Adv Data Anal Classif manuscript No.
(will be inserted by the editor)

Variable selection in model-based clustering using multilocus genotype data

Wilson Toussile · **Elisabeth Gassiat**

Received: May 21, 2009/ Accepted: date

Abstract We propose a variable selection procedure in model-based clustering multilocus genotype data. Indeed, it may happen that some loci are not relevant for clustering into statistically different populations. Inferring the number K of clusters and the relevant clustering subset S of loci is regarded as a model selection problem. The competing models are compared using penalized maximum likelihood criteria. Under weak assumptions on the penalty function, we prove the consistency of the resulting estimator $\left(\widehat{K}_n, \widehat{S}_n\right)$. An associated algorithm named *Mixture Model for Genotype Data (MixMoGenD)* was implemented using $c++$ programming language and is available on `www.math.u-psud.fr/~toussile`. To avoid an exhaustive research of the optimum model, we propose an adaptation of the Backward-Stepwise algorithm, which enables a better research of the optimum model among all possible cardinalities of S. We present numerical experiments on simulated and real datasets that highlight the interest of our loci selection procedure.

Keywords Model-Based Clustering · Penalized maximum likelihood criteria · Population Genetics · Variable Selection.

Subject classification JEL C89, AMS 62H30.

1 Introduction

A long standing issue in population genetics is the identification of genetically homogeneous populations. To give an answer to such a question using data coming from

W. Toussile
UR016 Institut de Recherche pour le Développement (IRD), Laboratoire de Mathématique d'Orsay (LMO), Ecole Nationale Supérieure Polytechnique de Yaoundé.
Bât 425, 91405 Orsay cedex, Tel. : +33 (0) 1 69 15 76 18
E-mail: wilson.toussile-fomazou@u-psud.fr

E. Gassiat
Laboratoire de Mathématique d'Orsay (LMO)
E-mail: Elisabeth.Gassiat@math.u-psud.fr

individuals for which there is no prior knowledge about the population they come from, one has to face the statistical problem of unsupervised clustering. A number of model based-clustering methods for multilocus genotype data have been developed in recent years. We can cite among others: *STRUCTURE*, *Bayesian Analysis of Population Structure (BAPS)*, *Geneland* and *Fastruct* proposed respectively by [19, J. K. Pritchard et al. (2000)], [8, J. Corander. et al. (2004)], [12, G. Guillot et al. (2005)] and [10, O. François et al. (2006)]. Multi-locus genotype datasets 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 of loci which cluster the population in the best way. None of the above methods perform automatically variable selection.

In this work, we propose a loci selection procedure in model-based clustering for multi-allelic loci data, and an associated algorithm named *Mixture Model for Genetic Data* (*MixMoGenD*). As almost all already proposed model-based clustering methods for multilocus genotype data, our procedure attempts to group samples into clusters of random mating individuals so that the *Hardy-Weinberg Disequilibrium* (HWD) and the Linkage Disequilibrium (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 proved to be useful in describing many population genetics attributes and serve as a simple model in the development of more realistic models of microevolution. Recall that in clustering, classification is not observed and there is no prior 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 method available to use. Thus it makes sense to include the loci selection procedure as a part of the clustering algorithm as recommended in [16, C. Maugis et al.] in a Gaussian setting.

Let K denote the (unknown) number of clusters and S the (unknown) subset of loci that are relevant for clustering. Inferring K and S is seen as a model selection problem. More precisely, let L and $\mathscr{P}^*(L)$ be respectively the number and the set of all nonempty subsets of the available loci. A specific collection

$$\mathscr{C} := \left(\mathscr{M}_{(K,\,S)}\right)_{(K,\,S)\in\mathbb{N}^*\times\mathscr{P}^*(L)}$$

of models is defined such that each model $\mathscr{M}_{(K,\,S)}$ corresponds to a particular structure situation with K clusters and a subset S of loci that are relevant for clustering. The observations are supposed to be realizations from an unknown probability distribution P_0 in some of the competing models $\mathscr{M}_{(K,\,S)}$. Consequently, inferring $(K,\,S)$ can be formulated as the choice of a model among the collection $\mathscr{C}$. This choice automatically leads to a data clustering and to a variable selection (the set of relevant loci). A data-driven criterion is thus needed to select the "best" model. There exists a huge literature on model selection via penalized criteria, see [15, P. Massart (2007)] and the references therein. We propose to use a penalized maximum likelihood criterion. Our analysis and our algorithm do not impose a particular choice of the penalization term. For the numerical experiments, we will use the BIC. In addition, we propose what we called a "backward-stepwise explorer" algorithm which

avoids an exhaustive research of the optimum model (which can be very painful in most situations). This algorithm enables the research of the optimum model among all possible cardinalities of S.

Although there exists a lot of works concerning the behaviour of BIC and other penalization methods in practice, theoretical results in a mixture framework are few. A general consistency theorem may be found in [11, E. Gassiat (2002)], applications to mixture models are developed for example in [4, J.-M. Azais, E. Gassiat and C. Mercadier] and [6, A. Chambaz, A. Garivier and E. Gassiat]. See also references therein. The consistency of the BIC estimator is shown in [16, Maugis et al.] for a variable selection problem with Gaussian mixture models when the number of components is known. But as far as we know, there is no consistency result for both a variable selection and clustering problem in a discrete distribution setting. Under weak assumptions on the penalty function, we prove that the probability to select the true number of populations and the true set of relevant variables tends to 1 as the size of the sample tends to infinity.

Our paper is organized as follows. In section 2, we describe the competing models and the model selection principle we will use. We then describe the modification of the Backward-Stepwise algorithm we propose to perform the model selection. In section 3, we first describe how the true model may be characterized as the "smallest" model. We then discuss identifiability properties of latent class models in our settings (mainly by presenting the result of *E.S. Allman et al.* (2008) [2]). We finally give our main consistency result for the estimation of the model using penalized criteria such as BIC type criteria. Section 4 is devoted to numerical experiments on both simulated and real datasets to highlight the practical interest of our variable selection method. In particular, the experiments show that our method performs well for unsupervised clustering of genetically homogeneous populations in situations where measures of population structure such as Wright's F statistics are in a range where it is thought that clustering is difficult. In such cases, the improvement is obviously due to the variable selection procedure.

2 Model and methods

2.1 Framework, notation and competing models

The data set we shall deal with consists of the genotypes of a sample of n diploid individuals at L loci that will be denoted by $1,\ldots,\ l,\ldots,\ L$. The observations are $x_1,\ldots,x_n$, where for each individual i, x_i contains the genotypes at the L loci, that is $x_i = (x_i^l)_{l=1,\ldots,L}$, where x_i^l is the genotype of the i^{th} individual at the l^{th} locus. The genotype x_i^l consists of a (non ordered) set $\left\{x_{i,1}^l,\ x_{i,2}^l\right\}$ of two (that may be equal) alleles in the set of distinct allele states at locus l. These allele states are labeled $1,\ 2,\ldots,\ A_l$, where A_l denotes their number. When $x_{i,1}^l = x_{i,2}^l$, individual i is said to be homozygous at locus l. The dataset $x_1,\ldots,x_n$ is assumed to be a realization of a n-sample (that is n independent identically distributed random variables) with the

same distribution as $X = \left(X^l\right)_{l=1,\ldots,L}$, where $X^l = \left\{X_1^l,\, X_2^l\right\}$, with X_1^l and X_2^l taking their values in the set $\{1,\ldots,\, l,\ldots,\, A_l\}$ of observed alleles. Let :

- Z be the non observed random variable indicating the population the individual comes from. We will denote z_i the (unobserved) population of origin of individual i;
- $\pi_k := P(Z = k)$, the probability that an individual comes from population k (the π_k's are called the mixing proportions);
- $\alpha_{k,l,j} := P\left(X_1^l = j | Z = k\right) = P\left(X_2^l = j |\, Z = k\right)$, the frequency of the j^{th} allele at locus l in population k;
- and $\mathbb{X}$, the set of all possible genotypes from the observed alleles.

Model-based clustering methods proceed by assuming that the observations from each cluster are drawn from some parametric model and the overall population is a finite mixture of these populations. As almost all already proposed model-based clustering methods using genotype data, we wish to group the sample into clusters of random mating individuals so that the *Hardy-Weinberg* (HW) and linkage disequilibria (LD) are minimized across the sample (see [14, E. K. Latch (2006)] and the references therein). 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 serve as a useful base model in the development of more realistic models of microevolution. Thus we assume *Hardy-Weinberg* and complete linkage equilibria in each cluster. *Hardy-Weinberg* means that the probability to observe a genotype x^l at locus l is given by

$$P\left(x^l |\, Z = k,\, \alpha_{k,l,\cdot}\right) = \left(2 - \mathbb{1}_{[x_1^l = x_2^l]}\right) \alpha_{k,l,x_1^l} \times \alpha_{k,l,x_2^l}, \tag{1}$$

whereas complete linkage equilibria in each cluster means that within populations, genotypes at different loci are independent random variables.

Let now S be the set of loci which are relevant for clustering. S^c is thus the set of loci that are not relevant for clustering ($S \cup S^c = \{1,\ldots,\, L\}$). Typically, the reason why the loci of S^c are not relevant for clustering is that their alleles are equally distributed across the clusters. This means that

($\mathscr{H}$): for every locus l in S^c and for every allele j in $\{1,\, 2,\ldots,\, A_l\}$, one has

$$\alpha_{1,l,j} = \alpha_{2,l,j} = \cdots = \alpha_{K,l,j} =: \beta_{l,j}. \tag{2}$$

Under conditional *Hardy-Weinberg* equilibrium, conditional complete linkage equilibrium, and assumption ($\mathscr{H}$), the observations are thus supposed to be independent and identically distributed random variables with probability distribution

$$\begin{aligned} P_{(K,\, S)}\left(x |\, \theta\right) &= P\left(x |\, K,\, S,\, \theta\right) \\ &= \left[\sum_{k=1}^{K} \pi_k \prod_{l \in S} P\left(x^l |\, Z = k,\, \alpha_{k,l,\,\cdot}\right)\right] \times \prod_{l \in S^c} P\left(x^l |\, \beta_{l,\,\cdot}\right), \end{aligned} \tag{3}$$

for $x = (x^l)_{l=1,\dots,L}$, where $\theta := \left(\pi,\ (\alpha_{\cdot,l,\ \cdot})_{l\in S},\ (\beta_{l,\ \cdot})_{l\in S^c}\right)$ is a multidimensional parameter ranging in some space $\Theta_{(K,\ S)}$. These parameters fulfill the following properties:

$$\begin{cases} 0 \leq \pi_k \leq 1,\ k = 1,\dots,\ K; \\ \sum_{k=1}^{K} \pi_k = 1. \end{cases} \tag{4}$$

$$\begin{cases} 0 \leq \alpha_{k,\ l,\ a} \leq 1,\ k = 1,\dots,\ K,\ l \in S,\ a = 1,\dots,\ A_l; \\ \sum_{a=1}^{A_l} \alpha_{k,\ l,\ a} = 1,\ k = 1,\dots,\ K,\ l = 1,\dots,\ L. \end{cases} \tag{5}$$

$$\begin{cases} 0 \leq \beta_{l,\ a} \leq 1,\ l \in S^c,\ a = 1,\dots,\ A_l; \\ \sum_{a=1}^{A_l} \beta_{l,\ a} = 1,\ l \in S^c. \end{cases} \tag{6}$$

The number K of clusters, the subset S of relevant loci for clustering, the mixing proportions $\pi = (\pi_k)_{k=1,\dots,\ K}$, the allelic frequencies $\alpha = \left(\alpha_{k,l,j}\right)_{k=1,\dots,K;\ l\in S;\ j=1,\dots,A_l}$ and $\beta := \left(\beta_{l,j}\right)_{l\in S^c;\ j=1,\dots,\ A_l}$ are treated as the parameters of the model, which have to be inferred. The assignment z_i of individual i to its population of origin is not observed and has to be predicted. The parameters K and S will be treated in a particular way.

For a given K and S, the parameter $\theta \equiv \theta_{(K,\ S)}$ is an element of the set $\Theta_{(K,\ S)}$ given by

$$\Theta_{(K,\ S)} := \mathbb{S}_{K-1} \times \left[\prod_{l\in S} \mathbb{S}_{A_l-1}\right]^K \times \prod_{l\in S^c} \mathbb{S}_{A_l-1}, \tag{7}$$

where $\mathbb{S}_{r-1} = \left\{p = (p_1,\ p_2,\dots,\ p_r) \in [0,\ 1]^r : \sum_{j=1}^{r} p_j = 1\right\}$ is the $r-1$ dimensional simplex. We then consider the parametric model $\mathscr{M}_{(K,\ S)}$ of probability distributions defined by

$$\mathscr{M}_{(K,\ S)} = \left\{P_{(K,\ S)}\left(\ \cdot \mid \theta_{(K,\ S)}\right);\ \theta_{(K,\ S)} \in \Theta_{(K,\ S)}\right\}. \tag{8}$$

Each model $\mathscr{M}_{(K,\ S)}$ corresponds to a particular structure situation with K clusters and a clustering relevant variable subset S. Thus the choice of a model among the collection $\mathscr{C} = \left(\mathscr{M}_{(K,\ S)}\right)_{(K,\ S)\in\mathbb{N}^*\times\mathscr{P}^*(L)}$ automatically leads to a data clustering (via the estimation of the parameter $\theta_{(K,\ S)}$ and the prediction of the z_i's, see below) and a variable selection (via the estimation of S).

In the following, we will refer to the number of free parameters of a model $\mathscr{M}_{(K,\ S)}$ given by

$$D_{(K,\ S)} = K - 1 + K\sum_{l\in S}(A_l - 1) + \sum_{l\notin S}(A_l - 1). \tag{9}$$

as the dimension of the model $\mathscr{M}_{(K,\ S)}$.

2.2 **Model selection principle**

We shall use penalized maximum likelihood as a model selection principle. Let $\mathscr{D}$ be the set of all discrete probability distributions on $\mathbb{X}$. Consider the empirical contrast γ_n defined for every $P \in \mathscr{D}$ by

$$\gamma_n(P) := -\frac{1}{n}\sum_{i=1}^{n} \ln P(x_i).$$

Now, consider the collection $\mathscr{C} = \left(\mathscr{M}_{(K,S)}\right)_{(K,S)\in\mathbb{N}^*\times\mathscr{P}^*(L)}$ of the competing models. For every $(K,S) \in \mathbb{N}^* \times \mathscr{P}^*(L)$, let $\widehat{P}_{(K,S)} := P_{(K,\,S)}\left(\cdot \mid \widehat{\theta}_{MLE,(K,\,S)}\right)$ be the maximum likelihood over $\mathscr{M}_{(K,S)}$, that is the probability distribution that minimizes $\gamma_n(P)$ for $P \in \mathscr{M}_{(K,S)}$. Consider moreover some penalty functions

$$\begin{aligned} pen_n : \mathbb{N}^* \times \mathscr{P}^*(L) &\longrightarrow \mathbb{R}_+ \\ (K,S) &\longmapsto pen_n(K,S). \end{aligned} \tag{10}$$

The estimator $\left(\widehat{K}_n,\, \widehat{S}_n\right)$ is defined as a minimizer of the penalized criterion (see *Massart* (2007) [15] for an overview of model selection via penalization)

$$crit(K,S) := \gamma_n\left(\widehat{P}_{(K,S)}\right) + pen_n(K,S). \tag{11}$$

We can then define the selected model $\mathscr{M}_{\left(\widehat{K}_n,\,\widehat{S}_n\right)}$ and the associated selected estimator $\widehat{P}_{\left(\widehat{K}_n,\,\widehat{S}_n\right)}$. The maximum likelihood estimate $\widehat{\theta}_{MLE,\left(\widehat{K}_n,\,\widehat{S}_n\right)}$ yields the Maximum a Posteriori (MAP) prediction rule defined by

$$\widehat{z}_i = \arg\max_{k\in\left\{1,\ldots,\,\widehat{K}_n\right\}} \widehat{\pi}_k P\left(x_i \middle|\, z_i = k,\, \widehat{\theta}_{MLE,\left(\widehat{K}_n,\,\widehat{S}_n\right)}\right). \tag{12}$$

One can notice that $\widehat{\theta}_{MLE,(K,S)} = \left(\widehat{\gamma}_{MLE,(K,S)},\, \widehat{\beta}_{MLE,(K,S)}\right)$, where $\gamma = (\pi,\, \alpha)$. The maximum likelihood estimate $\widehat{\gamma}_{MLE,(K,S)}$ is computed thanks to the Expectation Maximization (EM) algorithm (*Dempster et al.* (1977) [9]) (see Appendix for the EM equations), and the likelihood estimate $\widehat{\beta}_{MLE,(K,S)}$ is given by the observed frequencies of the alleles of the loci of S^c.

As shown below in subsection 3.1, assuming that the true density P_0 belongs to one of the competing models implies that there exists a "smallest" model $\mathscr{M}_{(K_0,\,S_0)}$ containing P_0. Thus, it makes sense to consider penalties pen_n that are increasing functions of the dimension $D_{(K,S)}$ such as the BIC type criteria. We prove below the consistency of the estimator $\left(\widehat{K}_n,\, \widehat{S}_n\right)$ under weak assumptions on the penalty functions.

2.3 **Selection 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. A two nested-step algorithm combined with a Backward-Stepwise algorithm is proposed in *C. Maugis et al.* [16] in a Gaussian framework to avoid and exhaustive research of the optimum model. This algorithm makes use of an exclusion and an inclusion steps. Starting from the exclusion with all the variables selected, Backward-Stepwise algorithm enables to take into account the possible interaction between variables. The algorithm proposed in *C. Maugis et al.* stops when there are no exclusion and no inclusion in two consecutive steps.

When performing numerical experiments, we found that this Backward-Stepwise algorithm could miss the optimum model in some cases, in particular in cases where the optimum subset of clustering loci is small. So we propose an adaptation of this Backward-Stepwise which forces to go down until the cardinality of S equals 1, so that sets S with small cardinality are always explored by the algorithm (see (18) below). The optimum model is then chosen between all the models explored by our proposed algorithm named "backward-stepwise explorer".

In addition, if the model is identifiable up to label switching, then the number of free parameters of the mixture part is at most equal to the cardinality of $\mathbb{X}^S := \left\{\left(x^l\right)_{l\in S} : \, x\in\mathbb{X}\right\}$:

$$K-1+K\sum\nolimits_{l\in S}(A_l-1)\leq\prod_{l\in S}\left(\binom{2}{A_l}+A_l\right)-1. \tag{13}$$

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 $K(S)$ is the smallest integer bigger than

$$\frac{\prod_{l\in S}\frac{A_l(A_l+1)}{2}}{1+\sum_{l\in S}(A_l-1)}. \tag{14}$$

Thus, the research of the best model can be done among the finite collection $\mathscr{C}_{K_{\max}}$ given by

$$\mathscr{C}_{K_{\max}} := \left(\mathscr{M}_{(K,\,S)}\right)_{K=1,\ldots,\,K_{\max};S\in\mathscr{P}^*(L)}. \tag{15}$$

The two nested-step algorithm is stated as follows.

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

$$\widehat{S}_n(K)=\arg\min_{S\in\mathscr{P}^*(L)} crit\,(K,\,S) \tag{16}$$

  by exploring competing models with K clusters using our proposed backward stepwise explorer procedure detailed hereafter.
- **Step 2.** We determine

$$\widehat{K}_n=\arg\min_{K\in\{1,\ldots,\,K_{\max}\}} crit\left(K,\,\widehat{S}_n(K)\right). \tag{17}$$

The selected model is then given by $\left(\widehat{K}_n, \widehat{S}_n\left(\widehat{K}_n\right)\right)$.

At each step, the following Backward-Stepwise explorer algorithm (18) searches for a locus in S to remove, and then assesses whether one of the current irrelevant loci in S^c can be selected. The decision of excluding a locus from or including a locus in the set of clustering loci is based on a penalized maximum likelihood criterion of the form given in equation (11). The proposed candidate locus c_{ex} for exclusion 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 submodels with $|S|-1$ loci. The proposed new clustering locus c_{in} 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.

BACKWARD-STEPWISE EXPLORER()
1 $S \leftarrow \{1,\ldots,L\}$, $S^c \leftarrow \emptyset$
2 $c_{ex} \leftarrow 0$, $c_{in} \leftarrow 0$
3 **repeat**
4 $c_{ex} \leftarrow \arg\min_{l\in S} crit\,(K, S\setminus\{l\})$
5 **if** $crit\,(K,\ S) - crit\,(K,\ S\setminus\{c_{ex}\}) \geq 0$ or $c_{in} = 0$
6 **then** $S \leftarrow S\setminus\{c_{ex}\}$
7 **else** $c_{ex} \leftarrow 0$
8 $c_{in} \leftarrow \arg\min_{l\in S^c} crit\,(K,\ S\cup\{l\})$
9 **if** $crit\,(K,\ S\cup\{c_{in}\}) - crit\,(K,\ S) < 0$ and the model $\mathscr{M}_{(K,\ S\cup\{c_{in}\})}$
10 never be "reference model"[1] in an exclusion step
11 **then** $S \leftarrow S\cup\{c_{in}\}$
12 **else** $c_{in} \leftarrow 0$
13 **until** $|S| = 1$.

(18)

3 Consistency

This section is devoted to the theoretical result of consistency of the estimator $\left(\widehat{K}_n,\ \widehat{S}_n\right)$ of parameter $(K_0,\ S_0)$ defined in subsection 3.1. Identifiability of the models $\mathscr{M}_{(K,\ S)}$ is discussed in subsection 3.2 using a result obtained by *E. S. Allman et al.* (2008) [2], and the main consistency result is given in subsection 3.3.

3.1 The "smallest" model $\mathscr{M}_{(K_0,\ S_0)}$

Let $\mathscr{M} = \bigcup_{(K,\ S)} \mathscr{M}_{(K,\ S)}$ be the set of all probability distributions defined by the models $\mathscr{M}_{(K,\ S)}$ in competition. We assume that the true probability distribution P_0 of the observations that we are dealing with is an element of $\mathscr{M}$. By lemma 1 stated

[1] What we call "reference model" is every model $(K,\ S)$ in line 4 of Algorithm (18).

hereafter there are more than one model $\mathscr{M}_{(K,\,S)}$ such that $P_0 \in \mathscr{M}_{(K,\,S)}$. But thanks to the lemma 2 below, there exists a "smallest" model $\mathscr{M}_{(K_0,\,S_0)}$ containing the true density P_0. This "smallest" model can be defined by $(K_0,\,S_0) := (K(P_0),\,S(P_0))$, where

$$K(P) = \min_{K} \left\{ K \,:\, P \in \bigcup_{S \in \mathscr{P}^*(L)} \mathscr{M}_{(K,\,S)} \right\}, \tag{19}$$

$$S(P) = \min_{S} \left\{ S \,:\, P \in \bigcup_{K \in \mathbb{N}^*} \mathscr{M}_{(K,\,S)} \right\}, \tag{20}$$

for every P in one of the competing models $\mathscr{M}_{(K,\,S)}$. In (20), min is in the sense of the partial order defined by the inclusion of sets. Consequently, we will refer to $\mathscr{M}_{(K_0,\,S_0)}$ as our uniquely defined true model.

Lemma 1 *For every $K_1,\ K_2$ in $\mathbb{N}^*$ and $S_1,\ S_2$ in $\mathscr{P}^*(L)$, if $K_1 \leq K_2$ and $S_1 \subseteq S_2$, then $\mathscr{M}_{(K_1,\,S_1)} \subseteq \mathscr{M}_{(K_2,\,S_2)}$.*

Lemma 2 *For every $K_1,\ K_2$ in $\mathbb{N}^*$ and $S_1,\ S_2$ in $\mathscr{P}^*(L)$, one has*

$$\mathscr{M}_{(K_1,\,S_1)} \cap \mathscr{M}_{(K_2,\,S_2)} = \mathscr{M}_{(K_1 \wedge K_2,\,S_1 \cap S_2)},$$

where $K_1 \wedge K_2 = \min\{K_1,\ K_2\}$.

The proofs of Lemmas 1 and 2 are given in Appendix A and B.

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

The classical definition of an identifiable model $\mathscr{M}_{(K,\,S)}$ 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 the 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 switching'. In such a case, the above map is always at least $K!$-to-one.

For a given K and S, assume that the frequencies of the genotypes in $\mathbb{X}$ are the parameters of interest. In this subsection, we refer to a finite mixture model $\mathscr{M}_{(K,\,S)}$ as the K-class, $|S|$-feature model, with state space $\prod_{l \in S}\{1,\ldots,\,G_l\}$, and denote it $\mathbb{M}\big(K\,;\,(G_l)_{l \in S}\big)$, where $G_l := \frac{A_l(A_l+1)}{2}$ is the number of distinct genotypes from observed allele states at locus l and $|S|$ the cardinality of S. *E. S. Allman et al.* (2008) [2] 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.* in our setting.

Theorem 1 *Consider model* $\mathbb{M}\left(K\,;\,(G_l)_{l\in S}\right)$ *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*

$$\min\left(K,\,\mathscr{G}_1\right)+\min\left(K,\,\mathscr{G}_2\right)+\min\left(K,\,\mathscr{G}_3\right)\geq 2\cdot K+2, \tag{21}$$

where $\mathscr{G}_i:=\prod_{l\in S_i}G_l$.

Then the model is generically identifiable, up to label switching. 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).

3.3 **The main result**

In this section, we prove that the probability of selecting the "smallest" model $(K_0,\,S_0)$ (see subsection 3.1) via a penalized maximum likelihood criterion tends to 1 as n tends to infinity, for penalty functions of the form

$$\begin{aligned} pen_n:\{1;\ldots;K_{max}\}\times\mathscr{P}^*(L)&\longrightarrow\mathbb{R}_+\\ (K,S)&\longmapsto pen_n(K,S)=\frac{1}{n}pen\left(D_{(K,S)},\,n\right)\end{aligned} \tag{22}$$

fulfilling the following properties:

- $(P1)$: for every integer D, $\lim_{n\to\infty}\frac{pen(D,\,n)}{n}=0$;
- $(P2)$: for every $(K_1,\,S_1)$ and $(K_2,\,S_2)$ such that $\mathscr{M}_{(K_1,\,S_1)}\subsetneq\mathscr{M}_{(K_2,\,S_2)}$, one has

$$\lim_{n\to\infty}\left(pen\left(D_{(K_2,\,S_2)},\,n\right)-pen\left(D_{(K_1,\,S_1)},\,n\right)\right)=\infty.$$

We need the following weak assumption:

$$(H)\;:\;\forall x\in\mathbb{X},\,P_0(x)>0, \tag{23}$$

where $\mathbb{X}$ is the set of distinct genotypes defined by the observed allele states, and P_0 the true probability distribution of the observations. Assumption (H) is not too strong since only observed alleles are considered. Recall that the research of the optimum model can be done among a finite sub collection $\mathscr{C}_{K_{\max}}$ (see equation 15)

Theorem 2 *Assume* (H) *and* $P_0\in\mathscr{M}:=\bigcup_{(K,\,S)\in\{1,\ldots,\,K_{\max}\}\times\mathscr{P}^*(L)}\mathscr{M}_{(K,\,S)}$. *Let*

$$\left(\widehat{K}_n,\,\widehat{S}_n\right):=\arg\min_{(K,\,S)\in\{1,\ldots,\,K_{\max}\}\times\mathscr{P}^*(L)}crit\,(K,\,S)$$

with $crit\,(K,\,S):=\gamma_n\left(\widehat{P}_{(K,S)}\right)+pen_n(K,S)$, *where* $pen_n(K,S)$ *has form (22) and fulfills* $(P1)$ *and* $(P2)$. *Then*

$$P_0\left[\left(\widehat{K}_n,\,\widehat{S}_n\right)=(K_0,\,S_0)\right]\underset{n\to\infty}{\to}1. \tag{24}$$

The proof is given in Appendix C.

The BIC is the most used of the asymptotic penalized maximum likelihood criteria fulfilling properties $(P1)$ and $(P2)$. Recall that for a given model $\mathscr{M}_{(K,\,S)}$, this criterion can be written as follows

$$BIC\,(K,\,S) := -\frac{1}{n}\sum_{i=1}^{n}\ln P_{(K,\,S)}\left(x_i|\;\widehat{\theta}_{MLE,(K,S)}\right)+\frac{D_{(K,\,S)}\ln n}{2n}, \tag{25}$$

where $\widehat{\theta}_{MLE,(K,S)} = \arg\max_{\theta\in\Theta_{(K,\,S)}}\sum_{i=1}^{n}\ln P_{(K,\,S)}\,(x_i|\;\theta)$. Thus the following corollary is a direct consequence of theorem 2.

Corollary 1 *Under assumption* (H)*, the estimator of* $(K_0,\,S_0)$ *given by* $\left(\widehat{K}_n,\,\widehat{S}_n\right) := \arg\min_{(K,\,S)} BIC\,(K,\,S)$ *is such that*

$$P\left[\left(\widehat{K}_n,\,\widehat{S}_n\right)=(K_0,\,S_0)\right]\xrightarrow[n\to\infty]{P_0}1. \tag{26}$$

This theoretical result on the consistency of the BIC holds empirically (see Subsection 4.1).

4 Numerical experiments

Our proposed method named *MixMoGenD* (*Mixture Model for Genotype Data*) has been implemented using $C++$ programming language. This section is devoted to the numerical experiments that illustrate its behavior and highlight the benefits of the loci selection procedure. In subsection 4.1, results of numerical experiments on simulated datasets are reported, and in subsection 4.2, the real dataset used in *N. A. Rosenberg et al.* (2001) [21, N. A. Rosenberg et al. (2001)] is considered[1]. Since some of the competing models are nested, we used the BIC for both simulated and real experiments as recommended by *Y. Wang and Q. Liu* (2006) [23].

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, EM algorithm is known to converge slowly in some situations and its solution can highly depend of its starting position and consequently produce sub-optimal maximum likelihood estimates. To act against this dependency of EM on its initial position, CEM (Classification EM) and SEM (Stochastic EM) have been proposed. We opt for the strategy of short runs of EM from random positions followed by a long run of EM from the solution maximizing the observed loglikelihood (See *C. Biernacki, G. Celeux and G. Govaert* (2001) [5]).

[1] This dataset is available on `http://rosenberglab.bioinformatics.med.umich.edu/jewishAut.html`

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

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

4.1 **Simulation examples**

4.1.1 ***First series***

The goal in the first series of simulated datasets is to see how the increase of the size of the sample improves the capacity of our clustering method to select the "smallest" model $\mathscr{M}_{(K_0,\ S_0)}$. We start with $n = 100$ individuals, and gradually increase this sample size to 400 by a step of 50. We assume a clustering structure with $K_0 = 2$ populations, $L = 4$ loci with 2 alleles per locus, the subset S_0 of clustering variables with cardinality $|S_0| = 2$. For each value n of the sample size, 100 datasets are generated using the parameters given in Table 1. As seen on Figure 1, *MixMoGenD* consistently identify the true model as $n \to \infty$. Other simulated datasets with $K_0 = 3$ clusters, $L = 6$ loci and cardinality of the subset set of clustering loci $|S_0| = 4$ confirmed these results. Thus, the theoretical result on the consistency that we showed in Section 3 holds empirically.

4.1.2 ***Second series***

Two other series of simulations are conducted to highlight the benefit of the variable selection procedure in our settings. First, we independently generated 100 datasets each with 1 000 individuals typed at $L = 6$ loci. We choose $K_0 = 3$ populations and $|S_0| = 4$ clustering loci. Simulation parameters are given in Table 2. Using all the 6 loci, the true model is selected **39** times against **61** for the model with $\widehat{K}_n = 2$ clusters. When including the variable selection procedure, *MixMoGenD* selects the true model $(K_0,\ S_0)$ **90 times** against **10** for $(K,\ S) = (2,\ S_0)$. Empirically, it appears that the number of populations can be under estimated when considering all available loci as relevant for clustering.

4.1.3 ***Third series***

We assume more variability in the third series. Here, each of the simulated datasets consists of 1 000 individuals structured into 5 subpopulations of equal proportions. We assume $L = 10$ loci each with 10 alleles, and four different cardinalities for S_0: 8, 6, 4 and 2. For each cardinality of S_0, we simulate 30 samples such that their Wright's parameter F_{ST}[2] are in $[0.0181,\ 0.0450]$. It is said in population genet-

[2] Wright's F statistics (Wright 1931) are the most widely used measures of population structure).

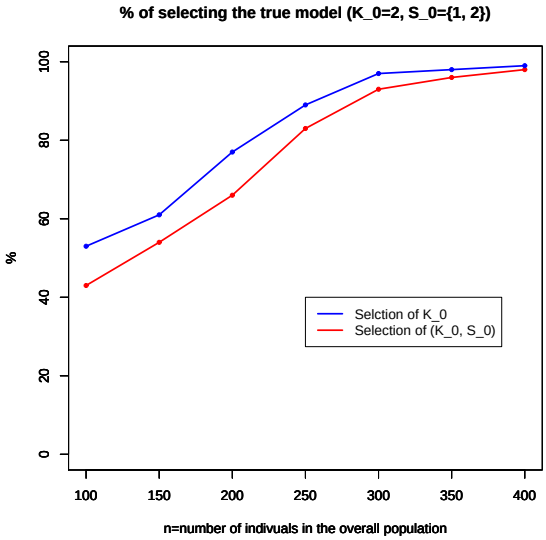

Fig. 1 % of selecting the true number K_0 of clusters and true model $(K_0,\ S_0)$ vs the sample size.

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

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

ics that unsupervised clustering is difficult with such a range of F_{ST} (*E. K. Latch et am.* (2006) [14]). We assume the uniform distribution for the alleles of the loci in S_0^c. These simulated datasets and their simulation parameters are available on `http://www.math.u-psud.fr/~toussile`. We used $K_{\max} = 10$ for all these simulations.

On these simulated samples, *MixMoGenD* gives three main results (see Tables 5, 6, 7 and 8). First, the true subset of clustering loci is systematically selected for all these simulations. Second, as expected, the variable selection procedure improves significantly the inference on the number K of clusters and the prediction capacity

Table 3 Example matrix of pairwise F_{ST}: the F_{ST} between population 4 and the others are all < 0.0260. MixMoGenD on this data set produces 4 clusters and we observed that Pop4 was uniformly distributed in the 4 clusters.

	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

measured by the percentage of missassigned individuals (% MA). In fact, the number of clusters can be underestimated when considering loci that are not relevant for clustering. Third, it appears that the benefit of the selection procedure is more important with the decrease of cardinality of the subset S_0. The more striking samples are the ones with 2 clustering variables (see Table 8). When using variable selection, the thresholds of F_{ST} for which *MixMoGenD* perfectly selects the true number K_0 of populations are 0.0342, 0.0307, 0.0316 and 0.0248 for $|S_0|$ equal to 8, 6, 4 and 2 respectively. These thresholds are more greater when using all loci as relevant for clustering (For example 0.0425 for $|S_0| = 8$). In addition, for each simulated sample for which $\widehat{K}_n < K_0$, we compute the square matrix of the pairwise F_{ST} between populations using the function *Fstat* of package *Geneland* [12] of **R** program. We observe that for each cardinality of S_0 we considered, 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 is approximately equal to 0.0270 on our simulated datasets with $|S_0| = 8$. The more striking example is the data 5 in Table 5 (d). The square matrix of pairwise F_{ST} is given in Table 3. The F_{ST} between population 4 and the others are all less than 0.0260. On this dataset, *MixMoGenD* produces 4 clusters and we observed that Pop4 was uniformly distributed in the 4 clusters.

4.2 **Real dataset example**

The dataset we considered consists of 159 males from 8 populations (6 Jewish and 2 non-Jewish populations): *Ashkenazi Jews* from Poland (20), *Druze* (20), *Ethiopian Jews* (19), *Iraqi Jews* (20), *Libyan Jews* (20), *Moroccan Jews* (20), *Palestinian Arabs* (20) and *Yemenite Jews* (20). Individuals were genotyped for 20 unlinked microsatellites spread across 14 autosomes. For this dataset, the question of interest is the relationship among these populations. See *N. A. Rosenberg et al.* (2001) [21] for a complete description of this dataset, in which the authors used several statistical analysis. To test the correspondence of genetic clusters with culturally labeled groups, they used the computer program *STRUCTURE* proposed by *J. K. Pritchard et al.* (2000) [19]. As *MixMoGenD*, this program implements a model-based clustering which identifies clusters of genetically similar diploid individuals from multilocus genotypes without prior knowledge of they population affinities. However, it does not contain a variable selection procedure which is the key point of *MixMoGenD*.

Table 4 Result from *MixMoGenD* with $K_{\max} = 10$. (a) Using the 8 populations: $\widehat{K}_n = 2$ clusters and the subset of clustering loci $\widehat{S}_n = \{D10S1426,\ D10S677\}$. This result indicates that the *Libyan Jewish* appelation labeled not only a cultural group, but also a genetic cluster. (b) The sample without the *Libyan Jewish*: $K_{\max} = 10$: $\widehat{K}_n = 2$ clusters and the subset of clustering loci $\widehat{S}_n = \{D1S1679\}$. This table suggests gene flow between these populations, particularly between *Ethiopian Jews*, *Moroccan Jews* and *Yemenite Jews* in one hand, and *Ashkenazi Jews*, *Druze* and *Palestinians* in the other hand.

	(a)		(b)	
	Cluster 1	Cluster 2	Cluster 1	Cluster 2
Ashkenazi	0	**20**	7	**13**
Druze	1	**19**	3	**17**
Ethiopian Jews	2	**17**	**17**	2
Iraqi Jews	3	**17**	9	**11**
Libyan Jews	**19**	1	-	-
Moroccan Jews	0	**20**	**14**	6
Palestinians	2	**18**	1	**19**
Yemenite Jews	0	**20**	**14**	6
%	0.17	0.83	0.47	0.53

MixMoGenD revealed a cluster that almost coincided with the sample of *Libyan Jews* (Table 4 (a). Of 20 *Libyan Jewish* individuals in the sample, 19 fell into cluster 1, while only 8 other individuals also fell into this cluster. Cluster 1 is similar to cluster 3 reported in [19] using *STRUCTURE*, indicating that the *Lybian Jewish* appelation labeled not only a cultural group, but also a genetic cluster. The additional important information obtained by *MixMoGenD* is the subset of clustering loci: only 2 loci $D10S1426$ and $D10S677$ suffice to distinguish *Libyan Jews* from the other populations. The other sampled individuals fell into cluster 2. Subclustering analysis showed that the sample whitout *Libyan Jews* could be divided in 2 clusters with the subset of clustering loci containing only one locus which is the tetranucleotide $D1S1679$ (Table 4 (b). This subclustering does not clearly separate any of the 7 populations felled in cluster 2 in the previous clustering analysis, but it suggests gene flow between these populations, particularly between *Ethiopian Jews*, *Moroccan Jews* and *Yemenite Jews* in one hand, and *Ashkenazi Jews*, *Druze* and *Palestinians* in the other hand.

5 Discussion

We believe that *MixMoGenD* will be useful for two main reasons. First, like *FASTRUCT*, our method 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].

More importantly, 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. 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. This improvement tends to be more important when the number of clustering variables decreases. We also found

that even in situations where measures of population structure such as F_{st} are in range where it is thought that clustering is difficult (*E. K. Latch* (2005) [14]), *MixMoGenD* perfectly identified the subset of clustering variables. In addition, due to the explosion of genomic projects, datasets are becoming increasingly large. The space of the models in competition 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 in *O. Francois et al.* (2006) [7]. 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. We also propose a modification of the Backward-Stepwise algorithm that we named *Backward-Stepwise explorer* (see 18), which enables not only to avoid an exhaustive research of the optimum model, but also the reseach of the optimum subset of clustering loci among all possible cardinalities.

Although the theoretical result on the consistency of the BIC was verified empirically, it is well known that this criterion is not uniformly the best one. We currently work on data dependent calibration of the penalty function in order to obtain an oracle inequality.

Acknowledgements This work was supported by a doctoral fellowship from "Institut de Recherche pour le Développement" (IRD). The authors thank Professor Henri Gwet for some helpful suggestions, Dr Isabelle Morlais for the explanations of the biological concepts we needed and Professor Gilles Celeux for a critical reading of the original version of the paper.

Table 5 Results given by *MixMoGenD* on 30 samples each with $n = 1\ 000$ individuals structured into $K_0 = 5$ populations of equal mixing proportions. We assume $L = 10$ loci typed and $|S_0| = 8$ clustering loci. The datasets are simulated so that the F_{st} are in $[0.0306,\ 0.0450]$. % MA and % MA^s = percentage of missassigned individuals without and with loci selection respectively; $\widehat{K}_n$ and $\widehat{K}_n^s$ = the estimates of the number of populations without and with loci selection respectively.

Data	F_{st}	$\widehat{K}_n$	% MA	$\widehat{K}_n^s$	% MA^s	Data	F_st	$\widehat{K}_n$	% MA	$\widehat{K}_n^s$	% MA^s
1	0.0306	3		3		16	0.0381	5	10.90	5	10.30
2	0.0318	3		3		17	0.0382	5	09.30	5	08.80
3	0.0328	3		3		18	0.0390	4		4	09.10
4	0.0331	3		3		19	0.0400	5	08.80	5	08.00
5	0.0335	3		4		20	0.0404	4		4	09.50
6	0.0337	3		3		21	0.0425	5	06.30	5	05.40
7	0.0340	4		4		22	0.0427	5	07.10	5	07.50
8	0.0342	3		5	11.80	23	0.0427	5	05.90	5	05.90
9	0.0348	3		5	12.40	24	0.0435	5	06.70	5	06.50
10	0.0362	3		5	09.10	25	0.0436	5	07.10	5	06.60
11	0.0373	4		5	08.90	26	0.0440	5	05.50	5	05.70
12	0.0373	5	08.50	5	07.60	27	0.0442	5	07.20	5	06.80
13	0.0377	5	11.40	5	10.40	28	0.0449	5	07.20	5	06.70
14	0.0377	5	10.50	5	10.20	29	0.0449	5	06.10	5	06.30
15	0.0377	5	10.30	5	10.20	30	0.0450	5	06.10	5	05.60

Table 6 Results given by *MixMoGenD* on 30 samples each with $n = 1\ 000$ individuals structured into $K_0 = 5$ populations of equal mixing proportions. We assume $L = 10$ loci typed and $|S_0| = 6$ clustering loci. The datasets are simulated so that the F_{st} are in $[0.0280,\ 0.0339]$. % MA and % MA^s = percentage of missassigned individuals without and with loci selection respectively; $\widehat{K}_n$ and $\widehat{K}_n^s$ = the estimates of the number of populations without and with loci selection respectively.

Data	F_{st}	$\widehat{K}_n$	% MA	$\widehat{K}_n^s$	% MA^s	Data	F_{st}	$\widehat{K}_n$	% MA	$\widehat{K}_n^s$	% MA^s
1	0.0280	2		4		16	0.0309	2		5	13.90
2	0.0284	1		5	15.20	17	0.0310	2		5	11.70
3	0.0285	1		5	14.30	18	0.0310	3		5	12.20
4	0.0287	2		5	14.70	19	0.0311	3		5	12.00
5	0.0289	2		5	13.40	20	0.0314	2		5	12.80
6	0.0289	2		5	13.60	21	0.0319	3		5	10.60
7	0.0290	1		5	14.20	22	0.0319	4		5	11.00
8	0.0291	3		4		23	0.0321	4		5	11.30
9	0.0296	2		4		24	0.0321	4		5	11.50
10	0.0299	2		5	12.20	25	0.0325	4		5	10.50
11	0.0303	2		4		26	0.0329	4		5	10.70
12	0.0305	3		4		27	0.0330	4		5	09.80
13	0.0307	2		5	14.80	28	0.0333	3		5	12.50
14	0.0307	2		5	12.10	29	0.0337	3		5	09.70
15	0.0308	2		5	15.10	30	0.0339	4		5	09.60

Table 7 Results given by *MixMoGenD* on 30 samples each with $n = 1\ 000$ individuals structured into $K_0 = 5$ populations of equal mixing proportions. We assume $L = 10$ loci typed and $|S_0| = 4$ clustering loci. The datasets are simulated so that the F_{st} are in $[0.0302,\ 0.0413]$. % MA and % MA^s = percentage of missassigned individuals without and with loci selection respectively; $\widehat{K}_n$ and $\widehat{K}_n^s$ = the estimates of the number of populations without and with loci selection respectively.

Data	F_{st}	$\widehat{K}_n$	% MA	$\widehat{K}_n^s$	% MA^s	Data	F_{st}	$\widehat{K}_n$	% MA	$\widehat{K}_n^s$	% MA^s
1	0.0302	2		4		16	0.0338	3		5	10.50
2	0.0303	1		5	12.80	17	0.0345	3		5	08.60
3	0.0309	2		4		18	0.0349	3		5	08.50
4	0.0316	3		5	12.90	19	0.0354	3		5	11.90
5	0.0317	2		5	15.10	20	0.0359	3		5	10.80
6	0.0320	3		5	13.30	21	0.0388	4		5	06.40
7	0.0322	2		5	10.80	22	0.0390	4		5	06.70
8	0.0323	3		5	09.70	23	0.0391	4		5	07.40
9	0.0326	2		5	13.80	24	0.0393	4		5	07.40
10	0.0327	2		5	12.10	25	0.0394	5	07.90	5	06.00
11	0.0327	3		5	14.10	26	0.0399	4		5	07.60
12	0.0327	3		5	09.90	27	0.0402	4		5	07.30
13	0.0329	2		5	13.10	28	0.0408	4		5	07.90
14	0.0332	3		5	13.50	29	0.0412	4		5	07.10
15	0.0332	3		5	11.10	30	0.0413	5	06.40	5	07.30

Table 8 Results given by *MixMoGenD* on 30 samples each with $n = 1\ 000$ individuals structured into $K_0 = 5$ populations of equal mixing proportions. We assume $L = 10$ loci typed and $|S_0| = 2$ clustering loci. The datasets are simulated so that the F_{st} are in $[0.0181,\ 0.0266]$. % MA and % MA^s = percentage of missassigned individuals without and with loci selection respectively; $\widehat{K}_n$ and $\widehat{K}_n^s$ = the estimates of the number of populations without and with loci selection respectively.

Data	F_{st}	$\widehat{K}_n$	% MA	$\widehat{K}_n^s$	% MA^s	Data	F_{st}	$\widehat{K}_n$	% MA	$\widehat{K}_n^s$	% MA^s
1	0.0181	1		4		16	0.0232	2		5	16.40
2	0.0186	1		4		17	0.0232	2		5	15.50
3	0.0193	1		4		18	0.0235	2		5	14.70
4	0.0195	1		4		19	0.0237	2		5	16.20
5	0.0195	1		4		20	0.0242	1		5	17.80
6	0.0199	1		4		21	0.0244	2		5	15.50
7	0.0199	1		4		22	0.0247	1		4	
8	0.0203	1		4		23	0.0248	1		5	16.60
9	0.0205	1		4		24	0.0249	1		5	19.30
10	0.0216	1		4		25	0.0251	1		5	16.40
11	0.0222	2		5	15.30	26	0.0252	1		5	15.00
12	0.0227	1		5	17.10	27	0.0252	1		5	15.40
13	0.0229	2		5	15.80	28	0.0254	1		5	14.70
14	0.0230	2		5	14.90	29	0.0263	1		5	18.00
15	0.0230	2		5	14.60	30	0.0266	1		5	16.30

References

1. H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.
2. E. S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of latent class models with many observed variables. *Annals of Stat.*, to appear.
3. Zeinab Annan, Patrick Durand, Francisco J Ayala, Céline Arnathau, Parfait Awono-Ambene, Frédéric Simard, Fabien G Razakandrainibe, Jacob C Koella, Didier Fontenille, and François Renaud. Population genetic structure of Plasmodium falciparum in the two main african vectors, anopheles gambiae and anopheles funestus. *Proc Natl Acad Sci U S A*, 104(19):7987–92, may 2007.
4. J.-M. Azais, E. Gassiat, and C. Mercadier. The likelihood ratio test for general mixture models with possibly structural parameter. *ESAIM P&S,* to appear.
5. C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the EM algorithm for getting highest likehood in multivariate Gaussian mixture models. *Computational Statistics and Data Analysis*, 41, 561-575, 2003.
6. A. Chambaz, A. Garivier, and E. Gassiat. A MDL approach to HMM with Poisson and Gaussian emissions: Application to order identification. *Journal of Statistical Planning and Inference*, to appear.
7. C. Chen, F. Forbes, and O. Francois. fastruct: model-based clustering made faster. *Molecular Ecology Notes*, 6(4):980–983, 2006.
8. J. Corander, P. Waldmann, P. Marttinen, and M.J. Sillanpaa. BAPS 2: enhanced possibilities for the analysis of genetic population structure, 2004.
9. 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.
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. E. Gassiat. Likelihood ratio inequalities with applications to various mixtures. In *Annales de l'Institut Henri Poincaré/Probabilités et statistiques*, volume 38, pages 897–906. Elsevier SAS, 2002.
12. G. Guillot, F. Mortier, and A. Estoup. Geneland: a computer package for landscape genetics. *Molecular Ecology Notes*, 5(3):712–715, 2005.
13. C. Keribin. Consistent estimation of the order of mixture models. *Sankhyā Ser. A*, 62(1):49–66, 2000.

14. 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.
15. P Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.
16. C. Maugis, G. Celeux, and M.L. Martin-Magniette. Variable Selection for Clustering with Gaussian Mixture Models. *Biometrics*, to appear.
17. Xiang-Feng Meng, Min Shi, and Xue-Xin Chen. Population genetic structure of Chilo suppressalis (Walker) (Lepidoptera: Crambidae): strong subdivision in china inferred from microsatellite markers and mtDNA gene sequences. *Mol Ecol*, 17(12):2880–2897, 2008.
18. Lisa Mirabello, Joseph H. Vineis, Stephen P. Yanoviak, Vera M. Scarpassa, Marinete M. Póvoa, Norma Padilla, Nicole L. Achee, and Jan E. Conn. Microsatellite data suggest significant population structure and differentiation within the malaria vector Anopheles darlingi in Central and South America. *BMC Ecol*, 8:3, 2008.
19. J K Pritchard, M Stephens, and P Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–59, 2000.
20. A.E. Raftery and N. Dean. Variable Selection for Model-Based Clustering. *Journal-American Statistical Association*, 101(473):168, 2006.
21. N A Rosenberg, E Woolf, J K Pritchard, T Schaap, D Gefel, I Shpirer, U Lavi, B Bonne-Tamir, J Hillel, and M W Feldman. Distinctive genetic signatures in the libyan jews. *Proc Natl Acad Sci U S A*, 98(3):858–63, 2001.
22. Philipp Trénel, Michael M. Hansen, Signe Normand, and Finn Borchsenius. Landscape genetics, historical isolation and cross-Andean gene flow in the wax palm, Ceroxylon echinulatum (Arecaceae). *Mol Ecol*, 2008.
23. Y. Wang and Q. Liu. Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of stock–recruitment relationships. *Fisheries Research*, 77(2):220–225, 2006.

A Proof of lemma 1

Let $P \in \mathscr{M}_{(K,\,S)}$ and let $\theta = (\pi,\ \alpha,\ \beta) \in \Theta_{(K,\,S)}$ be the parameter defining P. Assume without loss of generality that $\pi_K > 0$ (If not, recall that in the context of finite mixture, the latent classes can be freely relabeled without changing the distribution underlining the observations). Define for instance $\theta' = (\pi',\ \alpha',\ \beta') \in \Theta_{(K+1,\,S)}$ as follows

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

Obviously, One has $P(\cdot) = P_{(K+1,\,S)}(\,\cdot \mid \theta')$. So P is an element of model $\mathscr{M}_{(K+1,\,S)}$.

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

$$\alpha_{1,l,\,\cdot} = \ldots = \alpha_{K,l,\,\cdot}\ \forall l \in S' \setminus S. \tag{27}$$

B Proof of lemma 2

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

$$P(x) = \left[\sum_{k=1}^{K_1} \pi_k^1 \prod_{l \in S_1} P\left(x^l \mid (\alpha^1_{k,\,l,\,\cdot})\right)\right] \times \prod_{l \in S_1^c} P\left(x^l \mid (\beta^1_{l,\,\cdot})\right), \tag{28}$$

$$P(x) = \left[\sum_{k=1}^{K_2} \pi_k^2 \prod_{l \in S_2} P\left(x^l \mid (\alpha^2_{k,\,l,\,\cdot})\right)\right] \times \prod_{l \in S_2^c} P\left(x^l \mid (\beta^2_{l,\,\cdot})\right), \tag{29}$$

where $\theta^1 := \left(\pi^1,\ \alpha^1,\ \beta^1\right)$ and $\theta^2 := \left(\pi^2,\ \alpha^2,\ \beta^2\right)$ are in $\Theta_{(K_1,\,S_1)}$ and $\Theta_{(K_2,\,S_2)}$ respectively. Assume without loss 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 $C = L \setminus S_1 \cup S_2$, where L denotes the $\{1, \ldots,\ L\}$ of all typed loci. Using equation (28), the marginal probability distribution of the sub-vector $x^{S_2} := \left(x^l\right)_{l \in S_2}$ is given by

$$P\left(x^{S_2}\right) = \left[\sum_{k=1}^{K_1} \pi_k^1 \prod_{l \in S_1 \cap S_2} P\left(x^l \mid (\alpha^1_{k,\,l,\,\cdot})\right)\right] \times \prod_{l \in B} P\left(x^l \mid (\beta^1_{l,\,\cdot})\right), \tag{30}$$

and using equation (29) one has

$$\begin{aligned}
P(x) &= \left[\sum_{k=1}^{K_1} \pi_k^1 \prod_{l \in S_1 \cap S_2} P\left(x^l \mid (\alpha^1_{k,\,l,\,\cdot})\right)\right] \times \prod_{l \in B} P\left(x^l \mid (\beta^1_{l,\,\cdot})\right) \times \prod_{l \in A \cup C} P\left(x^l \mid (\beta^2_{l,\,\cdot})\right) \\
&= \left[\sum_{k=1}^{K_1} \pi_k^1 \prod_{l \in S_1 \cap S_2} P\left(x^l \mid (\alpha^1_{k,\,l,\,\cdot})\right)\right] \times \prod_{l \in A \cup B \cup C} P\left(x^l \mid \left(\beta^3_{l,\,\cdot}\right)\right),
\end{aligned}$$

where β^3 is defined as follows

$$\begin{aligned}
\beta^3_l &= \beta^1_l \text{ if } l \in B \\
\beta^3_l &= \beta^2_l \text{ if } l \in A \cup C.
\end{aligned}$$

Consequently, P is an element of model $\mathscr{M}_{(K_1 \wedge K_2,\ S_1 \cap S_2)}$.

C Proof of Theorem 2

For any $0 < \delta < 1$, define the compact set

$$\Theta^{\delta}_{(K,\,S)} = \left\{ \theta \in \Theta_{(K,\,S)} \,:\, \forall x \in \mathbb{X},\, P_{(K,\,S)}\left(x|\,\theta\right) \geq \delta \right\}. \tag{31}$$

We shall need the following proposition whose proof is given below in Appendix D.

Proposition 1 *Under assumption (H), there exists a real $\delta > 0$ such that for every $(K,\,S)$, one has*

$$-\gamma_n\left(\widehat{P}_{(K,\,S)}\right) = \sup_{\theta \in \Theta^{\delta}_{(K,\,S)}} \left\{ -\gamma_n\left(P_{(K,\,S)}\left(\,\cdot\mid\theta\right)\right) \right\} + o_{P_0}(1) \tag{32}$$

and

$$\sup_{\theta \in \Theta_{(K,\,S)}} E_{P_0}\left[\ln P_{(K,\,S)}\left(X|\,\theta\right)\right] = \sup_{\theta \in \Theta^{\delta}_{(K,\,S)}} E_{P_0}\left[\ln P_{(K,\,S)}\left(X|\,\theta\right)\right]. \tag{33}$$

One has the following upper bound

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

where the summation is for $(K,S) \in \{1,\ldots,\,K_{\max}\} \times \mathscr{P}^*(L)$ and has a finite number of terms. It thus suffices to prove that $\lim_{n\to\infty} P_0\left[\left(\widehat{K}_n,\,\widehat{S}_n\right) = (K,\,S)\right] = 0$ for every $(K,\,S) \neq (K_0,\,S_0)$.

Let $(K,\,S)$ be an element of $\{1,\ldots,\,K_{\max}\} \times \mathscr{P}^*(L)$ such that $(K,\,S) \neq (K_0,\,S_0)$. The probability $P_0\left[\left(\widehat{K}_n,\,\widehat{S}_n\right) = (K,\,S)\right]$ is bounded by

$$P_0\left[crit\,(K,\,S) < crit\,(K_0,\,S_0)\right] = P_0\left[\gamma_n\left(\widehat{P}_{(K_0,\,S_0)}\right) - \gamma_n\left(\widehat{P}_{(K,\,S)}\right) > pen_n\,(K,\,S) - pen_n\,(K_0,\,S_0)\right], \tag{34}$$

where $\gamma_n(P) := -\dfrac{1}{n}\sum_{i=1}^{n} \ln P(x_i)$ and $\widehat{P}_{(K,S)}$ is the maximum likelihood estimator (MLE) in $\mathscr{M}_{(K,\,S)}$. Two cases are considered: $P_0 \in \mathscr{M}_{(K,\,S)}$ and $P_0 \notin \mathscr{M}_{(K,\,S)}$.

• **Case 1:** $P_0 \in \mathscr{M}_{(K,\,S)}$, i.e there exists a parameter $\theta_{0,K,S}$ in $\Theta_{(K,\,S)}$ such that $P_0 = P_{(K,\,S)}\left(\cdot|\,\theta_{0,K,S}\right)$. Denote $\mathscr{D}$ the set of all possible probability distributions on the set $\mathbb{X}$ of the genotype states. Since $\mathscr{M}_{(K_0,\,S_0)} \subseteq \mathscr{M}_{(K,\,S)} \subseteq \mathscr{D}$, one has the following inequalities

$$-n\gamma_n(P_0) \leq -n\gamma_n\left(\widehat{P}_{(K_0,\,S_0)}\right) \leq -n\gamma_n\left(\widehat{P}_{(K,\,S)}\right) \leq \sup_{P\in\mathscr{M}}\left(-n\gamma_n(P)\right),$$

so that

$$0 \leq -n\gamma_n\left(\widehat{P}_{(K,\,S)}\right) + n\gamma_n\left(\widehat{P}_{(K_0,\,S_0)}\right) \leq \sup_{P\in\mathscr{M}}\left(-n\gamma_n(P)\right) + n\gamma_n(P_0).$$

But it is well known that $2\sup_{P\in\mathscr{M}}\left(-n\gamma_n(P)\right) + 2n\gamma_n(P_0)$ converges in distribution to a chi-square variable with $|\mathbb{X}|-1$ degrees of freedom, where $|\mathbb{X}|$ denote the cardinality of $\mathbb{X}$. Thus $-n\gamma_n\left(\widehat{P}_{(K,\,S)}\right) + n\gamma_n\left(\widehat{P}_{(K_0,\,S_0)}\right)$ is bounded in probability. But if P_0 is an element of model $\mathscr{M}_{(K,\,S)}$ and $(K,S) \neq (K_0,S_0)$, one has $\mathscr{M}_{(K_0,\,S_0)} \subsetneq \mathscr{M}_{(K,\,S)}$, and it follows from $(P3)$ that $pen\left(D_{(K,\,S)},\,n\right) - pen\left(D_{(K_0,\,S_0)},\,n\right)$ is positive and tends to infinity as n tends to infinity. Thus

$$P_0\left[n\gamma_n\left(\widehat{P}_{(K_0,\,S_0)}\right) - n\gamma_n\left(\widehat{P}_{(K,\,S)}\right) > pen\left(D_{(K,\,S)},\,n\right) - pen\left(D_{(K_0,\,S_0)},\,n\right)\right].$$

tends to 0 as n tends to infinity.

• **Case 2:** $P_0 \notin \mathscr{M}_{(K,\,S)}$, i.e for all θ in $\Theta_{(K,\,S)}$, one has $P_0 \neq P_{(K,\,S)}(\cdot|\,\theta)$. By equation (32) of proposition 1, there exists a positive real δ such that

$$-\gamma_n\left(\widehat{P}_{(K,\,S)}(\cdot|\,\theta)\right) = \sup_{\theta\in\Theta^{\delta}_{(K,\,S)}}\left\{-\gamma_n\left(P_{(K,\,S)}(\cdot|\,\theta)\right)\right\} + o_{P_0}(1).$$

The set of functions $\mathscr{F}^{\delta}_{(K,\,S)} := \left\{\ln P_{(K,\,S)}(\,\cdot\,|\,\theta), \theta \in \Theta^{\delta}_{(K,\,S)}\right\}$ is obviously P_0-Glivenko-Cantelli, so that

$$-\gamma_n\left(\widehat{P}_{(K,\,S)}(\cdot|\,\theta)\right) = \sup_{\theta\in\Theta^{\delta}_{(K,\,S)}} E_{P_0}\left[\ln P_{(K,\,S)}(X|\,\theta)\right] + o_{P_0}(1).$$

On the other hand, since P_0 is an element of $\mathscr{M}_{(K_0,\,K_0)}$, it is well known that

$$\inf_{\theta\in\Theta_{(K_0,\,S_0)}} E_{P_0}\left[\ln P_0(X) - \ln P_{(K_0,\,S_0)}(X|\,\theta)\right] = 0.$$

so that

$$\begin{aligned}\sup_{\theta\in\Theta^{\delta}_{(K_0,\,S_0)}} E_{P_0}\left[\ln P_{(K_0,\,S_0)}(X|\,\theta)\right] &= \sup_{\theta\in\Theta_{(K_0,\,S_0)}} E_{P_0}\left[\ln P_{(K_0,\,S_0)}(X|\,\theta)\right] \\ &\quad \text{(see equation (33) of proposition 1)} \\ &= E_{P_0}\left[\ln P_0(X)\right].\end{aligned}$$

Thus

$$\gamma_n\left(\widehat{P}_{(K_0,\,S_0)}\right) - \gamma_n\left(\widehat{P}_{(K,\,S)}\right) = -\inf_{\theta\in\Theta^{\delta}_{(K,\,S)}} E_{P_0}\left[\ln P_0(X) - \ln P_{(K,\,S)}(X|\,\theta)\right] + o_{P_0}(1).$$

In addition, the function $\theta \mapsto E_{P_0}\left[\ln P_0(X) - \ln P_{(K,\,S)}(X|\,\theta)\right]$ is continuous on the compact set $\Theta^{\delta}_{(K,\,S)}$ and recall that in this case P_0 is not in $\mathscr{M}_{(K,\,S)}$. Consequently, one has

$$-\inf_{\theta\in\Theta^{\delta}_{(K,\,S)}} E_{P_0}\left[\ln P_0(X) - \ln P_{(K,\,S)}(X|\,\theta)\right] < 0.$$

Also notice that by $(P3)$, $pen_n(K,\,S) - pen_n(K_0,\,S_0)$ tends to 0 as n tends to infinity. Then one has

$$\lim_{n\to\infty} P\left[\gamma_n\left(\widehat{P}_{(K_0,\,S_0)}\right) - \gamma_n\left(\widehat{P}_{(K,\,S)}\right) > pen_n(K,\,S) - pen_n(K_0,\,S_0)\right] = 0,$$

which is the desired result.

D Proof of proposition 1

Let n_x denote the observed frequency of genotype x. It is well known that one has $\frac{n_x}{n} = P_0(x) + o_{P_0}(1)$, so that

$$\begin{aligned}-\gamma_n\left(P_{(K,\,S)}(\,\cdot\,|\,\theta)\right) &= \sum_{x\in\mathbb{X}} \frac{n_x}{n} \ln P_{(K,\,S)}(x|\,\theta) \\ &= \sum_{x\in\mathbb{X}}\left[P_0(x) + o_{P_0}(1)\right] \times \ln P_{(K,\,S)}(x|\,\theta). \qquad (35)\end{aligned}$$

For every $(K,\ S)$, there exists at least one real $0<\widetilde{\delta}<1$ such that $\Theta^{\widetilde{\delta}}_{(K,\ S)}$ is not empty. Let $\widetilde{\delta}$ be such a real and $\widetilde{\theta}$ an element of $\Theta^{\widetilde{\delta}}_{(K,\ S)}$. By assumption (H) and using equation (35), one has the following inequality

$$-\gamma_n\left(P_{(K,\ S)}\left(\ \cdot\mid\widetilde{\theta}\right)\right)\geq\sum_{x\in\mathbb{X}}P_0\left(x\right)\ln\widetilde{\delta}+o_{P_0}\left(1\right)=\ln\widetilde{\delta}+o_{P_0}\left(1\right). \tag{36}$$

Since $\mathbb{X}$ is a finite set, one has $0<\inf_{x\in\mathbb{X}}P_0\left(x\right)\leq 1$. Let δ be a real such that

$$0<\delta<\min\left\{\widetilde{\delta}^{\frac{1}{\inf_{x\in\mathbb{X}}P_0(x)}},\ \inf_{x\in\mathbb{X}}P_0\left(x\right)\right\}.$$

Obviously, one has $\widetilde{\delta}^{\frac{1}{\inf_{x\in\mathbb{X}}P_0(x)}}\leq\widetilde{\delta}$, so that one has $\Theta^{\widetilde{\delta}}_{(K,\ S)}\subset\Theta^{\delta}_{(K,\ S)}$ and then the following inequalities

$$\sup_{\theta\in\Theta^{\delta}_{(K,\ S)}}\left\{-\gamma_n\left(P_{(K,\ S)}\left(\ \cdot\mid\theta\right)\right)\right\}\geq\sup_{\theta\in\Theta^{\widetilde{\delta}}_{(K,\ S)}}\left\{-\gamma_n\left(P_{(K,\ S)}\left(\ \cdot\mid\theta\right)\right)\right\}, \tag{37}$$

$$\sup_{\theta\in\Theta^{\delta}_{(K,\ S)}}E_{P_0}\left[\ln P_{(K,\ S)}\left(X\mid\theta\right)\right]\geq\sup_{\theta\in\Theta^{\widetilde{\delta}}_{(K,\ S)}}E_{P_0}\left[\ln P_{(K,\ S)}\left(X\mid\theta\right)\right]. \tag{38}$$

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

$$\begin{aligned}
-\gamma_n\left(P_{(K,\ S)}\left(\ \cdot\mid\theta\right)\right)&\leq\inf_{\mathbb{X}}P_0\left(x\right)\ln\delta+o_{P_0}\left(1\right)\\
&\leq\inf_{\mathbb{X}}P_0\left(u\right)\ln\widetilde{\delta}^{\frac{1}{\inf_{\mathbb{X}}P_0(x)}}+o_{P_0}\left(1\right)\ =\ln\widetilde{\delta}+o_{P_0}\left(1\right)\\
&\leq-\gamma_n\left(P_{(K,\ S)}\left(\ \cdot\mid\widetilde{\theta}\right)\right)+o_{P_0}\left(1\right)\\
&\leq\sup_{\theta\in\Theta^{\widetilde{\delta}}_{(K,\ S)}}\left\{-\gamma_n\left(P_{(K,\ S)}\left(\ \cdot\mid\theta\right)\right)\right\}+o_{P_0}\left(1\right)\\
&\leq\sup_{\theta\in\Theta^{\delta}_{(K,\ S)}}\left\{-\gamma_n\left(P_{(K,\ S)}\left(\ \cdot\mid\theta\right)\right)\right\}+o_{P_0}\left(1\right).
\end{aligned} \tag{39}$$

Consequently one has

$$\sup_{\theta\notin\Theta^{\delta}_{(K,\ S)}}\left\{-\gamma_n\left(P_{(K,\ S)}\left(\ \cdot\mid\theta\right)\right)\right\}\leq\sup_{\theta\in\Theta^{\delta}_{(K,\ S)}}\left\{-\gamma_n\left(P_{(K,\ S)}\left(\ \cdot\mid\theta\right)\right)\right\}+o_{P_0}\left(1\right)$$

so that

$$-\gamma_n\left(\widehat{P}_{(K,\ S)}\left(\ \cdot\mid\theta\right)\right)=\sup_{\theta\in\Theta^{\delta}_{(K,\ S)}}\left\{-\gamma_n\left(P_{(K,\ S)}\left(\ \cdot\mid\theta\right)\right)\right\}+o_{P_0}\left(1\right).$$

Using the same arguments, one gets

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

which are the disired results.

E EM equations

Here we describe the EM equations. To assign individual i to a cluster, we compute the posterior assignment probabilities $\tau_{ik} = P(z_i = k|\, x_i)$. 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

$$\tau_{ik}^{(r)} = \frac{\pi_k^{(r)} \prod_{l \in S} P\left(x_i^l |\, z_i = k,\, \alpha_{k,l,\cdot}^{(r)}\right)}{\sum_{h=1}^{K} \pi_h^{(r)} \prod_{l \in S} P\left(x_i^l |\, z_i = h,\, \alpha_{h,l,\cdot}^{(r)}\right)} \tag{40}$$

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

$$\pi_k^{(r+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{(r)} \tag{41}$$

and

$$\alpha_{k,l,j}^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)} \left(\mathbb{1}_{\left[x_{i,1}^l = j\right]} + \mathbb{1}_{\left[x_{i,2}^l = j\right]}\right)}{2 \sum_{i=1}^{n} \tau_{ik}^{(r)}}. \tag{42}$$

EM algorithm is known to converge slowly in some situations and its solution can highly depend of its starting position and consequently produce sub-optimal maximum likelihood estimates. To act against this high dependency of EM on its initial position, CEM (Classification EM) and SEM (Stochastic EM) have been proposed. We opt for the strategy of short runs of EM from random positions followed by a long run of EM from the solution maximizing the observed loglikelihood (See *C. Biernacki, G. Celeux and G. Govaert* (2001) [5]).