	0.30	0.75		2	0.15	0.15
2	1	0.35	0.70	4	1	0.50	0.50
	2	0.65	0.30		2	0.50	0.50

Table 1: Parameters of simulated data to show the consistency of the selection procedure. $K_0 = 2$, $S_0 = \{1,\ 2\}$, $\pi = (0.30,\ 0.70)$.

L	Allele	Pop1	Pop2	Pop3	L	Allele	Pop1	Pop2	Pop3
1	1	0.20	0.40	0.50	4	1	0.30	0.40	0.65
	2	0.30	0.40	0.20		2	0.60	0.40	0.15
	3	0.50	0.20	0.30		3	0.10	0.20	0.20
2	1	0.20	0.40	0.50	5	1	0.25	0.25	0.25
	2	0.20	0.40	0.10		2	0.30	0.30	0.30
	3	0.40	0.10	0.10		3	0.25	0.25	0.25
	4	0.20	0.10	0.30		4	0.20	0.20	0.20
3	1	0.15	0.25	0.50	6	1	0.40	0.40	0.40
	2	0.25	0.25	0.10		2	0.30	0.30	0.30
	3	0.60	0.50	0.40		3	0.30	0.30	0.30

Table 2: Parameters of simulated data to show the benefit of the selection procedure: $K_0 = 3$, $\pi = (0.20,\ 0.30,\ 0.50)$, $S_0 = \{1,\ 2,\ 3,\ 4\}$. L = locus, Pop=Population

Data	$\widehat{K}_n^s$	% MA	$\widehat{K}_n$	% MA
1	4		3	
2	5	09.10	3	
3	3		3	
4	3		3	
5	5	12.40	3	
6	4		4	
7	3		3	
8	3		3	
9	5	11.80	3	
10	3		3	

(a)

Data	$\widehat{K}_n^s$	% MA	$\widehat{K}_n$	% MA
1	5	08.00	5	08.80
2	5	08.90	4	
3	5	10.40	5	11.40
4	5	10.20	5	10.50
5	5	08.80	5	09.30
6	5	10.20	5	10.30
7	5	09.10	4	
8	5	07.60	5	08.50
9	5	09.50	4	
10	5	10.30	5	10.90

(b)

Data	$\widehat{K}_n^s$	% MA	$\widehat{K}_n$	% MA
1	5	07.50	5	07.10
2	5	05.40	5	06.30
3	5	06.50	5	06.70
4	5	05.90	5	05.90
5	5	06.70	5	07.20
6	5	05.60	5	06.10
7	5	06.60	5	07.10
8	5	05.70	5	05.50
9	5	06.80	5	07.20
10	5	06.30	5	06.10

(c)

Data	$\widehat{K}_n^s$	% MA	$\widehat{K}_n$	% MA
1	5	14.80	2	
2	5	14.20	1	
3	5	13.40	2	
4	5	13.60	2	
5	4		2	
6	5	14.30	1	
7	5	15.10	2	
8	5	13.90	2	
9	5	14.70	2	
10	5	15.20	1	

(d)

Data	$\widehat{K}_n^s$	% MA	$\widehat{K}_n$	% MA
1	5	10.60	3	
2	5	11.30	4	
3	5	09.70	3	
4	5	09.60	4	
5	5	11.00	4	
6	5	10.50	4	
7	5	09.80	4	
8	5	10.70	4	
9	5	11.50	4	
10	5	12.50	3	

(e)

Table 3: For all these simulations, $L = 10$ and $K = 5$. In the tables (a), (b) and (c), we assumed $|S| = 8$ and the F_{ST} for the loci in S were 0.0304, 0.0355 and 0.0407 respectivelly. As expected, the increase of F_{ST} for the loci in S improves the performances of MixMoGenD. In the Tables (d) and (e), we assumed $|S| = 6$, and the difference between running MixMoGenD with or without selection is clear. MA = Misassigned, $\widehat{K}_n^s$ and $\widehat{K}_n$ are the estimates of the number of populations with and without loci selection resprctively.

	Pop1	Pop2	Pop3	Pop4	Pop5
Pop1	0.00000000	0.04112990	0.03024947	0.02425668	0.03535726
Pop2	0.04112990	0.00000000	0.03831558	0.02255300	0.02756619
Pop3	0.03024947	0.03831558	0.00000000	0.02255183	0.03251246
Pop4	0.02425668	0.02255300	0.02255183	0.00000000	0.02509488
Pop5	0.03535726	0.02756619	0.03251246	0.02509488	0.00000000

Table 4: The F_{ST} between population 4 and the others are all < 0.026. MixMoGenD on this data set produces 4 clusters and we observed that Pop4 was uniformly distributed in the 4 clusters.

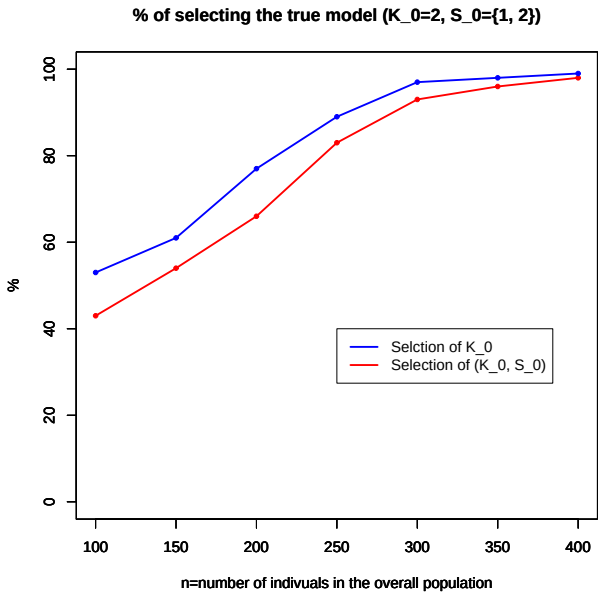

Figure 1: % of selecting the true model vs number of individuals