N° d'ordre : 9911

THÈSE

Présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES DES
UNIVERSITÉS PARIS-SUD XI ET YAOUNDE 1

Spécialité : Mathématiques

par

Wilson Toussile

Sélection de variable : structure génétique d'une population et transmission de *Plasmodium* à travers le moustique

Soutenue le 29 Septembre 2010 devant la Commission d'examen :

M.	Bitjong Ndombol	(Examinateur)
Mme.	Elisabeth Gassiat	(Directrice de thèse)
M.	Gilles Celeux	(Examinateur)
M.	Henri Gwet	(Co-Directeur de thèse)
Mme.	Isabelle Morlais	(Examinatrice)
M.	Jean-Louis Foulley	(Rapporteur)
M.	Pascal Massart	(Président du jury)

Après avis des rapporteurs :

M.	Geoff	McLachlan
M.	Jean-Louis	Foulley

Thèse préparée dans les institutions suivantes :

Département de Mathématiques d'Orsay
Laboratoire de Mathématiques (UMR 8628), Bât. 425
Université Paris-Sud 11
91 405 Orsay CEDEX

UR016, Institut de Recherche pour le Développement (IRD)
Centre IRD de Montpellier
911, Avenue Agropolis - BP 64 501
F - 34 394 Montpellier cedex 5

Département de Mathématiques de Yaoundé 1
Faculté des Sciences
Université de Yaoundé 1
Cameroun

Thèse financée par :

Bourse de thèse attribuée par le
Département Soutien et Formation (DSF)
Institut de Recherche pour le Développement (IRD)
Le Sextant, 44 bd de Dunkerque
CS 90 009 - 13 572 Marseille - Cedex 02

Bourse SETCI attribuée par la région
Île De France

Bourse du Gouvernement français du
Ministère Affaires Étrangères et Europèennes
attribuée par l'**Ambassade de France au Cameroun**.

Remerciements

Je remercie mes directeurs de thèse Elisabeth Gassiat et Henri Gwet pour m'avoir accueilli en Master puis en thèse au sein du Laboratoire de Mathématiques de l'Université Paris-Sud 11 et du Laboratoire de Mathématiques et d'Analyse des Systèmes de l'ENSP de Yaoundé : merci pour avoir su raffermir ma passion pour la recherche, pour l'intérêt que vous portez à mes travaux et pour avoir toujours été disponibles pendant tout mon travail. Je me souviens particulièrement des "Je ne suis pas inquiète" d'Elisabeth qui arrivaient toujours aux moments où j'en avais le plus besoin.

J'ai aussi eu la chance de profiter de l'encadrement de Isabelle Morlais, qui m'a accueilli à l'UR CCPV "Caractérisation et Contrôle des Populations de Vecteurs" (IRD 016) pour mon stage de M2 à partir duquel le sujet de ma thèse a été construit : merci pour ta disponibilité et ta patience ; pour les discussions toujours éclairantes au sujet des notions biologiques nécessaires à la compréhension des données qui ont soutendues ce travail de thèse. J'ai été ravi de travailler avec toi et souhaite d'ailleurs que notre collaboration continue. Je remercie aussi tous les membres de l'UR016 de l'IRD, en particulier Frédéric Simard, Costantini Carlo, et le Directeur de cette unité de recherche Didier Fontenille, qui m'a donné l'occasion de présenter une partie de mes travaux de thèse au Centre IRD de Montpellier. Je dis merci à Garde Xavier, Représentant de l'IRD au Cameroun.

Merci à Jean-Louis Foulley et Geoff McLachlan qui ont gracieusement accepté de rapporter cette thèse.

Je suis très honoré de pouvoir compter parmi les membres de mon jury Pascal Massart, Gilles Celeux et Bitjong Ndombol (venu spécialement du Cameroun). je me réjouis de leur participation à mon jury de thèse.

Pour réaliser cette thèse dans de bonnes conditions, j'ai tout d'abord bénéficié de l'appui financier du Département Soutien et Formation des communautés scientifiques du sud de l'IRD qui m'a accordé une allocation de recherche de 36 mois ; puis d'une allocation d'entretien de 4 mois de la part du Ministère des Affaires Étrangères de la France. J'ai aussi été soutenu financièrement par l'Île de France qui a bien voulu financer ma participation à des séminaires et conférences. Soyez en très sincèrement remerciés.

Je tiens à remercier l'ensemble de l'équipe Probabilités et Statistiques de l'Université Paris-Sud 11. Merci à Christine Keribin pour m'avoir initié au concept de programmation orienté objet sous le langage $C++$. Cette initiation a été d'une importance certaine dans le développement du logiciel `MixMoGenD`. Merci à Cécile Duro pour son cours bien construit de "Statistiques Asymptotiques" de M2.

Je remercie toute l'équipe du Laboratoire de lutte contre le paludisme de l'OCEAC de Yaoundé où est basée une partie de l'UR016 de l'IRD : Parfait Awono, Antonio Nkondjo, Josiane Etan et les autres doctorants Basile, Colince, Sandrine et Billy, Diana, sans oublier Sylvie et Elysé.

Merci aux membres du Groupe de Travail de l'ENSP de Yaoundé, en particulier Eugène-Patrice Ndong Nguéma et les autres doctorants Stafav.

Merci aux doctorants ou ex-doctorants : Dominique, Robin, Cyprien, Sébastien, Hatem, Richard, Nathalie, Cathy, Nicolas et tous les autres pour leur présence et leur amitié. Merci surtout à Dominique et Robin avec qui j'ai eu la chance de collaborer sur une partie de ma thèse, à Cathy pour nos discussions sur les algorithmes de sélection de modèle.

Grâce à la gentillesse et la patience des personnels administratifs des différents services de l'Université d'Orsay, J'ai pu rendre les bons documents au bon moment. Je pense notamment à Valérie Lavigne à qui je dis merci.

Enfin, merci à toute ma famille : mes parents Jacques Toussile et Jeanne Gaffo, mes filles, mes frères et soeurs, mes grands-parents pour leur soutien et leurs encouragements.

Merci particulier à Aline Yonta pour tellement.

Résumé

Dans cette thèse, nous considérons la question de sélection de variable dans deux problèmes pratiques. Le premier concerne la structure génétique d'une population en plusieurs populations, et le deuxième la transmission de *Plasmodium* à travers son vecteur moustique.

Le premier problème est motivé par la recherche de populations génétiquement homogènes sur la base de données génétiques multilocus, sans information a priori. Nous supposons que seul un sous-ensemble de variables est pertinent pour une classification optimale. Nous considérons ainsi le problème double de sélection de variable et de classification non supervisée. Ce problème est vu comme celui de sélection de modèle pour l'estimation de la densité des observations. La collection $\mathcal{C}$ des modèles candidats est défini de sorte que chaque modèle correspond à une classification particulière avec son sous-ensemble de variables associé. Nous sélectionnons le modèle optimal via des critères du maximum de vraisemblance pénalisé tels que BIC (Bayesian Information Criterion) et AIC (Akaike Information Criterion). Pour éviter une recherche exhaustive du modèle optimal qui peut être très coûteuse en temps de calcul, la procédure de sélection repose sur les stratégies Backward-Stepwise et Forward-Stepwise modifiées. Toutefois, une comparaison empirique montre qu'aucun des critères asymptotiques BIC et AIC n'est uniformément meilleur que l'autre par rapport à la taille de l'échantillon. Nous proposons alors de considérer les fonctions de pénalité de la forme $m \in \mathcal{C} \mapsto \mathbf{pen}(m) = \lambda \cdot d_m$, où d_m est la dimension du modèle m et λ une constante dépendant de la taille de l'échantillon et de la collection $\mathcal{C}$ de modèles. La constante λ est calibrée sur les données dans une procédure automatique basée sur l'"heuristique de la pente". Nous montrons empiriquement que le critère associé permet de répondre en partie à la question "quel critère pour quelle taille d'échantillon". Les méthodes proposées sont implémentées dans un logiciel nommé `MixMoGenD` (Mixture Model for Genotypic Data) sous le langage $C++$. Ce logiciel est disponible sur `www.math.u-psud.fr/~toussile`.

Sur le plan théorique, sous des hypothèses faibles sur la fonction de pénalité, nous montrons que la procédure de sélection est consistante à la fois pour la sélection de variable et l'estimation du nombre de populations. D'autre part, dans une approche non asymptotique, nous proposons un critère pénalisé et une inégalité oracle associée. Ce dernier résultat justifie en quelque sorte le choix de la forme de la fonction de pénalité $m \in \mathcal{C} \mapsto \mathbf{pen}(m) = \lambda \cdot d_m$ pour adapter la procédure de sélection aux données, en particulier à leur taille.

Le deuxième problème concerne les données d'une étude dont l'objectif est de mettre en évidence des facteurs qui influencent la transmission de *Plasmodium* à travers son vecteur moustique. De telles données sont décrites par des variables diverses dont leur nombre plus les interactions potentielles est au moins de l'ordre de la taille de l'échantillon. Nous nous servons de l'importance des variables obtenue des forêts aléatoires pour sélectionner les covariables les plus influentes. La procédure de sélection résultante est non paramétrique et répond aux deux principaux objectifs de la sélection de variables, à savoir : (1) sélectionner toutes les covariables reliées à la variable réponse et (2) sélectionner le plus petit sous-ensemble de covariables pour une bonne prédiction de la variable réponse. Les variables sélectionnées sont ensuite évaluées dans le modèle binomial négatif modifié en zéro. Ce modèle permet de prendre en compte à la fois la surdispersion (très fréquente en parasitologie) et les deux sources possibles des moustiques non infectés à l'issue de l'expérience.

Mots-clefs : Sélection de variable, Modèles de mélange, Maximum de vraisemblance pénalisé, Génétique des populations, Paludisme, Forêts aléatoires.

Variable selection: population structure and transmission of *Plasmodium* through its vector mosquito

Abstract

In this thesis, we consider the question of variable selection in two practical problems. The first problem concerns genetic structure in population genetics context, and the second one is related to *Plasmodium* transmission through its vector mosquito.

The first problem is motivated by a long standing issue in population genetics consisting of grouping the individuals of a sample into genetically homogeneous clusters on the basis of their genotypes at a certain number of loci. We assume a situation without prior information on the population the sample come from. In addition, it may happen that some loci are just noise for clustering purposes. We then have to face the two-fold problem of variable selection and classification. We consider a model selection procedure to simultaneously solve the variable selection and classification problems. A specific collection $\mathcal{C}$ of competing models is defined in such a way that each one corresponds to a particular classification with its associated relevant subset of variables. The competing models are compared using penalized maximum likelihood criteria such as Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC). To avoid an exhaustive search of the optimum model which is painful in most situations in practice, the selection procedure is based on modified backward-stepwise and forward-stepwise strategies. Empirical comparison of the above asymptotic criteria shows that none of them is uniformly better than the other with respect to the sample size. We then consider a family of penalized criteria associated to penalty functions of the shape $m \in \mathcal{C} \mapsto \mathbf{pen}(m) = \lambda \cdot d_m$, where d_m is the dimension of a model $m \in \mathcal{C}$, and λ a multiplicative term depending on the collection $\mathcal{C}$ of the models under competition and the data (typically via the sample size). The multiplicative term λ is calibrated in a data-driven automatic procedure based on the so called "slope heuristics". We found on simulated experiments that penalty calibration globally improves the selection procedure and gives an answer to the question "which criterion for which sample size". The proposed methods are implemented in a stand alone $C++$ package named `MixMoGenD` (Mixture Model for Genotypic Data)

On the theoretical aspect, under weak assumptions on the penalty function, we show that the selection procedure is consistent for both variable selection and the estimation of the number of populations. In the other hand, in a non-asymptotic approach, we propose a penalized criterion with an associated non-asymptotic oracle inequality. This result somehow justifies the shape $m \in \mathcal{C} \mapsto \mathbf{pen}(m) = \lambda \cdot d_m$ of penalty function we considered in practice.

The second problem is motivated by malaria control strategies aiming at reducing disease transmission intensity. We consider data from a study to investigate how density and genetic diversity of gametocytes impact on the success of transmission in the mosquito vector. In such data, the number of covariates plus attendant interactions is at least of order of the sample size, precluding usage of classical models such as General Linear Model. We then considered the variable importance from random forests to address the problem of selecting the most influent covariates. The selected covariates are assessed in the Zero Inflated Negative Binomial (ZINB) model which accommodates both over-dispersion and the possible sources of non infected mosquitoes. We find that the most important covariates related to infection prevalence and parasite intensity are gametocyte density and multiplicity of infection (MOI), respectively.

Keywords : Variable selection, Mixture models, Penalized maximum likelihood, Population genetics, Malaria, Random Forest.

Table des matières

Chapitre 1

Présentation générale

Dans cette thèse, nous nous intéressons à deux questions pratiques toutes liées à la sélection de variable. La première concerne la structure génétique de populations d'organismes vivants, et la deuxième la transmission du parasite responsable du paludisme à travers son vecteur moustique. Ce chapitre introductif présente les motivations, la solution proposée pour chacune des questions ainsi que les hypothèses sur lesquelles les méthodes utilisées reposent.

1.1 Structure génétique d'une population

Il s'agit d'une question récurrente en génétique des populations, qui consiste à regrouper les individus d'un échantillon dans des groupes génétiquement homogènes en un certain sens, sur la base de leurs génotypes à un certain nombre de marqueurs génétiques. Dans la plupart des cas, les biologistes disposent d'informations a priori sur la population cible. Ces informations a priori peuvent être par exemple la localisation géographique, les caractères linguistiques, physiques ou culturels. Il arrive que de telles informations induisent une classification erronée au sens génétique par exemple lorsque la population cible est issue d'un mélange récent de plusieurs populations ancestrales avec des jeux différents de fréquences alléliques, ou lorsqu'il y a des préférences dans le choix de partenaires de reproduction. Par exemple, il peut être intéressant de rechercher une classification au delà de la localisation géographique pour savoir s'il y a échange de gènes entre des populations. C'est par exemple le cas dans Rosenberg et al. (2001) où les auteurs montrent que la population juive libyenne se distingue génétiquement des juifs irakiens, marocains, éthiopiens et yéménites, entre lesquels il y aurait des échanges de gènes.

Dans cette thèse, nous supposons ne disposer d'aucune information sur la population cible. Nous avons alors à faire à un problème de classification non-supervisée.

Par ailleurs, les données auxquelles nous nous intéressons peuvent être décrites par un L-vecteur aléatoire $\mathbf{X} = \left(X^l\right)_{1 \leq l \leq L}$. Chaque composante X^l appelée locus ou gène, décrit les variants de gènes (appelés allèles) à une position donnée du génome. Par ex-

emple, pour les organismes diploïdes, on observe deux copies de gènes (pouvant être identiques) à chaque position. Dans ce cas, chaque composante X^l est une paire non ordonnée $\{X^{l,1}, X^{l,2}\}$ de variables décrivant les deux copies de gènes. Les chercheurs s'intéressent de plus en plus à un nombre L de loci de plus en plus grand. En principe, plus on dispose d'informations sur chaque individu, meilleures devraient être les performances des méthodes de classification. Mais il est possible que seul un sous-ensemble S de gènes discrimine les populations entre elles, les autres n'ajoutant que du bruit à la classification. En effet, l'exemple des données simulées dans l'Appendice 1.B montre que des gènes pour lesquels les allèles sont identiquement distribués dans toutes les populations peuvent détériorer la capacité de prédiction d'une méthode de classification.

Il est alors envisageable de maîtriser les informations nécessaires pour une classification optimale des observations par une procédure de sélection du sous-ensemble optimal parmi l'ensemble des gènes (variables) disponibles. L'objectif d'une telle procédure est d'améliorer les performances de la procédure de classification et faciliter l'interprétation des résultats obtenus.

Un certain nombre de méthodes de classification non-supervisée sur les données génotypiques multilocus sont proposées dans la littérature. Citons entre autres les procédures `Structure` de Pritchard et al. (2000), `Geneland` de Guillot et al. (2005), `Fastruct` de Chen et al. (2006), `BAPS` (Bayesian Application for Population Structure) de Corander et al. (2008) et `Admixture` de Alexander et al. (2009). A l'exception de `Fastruct`, les procédures ci-dessus reposent sur une approche bayésienne. Elles peuvent de ce fait être très gourmandes en ressources informatiques et très coûteuses en temps de calcul. De plus, aucune de ces méthodes n'intègre une procédure de sélection de variable.

Dans cette thèse, nous considérons simultanément le problème de recherche de la classification optimale et celui de sélection du sous-ensemble de variables produisant cette classification là.

1.1.1 Classification non supervisée

Les méthodes de classification non supervisée peuvent être rangées en deux catégories. La première comprend les méthodes basées sur des mesures de similarité ou de dissimilarité. Cette catégorie est connue sous le nom de "distance-based clustering" en anglais. Comme exemples, citons les méthodes de classification hiérarchique et les algorithmes "K-means" (voir Hastie et al. (2005)). Le principal inconvénient de ces méthodes est l'interprétation de la classification qu'elles produisent. En effet, les résultats que ces méthodes produisent dépendent du choix de la mesure de dissimilarité (ou de similarité). Par ailleurs, les procédures hiérarchiques nécessitent la définition d'une mesure de dissimilarité entre ensembles d'individus. La deuxième catégorie de méthodes de classification non supervisée part de l'idée que les observations sont issues d'une population constituée d'un mélange fini d'un certain nombre (inconnu) K de sous-populations. Les observations de chaque sous-population sont supposées provenir d'une loi de probabilité et l'ensemble des observations est alors décrit par le mélange des K lois de probabilité des différentes

sous-populations. Le principal avantage des modèles de mélange fini est de constitué un cadre rigoureux pour évaluer le nombre de composants de la population et le rôle des variables dans le processus de classification (Keribin, 2000; Maugis et al., 2009; Baudry, 2009). Dans un cadre paramétrique, chaque sous-population est caractérisée par un jeu de paramètres.

1.1.2 Sélection de variable

Comparée aux méthodes de classification supervisée et aux modèles de régression, la sélection de variable est un sujet assez récent en classification non supervisée. Les méthodes proposées reposent en général sur deux approches : "filter" et "wrapper" selon la terminologie introduite par Kohavi and John (1997) dans le cadre de la classification supervisée. Les méthodes "filter" traitent le problème de sélection de variable indépendamment du processus de classification. Cette indépendance constitue son principal inconvénient. Concernant ces méthodes en classification non supervisée, citons les travaux de Jouve and Nicoloyannis (2005) et ceux de Dash et al. (2002). À l'opposé, les méthodes "wrapper" sont des procédures de sélection de variable combinées au processus de classification. Ces méthodes ont l'avantage de permettre une meilleure interprétation des variables sélectionnées. Les premières méthodes "wrapper" ont été proposées dans une classification hiérarchique (Fowlkes et al., 1988) et dans un algorithme des K plus proches voisins (Brusco and Cradit, 2001). Combiner une procédure de sélection de variable à une méthode de classification non supervisée est une pratique récente ; citons en particulier les travaux de Raftery and Dean (2006) et de Maugis et al. (2009) dans un cadre Gaussien.

1.1.3 Sélection de variable en classification non supervisée par mélange fini

La première question à laquelle nous nous intéressons dans cette thèse est double : nous cherchons une classification optimale des observations en un nombre inconnu K de classes, en même temps que le sous-ensemble optimal S de variables qui produit cette classification. Nous nous intéressons particulièrement aux données génétiques multilocus, décrites par des variables qualitatives nominales. Pour répondre à cette question double pour de telles données, nous considérons les modèles de mélange fini dans un cadre multinomial et adoptons une approche de sélection de variable "wrapper".

Nous supposons que chaque classe recherchée correspond à un composant du mélange caractérisé par un jeu de paramètres. Pour un nombre K de composants du mélange et un sous-ensemble S de variables sélectionnées, nous définissons un modèle $\mathcal{M}_{K,S}$ de lois de probabilité dans lequel chaque loi $P_{K,S,\theta_{K,S}}(\cdot)$ donnée par

$$P_{K,S,\theta_{K,S}}(x) = \sum_{k=1}^{K} \pi_k P_{\alpha_{k,S}}(x),$$

est un mélange fini à K composants. Le vecteur des paramètres $\theta_{K,S} := (\pi, \alpha)$

est constitué du vecteur $\pi = (\pi_k)_{1\leq k\leq K}$ des proportions du mélange et du vecteur $\alpha := (\alpha_{k,S})_{1\leq k\leq K}$ des paramètres des lois de probabilité des observations de chaque composant. L'hypothèse d'indépendance conditionnelle complète permet ensuite de factoriser chaque terme $P_{\alpha_{k,S}}(\cdot)$. Le bien fondé de cette hypothèse et de celle relative à l'Équilibre de Hardy-Weinberg (EHW) est discuté dans le paragraphe suivant. La classification $\widehat{\mathbf{z}} = MAP\left(\widehat{\theta}_{K,S}\right)$ des observations est obtenue par la règle du Maximum A Posteriori (MAP) à partir de l'estimateur du maximum de vraisemblance $\widehat{\theta}_{K,S}$ du vecteur des paramètres. Chaque modèle $\mathcal{M}_{K,S}$ correspond ainsi à une classification des observations en K composants sur la base du sous-ensemble de variable S. Notons $\mathcal{P}^*(L)$ l'ensemble des sous-ensembles non vides de l'ensemble $\{1, \ldots, L\}$ des L gènes considérés. Le choix d'un modèle parmi la collection

$$\mathcal{C} = \left\{\mathcal{M}_{K,S} \ : \ (K,\ S) \in \{1\} \times \{\emptyset\} \bigcup [\mathbb{N} \setminus \{0,\ 1\}] \times \mathcal{P}^*(L)\right\}$$

des modèles candidats équivaut alors à celui de la "meilleure" classification en K classes et du sous-ensemble S de variables associé.

1.1.4 Sélection de modèle

Notre problème simultané de sélection de variable et de classification est devenu celui de sélection de modèle pour l'estimation de la densité des observations. Dans cette thèse, nous adoptons une vieille recette qui consiste à sélection le modèle minimisant un critère pénalisé. Nous considérons pour cela les critères du maximum de vraisemblance pénalisé.

Dans la pratique, nous nous sommes tout d'abord intéressés aux critères asymptotiques BIC (Bayesian Information Criterion) (Schwarz, 1978) et AIC (Akaike Information Criterion) (Akaike, 1973). Nous avons constaté sur des simulations qu'aucun de ces deux critères n'est uniformément meilleur que l'autre par rapport à la taille de l'échantillon. Il se pose alors la question "Quel critère pour quelle taille de l'échantillon ?". La solution pratique que nous proposons est de considérer une fonction de pénalité de la forme $\mathbf{pen}(K,\ S) = \lambda d_{K,S}$, où $d_{K,S}$ est le nombre de paramètres indépendants du modèle $\mathcal{M}_{K,S}$ et $\lambda = \lambda(n,\ \mathcal{C})$, un terme multiplicatif dépendant des données et de la collection $\mathcal{C}$ des modèles en compétition. Le terme multiplicatif λ est ensuite calibré sur les données grâce "l'heuristique de la pente" proposée par Birgé and Massart (2007). Nous considérons la version "détection du plus grand saut de dimension" de cette méthode, à laquelle nous ajoutons une fenêtre glissante (Chapitre 4). La grille des valeurs candidates de λ est choisie de sorte que la famille de critères associée contienne les critères AIC et BIC. Les résultats empiriques obtenus sur les données simulées montrent que cette calibration de la fonction de pénalité améliore globalement la procédure de sélection de modèle.

Bien qu'il existe de nombreux travaux sur le comportement pratique des critères AIC et BIC, les résultats théoriques sur ces critères ne sont pas nombreux. Sous des hypothèses faibles sur la fonction de pénalité, nous montrons que la procédure de sélection de modèle basée sur le critère du maximum de vraisemblance est consistante. Ensuite, partant d'un théorème général de sélection de modèle pour l'estimation de densité dû à Massart (2007),

nous obtenons une borne inférieure de la fonction de pénalité et une inégalité de type oracle associée. Bien que cette borne inférieure de la pénalité ne soit pas directement utilisable du fait qu'elle dépend d'une constante multiplicative inconnue, elle justifie en quelque sorte la forme $\mathbf{pen}(K,\ S) = \lambda d_{K,S}$ de la pénalité adoptée plus haut.

1.1.5 Équilibre de Hardy-Weinberg et Équilibre de liaison

Les modèles considérés dans cette thèse sont basés sur les hypothèses conditionnelles de Hardy-Weinberg et d'équilibre de liaison. Le modèle de Hardy-Weinberg constitue un ensemble d'hypothèses pour développer la prédiction des fréquences génotypiques à partir des fréquences alléliques. Il se présente sous la forme de 8 lois :

1. Les organismes considérés sont diploïdes.
2. La reproduction est sexuée.
3. Les générations sont non chevauchantes.
4. Les croisements se font au hasard (panmixie).
5. La taille de la population est très grande.
6. Les migrations sont négligeables.
7. On peut ignorer les mutations.
8. La sélection naturelle n'a pas d'effet sur les allèles considérés.

Le modèle de Hardy-Weinberg concerne les organismes diploïdes. Si on considère L loci (ou gènes), le génotype d'un organisme i est décrit par un vecteur $\mathbf{x}_i = \left(x_i^l\right)_{1\leq l\leq L}$, où chaque gène x_i^l est une paire non ordonnée $\left\{x_i^{l,1},\ x_i^{l,2}\right\}$ décrivant les variants génétiques (ou allèles) observés à une position l du génome. Les deux allèles $x_i^{l,1}$ et $x_i^{l,2}$ peuvent être identiques : l'organisme en question est alors dit homozygote à ce locus là. Dans le cas contraire, il est dit hétérozygote. Sous le modèle de Hardy-Weinberg, l'association des allèles $x_i^{l,1}$ et $x_i^{l,2}$ est aléatoire, et les probabilités des allèles ne changent pas d'une génération à la suivante. La probabilité du génotype x_i^l est alors donnée par

$$P_{\alpha_l}\left(x_i^l\right) = \left(2 - \mathbb{1}_{[x_i^{l,1}=x_i^{l,2}]}\right)\alpha_{l,x_i^{l,1}}\alpha_{l,x_i^{l,2}},$$

où $\alpha_{l,j}$ est la probabilité de l'allèle j au locus l, et $\alpha_{l,\cdot} = (\alpha_{l,j})_{1\leq j\leq A_l}$ le vecteur de toutes les probabilités alléliques de ce locus là, en supposant qu'on ait A_l allèles distincts.

On parle d'équilibre de liaison (EL) quand les allèles de gènes différents s'associent au hasard. Une telle hypothèse permet de factoriser la probabilité de $\mathbf{x}_i$

$$P_\alpha(\mathbf{x}_i) = \prod_{l=1}^{L} P_{\alpha_{l,\cdot}}\left(x_i^l\right).$$

Bien que les modèles de Hardy-Weinberg (HW) et d'équilibre de liaison (EL) semblent simplistes, ils servent de modèles de base pour l'élaboration d'autres modèles plus

complexes de micro-évolution en génétique des populations. Ils constituent un équilibre génétique difficilement observable et ce sont les écarts à cet équilibre qui sont porteurs d'informations. En effet, comme le montrent les Exemples 1.1.1 et 1.1.2 ci-dessous, la structuration d'une population en sous-populations amplifie l'écart à l'équilibre génétique. Par ailleurs, on sait que la dépendance entre deux gènes est inversement proportionnelle à la distance qui les sépare sur le génome. Elle est en principe très faible pour des gènes situés sur des chromosomes différents. Il est donc raisonnable de penser que lorsque les positions considérées sur le génome sont suffisamment éloignées les unes des autres (par exemple sur des chromosomes différents), l'écart à l'équilibre génétique de la population globale est expliqué par la structure de la population en plusieurs populations. La classification optimale d'un échantillon qui a un sens génétique en terme d'unité de reproduction est alors celle qui regroupe les observations de sorte à minimiser le déséquilibre génétique intra-population.

Dans la suite de cette thèse, nous utilisons aussi l'expression "équilibre de Hardy-Weinberg" pour désigner l'indépendance de l'association des allèles d'un gène pour les organismes polyploïdes.

Exemple 1.1.1. *Considérons une population constituée de deux sous-populations en proportions égales et en équilibre de Hardy-Weinberg chacune. Considérons un locus biallélique* $(0|1)$ *et notons* $\alpha_{k,j}$ *la probabilité de l'allèle* j $(\in \{0,\ 1\})$ *dans la sous-population* k $(\in \{1,\ 2\})$. *Dans la population totale, les probabilités alléliques sont données par* $\alpha_{\cdot,j} = \frac{\alpha_{1,j}+\alpha_{2,j}}{2}$ *pour l'allèle* j. *L'hétérozygotie est alors* $H = 2\sum_{k=1}^{2}\prod_{j=0}^{1}\alpha_{k,j}$, *qui est en général différent de* $2\alpha_{\cdot,0}\alpha_{\cdot,1}$, *la valeur attendue sous HW. Cet exemple montre que la structure en sous-populations peut créer le déséquilibre de Hardy-Weinberg.*

Exemple 1.1.2. *Considérons* 2 *loci bi-alléliques* L_1 *et* L_2 *et soient* A *un allèle du locus* L_1 *et* B *un allèle du locus* L_2. *Le déséquilibre de liaison peut être mesuré par*

$$D = \alpha_{AB} - \alpha_{A,\cdot}\alpha_{\cdot,B},$$

où $\alpha_{A,\cdot}$, $\alpha_{\cdot,B}$ *et* α_{AB} *sont les probabilités des allèles* A *et* B, *et de l'haplotype* AB *respectivement.*

Considérons deux populations et pour chaque population k $(\in \{1,\ 2\})$, *notons* $\alpha_{A,\cdot}^{(k)}$, $\alpha_{\cdot,B}^{(k)}$ *et* $\alpha_{AB}^{(k)}$ *les probabilités des allèles* A *et* B, *et de l'haplotype* AB. *Supposons que*

- *dans la population* 1, $\alpha_{A,\cdot}^{(1)} = 1$ *et* $\alpha_{\cdot,B}^{(1)} = 0$;
- *dans la population* 2, $\alpha_{A,\cdot}^{(2)} = 0$ *et* $\alpha_{\cdot,B}^{(2)} = 1$.

Alors la mesure D *du déséquilibre de liaison est nulle dans chaque population. Par ailleurs, les probabilités des allèles* A *et* B *dans la population globale sont données par* $\alpha_{A,\cdot} = \alpha_{\cdot,B} = \frac{1}{2}$, *et celle de l'haplotype* AB *par* $\alpha_{AB} = 0$. *Dans la population globale, la mesure* D *du déséquilibre de liaison est alors non nulle* $(|D| = \frac{1}{4})$. *Ce qui implique que les deux loci considérés ne sont pas en équilibre de liaison dans la population globale.*

1.2 Transmission de *Plasmodium*

Le paludisme, aussi appelé malaria, est une maladie infectieuse due à un parasite du genre *Plasmodium*, transmis d'un vertébré à un autre par la piqûre de certaines espèces de moustiques *anophèles*. Cette maladie a été éradiquée de certaines régions du monde, mais elle continue de peser de façon importante sur la morbidité et la mortalité en régions tropicales, et concerne majoritairement les enfants de moins de 5 ans et les femmes enceintes. L'Afrique subsaharienne est la région la plus touchée avec près de 80 % de cas enregistrés parmi lesquels près de 2 Millions de morts chaque année. Plusieurs stratégies sont mises en œuvre pour tenter de limiter la transmission de *Plasmodium*, notamment le contrôle des populations de moustiques à l'aide d'insecticides. Si cette solution a été efficace en régions tempérées, ses résultats restent mitigés en régions tropicales, et le problème risque de s'aggraver avec l'émergence de résistances des moustiques aux insecticides (Miller and Greenwood, 2002) et du parasite aux antipaludiques ; sans oublier le réchauffement climatique qui favorise l'extension de l'aire de répartition des espèces *anophéliennes* vectrices. D'autres initiatives de lutte sont concentrées sur la prévention par la vaccination, mais aucun vaccin efficace n'a encore été mis au point à ce jour. Le contrôle génétique du vecteur moustique s'inscrit aussi parmi les moyens de lutte envisageables, l'objectif étant d'interrompre le cycle de transmission chez le vecteur en le rendant réfractaire au parasite. Cette solution nécessite de comprendre les mécanismes génétiques associés à la compétence vectorielle. Par exemple, des travaux sur l'expression de gènes de l'immunité suite à l'infection par *Plasmodium* ou par des bactéries ont montré que le moustique est capable de développer une réponse immunitaire contre certains agents pathogènes (Dimopoulos et al., 1997; Osta et al., 2004; Riehle et al., 2006). On peut imaginer que le succès du développement du parasite chez le moustique dépend aussi de facteurs intrinsèques au parasite, illustrant le concept d'"Extended phenotype" introduit par Dawkins (1999) : le phénotype du vecteur (resistance, capicité vectorielle, ...) et celui du parasite (virulence du parasite, ...) ne résultent pas seulement de leur propre génotype, mais aussi de celui de l'autre.

Dans le cas spécifique de *Plasmodium*, bien qu'il ait été démontré à plusieurs reprises qu'une forte densité parasitaire dans le repas de sang ne garantit pas le succès du développement du parasite chez le moustique, les résultats obtenus par Paul et al. (2007) suggèrent qu'il existe un seuil de la densité parasitaire dans le repas de sang au delà duquel la prévalence chez les moustiques est croissante, et aussi qu'il existe un seuil en deçà duquel la prévalence est presque nulle. Les données utilisées par Paul et al. sont constituées d'une part de la densité parasitaire dans le repas de sang du moustique, et d'autre part du nombre de parasites sous la forme oocystes (voir Sous-section 1.2.1) quelques jours après le repas de sang. Sur le plan méthodologique, la modélisation de la charge parasitaire des moustiques repose sur le modèle binomial négatif qui est le modèle type permettant de prendre en compte le phénomène de sur-dispersion dans de telles données en parasitologie (Pichon et al., 2000, et les références qui y sont).

En plus de la densité parasitaire dans le repas de sang des moustiques et de la charge parasitaire des moustiques, nous disposons de la diversité génétique du parasite présent dans le repas de sang des moustiques. Cette diversité génétique est décrite par les géno-

types des parasites pour 7 marqueurs génétiques. Les marqueurs génétiques considérés sont très polymorphes avec une moyenne de 12 allèles par marqueur. L'ensemble des covariables comprend alors la densité parasitaire dans le repas de sang des moustiques, les indicatrices des allèles observés et leurs éventuelles interactions. Nous sommes dans une situation où les covariables sont diverses (qualitatives et quantitatives) et leur nombre est au moins de l'ordre de la taille de l'échantillon. Dans une telle situation, les modèles classiques de régression tels que le GLM (Generalized Linear Model) aboutissent à un sur-ajustement des données. Il se pose alors un problème de sélection de variable. S'inspirant des recommandations de Segal et al. (2001) qui préconisent l'utilisation des arbres de décision dans de telles situations, nous proposons de réaliser la sélection des covariables les plus importantes pour la question à l'aide des indices d'importance des variables obtenus des forêts aléatoires. La procédure de sélection de variable résultante est non paramétrique et convient aux covariables diverses (qualitatives ou quantitatives). Elle répond à deux objectifs principaux : (1) sélectionner toutes les covariables reliées à la variable réponse et (2) sélectionner le plus petit sous-ensemble de covariables permettant de faire une bonne prédiction de la variable réponse. L'évaluation de l'effet des variables sélectionnées est ensuite faite dans le modèle binomial négatif modifié en zéro (Zero Inflated Negative Binomial en anglais). Nous aboutissons à ce modèle grâce à une modélisation de la charge parasitaire des moustiques qui tient compte des deux sources possibles de moustiques non infectés.

1.2.1 Développement sporogonique

Plasmodium a un cycle biologique complexe qui se déroule successivement chez un hôte vertébré et chez le moustique. La phase de ce cycle biologique se déroulant chez le moustique est nommée développement sporogonique.

Le développement sporogonique de *Plasmodium* commence chez le moustique après que ce dernier ait ingéré des gamétocytes, qui représentent les formes sexuées du parasite présentes chez l'hôte vertébré, les autres formes étant les trophozoïtes et les schizontes. Ces autres formes sont digérées par le moustique, seuls les gamétocytes poursuivent leur développement. Ils se divisent en gamètes mâles et femelles. Il s'ensuit une fécondation entre les deux genres de gamètes aboutissant à la formation de zygotes qui évoluent vers une forme mobile appelée ookinètes. Les ookinètes traversent la paroi de l'estomac et se transforment en oocystes. Après une phase de maturation, chaque oocyste libère dans l'hémocoele du moustique plusieurs milliers de sporozoïtes qui vont alors migrer vers les glandes salivaires. Le moustique alors infectant, transmettra des sporozoïtes à un vertébré lors du prochain repas de sang, pérennisant ainsi le cycle parasitaire. Les stades zygotes, ookinètes et oocystes sont des formes diploïdes, tandis que les stades gamétocytes et sporozoïtes sont haploïdes.

La transmission du paludisme repose ainsi sur le succès du développement sporogonique au cours duquel le parasite accomplit une reproduction sexuée, qui est connue pour accroître les capacités adaptatives des êtres vivants. Au cours de ce cycle sporogonique, de fortes réductions parasitaires s'opèrent lors de la fertilisation des gamètes, pendant la

la phase de la différentiation des zygotes en ookinètes et lors du passage des ookinètes au travers de l'épithélium intestinal (Sinden et al., 2007; Vaughan, 2007).

1.3 Plan de la thèse

La suite de cette thèse est divisée en 6 chapitres qui peuvent être lus de façon indépendante. Pour faciliter la lecture du manuscrit, les preuves et les calculs techniques sont reportés dès que possible en appendices du chapitre correspondant. Les chapitres 3-6 sont en anglais, ce sont des articles déjà publiés ou soumis.

Chapitre 2 : Classification non supervisée par mélange fini de lois multinomiales : Application aux données génétiques multilocus

Ce chapitre est consacré à une présentation générale des modèles de mélange fini et leur application à la classification non supervisée. L'accent est mis sur les données génétiques multilocus. Au nombre de composants du mélange fixé, nous décrivons l'algorithme EM pour approcher l'estimateur du maximum de vraisemblance. La classification des observations est ensuite obtenue par la règle du Maximum A Posteriori (MAP). Nous décrivons ensuite la sélection du nombre de composants du mélange via des critères du maximum de vraisemblance pénalisés tels que BIC et ICL, et via une méthode inspirée de la statistique *Gap* de Hastie et al. (2001).

Chapitre 3 : Variable selection in model-based clustering using multilocus genotypic data

Ce chapitre est un travail réalisé en collaboration avec Elisabeth Gassiat (Université Paris-Sud 11). Il s'agit d'une version légèrement modifiée de l'article portant le même titre paru en 2009 dans le volume 3 de la revue "Advances in Data Analysis and Classification". Ce chapitre traite de la sélection de variable dans la classification non supervisée par mélange fini sur données génétiques multilocus. Le problème simultané de sélection de variable et de classification est vu comme un problème de sélection de modèle pour l'estimation de la densité des observations. Sous des hypothèses faibles sur la fonction de pénalité, nous montrons que les critères du maximum de vraisemblance pénalisé sélectionnent le vrai modèle avec une probabilité qui tend vers 1, le vrai modèle étant défini de manière identifiable. Cette consistance est vérifiée empiriquement sur des données simulées. Nous appliquons ensuite cette procédure sur un jeu de données réelles.

Chapitre 4 : A data-driven penalized criterion for variable selection and clustering in multivariate multinomial mixtures

Ce chapitre est le fruit d'un travail en collaboration avec Dominique Bontemps (Université Paris-Sud 11). Il concerne le même problème qu'au Chapitre 3. Il est motivé

par la recherche d'un critère de sélection qui s'adapte à la taille de l'échantillon. En effet, une comparaison empirique des critères BIC et AIC montre qu'aucun de ces deux critères asymptotiques n'est uniformément meilleure que l'autre par rapport à la taille de l'échantillon. Nous adoptons une approche non asymptotique et proposons une fonction de pénalité et une inégalité oracle associée. Ce résultat est une application d'un théorème général de sélection de modèle pour l'estimation de densité par maximum de vraisemblance pénalisé, théorème dû à Massart (2007). L'application de ce théorème dans notre cas sépécifique nécessite le contrôle des entropies à crochets des modèles de mélange fini de lois multinomiales dont les paramètres sont donnés par le modèle de Hardy-Weinberg. Nous nous servons pour cela des résultats de Genoveve and Wasserman (2000) donnant les entropies à crochets des simplexes et des modèles de mélange fini.

Dans la pratique, nous avons adopté une fonction de pénalité proportionnelle à la dimension des modèles. Le terme multiplicatif est calibré automatiquement sur les données par la version "détection du plus grand saut de dimension" de la méthode "heuristique de la pente" proposée par Birgé and Massart (2007). Nous nous servons d'une fenêtre glissante pour améliorer la sélection du critère optimal. Nous montrons sur des simulations qu'une telle calibration de la fonction de pénalité répond en partie à la question "Quel critère pour quelle taille d'échantillon ?"

Chapitre 5 : MixMoGenD, a software for both loci selection and clustering on genotypic data

Nous proposons dans ce chapitre un logiciel autonome nommé `MixMoGenD` (Mixture Model for Genotypic Data), qui implémente les méthodes décrites dans les chapitres 3 et 4. Ce logiciel est implémenté dans une approche orienté objet, sous le langage $C++$. La gestion de la mémoire est entièrement dynamique. Les résultats des analyses sont consignés dans des fichiers. Pour éviter une recherche exhaustive du modèle optimal qui serait très coûteuse en temps de calcul, la procédure de sélection de `MixMoGenD` repose sur des versions modifiées des algorithmes "Backward-Stepwise" et "Forward-Stepwise". Les modifications que nous avons apportées à ces algorithmes permettent d'explorer tous les cardinaux possibles du sous-ensemble optimal de variables.

Chapitre 6 : Gametocytes infectiousness to mosquitoes : variable selection using random forests, and zero inflated models

Ce chapitre est un travail en collaboration avec Robin Genuer (Université Paris-Sud 11) et Isabelle Morlais (CR1 à l'Institut de Recherche pour le Développement). Notre contribution est d'aider à la compréhension de la transmission de *Plasmodium* à travers son vecteur moustique. Le nombre de covariables dans le jeu de données à analyser est de l'ordre de la taille de l'échantillon. Nous sélectionnons tout d'abord les covariables les plus influentes vis-à-vis de la variable réponse, puis nous évaluons l'effet des covariables sélectionnées via le modèle binomial négatif modifié en zéro. La sélection de variable est basée sur l'importance des variables obtenue par les forêts aléatoires. La procédure résultante

est non paramétrique et s'applique aux variables diverses (qualitatives, quantitatives). Elle répond aux principaux objectifs en matière de sélection de variable, à savoir : (1) sélectionner toutes les covariables reliées à la variable réponse, (2) sélectionner le plus petit sous-ensemble de covariables suffisant pour une bonne prédiction de la réponse. Par ailleurs, vu comme mélange de la loi empirique en zéro et d'un modèle de comptage, le modèle modifié en zéro permet de prendre en compte les deux sources possibles de moustiques non infectés dans l'expérience considérée.

Appendices

1.A Accord entre deux partitions : Indice de Rand indice

Lorsqu'on dispose de deux partitions d'un même ensemble d'objets, il se pose la question de savoir si elles sont en accord ou pas en un certain sens. Une manière d'aborder cette question consiste à calculer un indice de concordance entre partitions. Nous considérons l'indice de Rand à cet effet.

Considérons deux partitions $\mathcal{P}_1$ et $\mathcal{P}_2$ d'un même ensemble de n objets. Chaque partition $\mathcal{P}_h$ est représentée par un tableau relationnel (matrice de comparaison des paires) $C^{(h)}$ dans l'espace des objets de dimension $n \times n$, dont le terme général $c_{i,j}^{(h)}$ est donné par

$$c_{i,j}^{(h)} = \begin{cases} 1 & \text{si les objets } i \text{ et } j \text{ sont dans la même classe de la partition } \mathcal{P}_h \\ 0 & \text{sinon.} \end{cases} \tag{1.1}$$

Puisqu'une partition est une relation d'équivalence, les $c_{i,j}^{(h)}$ vérifient les relations suivantes :

- Réflexivité : $c_{i,i}^{(h)} = 1$;
- Symétrie : $c_{i,j}^{(h)} = c_{j,i}^{(h)}$;
- Transitivité : $c_{i,j}^{(h)} + c_{j,k}^{(h)} - c_{i,k}^{(h)} \leq 1$.

On a de plus les formules suivantes :

- le nombre K_h de classes de la partition $\mathcal{P}_h$ est donné par

$$K_h = \sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{n} c_{i,j}^{(h)}};$$

- Si $n_{u,\cdot}$ désigne le nombre d'objets dans la classe u de la partition $\mathcal{P}_1$, et $n_{\cdot,v}$ celui de la classe v de $\mathcal{P}_2$, alors

$$\begin{aligned} \sum_{i=1}^{n}\sum_{j=1}^{n} c_{i,j}^{(1)} &= Trace\left(C^{(1)}C^{(1)'}\right) = \sum_{u=1}^{K_1} n_{u,\cdot}^2 \\ \sum_{i=1}^{n}\sum_{j=1}^{n} c_{i,j}^{(2)} &= Trace\left(C^{(2)}C^{(2)'}\right) = \sum_{v=1}^{K_2} n_{\cdot,v}^2. \end{aligned}$$

Considérons les $n \times n$ paires d'objets et notons :

- A, le nombre de paires d'objets qui sont dans une même classe de $\mathcal{P}_1$ et dans une même classe de $\mathcal{P}_2$;
- B, le nombre de paires d'objets qui sont dans une même classe de $\mathcal{P}_1$ et dans des classes distinctes de $\mathcal{P}_2$;
- C, le nombre de paires d'objets qui sont dans des classes distinctes de $\mathcal{P}_1$ et dans une même classes de $\mathcal{P}_2$;
- D, le nombre de paires d'objets qui sont dans des classes distinctes de $\mathcal{P}_1$ et dans des classes distinctes de $\mathcal{P}_2$.

La proportion des paires concordantes est donnée par $\frac{A}{n^2}$. Mais il est courant d'utiliser l'indice Rand $R = \frac{A+D}{n^2}$ tel que défini par Marcotorchino and Michaud (1982), si l'on donne la même importance à l'appartenance à une classe et à son complémentaire. Si on note $C^{(1)}$ et $C^{(2)}$ les matrices de comparaisons de paires associées aux deux partitions $\mathcal{P}_1$ et $\mathcal{P}_2$ respectivement, on trouve

–
$$A = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}^{(1)} c_{ij}^{(2)} = Trace\left(C^{(1)}C^{(2)}\right) = \sum_{u=1}^{K_1}\sum_{v=1}^{K_2} n_{uv}$$
où n_{uv} désigne le terme général du tableau de contingence croisant les deux partitions ;

–
$$D = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(1 - c_{ij}^{(1)}\right)\left(1 - c_{ij}^{(2)}\right).$$

L'indice de Rand est alors donné par

$$\begin{aligned} R &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}^{(1)} c_{ij}^{(2)} + \sum_{i=1}^{n}\sum_{j=1}^{n}\left(1 - c_{ij}^{(1)}\right)\left(1 - c_{ij}^{(2)}\right)}{n^2} \\ &= \frac{2\sum_{u=1}^{K_1}\sum_{v=1}^{K_2} n_{uv}^2 - \sum_{u=1}^{K_1} n_{u\cdot}^2 - \sum_{v=1}^{K_2} n_{\cdot v}^2 + n^2}{n^2}, \end{aligned}$$

où $n_{u\cdot}$ et $n_{\cdot v}$ désignent les effectifs des classes u et v des partitions $\mathcal{P}_1$ et $\mathcal{P}_2$ respectivement.

1.B La sélection de variable améliore les performances de la classification

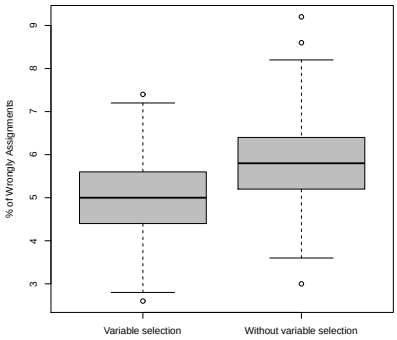
Cet exemple vise à montrer empiriquement que certaines variables peuvent ajouter du bruit à la classification, détériorant ainsi les performances de certaines méthodes de classification. Nous considérons pour cela $L = 10$ gènes (variables) avec $A_l = 10$ allèles chacun. Nous simulons 100 jeux de données selon un modèle de mélange à $K_0 = 5$ composants en proportions égales. Les paramètres alléliques sont choisis de sorte que les allèles des 5 premiers gènes soient différentiellement distribués dans les différentes populations. Notons S_0 le sous-ensemble de ces 5 premiers gènes. Les allèles des autres gènes sont identiquement distribués à travers les $K_0 = 5$ populations. Pour chaque jeu de données, nous considérons la partition simulée. Nous obtenons ensuite deux classifications : la première à partir des gènes de S_0 et à partir de tous les $L = 10$ gènes.

Dans un premier temps, la sélection de la meilleure partition est faite parmi celles à K classe avec $K \in \{1, \ldots, K_{\max} = 10\}$ par le critère BIC. Le Tableau 1.1 montre que lorsqu'on utilise tous les gènes, le critère BIC sous-estime le nombre de composants du mélange.

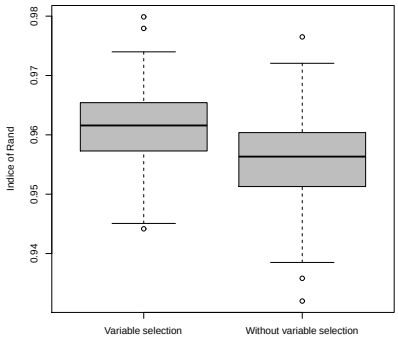
	$\widehat{K}_n$					
	1	2	3	4	5	6
Avec tous les gènes	0	3	93	4	0	0
Seulement les gènes dans S_0	0	0	0	8	92	0

TABLE 1.1 – Nombre de jeux de données pour lesquels $\widehat{K}_n$ est sélectionné avec tous les gènes et avec seulement les 5 gènes de S_0.

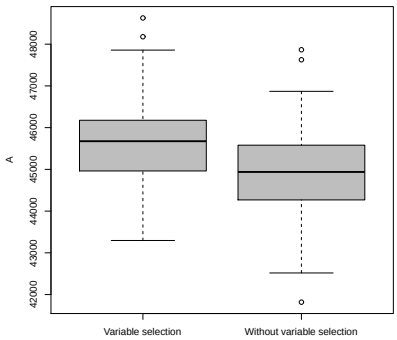
Dans un deuxième temps, nous fixons le nombre de composants du mélange à $K = K_0 = 5$. Nous produisons deux partitions par la règle du Maximum A Posteriori à partir du maximum de vraisemblance estimé par l'algorithme Expectation-Maximisation (EM). La première partition est basée sur tous les gènes et la deuxième sur les 5 gènes de S_0. Chacune des deux partitions est ensuite comparée à la partition simulée par le pourcentage des mal classés et les indices Rand R, A, B, C et D définis dans l'Appendice 1.A). Les Figures 1.1 montrent les box plots de ces indices de comparaison de partitions. Il apparaît clairement que les gènes qui ne sont pas dans S_0 détériorent les capacités de prédiction de la règle du MAP.



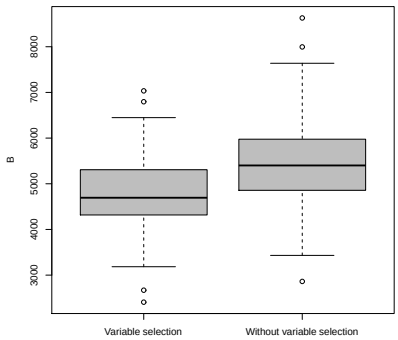
(a) Percentage of wrongly assignments



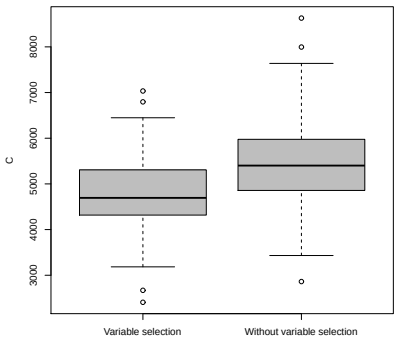
(b) Indice of Rand



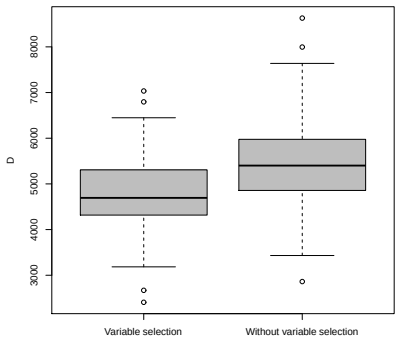
(c) A



(d) B



(e) C



(f) D

Figure 1.1 – These box plots compare percentages of wrongly assignments, the values of R, A, B, C and D defined in Appendix. 1.A

Chapitre 2

Classification non supervisée par mélange fini de lois multinomiales : Application aux données génétiques multilocus

Résumé

Dans ce chapitre, nous rappelons les modèles de mélange fini de lois de probabilité, en particulier dans un cadre multinomial pour résoudre le problème de classification non-supervisée à partir de données génétiques multilocus. La classification est obtenue par Maximum A Posteriori (MAP) après sélection du nombre de composants par des critères du maximum de vraisemblance pénalisés ou par une méthode inspirée de la statistique *Gap* de Hastie et al. (2001). Pour ce qui est du maximum de vraisemblance pénalisé, en plus de considérer les critères BIC (Bayesian Information Criterion) (Schwarz, 1978) et ICL (Integrated Completed Likelihood) (Biernacki et al., 2000), nous nous intéressons aussi à une forme plus générale de la fonction de pénalité. Cette pénalité est calibrée automatiquement sur les données grâce à la version "détection du plus grand saut de dimension" de la méthode nommée "heuristique de la pente" proposée par Birgé and Massart (2007). Nous donnons quelques résultats empiriques obtenus sur des données simulées.

2.1 Introduction

La classification non supervisée a pour but de partitionner les observations auxquelles on s'intéresse dans des groupes "homogènes". Les observations sont regroupées de sorte que les mesures de dissimilarité entre les observations prises deux à deux sont en général plus petites dans un même groupe qu'entre des groupes différents. On distingue en gros deux catégories de méthodes de classification non-supervisée. La première est basée sur la donnée d'une distance ou d'une mesure de dissimilarité. Les méthodes de cette catégorie utilisent les observations sans directement se référer au modèle sous-jacent de lois de probabilités. Comme exemples, citons les algorithmes K-means et de classification hiérarchique. Contrairement aux algorithmes de classification hiérarchique, les résultats de "K-means" dépendent du choix du nombre K de composants et de la classification initiale. Cependant, les méthodes de classification hiérarchique nécessitent que l'utilisateur définisse la mesure de dissimilarité entre des groupes (disjoints) d'observations. La deuxième catégorie de méthodes de classification non supervisée est constituée de méthodes basées sur les mélanges finis de distributions de probabilité. Dans cette thèse, nous nous intéressons à la classification non supervisée par les mélanges finis de lois multinomiales dans le cadre spécifique de données génétique multilocus.

Les mélanges finis de distributions de probabilité ont fait l'objet de nombreux travaux depuis l'article de Newcomb (1886) pour la détection des points aberrants, et celui de Pearson (1894) pour l'estimation des paramètres d'un mélange de deux lois normales. Ces dernières années, le regain de popularité à l'égard de ces modèles est dû à leur flexibilité qui permet de modéliser une large variété de phénomènes aléatoires. Ils reflètent l'idée intuitive que l'échantillon dont on dispose provient d'une population structurée en plusieurs classes homogènes dans le sens où chacune d'elle est caractérisée par une distribution de probabilité. Dans un cadre paramétrique, chaque classe est caractérisée par un jeu de paramètres. Les mélanges finis ont suscité un intérêt aussi bien théorique que pratique comme en témoignent les livres de Everitt and Hand (1981); Titterington et al. (1985); McLachlan and Basford (1988) ou encore celui plus récent de McLachlan and Peel (2000).

Dans cette thèse, nous considérons les données décrites par un vecteur $\mathbf{X} = \left(X^l\right)_{1 \leq l \leq L}$ d'un nombre L de variables aléatoires qualitatives. Le nombre K de composants du mélange est supposé inconnu et est traité comme un paramètre du modèle, les autres paramètres étant les probabilités des différentes modalités des variables X^l. Tout au long de ce travail, l'accent est mis sur la modélisation des données génétiques multilocus. Pour de telles données, chaque composante X^l du L-vecteur aléatoire $\mathbf{X}$ est un ensemble (non ordonné) $\left\{X^{l,1}, \ldots, X^{l,n_{nchr}}\right\}$ d'un nombre n_{chr} de variables qualitatives à valeurs dans un même ensemble de labels $\{1, \ldots, A_l\}$. n_{chr} désigne le nombre de copies de chromosomes de chaque type ; il vaut 2 pour les organismes diploïdes. Nous supposons l'indépendance conditionnelle complète des variables X^l d'une part, et d'autre part de celle des variables $X^{l,1}, \ldots, X^{l,n_{chr}}$ pour tout l. Ces hypothèses sont courantes en génétique des populations : elles servent de modèles de base à l'élaboration de modèles plus complexes, et contribuent à donner un sens biologique aux groupes obtenus. Plus précisément, elles permettent d'obtenir des groupes d'individus génétiquement homogènes

(Pritchard et al., 2000; François et al., 2006) qui constituent dans certains cas des unités de reproduction.

La classification non supervisée par mélange fini se déroule en deux étapes. La première consiste à déterminer un estimateur $\widehat{P}_K$ de la vraie loi P_0 des observations pour chaque valeur de K dans un ensemble de valeurs raisonnables. La deuxième phase consiste à sélectionner un estimateur parmi la collection $\left\{\widehat{P}_K\right\}_K$ et à en déduire une classification par la règle du Maximum A Posteriori (MAP). Ainsi, chaque estimateur $\widehat{P}_K$ est associé à une classification des données en K populations. Dans la suite de ce chapitre, nous considérons la collection des estimateurs du maximum de vraisemblance que nous approchons via l'algorithme Espérance et Maximisation (EM) (Dempster et al., 1977). La phase de sélection est ensuite basée sur un critère du maximum de vraisemblance pénalisé ou sur une méthode qui reprend l'idée de la statistique *Gap* proposée par Hastie et al. (2001). Dans le cas de la sélection via pénalisation du maximum de vraisemblance, la fonction de pénalité est typiquement une fonction de la taille de l'échantillon et du nombre de paramètres libres. Sous des hypothèses faibles sur la fonction de pénalité, la procédure de sélection est consistante pour le nombre de composants du mélange.

La suite du chapitre est organisée comme suit. Les modèles de mélanges finis de lois de probabilité sont présentés dans la Section 2.2, en mettant l'accent sur les mélanges de lois multinomiales dans le contexte de données génétiques multilocus. Le principe de la classification non supervisée par mélange fini est donnée dans la Section 2.3. La procédure de sélection du nombre de composants du mélange est décrite dans la Section 2.5. Des résultats empiriques sur les données simulées sont donnés dans la Section 2.6

2.2 Mélanges finis de lois de probabilité

Une variable aléatoire $\mathbf{X}$ à valeurs dans un espace $\mathbb{X}$ suit une loi de mélange fini si sa densité de probabilité f est une combinaison convexe d'un nombre fini K de densités

$$f(\mathbf{x}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}), \ \mathbf{x} \in \mathbb{X}, \tag{2.1}$$

où les $(f_k(\cdot))_{1\leq k\leq K}$ sont les densités des composants du mélange, et $\pi = (\pi_k)_{1\leq k\leq K}$ le vecteur des proportions du mélange appartenant au $(K-1)$-simplexe $\mathbb{S}_{K-1} = \left\{p = (p_1, \ldots, p_K) \in [0, 1]^K : \sum_{k=1}^{K} p_k = 1\right\}$. Nous nous plaçons dans un cadre paramétrique, ce qui nous permet de caractériser chaque composant k du mélange par un jeu de paramètres α_k. La densité du mélange fini s'écrit alors

$$f_{\theta_K}(\mathbf{x}) = \sum_{k=1}^{K} \pi_k f_{\alpha_k}(\mathbf{x}), \ \forall \mathbf{x} \in \mathbb{X}, \tag{2.2}$$

où $\theta_K = (\pi_k, \ldots, \pi_K, \alpha_1, \ldots, \alpha_K)$ est le vecteur des paramètres du mélange.

Nous supposons que les observations sont décrites par un nombre L de variables. Si les L variables sont continues, il est de coutume de se placer dans un cadre gaussien et de modéliser chaque composant par une densité gaussienne L-dimensionnelle. La loi des observations est alors donnée par

$$f_{\theta_K}(\mathbf{x}) = \sum_{k=1}^{K} \mathcal{N}_{(\mu_k, \Sigma_k)}(\mathbf{x}), \tag{2.3}$$

où $\theta_K = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$ est le vecteur des paramètres à estimer et $\mathcal{N}_{(\mu_k, \Sigma_k)}(\cdot)$ la densité de la loi gaussienne L-dimensionnelle de moyenne μ_k et de matrice de variance-covariance Σ_k. Son expression est donnée par

$$\mathcal{N}_{(\mu_k, \Sigma_k)}(\mathbf{x}) = |2\pi_k\Sigma_k|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu_k)' \Sigma_k^{-1}(x - \mu_k)\right], \ \forall \mathbf{x} \in \mathbb{R}^L. \tag{2.4}$$

Selon Banfield and Raftery (1993) et Celeux and Govaert (1995), chaque matrice Σ_k est décomposable en

$$\Sigma_k = V_k H_k D_k H_k'. \tag{2.5}$$

Dans cette décomposition, $V_k = |\Sigma_k|^{1/L}$, H_k est la matrice orthogonale des vecteurs propres de Σ_k et D_k la matrice diagonale des valeurs propres normalisées de Σ_k, rangées par ordre décroissant et telle que $|D_k| = 1$. L'intérêt d'une telle décomposition est d'avoir une interprétation géométrique des paramètres. En effet, les densités gaussiennes associées à chaque composant du mélange correspondent géométriquement à des ellipsoïdes d'inerties centrés en les moyennes μ_k. Dans la décomposition (2.5), V_k caractérise le volume du composant k, H_k son orientation et D_k sa forme. En faisant varier ou non les volumes, les formes et les orientations, on obtient une collection de 28 modèles présentés dans Celeux and Govaert (1995).

Application aux données génétiques multilocus

Dans cette thèse, nous nous intéressons aux observations décrites par L variables qualitatives. En particulier, les variables auxquelles nous nous intéressons décrivent les données génétiques multilocus. Ces données sont constituées des variants génétiques observés à des positions précises du génome d'organismes vivants. Ces positions sont appelées marqueurs génétiques ou loci. Le génome des organismes vivants est organisé en un certain nombre de chromosomes. Chez la plupart d'espèces d'organismes vivants, chaque cellule contient un certain nombre de paires de chromosomes identiques (à l'exception du chromosome sexuel) : on dit que ces organismes là sont diploïdes. On rencontre aussi des organismes qui, à un stade donné de leur développement ne possèdent qu'un seul exemplaire de chaque type de chromosome : ces organismes sont dits haploïdes. Plus rarement, on peut rencontrer des cellules ou des stades de développement avec 3 exemplaires par type de chromosome : ces organismes sont dits triploïdes. On parle d'organismes polyploïdes lorsque le patrimoine chromosomique est le double de la normale (4, 6 voire plus d'exemplaires de chaque chromosome).

Plaçons nous dans le cas plus général d'organismes ayant n_{chr} exemplaires de chaque type de chromosome, et supposons que l'on s'intéresse à un nombre L de marqueurs génétiques. Par exemple, $n_{chr} = 2$ pour les organismes diploïdes, et $n_{chr} = 1$ pour les organismes haploïdes. Les observations sont alors décrites par un L-vecteur $\mathbf{X} = \left(X^l\right)_{=1}^{L}$ de variables où chaque composante X^l décrit les variants génétiques observés au locus (ou marqueur génétique) l. Chaque locus X^l est décrit par un ensemble non ordonné $\left\{X^{l,1}, \ldots,\ X^{l,n_{chr}}\right\}$ de n_{chr} variables qualitatives à valeurs dans un même ensemble de modalités appelées allèles. Nous noterons $1, \ldots,\ j, \ldots,\ A_l$ les modalités distinctes à la position l, où A_l désigne leur nombre.

Il est de coutume de supposer une indépendance conditionnelle complète des composantes du vecteur aléatoire $\mathbf{X}$. L'expression consacrée en génétique des populations est l'"Équilibre de Liaison" (EL). On parle de "Déséquilibre de Liaison" (DL) entre deux variables (loci) lorsqu'il n'y a pas EL. Cette hypothèse est raisonnable en particulier dans les cas où les positions considérées sur le génome sont suffisamment éloignées les unes des autres pour que la stratification de la population soit la cause principale du DL observée dans la population globale. En effet, la dépendance entre 2 gènes est proportionnelle à la distance qui les sépare sur le génome. Nous nous plaçons dans un cadre paramétrique et considérons les proportions du mélange et les probabilités des différents états $1, \ldots,\ j, \ldots,\ A_l$ des variables $X^{l,a}$, $a = 1, \ldots,\ n_{chr}$ aux différents loci $l = 1, \ldots,\ L$, comme les paramètres du modèle. Nous notons $\alpha_{k,l,j}$ la probabilité de la modalité j du locus l dans le composant k de la population. Autrement dit, $\alpha_{k,l,j} = P\left(X^{l,a} = j |\ j \in \mathcal{P}_k\right)$, où $\mathcal{P}_k$ désigne le composant k de la population. Sous l'hypothèse d'EL,

$$P_{\theta_K}(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \prod_{l=1}^{L} P_{\alpha_{k,l,\cdot}}\left(x^l\right),\ \forall \mathbf{x} = \left(x^l\right)_{1 \leq l \leq L} \in \mathbb{X}, \tag{2.6}$$

avec $\theta_K = \left(\pi_1, \ldots,\ \pi_K,\ (\alpha_{k,l,\cdot})_{1 \leq k \leq K;\ 1 \leq l \leq L}\right)$ où $\alpha_{k,l,\cdot} = (\alpha_{k,l,j})_{1 \leq j \leq A_l}$. La forme de $P_{\alpha_{k,l,\cdot}}(\cdot)$ dépend de la ploïdie des organismes considérés. Dans le cas où les organismes auxquels on s'intéresse sont haploïdes ($n_{chr} = 1$), on a $P_{\alpha_{k,l,\cdot}}(j) = \alpha_{k,l,j}$ et $P_\theta(\cdot)$ est tout simplement un mélange de produit de lois multinomiales.

Maintenant, intéressons nous en général aux organismes polyploïdes avec n_{chr} exemplaires de chaque type de chromosome. En plus de l'hypothèse d'EL, il est courant de supposer que chaque composant du mélange est en Équilibre de Hardy-Weinberg (EHW). Ce qui implique que les croisements se font au hasard dans chaque population. Autrement dit pour chaque locus X^l, les variables $X^{l,a}$, $a = 1, \ldots,\ n_{chr}$ sont indépendantes les unes des autres dans chaque composant du mélange. Les observations d'un individu au locus X^l peuvent être décrites par un A_l-uplet $\mathbf{n}^l = \left(n^{l,j}\right)_{1 \leq j \leq A_l}$ où $n^{l,j}$ donne le nombre de copies de la modalité (allèle) j. On a évidemment $\sum_{j=1}^{A_l} n^{l,j} = n_{chr}$. La loi de Hardy-Weinberg permet d'écrire

$$P_{\alpha_{k,l,\cdot}}\left(x^l\right) = \frac{n_{chr}!}{\prod_{j=1}^{A_l} n^{l,j}!} \prod_{j=1}^{A_l} \alpha_{k,l,j}^{n^{l,j}}.$$

La loi des observations est alors donnée par

$$P_{\theta_K}(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \prod_{l=1}^{L} \left(\frac{n_{chr}!}{\prod_{j=1}^{A_l} n^{l,j}!} \prod_{j=1}^{A_l} \alpha_{k,l,j}^{n^{l,j}} \right). \tag{2.7}$$

Dans les modèles présentés ci-dessus, à K fixé, le vecteur θ_K des paramètres est considéré comme un paramètre à estimer.

2.3 Classification par mélange fini

Considérons un échantillon $\mathbf{x} = (\mathbf{x_1}, \ldots, \mathbf{x_n})$ où les observations $\mathbf{x_i} = \left(x_i^1, \ldots, x_i^L\right)$ de l'individu i sont décrites par L variables. On désire obtenir une classification de ces données en un nombre fini K de groupes, K étant inconnu. Cela revient à rechercher une partition $(\mathcal{P}_k)_{1 \leq k \leq K}$ des observations en K ensembles (disjoints). Cette partition inconnue peut être formalisée par un vecteur $\mathbf{z} = (\mathbf{z_i})_{1 \leq i \leq n}$ où chaque $\mathbf{z_i}$ est un K-uplet $(z_{i,k})_{1 \leq k \leq K}$ donné par

$$z_{i,k} = \begin{cases} 1 & \text{si } \mathbf{x}_i \in \mathcal{P}_k \\ 0 & \text{sinon.} \end{cases} \tag{2.8}$$

La résolution du problème de classification se fait en deux étapes. Le nombre de composants K étant inconnu, on se donne tout d'abord un ensemble raisonnable de valeurs de K, par exemple en se fixant une valeur maximale $K_{\max}$. La première étape consiste à chercher la "meilleure" classification des données en K composants pour chaque valeur K dans cet ensemble. Puis on résout le problème de sélection du nombre de composants (par conséquent de la meilleure classification qui va avec) par exemple par une procédure de sélection de modèle. L'étape de sélection du nombre de composants est abordée dans les sections suivantes. Nous nous concentrons ici sur la classification à K fixé.

Le problème de classification des données peut être résolu de deux manières différentes. La première dite "classifiante", consiste à considérer le vecteur $\mathbf{z}$ comme un paramètre et alors estimer $\mathbf{z}$ en même temps que le vecteur θ_K des paramètres du mélange. Cette approche est utilisée par Dawson and Belkhir (2001) pour à la fois identifier les populations panmictiques et classer les individus. Dans cette thèse, nous nous intéressons à la deuxième approche dite "par mélange". Elle consiste à estimer les paramètres du mélange dans un premier temps, puis à en déduire une classification par Maximum A Posteriori (MAP), c'est-à-dire en affectant chaque observation à la classe dont la probabilité d'appartenance est la plus grande. Notons $\widehat{\theta}_K$, une estimation du vecteur des paramètres du mélange obtenue par exemple par maximisation de la vraisemblance, et $\tau_{i,k}\left(\widehat{\theta}_K\right) := P_{\widehat{\theta}_K}(\mathbf{x}_i \in \mathcal{P}_k | \mathbf{x})$ la probabilité que l'observation $\mathbf{x}_i$ de l'individu i appartienne à la classe $\mathcal{P}_k$ conditionnellement aux observations $\mathbf{x}$ sous la loi de probabilité $P_{\widehat{\theta}_K}(\cdot)$ des observations. La règle du MAP est donnée par

$$\widehat{z}_{i,k} = \begin{cases} 1 & \text{si } \tau_{i,k}\left(\widehat{\theta}_K\right) > \tau_{i,h}\left(\widehat{\theta}_K\right),\ \forall h \neq k \\ 0 & \text{sinon.} \end{cases} \tag{2.9}$$

Nous considérons l'estimateur du maximum de vraisemblance $\widehat{\theta}_K^{MLE}$ que nous noterons tout simplement $\widehat{\theta}_K$ pour simplifier. L'algorithme Espérance et Maximisation (EM) proposé par Dempster et al. (1977) permet d'obtenir une approximation de $\widehat{\theta}_K$.

2.4 Algorithme EM pour l'estimation de θ_K

Pour résoudre le problème de classification, nous avons considéré une approche par mélange fini qui nécessite tout d'abord d'estimation le vecteur θ_K des paramètres du mélange au nombre K fixé de composants. Nous souhaitons déterminer le vecteur des paramètres donnant un mélange à K composants le plus proche au sens de la divergence de Küllback-Leibler de la vraie densité inconnue P_0 des données. Cela revient à chercher le vecteur $\widehat{\theta}_K$ qui maximise la log-vraisemblance observée

$$\mathcal{L}_n\left(\theta;\ \mathbf{x}\right) = \sum_{i=1}^{n} \ln \left\{ \sum_{k=1}^{K} \pi_k \prod_{l=1}^{L} P_{\alpha_{k,l,\cdot}}\left(x_i^l\right) \right\}.$$

Le nombre de paramètres indépendants à estimer est $d_K = K - 1 + K\sum_{l=1}^{L}\left(A_l - 1\right)$. Ce nombre peut être très grand et la fonction à maximiser n'est pas linéaire. La maximisation de la log-vraisemblance selon θ_K est alors très complexe. Si on connaissait le composant d'origine de chaque observation, alors le problème serait un problème d'estimation tout à fait simple et très classique.

2.4.1 Algorithme EM pour approcher $\widehat{\theta}_K$

L'algorithme EM proposé par Dempster et al. (1977) est couramment utilisé pour approcher $\widehat{\theta}_K$ dans un contexte de données incomplètes comme le notre. La force de cet algorithme est justement de s'appuyer sur les données non observées pour réaliser l'estimation du maximum de vraisemblance. Il s'agit d'un algorithme itératif qui procède par maximisations successives de l'espérance de la log-vraisemblance des données complétées conditionnellement aux observations $\mathbf{x}$ et à une valeur courante $\theta^{(r)}$ du vecteur des paramètres

$$Q\left(\theta|\ \theta^{(r)},\ \mathbf{x}\right) = \mathbf{E}_Z\left[\mathcal{L}_n\left(\theta;\ \mathbf{x},\ \mathbf{z}\right) |\ \mathbf{x},\ \theta^{(r)}\right]$$

où

$$\mathcal{L}_n\left(\theta;\ \mathbf{x},\ \mathbf{z}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{i,k} \ln \left\{ \pi_k \prod_{l=1}^{L} \frac{n_{chr}!}{\prod_{j=1}^{A_l} n_i^{l,j}!} \prod_{j=1}^{A_l} \alpha_{k,l,j}^{n_i^{l,j}} \right\}.$$

Il vient alors

$$Q\left(\theta|\ \theta^{(r)},\ \mathbf{x}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K} \underbrace{\mathbf{E}\left[Z_{i,k}|\ \mathbf{x},\ \theta^{(r)}\right]}_{\tau_{i,k}^{(r)}} \ln \left\{ \pi_k^{(r)} \prod_{l=1}^{L} \frac{n_{chr}!}{\prod_{j=1}^{A_l} n_i^{l,j}!} \prod_{j=1}^{A_l} \left(\alpha_{k,l,j}^{(r)}\right)^{n_i^{l,j}} \right\},$$

où le paramètre courant $\theta^{(r)}$ à l'itération r est donné par $\theta^{(r)} = \left(\pi^{(r)},\ \alpha^{(r)}\right)$. Notons que $\tau_{i,k}^{(r)} := \mathbf{E}\left[Z_{i,k}|\ \mathbf{x},\ \theta^{(r)}\right]$ est la probabilité à posteriori que la donnée $\mathbf{x_i}$ provienne du composant k du mélange sous le paramètre courant $\theta^{(r)}$. Partant d'un paramètre initiale $\theta^{(0)}$ donné par l'utilisateur, l'algorithme alterne les deux étapes suivantes jusqu'à convergence vers un point stationnaire de la log-vraisemblance. À l'itération r,

- **Étape E :** Elle consiste à calculer l'espérance $Q\left(\theta|\ \theta^{(r)},\ \mathbf{x}\right)$, ce qui revient à calculer les probabilités $\tau_{i,k}^{(r)}$. Par la règle de Bayes, elles s'écrivent

$$\tau_{i,k}^{(r)} = P\left(Z_{i,k} = 1|\ \mathbf{x},\ \theta^{(r)}\right) = \frac{\pi_k^{(r)} \prod_{l=1}^{L} \frac{n_{chr}!}{\prod_{j=1}^{A_l} n_i^{l,j}!} \prod_{j=1}^{A_l} \left(\alpha_{k,l,j}^{(r)}\right)^{n_i^{l,j}}}{\sum_{h=1}^{K} \pi_h^{(r)} \prod_{l=1}^{L} \frac{n_{chr}!}{\prod_{j=1}^{A_l} n_i^{l,j}!} \prod_{j=1}^{A_l} \left(\alpha_{h,l,j}^{(r)}\right)^{n_i^{l,j}}}.$$

- **Étape M :** L'étape de maximisation consiste à mettre $\theta^{(r)}$ à jour par la valeur $\theta^{(r+1)}$ du paramètre θ qui maximise $Q\left(\theta|\ \theta^{(r)},\ \mathbf{x}\right)$. Le paramètre $\theta^{(r+1)}$ est donné par :

$$\begin{cases} \pi_k^{(r+1)} &= \frac{1}{n} \sum_{i=1}^{n} \tau_{i,k}^{(r)} \\ \alpha_{k,l,j}^{(r+1)} &= \frac{1}{\sum_{i=1}^{n} \sum_{j=1}^{A_l} n_i^{l,j} \tau_{i,k}^{(r)}} \sum_{i=1}^{n} n_i^{l,j} \tau_{i,k}^{(r)}. \end{cases} \tag{2.10}$$

Remarque 2.4.1. *Les formules d'actualisation (2.10) peuvent s'interpréter de la façon suivante. Partant d'un paramètre courant* $\theta^{(r)}$, $\sum_{i=1}^{n} \tau_{i,k}^{(r)}$ *peut être vu comme l'effectif attendu des individus de l'échantillon provenant du composant* k *du mélange. De même,* $\sum_{i=1}^{n} n_i^{l,j} \tau_{i,k}^{(r)}$ *et* $\sum_{i=1}^{n} \sum_{j=1}^{A_l} n_i^{l,j} \tau_{i,k}^{(r)}$ *représentent le nombre attendu de copies d'un allèle* j *du locus* l *dans le composant* k*, et le nombre total d'allèles attendu dans ce même composant respectivement. De sorte que* $\pi_k^{(r+1)}$ *est tout simplement la proportion attendue du composant* k*, et* $\alpha_{k,l,j}^{(r+1)}$ *la fréquence attendue de l'allèle* j *du locus* l *dans le composant* k*.*

D'une itération à l'autre, l'espérance $Q\left(\theta|\ \theta^{(r)},\ \mathbf{x}\right)$ croît, et par conséquent la log-vraisemblance observée $\mathcal{L}_n\left(\theta;\ \mathbf{x}\right)$ aussi. En effet,

$$Q\left(\theta|\ \theta^{(r)},\ \mathbf{x}\right) = \mathcal{L}_n\left(\theta;\ \mathbf{x}\right) + H\left(\theta|\ \theta^{(r)},\ \mathbf{x}\right)$$

avec

$$H\left(\theta|\ \theta^{(r)},\ \mathbf{x}\right) := \mathbf{E}\left[\mathcal{L}_n\left(\mathbf{Z}|\ \mathbf{x},\ \theta\right) |\ \mathbf{x},\ \theta^{(r)}\right],$$

l'espérance conditionnelle de la log-vraisemblance à posteriori de $\mathbf{z}$, sachant les observations $\mathbf{x}$ et le paramètre courant $\theta^{(r)}$. Et on montre par l'inégalité de Jensen que

$$H\left(\theta|\ \theta^{(r)},\ \mathbf{x}\right) \leq H\left(\theta^{(r)}|\ \theta^{(r)},\ \mathbf{x}\right).$$

Sous certaines conditions de régularité, l'estimateur du maximum de vraisemblance obtenu par cet algorithme converge vers un maximum local de la log-vraisemblance (Dempster et al., 1977).

En pratique, l'algorithme EM converge parfois lentement et son résultat dépend de la valeur initiale $\theta^{(0)}$ du vecteur des paramètres. La sous-section suivante présente quelques variants de cet algorithme qui ont été aussi conçus dans le cadre de mélange fini et qui permettent d'améliorer les performances de l'algorithme EM. Le livre de McLachlan and Krishnan (1997) présente une vue d'ensemble des travaux sur l'algorithme EM.

2.4.2 Quelques variants de l'algorithme EM

Celeux and Diebolt (1991) proposent l'algorithme SEM (Stochastic EM) qui consiste à intercaler une étape stochastique S de classification entre les étapes E et M afin de limiter les risques de converger vers un maximum local de la log-vraisemblance observée.

- **Étape E :** Cette étape est identique à celle de l'algorithme classique.
- **Étape S :** À l'itération r, cette étape consiste à simuler le label $z_{i,k}^{(r)}$ de chaque individu selon la loi multinomiale $\mathcal{M}\left(1,\ \tau_{i,k}^{(r)}, \ldots,\ \tau_{i,k}^{(r)}\right)$.
- **Étape M :** Les paramètres sont actualisés en maximisant la log-vraisemblance des observations complétées par les labels obtenus à l'étape S. Dans un cadre multinomial comme le notre, le paramètre qui maximise cette log-vraisemblance est donné par les fréquences observées.

Dans une approche classifiante, Celeux and Govaert (1992) proposent la variante nommée CEM (Classification EM). Ils considèrent les labels inconnus comme des paramètres et visent à maximiser la log-vraisemblance complétée. L'étape C de classification intercalée entre les étapes E et M identiques à celles de SEM, consiste à affecter des labels aux observations par la règle de MAP.

2.5 Sélection du nombre K de composants

Nous sommes maintenant capables d'obtenir la "meilleure" classification au nombre fixé K de composants. Ce nombre est supposé être inconnu et on aimerait que K, aussi bien que la classification, soit estimé à partir des données.

2.5.1 Sélection par critère pénalisé

Une des stratégies de sélection du nombre de composants consiste à considérer les modèles de lois de probabilité définis par les différentes valeurs de K et à sélectionner le modèle minimisant un critère pénalisé. En effet, à K fixé, le vecteur des paramètres θ_K appartient à un ensemble Θ_K qui, dans notre cas est un produit de simplexes. Chaque valeur de K définit ainsi un modèle

$$m_K = \{P_{K,\theta_K} :\ \theta_K \in \Theta_K\}$$

de lois de probabilité correspondant à une situation particulière avec K composants dans le mélange.

Le critère BIC

Remarquons que les modèles m_K sont emboîtés. Wang and Liu (2006) suggèrent le choix du critère BIC (Bayesian Information Criterion) de Schwarz (1978), qui est le critère asymptotique le plus utilisé pour une telle collection de modèles emboîtés. Le critère BIC se place dans un contexte bayésien : θ_K et m_K sont vus comme des variables aléatoires et sont munis de distributions a priori que nous notons $\pi\left(\theta_K|\ m_K\right)$ et $P\left(m_K\right)$ respectivement. Ce critère cherche à sélectionner le modèle le plus vraisemblable au vu des données. Si on considère un prior non informatif uniforme sur les modèles en compétition, sélectionner le modèle le plus vraisemblable revient à sélectionner celui maximisant la vraisemblance intégrée donnée par

$$P(\mathbf{x}|\ m_K) = \int_{\Theta_K} P\left(\mathbf{x},\ \theta_K\right) \pi\left(\theta_K|\ m_K\right) d\theta_K,$$

où $P\left(\mathbf{x},\ \theta_K\right)$ est la distribution jointe du vecteur des paramètres θ_K et des observations $\mathbf{x}$, et $\pi\left(\theta|\ K\right)$ la distribution a priori du vecteur des paramètres. Le calcul exact de cette vraisemblance intégrée est rarement possible. Le critère BIC est alors une approximation de Laplace de sa log-transformée. Il peut être donné par

$$BIC\left(K\right) = -\frac{1}{n} \ln\left[P\left(\mathbf{x}|\ K,\ \widehat{\theta}_K\right)\right] + \frac{\ln n}{2n} d_K,$$

où d_K est le nombre de paramètres libres du modèle m_K et $\widehat{\theta}_K$ l'estimateur du maximum de vraisemblance de P_0 dans le modèle m_K. Le modèle sélectionné par BIC est alors donné par

$$\widehat{K}_n \in \min_{K \in \mathbb{N}^*} BIC\left(K\right).$$

Bien que les conditions classiques de régularité ne soient pas satisfaites par les mélanges pour justifier l'approximation de Laplace, BIC converge pour de nombreux modèles (Keribin, 2000).

Précisons quel est le modèle recherché par BIC dans notre contexte. Le critère BIC est en général consistant pour la dimension. Il semblait difficile de concevoir que ce critère permette la convergence vers un modèle autre que celui qui a engendré les données que nous appellerons vrai modèle. C'est probablement pour cette raison que certains auteurs ont choisi de supposer que le vrai modèle appartient à la collection des modèles en compétition (Schwarz, 1978; Raftery, 1995), bien que nulle part cette hypothèse n'apparaisse comme nécessaire dans la construction de BIC. Les modèles $m_1, \ldots,\ m_K, \ldots$ sont emboîtés. La divergence de Küllback-Leibler (**KL**) est donnée par

$$\mathbf{KL}\left(P, Q\right) = \sum_{\mathbf{x} \in \mathbb{X}} \ln\left(\frac{P\left(\mathbf{x}\right)}{Q\left(\mathbf{x}\right)}\right) P\left(\mathbf{x}\right),$$

où $\mathbb{X}$ est l'ensemble des états possibles de la variable d'intérêt $\mathbf{X}$. Définissons

$$\mathbf{KL}\left(P,\ m\right) := \inf_{Q \in m} \mathbf{KL}\left(P,\ Q\right),$$

où m désigne un ensemble de distributions de probabilité sur $\mathbb{X}$, et considérons la sous-collection

$$\mathcal{C}_{\max} = \{m_K\}_{1 \leq K \leq K_{\max}}$$

associée à une valeur maximale $K_{\max}$ du nombre de composants du mélange. Puisque les modèles m_K, $K = 1, \ldots,\ K_{\max}$ sont emboîtés, la fonction $K \mapsto \mathbf{KL}\left(P_0,\ m_K\right)$ est décroissante. Il existe un modèle m_{K^*} à partir duquel $\mathbf{KL}\left(P_0,\ m_K\right)$ ne diminue plus. Du point de vue de la divergence $\mathbf{KL}$, le modèle m_{K^*} doit être préféré aux sous-modèles m_K, $K = 1, \ldots,\ K^*-1$ puisqu'il est plus proche de la vrai loi P_0 des observations. Par ailleurs, m_{K^*} doit aussi être préféré à tous les sur-modèles m_K, $K = K^*+1, \ldots,\ K_{\max}$ puisqu'ils sont plus complexes sans pour autant être plus proche de P_0. Burnham and Anderson (2002) définissent m_{K^*} comme le "quasi-vrai" modèle. Dans notre contexte précis, nous avons le résultat de consistance suivant.

Proposition 2.5.1. *Soit K_{max} une valeur maximale du nombre de composants du mélange donnée par l'utilisateur. Si la vraie loi des observations est strictement positive sur $\mathbb{X}$, alors tout critère de type BIC défini ci-dessous sélectionne le quasi-vrai modèle associé à la sous-collection $\mathcal{C}_{K_{\max}}$ avec une probabilité qui tend vers 1 lorsque la taille n de l'échantillon tend vers l'infini.*

Démonstration. La preuve de ce résultat est semblable à celle du résultat de consistance donnée dans le Chapitre 3 relatif aux critères de type BIC pour la sélection simultanée du nombre de composants et du sous-ensemble de variables pertinent pour la classification. Il faut juste remplacer le vrai modèle par le "quasi-vrai" modèle défini plus haut. □

Le critère BIC entre dans la famille des critères du maximum de vraisemblance pénalisée qui se mettent sous la forme

$$\mathbf{crit}\left(K\right) = \gamma_n\left(P_{\widehat{\theta}_K}\right) + \mathbf{pen}_n\left(K\right), \tag{2.11}$$

où

$$\gamma_n\left(P\right) := -\frac{1}{n}\sum_{i=1}^{n} \ln P\left(\mathbf{x}_i\right)$$

est le contraste empirique log-vraisemblance mesurant l'ajustement du modèle aux données, et $\mathbf{pen}_n\left(\cdot\right)$ la fonction de pénalité dépendant typiquement de la taille n de l'échantillon et de la complexité des modèles en compétition via leur dimension. La dimension d'un modèle m_K est donné par

$$d_K = K - 1 + K\sum_{l=1}^{L}\left(A_l - 1\right),$$

qui est le nombre de paramètres libres dans l'espace Θ_K.

Définition 2.5.1. *Un critère du maximum de vraisemblance pénalisé de la forme (2.11) est dit de type BIC si son terme de pénalité est sous la forme* $\mathbf{pen}_n(K) = \frac{\lambda_n}{n} d_K$ *où* λ_n *est une fonction de la taille* n *de l'échantillon vérifiant* $\lim_{n\to\infty} \frac{\lambda_n}{n} = 0$ *et* $\lim_{n\to\infty} \lambda_n = \infty$.

BIC est évidemment le prototype le plus utilisé d'un tel critère, avec $\lambda_n = \frac{\ln n}{2}$.

Critère avec calibration du terme de pénalité

La famille de critères de type BIC que nous avons défini dans la sous-section précédente entre dans une famille beaucoup plus large de critères pour lesquels la fonction de pénalité est proportionnelle à la dimension des modèles :

$$\mathbf{pen}(K) = \lambda d_K,$$

où λ dépend du jeu de données à analyser. Une telle forme de la pénalité est parfois obtenue dans le but d'établir une inégalité de type oracle. C'est par exemple le cas dans le Chapitre 4 où la pénalité obtenue est aussi consistance. Une méthode nommé "heuristique de la pente" pour la calibration de λ est proposée dans Birgé and Massart (2007). Une mise en oeuvre pratique est décrite dans Arlot and Massart (2008). Cette méthode est basée sur la conjecture qu'il existe une pénalité minimale définie par une valeur $\lambda_{\min}$ de λ pour que la procédure de sélection marche. La pénalité optimale est alors deux fois la pénalité minimale :

$$\mathbf{pen}_{opt}(K) = 2\lambda_{\min} d_K.$$

Nous avons adopté la version “détection du plus grand saut de dimension“ combinée à une fenêtre glissante. Elle est décrite dans le Chapitre 4 où nous montrons aussi sur des simulations qu'une telle méthode de calibration de la fonction de pénalité adapte la procédure de sélection à la taille de l'échantillon.

Le critère ICL (Integrated Completed Likelihood)

Le critère BIC est une approximation de la log-vraisemblance intégrée observée et se place dans un cadre d'estimation de la densité. Le critère ICL proposé par Biernacki et al. (2000) est quant à lui construit dans une approche classifiante. Il est obtenu à partir d'une approximation de Laplace de la vraisemblance complétée intégrée

$$P(\mathbf{x},\ \mathbf{z}|\ K) = \int_{\Theta_K} P(\mathbf{x},\ \mathbf{z}|\ K,\ \theta)\, \pi(\theta|\ K)\, d\theta.$$

Cette approximation est donnée par

$$\ln P(\mathbf{x},\ \mathbf{z}|\ K) \approx -\ln P\left(\mathbf{x},\ \mathbf{z}|\ K,\ \widehat{\theta}^*\right) + \frac{\ln n}{2} d_K,$$

où $\widehat{\theta}^*$ est le vecteur des paramètres maximisant la vraisemblance complétée. Mais les labels $\mathbf{z}$ ne sont pas observés. Dans la pratique, on remplace $\widehat{\theta}^*$ par l'estimateur du maximum de vraisemblance $\widehat{\theta}_K$ des observations $\mathbf{x}$, et le vecteur des labels $\mathbf{z}$ par $\widehat{\mathbf{z}} = MAP\left(\widehat{\theta}_K\right)$. Le critère ICL est alors donné par

$$ICL\left(K\right) = -\frac{1}{n}\ln P\left(\mathbf{x},\ \widehat{\mathbf{z}}|\ K,\ \widehat{\theta}_K\right) + \frac{\ln n}{2n}d_K, \tag{2.12}$$

et peut être décomposé de la manière suivante :

$$ICL\left(K\right) = BIC\left(K\right) + \frac{1}{n}ENT\left(K\right),$$

où le terme d'entropie est donné par

$$ENT\left(K\right) = -\sum_{i=1}^{n}\sum_{k=1}^{K}\widehat{z}_{i,k}\ln\left[\tau_{i,k}\left(\widehat{\theta}_K\right)\right].$$

En effet,

$$\begin{aligned}
ICL\left(K\right) =& -\frac{1}{n}\ln P\left(\mathbf{x}|\ K,\ \widehat{\theta}_K\right) + \frac{\ln n}{2n}d_K \\
&-\frac{1}{n}\ln P\left(\mathbf{x},\ \widehat{\mathbf{z}}|\ K,\ \widehat{\theta}_K\right) + \frac{1}{n}\ln P\left(\mathbf{x}|\ K,\ \widehat{\theta}_K\right) \\
=& BIC\left(K\right) - \frac{1}{n}\ln\left[\frac{P\left(\mathbf{x},\ \widehat{\mathbf{z}}|\ K,\ \widehat{\theta}_K\right)}{P\left(\mathbf{x}|\ K,\ \widehat{\theta}_K\right)}\right] \\
=& BIC\left(K\right) - \frac{1}{n}P\left(\widehat{\mathbf{z}}|\ \mathbf{x},\ \widehat{\theta}_K\right) \\
=& BIC\left(K\right) - \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\widehat{z}_{i,k}\ln\left[\tau_{i,k}\left(\widehat{\theta}_K\right)\right].
\end{aligned}$$

On peut ainsi voir le critère ICL comme BIC auquel on a ajouter une pénalité sous forme d'entropie. Le terme d'entropie mesure la capacité du mélange à bien séparer les classes : si les classes obtenues sont bien distinctes, le terme d'entropie $ENT\left(K\right)$ est proche de zéro, alors qu'il est grand lorsque les classes ne sont pas bien séparées.

2.5.2 Sélection inspirée par la statistique *Gap*

La méthode d'estimation du nombre K de composants du mélange que nous décrivons ici reprend l'idée de la statistique *Gap* proposée par Hastie et al. (2001) pour la méthode de classification K-means. L'idée est la suivante : pour les modèles de suffisamment grandes dimensions, le biais de l'estimateur du minimum de contraste se stabilise et ne peut être significativement amélioré. Pour être plus précis, si K^* défini le "quasi-vrai" modèle, alors en passant de K à $K+1$, l'augmentation du maximum de la log-vraisemblance

$$\mathcal{L}_n\left(\widehat{\theta}_{K+1};\ \mathbf{x}\right) - \mathcal{L}_n\left(\widehat{\theta}_K;\ \mathbf{x}\right)$$

est beaucoup plus grande pour $K < K^*$ que pour $K \geq K^*$:

$$\left\{\mathcal{L}_n\left(\widehat{\theta}_{K+1};\ \mathbf{x}\right) - \mathcal{L}_n\left(\widehat{\theta}_K;\ \mathbf{x}\right)\ :\ K < K^*\right\} >> \left\{\mathcal{L}_n\left(\widehat{\theta}_{K+1};\ \mathbf{x}\right) - \mathcal{L}_n\left(\widehat{\theta}_K;\ \mathbf{x}\right)\ :\ K \geq K^*\right\}.$$

Une méthode pratique pour estimer K^* est basée sur la comparaison du graphe du maximum de la log-vraisemblance $\mathcal{L}_n\left(\widehat{\theta}_K;\ \mathbf{x}\right)$ en fonction de K à celui de son espérance $\mathbf{E}_n^*\left[\mathcal{L}_n\left(\widehat{\theta}_K;\ \mathbf{x}\right)\right]$ sous un modèle nul de référence. Hastie et al. (2001) recommandent de commencer avec l'hypothèse d'un modèle avec une seule population (c'est-à-dire $K = 1$) qui doit être rejetée en faveur d'un modèle à $K > 1$ populations. La quantité $\mathbf{E}_n^*\left[\mathcal{L}_n\left(\widehat{\theta}_K;\ \mathbf{x}\right)\right]$ peut être estimée par une procédure de Monte Carlo, c'est-à-dire par la moyenne empirique $\overline{\mathcal{L}^{K*}}$ des valeurs $\left\{\mathcal{L}_n\left(\widehat{\theta}_K^{*b};\ \mathbf{x}^{*b}\right)\right\}_{b=1,\ldots,\ B}$ obtenues à partir de B jeux de données $\left\{\mathbf{x}^{*b}\right\}_{b=1,\ldots,\ B}$ de taille n simulés à partir de $\widehat{\theta}_1$

$$\overline{\mathcal{L}_n^{K*}} = \frac{1}{B}\sum_{b=1}^{B}\mathcal{L}_n\left(\widehat{\theta}_K^{*b};\ \mathbf{x}^{*b}\right). \tag{2.13}$$

On estime alors K^* par la valeur $\widehat{K}^*$ de K pour laquelle la quantité

$$\widehat{G}\left(K\right) := \mathcal{L}_n\left(\widehat{\theta}_K|\ \mathbf{x}\right) - \overline{\mathcal{L}_n^{K*}}$$

est maximale. La procédure d'estimation est décrite dans l'algorithme (1) suivant.

Algorithm 1 Estimation de K^* par le méthode de *Gap*

1: Choisir une distribution nulle sur $\mathbb{X}$;
2: Simuler B jeux de données $\mathbf{x}^{*b}$, $b = 1, \ldots,\ B$ distribuées selon la distribution nulle choisie ;
3: Calculer $\mathcal{L}_n\left(\widehat{\theta}_K^{*b}|\ \mathbf{x}^{*b}\right)$, $K = 1, \ldots,\ K_{max}$ et $b = 1, \ldots,\ B$;
4: Pour $K = 1, \ldots,\ K_{\max}$, calculer $\overline{\mathcal{L}_n^{K*}}$ et la variance $\sigma_K^2 = \frac{1}{B-1}\sum_{b=1}^{B}\left(\mathcal{L}_n\left(\widehat{\theta}_K^{*b}|\ \mathbf{x}^{*b}\right) - \overline{\mathcal{L}_n^{K*}}\right)^2$;
5: Calculer $\widehat{G}\left(K\right) = |\mathcal{L}_n\left(\widehat{\theta}_K|\ \mathbf{x}\right) - \overline{\mathcal{L}_n^{K*}}|$, $K = 1, \ldots,\ K$;
6: $\widehat{K}^* = \arg\max_K\left\{\widehat{G}\left(K\right)\right\}$;
7: $\widehat{K} = \arg\min_K\left\{K|\ \widehat{G}\left(K\right) \geq \widehat{G}\left(K+1\right) - \delta\sigma_{K+1}\right\}$ où $\delta = \sqrt{1 + \frac{1}{B}}$ est un coefficient de correction ;
8: **return** $\widehat{K}^*$, $\widehat{K}$;

Selon Hastie et al., multiplier σ_K par $\delta = \sqrt{1 + \frac{1}{B}}$ augmente la puissance du test contre l'hypothèse nulle d'une seule population.

Les Figures 2.1 montrent les variations de la log-vraisemblance et de la statistique *Gap* dans un exemple de jeu de données génétiques simulées selon un mélange de $K_0 = 5$ populations. Dans cette exemple, la valeur $\widehat{K}^* = 6$ est associée à la plus grande valeur de la statistique *Gap*, mais le nombre de composants sélectionné est bien $\widehat{K}_n = 5$.

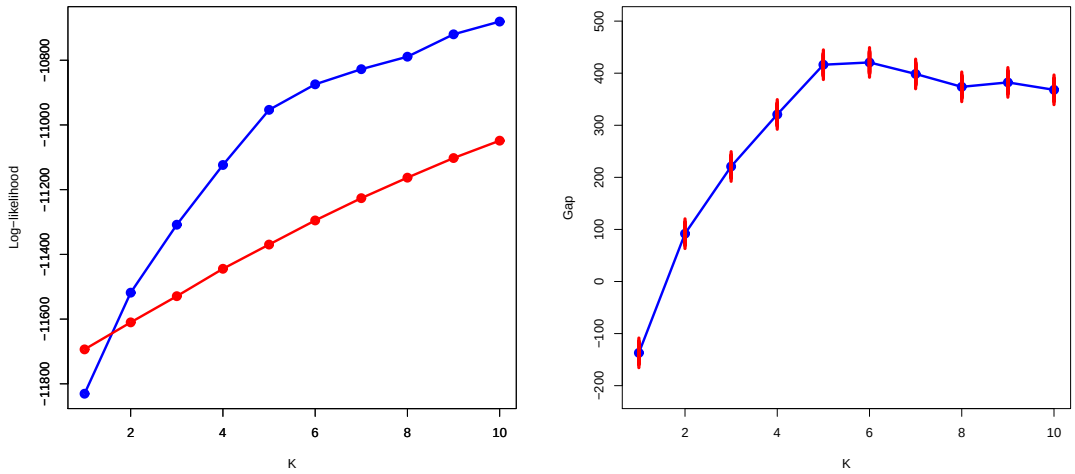


FIGURE 2.1 – Exemple de variation de la log-vraisemblance et de la statistique $\widehat{G}$ en fonction de K. Dans ce exemple, $B = 20$ jeux de données sont simulées soue le modèle à un composant. Les valeurs estimées de K sont $\widehat{K}^* = 6$ et $\widehat{K} = 5$.

2.6 Mise en pratique et simulations

2.6.1 Mise en pratique

Il existe un certains nombre de logiciels dédiés aux modèles de mélange. Pour l'analyse de données génétiques multilocus, citons entre autres les logiciels `Structure` de Pritchard et al. (2000), `Fastruct` de Chen et al. (2006) et `BAPS` (Bayesian Application for Population Structure) de Corander et al. (2008). Le premier considère que le génome de chaque individu de la population cible est un mélange. Ainsi, les probabilités alléliques sont modélisées comme des mélanges finis. Cela permet de prendre en compte les individus issus de parents originaires de populations différentes. Ainsi, chaque individu a ses propres proportions du mélange, ce qui augmente considérablement le nombre de paramètres à estimer. Tout comme le logiciel `BAPS`, la procédure d'estimation des paramètres de `Structure` repose sur la méthode de Monte Carlo par chaînes de Markov (MCMC), ce qui peut être très coûteux en temps de calcul. Le logiciel `Fastruct` quant à lui est basé sur l'algorithme EM et ne réalise pas la sélection du nombre de populations. Comme le disent Chen et al., il peut servir de point de départ pour `Structure` en fournissant les paramètres initiaux associés à une vraisemblance assez élevée.

Nous proposons un nouveau logiciel nommé `MixMoGenD` pour "Mixture Model for Genotypic Data". Comme `Fastruct`, il repose sur l'algorithme EM. Mais en plus de réaliser la sélection du nombre K de populations, ce logiciel s'intéresse aussi aux marqueurs génétiques les plus pertinents qui discriminent les différentes populations. La question de la sélection de variable dans la classification non-supervisée par mélange fini est abordée dans les chapitres 3 et 4. La description de la procédure de sélection de `MixMoGenD` est décrite dans le chapitre 5, qui présente en même temps comment se servir de ce logiciel. Ce logiciel est implémenté sous le langage $C++$ dans une approche orienté objet. La gestion de la mémoire est entièrement dynamique de sorte que seules la capacité mémoire et la puissance de calcul de l'ordinateur de l'utilisateur déterminent la taille limite des données à analyser.

2.6.2 Simulations

Dans cette partie, la sélection du nombre K de composants est réalisée par les critères du maximum de vraisemblance pénalisés BIC, ICL et un critère avec une pénalité de la forme $\mathbf{pen}(K) = \lambda d_K$ que nous notons $\lambda \cdot d_K$. Nous considérons aussi la méthode de sélection inspirée de la statistique *Gap* décrite dans la sous section 2.5.2. Les données simulées sont supposées être celles d'une population d'individus diploïdes structurée en $K_0 = 5$ sous-populations. Ces données concernent $L = 10$ marqueurs génétiques avec $A_l = 10$ allèles par marqueur. Les paramètres alléliques sont choisis de sorte que le niveau de différentiation génétique mesuré par le F_{st} [1] est de 0.0387. Un tel niveau de différentiation génétique est jugé critique pour une classification non-supervisée (Latch et al., 2006). Nous considérons différentes tailles de l'échantillon ($n \in \{300,\ 400,\ 500,\ 600\}$). Dans la méthode *Gap*, pour chaque jeu de données, $B = 20$ jeux de données sont simulés sous le modèle à une population. Le Tableau 2.1 résume les résultats obtenus sur dix jeux de données simulées. Ce tableau donne le nombre de jeux de données pour lesquelles on a $\widehat{K}_n$ composants.

Il apparaît que les méthodes *Gap* et $\lambda \cdot d_K$ ont des comportements semblables pour la sélection du nombre de composants du mélange. Elles se comportent globalement bien pour les différentes tailles de l'échantillon choisies. Notons par exemple que, déjà pour une taille d'échantillon de $n = 300$, elles ont sélectionné le vrai nombre de composants neuf fois sur dix jeux de données simulées. Nous sommes dans une situation critique non seulement de différentiation génétique, mais aussi de taille de l'échantillon au regard du nombre de paramètres libres à estimer qui est de $d_{K_0} = 4 + 5 \times 9 \times 10 = 454$. D'autre part, les critères BIC et ICL donnent aussi des résultats semblables et font moins bien sur les données de petites tailles. Cela est probablement dû au terme de pénalité de BIC qui dans une telle situation sur-pénalise les modèles. Dans notre cas particulier de données génétiques multilocus, BIC et ICL donnent des résultats semblables probablement parce que le terme d'entropie de ICL des données simulées est trop faible (de l'ordre de 10) pour compenser les variations (très grandes) des dimensions lorsqu'on passe d'une valeur de K à la suivante $K + 1$ (de l'ordre de 90).

1. Le F_{st} est un indice appartenant à $[0, 1]$ qui mésure la part de la stratification de la population dans la variabilité génétique totale

		$\widehat{K}_n$						
n	**crit**	1	2	3	4	5	6	7
300	BIC	3	7	0	0	0	0	0
	ICL	4	6	0	0	0	0	0
	Gap	0	0	0	1	9	0	0
	$\lambda \cdot d_K$	0	0	0	1	9	0	0
400	BIC	1	5	4	0	0	0	0
	ICL	1	6	3	0	0	0	0
	Gap	0	0	0	0	10	0	0
	$\lambda \cdot d_K$	0	0	0	0	10	0	0
500	BIC	0	4	4	2	0	0	0
	ICL	0	5	4	1	0	0	0
	Gap	0	0	0	0	10	0	0
	$\lambda \cdot d_K$	0	0	0	0	10	0	0
600	BIC	0	0	0	2	8	0	0
	ICL	0	0	0	2	8	0	0
	Gap	0	0	0	0	10	0	0
	$\lambda \cdot d_K$	0	0	0	0	10	0	0

TABLE 2.1 – Les jeux de données sont simulées dans une configuration avec $K_0 = 5$ populations, $L = 10$ loci avec $A_l = 10$ allèles par locus. Les paramètres de simulation sont choisis de sorte que le niveau de différentiation génétique est $F_{st} = 0.0387$. 10 jeux de données sont simulées par taille n de l'échantillon. Le tableau donne le nombre de jeux de données pour lesquelles chaque valeur $\widehat{K}_n$ du nombre de composants du mélange est sélectionnée : BIC=critère BIC, ICL=critère ICL, *Gap*=inspirée de la statistique *Gap*, $\lambda \cdot d_K$=critère du maximum de vraisemblance pénalisé avec calibration du terme de pénalité par la détection du plus grand saut de dimension.

2.7 Discussion

Dans ce chapitre, nous avons considéré les modèles de mélange fini de lois de probabilité pour résoudre le problème de classification non-supervisée. Nous nous sommes particulièrement intéressés aux données génétiques multilocus, ce qui nous a placé dans un cadre multinomial. Le problème de sélection du nombre de composants du mélange est transformé en un problème de sélection de modèle pour l'estimation de la densité de probabilité des observations. La procédure de sélection s'appuie sur les critères du maximum de vraisemblance pénalisés tels que BIC, ICL et une famille de critères associés à une fonction de pénalité de la forme λd_K, où d_K désigne la dimension du modèle à évaluer et λ une constante dépendant des données. La constante λ est calibré grâce à la détection du plus grand saut de dimension (voir Chapitre 4). Nous nous sommes aussi intéressés à une méthode de sélection de modèle inspirée de la statistique *Gap* proposée par Hastie et al. (2001). Sur les données simulées, il ressort que la méthode *Gap* et celle avec calibration automatique de la constante λ de la pénalité sont globalement meilleures que celles basées sur les critères BIC et ICL, du moins empiriquement. La différence entre les deux groupes de méthodes est plus prononcée pour les jeux de données de petites tailles. Nous avons constaté sur ces simulations que le terme d'entropie de ICL est de

l'ordre de 10, ce qui est négligeable au regard du saut de dimension lorsqu'on passe d'une valeur de K à $K+1$ (de l'ordre de 90). C'est probablement ce qui explique les comportements semblables de BIC et ICL sur ces données là. Ces simulations montrent de façon empirique que "bien" estimer la densité de probabilité des observations sous forme de mélange fini permet de bien estimer le nombre de composants lorsqu'il s'agit de données génétiques multilocus.

Vu l'explosion de projets génomiques, le nombre de marqueurs génétiques auxquels on s'intéresse est de plus en plus grand. Il peut arriver que seul un sous-ensemble de marqueurs génétiques soient pertinents pour discrimer les populations entre-elles, les autres n'apportant que du bruit à la classification. Il se pose alors la question supplémentaire de sélection des loci les plus pertinents pour discriminer les différentes populations. Dans les deux chapitres suivants, nous nous intéressons au problème double de sélection de variable et de la classification non supervisée sur les données génétiques multilocus.

Chapitre 3

Variable selection in a model-based clustering using multilocus genotypic data

This chapter is a slightly modified version of the article of the same title by the author in collaboration with Elisabeth Gassiat, appeared in *Advances in Data Analysis, 3 (2) 109-134*.

Abstract

We propose a variable selection procedure in model-based clustering using 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 seen 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 $\left(\widehat{K}_n,\ \widehat{S}_n\right)$. We also propose an associated stand alone $C++$ program named `MixMoGenD` (Mixture Model for Genotypic Data). It is available free of charge on `www.math.u-psud.fr/~toussile`. To avoid an exhaustive search of the optimum model, the selection procedure of `MixMoGenD` is based on a modified Backward-Stepwise algorithm, which enables a better search 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.

3.1 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 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`, `BAPS` (Bayesian Analysis of Population Structure), `Geneland` and `Fastruct` proposed by Pritchard et al. (2000), Corander et al. (2008), Guillot et al. (2005) and François et al. (2006) respectively. Multilocus 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 performs 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 Genotype Data* (`MixMoGenD`). As almost all already proposed model-based clustering for multilocus genotype data, our procedure attempts to group samples into clusters of randomly 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 micro-evolution. Recall that in clustering, classification is not observed and there is no prior knowledge on the cluster 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 Maugis et al. (2009) 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 $\mathcal{P}^*(L)$ be the number of available loci and the set of all nonempty subsets of these L loci respectively. Denote by $\mathbb{N}^*$ the set of positive integers. A specific collection

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

of models is defined such that each model $\mathcal{M}_{(K,\ S)}$ 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 independent realizations from an unknown probability distribution P_0 in some of the competing models $\mathcal{M}_{(K,\ S)}$. Consequently, inferring $(K,\ S)$ can be formulated as the choice of a model among the collection $\mathcal{C}$. 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. We propose to

use a penalized maximum likelihood criterion. There exists a huge literature on model selection via penalized criteria, see Massart (2007) and the references therein. Our analysis and our algorithm do not impose a particular choice of the penalization term. For the numerical experiments, we will use the BIC.

Although there exists a lot of articles concerning the behavior of the BIC and other penalization methods in practice, theoretical results in a mixture framework are few. The consistency of the BIC estimator is shown in Maugis et al. (2009) for a variable selection problem with Gaussian mixture models when the number of components is known. A general consistency theorem may be found in Gassiat (2002), applications to mixture models are developed for example in Azais et al. (2009) and Chambaz et al. (2008) (see also references therein). 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.

We also propose what we called "Backward-Stepwise Explorer" algorithm which avoids an exhaustive search of the optimum model (which can be very painful in most situations) and enables the search of the optimum set of clustering variables among all possible cardinalities of S.

Our paper is organized as follows. In section 3.2, we describe the competing models and the model selection principle we will use. We then describe the "Backward-Stepwise Explorer" algorithm we propose to perform the model selection. In section 3.3, 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 obtained by Allman et al. (2008). We finally give our main consistency result for the estimation of the model using penalized criteria such as BIC type criteria. Section 3.4 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 on the estimation of the number of clusters and the prediction capacity is obviously due to the variable selection procedure.

3.2 Model and methods

3.2.1 Framework, notation and competing models

The dataset we shall deal with consists of the genotypes at L loci $l = 1, \ldots, L$ of a sample of n diploid individuals $i = 1, \ldots, n$. The observations are denoted by $x_1, \ldots, x_n$, with x_i containing the genotypes of individual i at the L loci, that is $x_i = (x_i^l)_{l=1,\ldots,L}$, where x_i^l is the genotype of the individual i at the l^{th} locus. The genotype x_i^l consists

of a (unordered) 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, otherwise as heterozygous. The dataset $x_1, \ldots, x_n$ is assumed to be a realization of a n-sample (that is n independent and identically distributed random vectors) 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 $\{1, \ldots,\ j, \ldots,\ A_l\}$ of observed alleles at locus l. Let :

- Z be the unobserved random variable indicating the population an individual comes from. We will denote by z_i the (unobserved) population of origin of the sampled individual i, with $i = 1, \ldots,\ n$;
- $\pi_k := P\left(Z = k\right)$, the probability that an individual comes from population k with $k = 1, \ldots,\ K$ (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, with $j = 1, \ldots,\ A_l$;
- and $\mathbb{X}$, the set of all possible genotypes from the observed alleles, that is

$$\mathbb{X} = \prod_{l=1}^{L} \left\{ \{a,\ b\} \mid a,\ b \in \{1, \ldots,\ A_l\} \right\}. \tag{3.1}$$

Model-based clustering methods proceed by assuming that the observations from each population 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 randomly mating individuals so that the *Hardy-Weinberg* (HW) and linkage disequilibria (LD) are minimized across the sample (see Latch et al. (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 micro-evolution. Thus we assume *Hardy-Weinberg* and complete linkage equilibria in each cluster. *Hardy-Weinberg* equilibrium in each cluster means that conditionally to Z, for any locus l the random variables X_1^l and X_2^l are independent so that the probability to observe a genotype x^l at locus l is given by

$$P\left(x^l |\ Z = k\right) = \left(2 - \mathbb{1}_{\left[x_1^l = x_2^l\right]}\right) \alpha_{k,l,x_1^l} \times \alpha_{k,l,x_2^l}. \tag{3.2}$$

Complete linkage equilibrium in each cluster means that within each population, the genotypes at different loci are independent random vectors.

Now assume that there exists an (unknown) subset S of loci such that the individuals are clustered into an unknown number K of clusters on the basis of the data from the corresponding loci. Thus the set S^c such that $S \cup S^c = \{1, \ldots,\ L\}$ is the one of the loci that are just noise or irrelevant for clustering purposes. Typically, the reason why the loci of S^c are not relevant for clustering purposes is that their alleles are equally distributed across the clusters. This means that

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

$$\alpha_{1,l,j} = \alpha_{2,l,j} = \cdots = \alpha_{K,l,j} =: \beta_{l,j}. \tag{3.3}$$

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

$$\begin{aligned} P_{(K,\,S)}\left(x|\,\theta\right) &= P\left(x|\,K,\,S,\,\theta\right) \\ &= \left[\sum_{k=1}^{K}\pi_k\prod_{l\in S}P\left(x^l|\,\alpha_{k,l,\,.}\right)\right]\times\prod_{l\in S^c}P\left(x^l|\,\beta_{l,\,.}\right), \end{aligned} \tag{3.4}$$

for any $x=(x^l)_{l=1,\ldots,L}\in\mathbb{X}$ (see (3.1)), where

$$\theta := \left(\pi,\ (\alpha_{.,l,\,.})_{l\in S},\ (\beta_{l,\,.})_{l\in S^c}\right)$$

is considered as a multidimensional parameter ranging in the set $\Theta_{(K,\,S)}$ defined in equation (3.5) below for a given K and S, with

$$\begin{aligned} \alpha_{k,l,.} &= (\alpha_{k,l,j})_{j=1,\ldots,A_l} \\ \alpha_{.,l,.} &= (\alpha_{k,l,j})_{k=1,\ldots,K;\ j=1,\ldots,A_l} \\ \beta_{l,.} &= (\beta_{l,j})_{j=1,\ldots,A_l}\,. \end{aligned}$$

The number K of clusters, the subset S of clustering loci, the mixing proportions $\pi=(\pi_k)_{k=1,\ldots,\,K}$, the allelic frequencies $\alpha=(\alpha_{k,l,j})_{k=1,\ldots,K;\ l\in S;\ j=1,\ldots,A_l}$ 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 inferred. The assignment z_i of the 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

$$\Theta_{(K,\,S)} := \mathbb{S}_{K-1}\times\left[\prod_{l\in S}\mathbb{S}_{A_l-1}\right]^K\times\prod_{l\in S^c}\mathbb{S}_{A_l-1}, \tag{3.5}$$

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

$$\mathcal{M}_{(K,\,S)}=\left\{P_{(K,\,S)}\left(\,\cdot\,|\,\theta_{(K,\,S)}\right);\ \theta_{(K,\,S)}\in\Theta_{(K,\,S)}\right\}. \tag{3.6}$$

Each model $\mathcal{M}_{(K,\,S)}$ 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 $\mathcal{C}=\left(\mathcal{M}_{(K,\,S)}\right)_{(K,\,S)\in\mathbb{N}^*\times\mathcal{P}^*(L)}$ 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 $\mathcal{M}_{(K,\,S)}$ given by

$$D_{(K,\,S)}=K-1+K\sum_{l\in S}\left(A_l-1\right)+\sum_{l\notin S}\left(A_l-1\right). \tag{3.7}$$

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

3.2.2 Model selection principle

We have at hand a collection $\mathcal{C} = \left(\mathcal{M}_{(K,S)}\right)_{(K,S)\in\mathbb{N}^*\times\mathcal{P}^*(L)}$ of competing models. A classical idea in model selection is to choose a model indexed by $\left(\widehat{K}_n,\ \widehat{S}_n\right)$ that minimizes a penalized criterion of the form

$$\mathbf{crit}(K,S) := \gamma_n\left(\widehat{P}_{(K,S)}\right) + \mathbf{pen}_n(K,S), \tag{3.8}$$

where

$$\begin{array}{rccc} \mathbf{pen}_n: & \mathbb{N}^*\times\mathcal{P}^*(L) & \longrightarrow & \mathbb{R}_+ \\ & (K,S) & \longmapsto & \mathbf{pen}_n(K,S) \end{array} \tag{3.9}$$

is the penalty function,

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

the log-likelihood contrast defined for any probability distribution P on $\mathbb{X}$ (see equation (3.1)), and $\widehat{P}_{(K,S)} := P_{\left(\widehat{K}_n,\ \widehat{S}_n\right)}\left(\cdot \mid \widehat{\theta}_{MLE,\left(\widehat{K}_n,\ \widehat{S}_n\right)}\right)$ the maximum likelihood over the model $\mathcal{M}_{(K,S)}$: that is the probability distribution of X that is obtained for the maximum likelihood parameter estimate $\widehat{\theta}_{MLE,\left(\widehat{K}_n,\ \widehat{S}_n\right)}$. Thus this parameter estimate minimizes $\gamma_n(P)$ with respect to $P\in\mathcal{M}_{(K,S)}$. There exists a huge literature on model selection via penalized criteria (see Massart (2007) and the references therein for an overview).

We can then define the selected model $\mathcal{M}_{\left(\widehat{K}_n,\ \widehat{S}_n\right)}$ and the associated selected distribution estimator $\widehat{P}_{\left(\widehat{K}_n,\ \widehat{S}_n\right)}$, and 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

$$\widehat{z}_i = \arg\max_{k\in\left\{1,\ldots,\ \widehat{K}_n\right\}}\widehat{\pi}_k P\left(x_i|\ z_i=k,\ \widehat{\theta}_{MLE,\left(\widehat{K}_n,\ \widehat{S}_n\right)}\right). \tag{3.10}$$

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) (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 in S^c.

As shown below in subsection 3.3.1, assuming that the true density P_0 belongs to one of the competing models implies that there exists a "smallest" model $\mathcal{M}_{(K_0,\ S_0)}$ containing P_0. Thus, it makes sense to consider penalties $\mathbf{pen}_n$ 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.

3.2.3 Selection procedure

The space of competing models can be very large, consisting of all combinations of all $\left(2^L - 1\right)$ nonempty subsets of the available loci with each possible number of populations. Thus an exhaustive search of an optimum model is very painful in most situations. A two nested-step algorithm combined with a Backward-Stepwise algorithm is proposed in Maugis et al. (2009) in a Gaussian framework. This algorithm makes use of an exclusion and an inclusion step. Starting from the exclusion with all the variables selected, Backward-Stepwise algorithm takes into account the possible interaction between variables. The algorithm proposed in 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 a modified Backward-Stepwise named "Backward-Stepwise Explorer" 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 (2) below). The optimum model is then chosen between all the explored models.

In addition, if the model is identifiable up to label switching, then the number of free parameters of the mixture part (given S) is at most equal to the number of free probabilities of an arbitrary, unconstrained probability distribution on the set $\mathbb{X}^S := \left\{\left(x^l\right)_{l\in S} : \ x \in \mathbb{X}\right\}$ where $\mathbb{X}$ is given by equation (3.1) :

$$K - 1 + K\sum\nolimits_{l\in S}\left(A_l - 1\right) \leq \prod_{l\in S}\left(\frac{A_l\left(A_l+1\right)}{2}\right) - 1. \tag{3.11}$$

Even if this condition is not sufficient, it gives an upper bound on $K_{\max} = \max_S K\left(S\right)$ of the number of populations where $K\left(S\right)$ is the smallest integer exceeding

$$\frac{\prod_{l\in S}\frac{A_l(A_l+1)}{2}}{1+\sum_{l\in S}\left(A_l-1\right)}. \tag{3.12}$$

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

$$\mathcal{C}_{K_{\max}} := \left(\mathcal{M}_{(K,\ S)}\right)_{K=1,\ldots,\ K_{\max};S\in\mathcal{P}^*(L)}. \tag{3.13}$$

The two nested-step algorithm for optimizing $(K,\ S)$ is stated as follows.

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

$$\widehat{S}_n\left(K\right) = \arg\min_{S\in\mathcal{P}^*(L)} \mathbf{crit}\left(K,\ S\right) \tag{3.14}$$

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

$$\widehat{K}_n = \arg\min_{K\in\{1,\ldots,\ K_{\max}\}} \mathbf{crit}\left(K,\ \widehat{S}_n\left(K\right)\right). \tag{3.15}$$

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 (2) 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 (3.8). The proposed candidate locus c_{ex} for exclusion from the currently selected clustering loci S is chosen to be the one whose exclusion yields the maximum decrease in the penalized clustering criterion (3.8) and insofar attains the best model among all possible sub-models with one locus less than S. 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 unselected loci which shows most evidence of multivariate clustering including the previous selected loci.

Algorithm 2 Backward-Stepwise Explorer(**crit**, K)

1: $S \leftarrow \{1, \ldots, L\}$, $c_{ex} \leftarrow 0$, $c_{in} \leftarrow 0$;
2: **repeat**
3: *EXCLUSION* (K, S);
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 $c_{in} = 0$ **then**
6: $S \leftarrow S \setminus \{c_{ex}\}$
7: **else**
8: $c_{ex} \leftarrow 0$;
9: **end if**
10: *INCLUSION* (K, S)
11: $c_{in} \leftarrow \arg\min_{l \notin S}$ **crit** $(K, S \cup \{l\})$;
12: **if** $\Big($**crit** $(K, S \cup \{c_{in}\})$ − **crit** $(K, S) < 0$ and $S \cup \{c_{in}\}$ has never been the current set in an EXCLUSION step$\Big)$ **then**
13: $S \leftarrow S \cup \{c_{in}\}$
14: **else**
15: $c_{in} \leftarrow 0$;
16: **end if**
17: **until** $|S| = 1$

3.3 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 3.3.1. Identifiability of the models

1. What we call "reference model" is any model (K, S) in line 4 in Algorithm (2).

$\mathcal{M}_{(K,\ S)}$ is discussed in subsection 3.3.2 using a result obtained by Allman et al. (2008), and the main consistency result is given in subsection 3.3.3.

3.3.1 The "smallest" model $\mathcal{M}_{(K_0,\ S_0)}$

Let $\mathcal{M} = \bigcup_{(K,\ S)} \mathcal{M}_{(K,\ S)}$ 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 $\mathcal{M}$. By Lemma 3.3.1 stated hereafter there is always more than one model $\mathcal{M}_{(K,\ S)}$ such that $P_0 \in \mathcal{M}_{(K,\ S)}$. But thanks to the Lemma 3.3.2 below, there exists a "smallest" model $\mathcal{M}_{(K_0,\ S_0)}$ containing the true density P_0. This "smallest" model is defined by $(K_0,\ S_0) := (K(P_0),\ S(P_0))$, where

$$K(P) = \min\left\{K |\ P \in \bigcup_{S \in \mathcal{P}^*(L)} \mathcal{M}_{(K,\ S)}\right\}, \tag{3.16}$$

$$S(P) = \min\left\{S |\ P \in \bigcup_{K \in \mathbb{N}^*} \mathcal{M}_{(K,\ S)}\right\}, \tag{3.17}$$

for every P in one of the competing models $\mathcal{M}_{(K,\ S)}$. In (3.17), min is in the sense of the partial order defined by the inclusion of sets. Consequently, we will refer to $\mathcal{M}_{(K_0,\ S_0)}$ as our uniquely defined "true" model.

Lemme 3.3.1. *For every $K_1,\ K_2$ in $\mathbb{N}^*$ and $S_1,\ S_2$ in $\mathcal{P}^*(L)$, if $K_1 \leq K_2$ and $S_1 \subseteq S_2$, then $\mathcal{M}_{(K_1,\ S_1)} \subseteq \mathcal{M}_{(K_2,\ S_2)}$.*

Lemme 3.3.2. *For every $K_1,\ K_2$ in $\mathbb{N}^*$ and $S_1,\ S_2$ in $\mathcal{P}^*(L)$, one has*

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

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

The proofs of Lemmas 3.3.1 and 3.3.2 are given in Appendix 3.A and 3.B.

3.3.2 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 parametrization 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 multinomial finite mixture model $\mathcal{M}_{(K,\ S)}$ for $X = \left(X^l\right)_{l\in S}$ as the K-class, $|S|$-feature model, with state space $\prod_{l\in S}\{1,\ldots,\ G_l\}$, and denote it by $\mathbb{M}\left(K\ ;\ (G_l)_{l\in S}\right)$, 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. Allman et al. (2008) have proved that finite mixtures of multinomial distributions are *generically* identifiable. In the case of parametric setting, 'generic' means that the set of parameter values for which identifiability does not hold has Lebesgue measure zero. Here is the result of Allman et al. (2008) in our setting.

Théorème 3.3.1. *Consider a 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 nonempty subsets* S_1, S_2 *and* S_3, *such that*

$$\min\left(K,\ \mathcal{G}_1\right) + \min\left(K,\ \mathcal{G}_2\right) + \min\left(K,\ \mathcal{G}_3\right) \geq 2\cdot K + 2, \tag{3.18}$$

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

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 Polymorphism (SNP).

3.3.3 The main result

In this subsection, we assume that the true probability distribution P_0 is in one of the competing models in a sub-collection $\mathcal{C}_{K_{\max}}$ associated to a maximum number $K_{\max}$ of populations provided by the user. And we prove that the probability of selecting the "smallest" model $(K_0,\ S_0)$ (see subsection 3.3.1) tends to 1 as n tends to infinity. We consider penalty functions of the following form

$$\begin{array}{rccl} \mathbf{pen}_n: & \{1;\ldots;K_{max}\}\times\mathcal{P}^*(L) & \longrightarrow & \mathbb{R}_+ \\ & (K,S) & \longmapsto & \mathbf{pen}_n(K,S) = \dfrac{1}{n}\,\mathbf{pen}\left(D_{(K,S)},\ n\right), \end{array} \tag{3.19}$$

where $\mathbf{pen}\left(D,\ n\right)$ is a function with the following properties :

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

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

We need the following weak assumption :

$$(H)\ :\ \forall x\in\mathbb{X},\ P_0(x) > 0, \tag{3.20}$$

where $\mathbb{X}$ is the set of distinct genotypes defined by the observed allele states (see equation 3.1), and P_0 the true probability distribution of the observations. Assumption (H) is reasonable since only observed alleles are considered in practice.

Théorème 3.3.2. *Assume that P_0 is in one of the competing models and fulfills assumption (H). Let $\left(\widehat{K}_n,\ \widehat{S}_n\right)$ be the minimizer of a penalized criterion* $\mathbf{crit}\,(K,\ S)$ *of the form given in equation (3.8). If the penalty function* $\mathbf{pen}_n\,(K,S)$ *has the form given in equation (3.19) and fulfills $(P1)$ and $(P2)$, then*

$$\lim_{n\to\infty} P_0\left[\left(\widehat{K}_n,\ \widehat{S}_n\right) = (K_0,\ S_0)\right] = 1, \tag{3.21}$$

where $(K_0,\ S_0) = (K\,(P_0)\,,\ S\,(P_0))$ (see equations (3.16) and (3.17)).

The proof of this theorem is given in Appendix 3.C.

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

$$BIC\,(K,\ S) := -\frac{1}{n}\sum_{i=1}^{n} \ln P_{(K,\ S)}\left(x_i|\ \widehat{\theta}_{MLE,(K,S)}\right) + \frac{D_{(K,\ S)}\ln n}{2n}, \tag{3.22}$$

where $\mathbf{pen}\,(D,\ n) = D\frac{\ln n}{2}$ and $\widehat{\theta}_{MLE,(K,S)} = \arg\max_{\theta\in\Theta_{(K,\ S)}} \sum_{i=1}^{n} \ln P_{(K,\ S)}\,(x_i|\ \theta)$. Thus the following corollary is a direct consequence of theorem 3.3.2.

Corollaire 3.3.1. *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

$$\lim_{n\to\infty} P_0\left[\left(\widehat{K}_n,\ \widehat{S}_n\right) = (K_0,\ S_0)\right] \underset{n\to\infty}{=} 1. \tag{3.23}$$

Note that these theoretical results are also valid for the variable selection problem in clustering with multinomial mixture models. We also found that the consistency of the BIC holds empirically (see Subsection 3.4.1).

3.4 Numerical experiments

Our proposed method named `MixMoGenD` (*Mixture Model for Genotype Data*) has been implemented using $C++$ programming language. This program is available on `www.math.u-psud.fr/~toussile`. In this section, we conduct numerical experiments on simulated and real datasets that illustrate the behavior of `MixMoGenD` and highlight the benefits of its loci selection procedure. The results obtained on simulated datasets are reported in subsection 3.4.1. In subsection 3.4.2, the real dataset used in Rosenberg et al. (2001) is considered [1]. Since some of the competing models are nested, we consider the

1. Dataset available on `http://rosenberglab.bioinformatics.med.umich.edu/jewishAut.html`

TABLE 3.1 – 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

BIC for both simulated and real experiments as recommended by Wang and Liu (2006).

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 on 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 decided for the strategy of short runs of EM from random positions followed by a long run of EM from the solution maximizing the observed log-likelihood (see Biernacki et al. (2001)).

3.4.1 Simulation examples

First examples The goal in the first examples of simulated datasets is to see how the increase of the sample size improves the capacity of `MixMoGenD` to select the "smallest" model $\mathcal{M}_{(K_0,\ S_0)}$. 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, and $|S_0| = 2$ clustering loci. For each value n of the sample size, 100 datasets are generated using the parameters given in Table 3.1. As seen on Figure 3.1, `MixMoGenD` consistently identify the true model as $n \to \infty$. Other simulated datasets with $K_0 = 3$ clusters, $L = 6$ loci and $|S_0| = 4$ clustering loci confirmed these results. Thus, the theoretical result on the consistency that we showed in Section 3.3 holds empirically.

Second examples The aim in these examples is to compare the inference on the number of clusters when the variable selection is not included with when it is included. We independently generate 100 datasets each with 1 000 individuals. We choose $K_0 = 3$ populations, $L = 6$ loci and $|S_0| = 4$ clustering loci. Simulation parameters are given in Table 3.2. 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** with $\widehat{K}_n = 2$ clusters from $\widehat{S}_n = S_0$. Empirically, it appears that the number of populations can be underestimated when

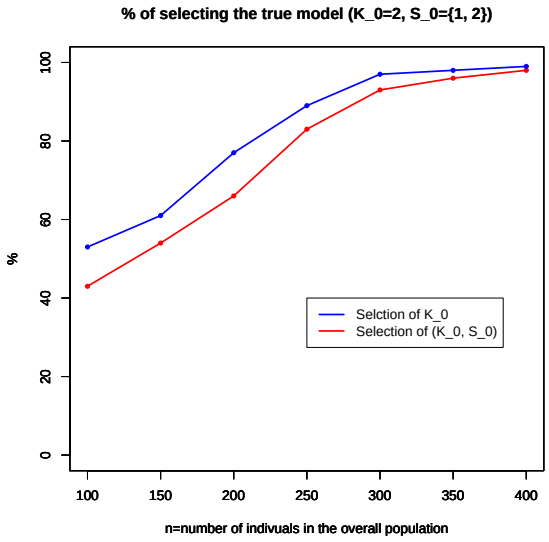


FIGURE 3.1 – % of selecting the true number K_0 of clusters and true model $(K_0,\ S_0)$ vs the sample size.

considering all available loci as relevant for clustering.

Third examples In the third examples, we assume more variability in the simulated datasets. Here, each of the simulated datasets consists in 1 000 individuals structured into 5 subpopulations with equal proportions. We assume $L = 10$ loci each with 10 alleles, and four different cardinalities for S_0 : 8, 6, 4 and 2. We assume the uniform distribution for the alleles of the loci in S_0^c. 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 genetics that unsupervised clustering is difficult with such a range of F_{ST} (Latch et al., 2006). Some of these simulated datasets and their simulation parameters are available on the following address `http://www.math.u-psud.fr/~toussile`. We used $K_{\max} = 10$ for the analysis of all these simulations.

On these simulated samples, `MixMoGenD` provides three main conclusions (see Tables 3.3, 3.4, 3.5 and 3.6). 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 measured by the percentage of wrongly assigned individuals (% WA). We determine % WA only in cases where the estimate $\widehat{K}_n$ of the number of clusters is equal to the true number K_0. We observe that 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

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

TABLE 3.2 – 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

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 3.6). 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* Guillot et al. (2005) of **R** program package R Development team (2009). 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 3.3. The square matrix of pairwise F_{ST} is given in Table 3.7. 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.

3.4.2 Real dataset example

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

TABLE 3.3 – 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]. % WA and % WA^s = percentage of wrongly assigned 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}	$\widehat{K}_n$	% WA	$\widehat{K}_n^s$	% WA^s	Data	F_ST	$\widehat{K}_n$	% WA	$\widehat{K}_n^s$	% WA^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	-	5	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	-	5	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

`MixMoGenD` revealed a cluster that was almost identical to the sample of *Libyan Jews* (Table 3.8 (a)). From 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 Rosenberg et al. (2001) using STRUCTURE, indicating that the Libyan Jewish appellation 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 felled in cluster 2. Subclustering analysis showed that the sample without Libyan Jews could be divided in 2 clusters with the subset of clustering loci containing only one locus which is the tetranucleotide $D1S1679$ (Table 3.8 (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.

3.5 Discussion

Theoretical results concerning the behavior of the BIC and other penalization methods in a mixture framework are few. As far as we know, there is no consistency result for both variable selection and clustering problem in multinomial setting. Under weak assumptions on the penalty function, we have proved that the probability to select the "smallest" model tends to 1 as the sample size tends to infinity. In numerical experiments, we have used the BIC. It is well known that this criterion is not uniformly the

TABLE 3.4 – 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]. % WA and % WA^s = percentage of wrongly assigned 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}	$\widehat{K}_n$	% WA	$\widehat{K}_n^s$	% WA^s	Data	F_{ST}	$\widehat{K}_n$	% WA	$\widehat{K}_n^s$	% WA^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

best one. We currently work on a data dependent calibration of the penalty function, so that the method does not require ad-hoc choice of penalty parameters, and adapts automatically to the data. We also work on oracle inequalities in order to obtain non asymptotic bounds for the risk of the estimated model.

In a practical point of view, we propose a modified Backward-Stepwise algorithm that we named *Backward-Stepwise Explorer* (see 2), which does not only avoid an exhaustive search of the optimum model, but also enables the search for the optimum subset of clustering loci among all possible cardinalities. In fact, due to the explosion of genomic projects, datasets are becoming increasingly large. The space of the models under competition can then be very large so that an exhaustive research of an optimum model is very painful in most situations. In the other hand, we notice that the classical Backward-Stepwise algorithm could miss the optimum model in some cases, in particular in cases where the optimum subset of clustering loci is small.

In addition, 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 (François et al., 2006). Second, our method is combined with a loci selection which is its key point. 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 a range where it is thought that clustering is difficult (Latch et al., 2006), `MixMoGenD` perfectly identified the subset of clustering variables.

TABLE 3.5 – 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 $|S_0| = 4$ clustering loci. The datasets are simulated so that the F_{ST} are in [0.0302, 0.0413]. % WA and % WA^s = percentage of wrongly assigned 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}	$\widehat{K}_n$	% WA	$\widehat{K}_n^s$	% WA^s	Data	F_{ST}	$\widehat{K}_n$	% WA	$\widehat{K}_n^s$	% WA^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

acknowledgements

This work was supported by a doctoral fellowship from "Institut de Recherche pour le Développement" (IRD). We 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 3.6 – 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 $|S_0| = 2$ clustering loci. The datasets are simulated so that the F_{ST} are in [0.0181, 0.0266]. % WA and % WA^s = percentage of wrongly assigned 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}	$\widehat{K}_n$	% WA	$\widehat{K}_n^s$	% WA^s	Data	F_{ST}	$\widehat{K}_n$	% WA	$\widehat{K}_n^s$	% WA^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

TABLE 3.7 – 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

TABLE 3.8 – 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* appellation 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

Appendices

3.A Proof of lemma 3.3.1

First, we prove that any P in $\mathcal{M}_{(K,\ S)}$ is also in $\mathcal{M}_{(K+1,\ S)}$. Let $P \in \mathcal{M}_{(K,\ S)}$ 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 underlying the observations). Define for instance $\theta' = (\pi',\ \alpha',\ \beta') \in \Theta_{(K+1,\ S)}$ as follows

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

Obviously, One has $P(\cdot) = P_{(K+1,\ S)}(\ \cdot \mid \theta')$. So P is an element of model $\mathcal{M}_{(K+1,\ S)}$. We have just showed that $\mathcal{M}_{(K,\ S)} \subseteq \mathcal{M}_{(K+1,\ S)}$. It remains to show that $\mathcal{M}_{(K,\ S)} \subseteq \mathcal{M}_{(K,\ S')}$ 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 :

$$\alpha_{1,l,\ .} = \ldots = \alpha_{K,l,\ .}\ \forall l \in S' \setminus S. \tag{3.24}$$

3.B Proof of lemma 3.3.2

Let P be a probability distribution in $\mathcal{M}_{(K_1,\ S_1)} \cap \mathcal{M}_{(K_2,\ S_2)}$ and $\mathbb{X}$ given by equation (3.1). Then for every x in $\mathbb{X}$, $P(x)$ is given by the following two equations.

$$P(x) = \left[\sum_{k=1}^{K_1} \pi_k^1 \prod_{l \in S_1} P\left(x^l \mid \left(\alpha^1_{k,\ l,\ .}\right)\right)\right] \times \prod_{l \in S_1^c} P\left(x^l \mid \left(\beta^1_{l,\ .}\right)\right), \tag{3.25}$$

$$P(x) = \left[\sum_{k=1}^{K_2} \pi_k^2 \prod_{l \in S_2} P\left(x^l \mid \left(\alpha^2_{k,\ l,\ .}\right)\right)\right] \times \prod_{l \in S_2^c} P\left(x^l \mid \left(\beta^2_{l,\ .}\right)\right), \tag{3.26}$$

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 $C = \{1, \ldots,\ L\} \setminus S_1 \cup S_2$. Using equation (3.25), the marginal probability distribution of the subvector $x^{S_2} := \left(x^l\right)_{l \in S_2}$ is given by

$$P\left(x^{S_2}\right) = \left[\sum_{k=1}^{K_1} \pi_k^1 \prod_{l \in S_1 \cap S_2} P\left(x^l \mid \left(\alpha^1_{k,\ l,\ .}\right)\right)\right] \times \prod_{l \in B} P\left(x^l \mid \left(\beta^1_{l,\ .}\right)\right), \tag{3.27}$$

and using equation (3.26) one has

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

where β^3 is defined as follows

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

Consequently, P is an element of model $\mathcal{M}_{(K_1 \wedge K_2,\ S_1 \cap S_2)}$. We have just proved that $\mathcal{M}_{(K_1,\ S_1)} \cap \mathcal{M}_{(K_2,\ S_2)} \subseteq \mathcal{M}_{(K_1 \wedge K_2,\ S_1 \cap S_2)}$. In addition, by Lemma 3.3.1 one has $\mathcal{M}_{(K_1 \wedge K_2,\ S_1 \cap S_2)} \subseteq \mathcal{M}_{(K_1,\ S_1)} \cap \mathcal{M}_{(K_2,\ S_2)}$. We then have the desired result.

3.C Proof of Theorem 3.3.2

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

$$\Theta_{(K,\ S)}^{\delta} = \left\{\theta \in \Theta_{(K,\ S)} \ :\ \forall x \in \mathbb{X},\ P_{(K,\ S)}(x|\ \theta) \geq \delta\right\}, \tag{3.28}$$

where $\mathbb{X}$ is given by equation (3.1). We shall need the following theorem whose proof is given below in Appendix 3.D.

Théorème 3.C.1. *Under assumption (H), there exists a real $\delta > 0$ such that for every $(K,\ S)$, one has*

$$-\gamma_n\left(\widehat{P}_{(K,\ S)}\right) = \sup_{\theta \in \Theta_{(K,\ S)}^{\delta}} \left\{-\gamma_n\left(P_{(K,\ S)}(\ \cdot\ |\ \theta)\right)\right\} + o_{P_0}(1) \tag{3.29}$$

and

$$\sup_{\theta \in \Theta_{(K,\ S)}} E_{P_0}\left[\ln P_{(K,\ S)}(X|\ \theta)\right] = \sup_{\theta \in \Theta_{(K,\ S)}^{\delta}} E_{P_0}\left[\ln P_{(K,\ S)}(X|\ \theta)\right]. \tag{3.30}$$

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 $(K, S) \in \{1, \ldots,\ K_{\max}\} \times \mathcal{P}^*(L)$ 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 $\{1,\ldots,\ K_{\max}\}\times\mathcal{P}^{*}(L)$ 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

$$P_0\left[\mathbf{crit}\left(K,\ S\right)<\mathbf{crit}\left(K_0,\ S_0\right)\right]=P_0\Bigg[\gamma_n\left(\widehat{P}_{(K_0,\ S_0)}\right)-\gamma_n\left(\widehat{P}_{(K,\ S)}\right)>\mathbf{pen}_n\left(K,\ S\right)-\mathbf{pen}_n\left(K_0,\ S_0\right)\Bigg],\tag{3.31}$$

where $\gamma_n\left(P\right):=-\frac{1}{n}\sum_{i=1}^{n}\ln P\left(x_i\right)$ and $\widehat{P}_{(K,S)}$ is the maximum likelihood estimator (MLE) in $\mathcal{M}_{(K,\ S)}$. Recall that $x_1,\ldots,\ x_i,\ldots,\ x_n$ are the observations. Two cases are considered : $P_0\in\mathcal{M}_{(K,\ S)}$ and $P_0\notin\mathcal{M}_{(K,\ S)}$.

• **Case 1 :** $P_0\in\mathcal{M}_{(K,\ S)}$, 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 by $\mathcal{D}$ the set of all possible probability distributions on the set $\mathbb{X}$ of the genotype states. Since $\mathcal{M}_{(K_0,\ S_0)}\subseteq\mathcal{M}_{(K,\ S)}\subseteq\mathcal{D}$, one has the following inequalities

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

so that

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

But it is well known that $2\sup_{P\in\mathcal{M}}\left(-n\gamma_n\left(P\right)\right)+2n\gamma_n\left(P_0\right)$ converges in distribution to a χ^2 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 $\mathcal{M}_{(K,\ S)}$ and $(K,S)\neq(K_0,S_0)$, one has $\mathcal{M}_{(K_0,\ S_0)}\subsetneq\mathcal{M}_{(K,\ S)}$, and it follows from $(P2)$ that $\mathbf{pen}\left(D_{(K,\ S)},\ n\right)-\mathbf{pen}\left(D_{(K_0,\ S_0)},\ n\right)$ tends to infinity as n tends to infinity. Thus

$$\lim_{n\to\infty}P_0\left[n\gamma_n\left(\widehat{P}_{(K_0,\ S_0)}\right)-n\gamma_n\left(\widehat{P}_{(K,\ S)}\right)>\mathbf{pen}\left(D_{(K,\ S)},\ n\right)-\mathbf{pen}\left(D_{(K_0,\ S_0)},\ n\right)\right]=0$$

• **Case 2 :** $P_0\notin\mathcal{M}_{(K,\ S)}$, i.e for all θ in $\Theta_{(K,\ S)}$, one has $P_0\neq P_{(K,\ S)}\left(\cdot|\ \theta\right)$. By equation (3.29) of Theorem 3.C.1, there exists a positive real δ such that

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

The set of functions $\mathcal{F}^{\delta}_{(K,\ S)}:=\left\{\ln P_{(K,\ S)}\left(\ \cdot\ |\ \theta\right),\theta\in\Theta^{\delta}_{(K,\ S)}\right\}$ is obviously P_0-Glivenko-Cantelli (see Appendix 3.E below), so that

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

On the other hand, since P_0 is an element of $\mathcal{M}_{(K_0,\ K_0)}$, it is obvious 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

$$\begin{aligned}\sup_{\theta\in\Theta^{\delta}_{(K_0,\ S_0)}} E_{P_0}\left[\ln P_{(K_0,\ S_0)}(X|\ \theta)\right] &= \sup_{\theta\in\Theta_{(K_0,\ S_0)}} E_{P_0}\left[\ln P_{(K_0,\ S_0)}(X|\ \theta)\right] \\ &\quad \text{(see equation (3.30) of Theorem 3.C.1)} \\ &= E_{P_0}\left[\ln P_0(X)\right].\end{aligned}$$

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 $\mathcal{M}_{(K,\ S)}$. 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 $(P1)$, $\mathbf{pen}_n(K,\ S) - \mathbf{pen}_n(K_0,\ S_0)$ tends to 0 as n tends to infinity. Then one has

$$\lim_{n\to\infty} P_0\left[\gamma_n\left(\widehat{P}_{(K_0,\ S_0)}\right) - \gamma_n\left(\widehat{P}_{(K,\ S)}\right) > \mathbf{pen}_n(K,\ S) - \mathbf{pen}_n(K_0,\ S_0)\right] = 0,$$

which is the desired result.

3.D Proof of Theorem 3.C.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

$$\begin{aligned}-\gamma_n\left(P_{(K,\ S)}(\ \cdot\ |\ \theta)\right) &= \sum_{x\in\mathbb{X}} \frac{n_x}{n} \ln P_{(K,\ S)}(x|\ \theta) \\ &= \sum_{x\in\mathbb{X}} \left[P_0(x) + o_{P_0}(1)\right] \times \ln P_{(K,\ S)}(x|\ \theta). \qquad (3.32)\end{aligned}$$

For any $(K,\ S)$, 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 (3.32), one has the following inequality

$$-\gamma_n\left(P_{(K,\ S)}\left(\ \cdot\ |\ \widetilde{\theta}\right)\right) \geq \sum_{x\in\mathbb{X}} P_0\left(x\right)\ln\widetilde{\delta} + o_{P_0}\left(1\right) = \ln\widetilde{\delta} + o_{P_0}\left(1\right). \tag{3.33}$$

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

$$\sup_{\theta\in\Theta^{\delta}_{(K,\ S)}}\left\{-\gamma_n\left(P_{(K,\ S)}\left(\ \cdot\ |\ \theta\right)\right)\right\} \geq \sup_{\theta\in\Theta^{\widetilde{\delta}}_{(K,\ S)}}\left\{-\gamma_n\left(P_{(K,\ S)}\left(\ \cdot\ |\ \theta\right)\right)\right\}, \tag{3.34}$$

$$\sup_{\theta\in\Theta^{\delta}_{(K,\ S)}} E_{P_0}\left[\ln P_{(K,\ S)}\left(X|\ \theta\right)\right] \geq \sup_{\theta\in\Theta^{\widetilde{\delta}}_{(K,\ S)}} E_{P_0}\left[\ln P_{(K,\ S)}\left(X|\ \theta\right)\right]. \tag{3.35}$$

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

$$\begin{aligned}
-\gamma_n\left(P_{(K,\ S)}\left(\ \cdot\ |\ \theta\right)\right) &\leq \inf_{\mathbb{X}} P_0\left(x\right)\ln\delta + o_{P_0}\left(1\right)\\
&\leq \inf_{\mathbb{X}} P_0\left(u\right)\ln\widetilde{\delta}^{\frac{1}{\inf_{\mathbb{X}} P_0(x)}} + o_{P_0}\left(1\right) \ = \ln\widetilde{\delta} + o_{P_0}\left(1\right)\\
&\leq -\gamma_n\left(P_{(K,\ S)}\left(\ \cdot\ |\ \widetilde{\theta}\right)\right) + o_{P_0}\left(1\right)\\
&\leq \sup_{\theta\in\Theta^{\widetilde{\delta}}_{(K,\ S)}}\left\{-\gamma_n\left(P_{(K,\ S)}\left(\ \cdot\ |\ \theta\right)\right)\right\} + o_{P_0}\left(1\right)\\
&\leq \sup_{\theta\in\Theta^{\delta}_{(K,\ S)}}\left\{-\gamma_n\left(P_{(K,\ S)}\left(\ \cdot\ |\ \theta\right)\right)\right\} + o_{P_0}\left(1\right).
\end{aligned} \tag{3.36}$$

Consequently one has

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

so that

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

Using the same arguments, one gets

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

which are the desired results.

3.E Bracketing entropy and Glivenko-Cantelli Property

Let us recall the notions of Glivenko-Cantelli and entropy with bracketing of a family of functions with respect to a probability distribution P over some space Ω.

Définition 3.E.1. *A family $\mathcal{F}$ of measurable functions $f : \Omega \to \mathbb{R}$ is P-Glivenko-Cantelli if and only if :*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbf{E}[f(X)] \right\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbf{E}[f(X)] \right| \xrightarrow{a.s.} 0, \tag{3.37}$$

where $X_1, \ldots, X_n$ are independent and identically distributed as X with probability distribution P.

A sufficient condition for a family $\mathcal{F}$ to be P-Glivenko-Cantelli is proved in Vaart. (1998). This condition is based on the complexity of $\mathcal{F}$ measured by entropy with bracketing defined as follows.

Définition 3.E.2. *Let $p \in \mathbb{N}^*$ and $l,\ u \in L_p(P)$.*

The bracket $[l,\ u]$ is the set of all functions $f \in L_p(P)$ fulfilling $l \leq f \leq u$ (i.e. $\forall x \in \Omega,\ l(x) \leq f(x) \leq u(x)$).

A bracket $[l,\ u]$ is an ε-bracket if $\|l - u\|_p = (\mathbf{E}[|l-u|^p])^{\frac{1}{p}} \leq \varepsilon$.

The bracketing number $\mathcal{N}_{[]}(\varepsilon,\ \mathcal{F},\ L_p(P))$ is the minimum number of ε-brackets needed to covert $\mathcal{F}$.

The entropy with bracketing $\mathcal{H}_{[]}(\varepsilon,\ \mathcal{F},\ L_p(P))$ of $\mathcal{F}$ is the logarithm of the bracketing number.

The bracketing entropy is a L_p-measure of the complexity of a family $\mathcal{F}$ of measurable functions. It is quite natural to expect a family $\mathcal{F}$ that is not too complex to be P-Glivenko-Cantelli. Here is the sufficient condition for a family to be P-Glivenko-Cantelli.

Théorème 3.E.1 (Vaart. (1998))**.** *Every class $\mathcal{F}$ of measurable functions such that $\mathcal{H}_{[]}(\varepsilon,\ \mathcal{F},\ L_p(P)) < \infty$ for any $\varepsilon > 0$ is P-Glivenko-Cantelli.*

3.F EM equations

We have consider a model selection procedure based on penalized maximum likelihood criterion to solve our two-fold problem of loci selection and classification. Recall that we are in unsupervised classification settings. More precisely, the population of origin of the sample we deal with is missing. For the estimation of maximum likelihood in such a situation, Expectation and Maximization (EM) algorithm is widely used. It consists of iteratively maximizing the conditional expectation of the log-likelihood of the complete data, given the observations x and a current parameter $\theta^{(r)}$. Recall that we deal with a

n-sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ that we denote by $\mathbf{x}$. For a given density $P_{K,S,\theta}$ in a model $\mathcal{M}_{K,S}$, the log-likelihood is given by

$$\mathcal{L}_n(\theta;\ \mathbf{x}) = \sum_{i=1}^{n} \ln \left\{ \sum_{k=1}^{K} \pi_k P_{k,\alpha}\left(\mathbf{x}_i^S\right) \right\} + \sum_{i=1}^{n} \ln \left\{ P_\beta\left(\mathbf{x}_i^{S^c}\right) \right\},$$

where

- $\mathbf{x}_i^A = \left(x_i^l\right)_{l \in A}$ for a give subset A of the set $\{1, \ldots, L\}$ the considered loci ;
- π_k is the probability that an individual come from population k ;
- $P_{k,\alpha}$ is the density of the random vector $\mathbf{x}_i^S$ in population k, with α the allelic frequencies of the loci in S, in different populations ;
- P_β is the density of the random vector $\mathbf{x}_i^{S^c}$, with β the allelic frequencies of loci not in S, in the overall population ;
- $\theta = (\pi,\ \alpha,\ \beta)$.

Let $\gamma = (\pi,\ \alpha)$. The maximum likelihood estimator (MLE) $\widehat{\theta}_{MLE}$ of θ is obviously given by $\widehat{\theta}_{MLE} = \left(\widehat{\gamma}_{MLE},\ \widehat{\beta}_{MLE}\right)$. Since we are in multinomial settings, $\widehat{\beta}_{MLE}$ is given by the observed allelic frequencies of the loci not in S. Thus EM algorithm concerns only $\gamma = (\pi,\ \alpha)$.

The conditional expectation of the mixture part is given as follows,

$$Q\left(\gamma |\ \gamma^{(r)},\ \mathbf{x}^S\right) = \mathbf{E}_Z\left[\ln\left(P_\gamma\left(\mathbf{x}^S,\ Z\right)\right) \mid \mathbf{x}^S,\ \gamma^{(r)}\right], \tag{3.38}$$

where the completed log-likelihood is given by

$$\ln\left(P_\gamma\left(\mathbf{x}^S,\ z\right)\right) = \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{i,k} \ln\left(P_{k,\alpha}\left(\mathbf{x}_i^S\right)\right).$$

In our settings, the principle of the EM algorithm is to iteratively replace $Z_{i,k}$ by its conditional expectation for a given set $\mathbf{x}$ of observations and a current parameter $\gamma^{(r)}$. This expectation is the posterior assignment probability of individual i in population k. The algorithm starts with an initial parameter $\gamma^{(0)}$, and alternates between the two following steps. At the r^{th} iteration,

- **E step :** This step is to calculate $Q\left(\gamma |\ \gamma^{(r)},\ \mathbf{x}^S\right)$, which is to express the conditional probabilities $\tau_{ik}^{(r)}$ that individual i come from population k :

$$\tau_{ik}^{(r)} = \frac{\pi_k^{(r)} \prod_{l \in S} P_{\alpha_{k,l,\ .}^{(r)}}\left(x_i^l |\ z_i = k\right)}{\sum_{h=1}^{K} \pi_h^{(r)} \prod_{l \in S} P_{\alpha_{h,l,\ .}^{(r)}}\left(x_i^l |\ z_i = h\right)}, \tag{3.39}$$

  where $\alpha_{k,l,\ .}^{(r)}$ is the vector of allelic frequencies at locus l in population k, at iteration r.
- **M step :** This step consists of updating the parameters by estimating the parameter $\gamma^{(r+1)}$ that maximizes $Q\left(\gamma |\ \gamma^{(r)},\ \mathbf{x}^S\right)$. The update formula for the parameters can be derived using the standard method of the EM algorithm

$$\pi_k^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n} \tau_{ik}^{(r)} \tag{3.40}$$

and

$$\alpha_{k,l,j}^{(r+1)} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)} \left(\mathbb{1}_{[x_{i,1}^{l}=j]} + \mathbb{1}_{[x_{i,2}^{l}=j]} \right)}{2 \sum_{i=1}^{n} \tau_{ik}^{(r)}}. \tag{3.41}$$

Remarque 3.F.1. *The formulas (3.40) and (3.41) can be interpreted as follows. For a given current vector of parameters $\theta^{(r)}$ at iteration r,*

- $\sum_{i=1}^{n} \tau_{i,k}^{(r)}$ *is the expected size of cluster k ;*
- $\sum_{i=1}^{n} \left(\mathbb{1}_{[x_i^{l,1}=j]} + \mathbb{1}_{[x_i^{l,2}=j]} \right) \tau_{i,k}^{(r)}$ *is the expected number of copies of the j-th allele in cluster k ;*
- $\sum_{i=1}^{n} \sum_{j=1}^{A_l} \left(\mathbb{1}_{[x_i^{l,1}=j]} + \mathbb{1}_{[x_i^{l,2}=j]} \right) \tau_{i,k}^{(r)}$ *is the expected total number of copies of observed alleles in cluster k.*

Thus, $\pi_k^{(r+1)}$ and $\alpha_{k,l,j}^{(r+1)}$ are just the expected proportion of cluster k and the expected frequency of allele j of locus l in cluster k, respectively.

The growth of $Q\left(\gamma |\, \gamma^{(r)},\, \mathbf{x}^S\right)$ at each iteration of the algorithm implies that of observed log-likelihood $\mathcal{L}_n\left(\gamma;\, \mathbf{x}^S\right)$, since

$$Q\left(\gamma |\, \gamma^{(r)},\, \mathbf{x}^S\right) = \mathcal{L}_n\left(\gamma;\, \mathbf{x}^S\right) + H\left(\gamma |\, \gamma^{(r)},\, \mathbf{x}^S\right),$$

where

$$H\left(\gamma |\, \gamma^{(r)},\, \mathbf{x}^S\right) := \mathbf{E}\left[\ln\left(P_\gamma\left(Z |\, \mathbf{x}^S\right)\right) \mid \mathbf{x}^S,\, \gamma^{(r)}\right]$$

satisfies

$$H\left(\gamma |\, \gamma^{(r)},\, \mathbf{x}^S\right) \leq H\left(\gamma^{(r)} |\, \gamma^{(r)},\, \mathbf{x}^S\right),$$

from the Jensen inequality. Under certain regularity conditions, EM algorithm is known to converge slowly in some situations and its solution can highly depend on 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) (Celeux and Govaert, 1992) and SEM (Stochastic EM) (Celeux and Diebolt, 1991) have been proposed. We recommend here the strategy of short runs of EM from random positions followed by a long run of EM from the solution maximizing the observed log-likelihood (Biernacki et al., 2001).

Chapitre 4

A data-driven penalized criterion for variable selection and clustering in multivariate multinomial mixtures

This chapter presents a work in collaboration with Dominique Bontemps.

Abstract

We consider the problem of estimating the number of components and the relevant variables in a multivariate multinomial mixture. This kind of models arise in particular when dealing with multilocus genotypic data. A new penalized maximum likelihood criterion is proposed, and a non-asymptotic oracle inequality is obtained. Further, under weak assumptions on the true probability underlying the observations, the selected model is asymptotically consistent. On a practical aspect, the shape of our proposed penalty function is defined up to a multiplicative parameter which is calibrated thanks to the slope heuristics, in an automatic data-driven procedure. Using simulated data, we found that this procedure improves the performances of the selection procedure with respect to classical criteria such as **BIC** and **AIC**. The new criterion gives an answer to the question "Which criterion for which sample size ?".

4.1 Introduction

This article is concerned with the unsupervised classification on categorical multivariate data. The model-based clustering, which uses finite mixture models, is an intuitive and rigorous framework for the unsupervised classification. However there is no clear consensus on the way to gather individuals in general : on the basis of well separated clusters, or on the basis of the components of the mixture distribution ? We refer to **?** for a general discussion on this topic. Finite mixture models are specially adapted when each class is supposed to be characterized by a set of parameters, for instance in population genetics : in this case the populations that the biologists look for are characterized by their allelic frequencies and a genetic equilibrium ; this corresponds to the notion of population as a reproduction unit, or a group of individuals sharing the same genetic structure. Finite mixture models are also known in the literature as the latent class models.

The observations are n independent realizations of a random vector, whose number L of coordinates (variables) may be large. The individuals of the sample are clustered into a certain unknown number K of populations on the basis of the frequencies of apparition of the possible states of each variable. It may happen that only a subset S of the variables are relevant for clustering purposes, and the others are just noise. Thus, in addition to the number K of populations and the frequencies of the different states, we are also interested in the subset S, which may have significance in the interpretation of the results.

A number of clustering methods for categorical multivariate data have been proposed in recent years in the context of genomics (see (Pritchard et al., 2000; Chen et al., 2006; Corander et al., 2008)). But the problem of variable selection for clustering using such data was first addressed in (Toussile and Gassiat, 2009), where the question is regarded as a model selection problem in a density estimation framework. First the components of a finite mixture distribution are identified, then the individuals are clustered into these components using the Maximum A Posteriori (MAP) method.
Using simulated data, that article shows that the variable selection procedure based on the Bayesian Information Criterion (**BIC**) significantly improves clustering and prediction capacities in our framework. It also gives a theoretical consistency result : when the true density P_0 underlying the observations belongs to one of the competing models, then there exists a smallest model $\mathcal{M}_{(K_0,\, S_0)}$ containing P_0 ; further, the **BIC** type criteria select $\mathcal{M}_{(K_0,\, S_0)}$ with probability tending to one as the sample size n goes to infinity. This consistency approach requires large sample sizes which may be difficult to obtain. However the knowledge of the true model, aside the frequencies of the states, is an important information for the interpretation of the results.

In the present paper we adopt an oracle approach. We do not aim at choosing the true model underlying the data, even if our procedure performs well also for that. The criteria are rather designed to minimize some risk function of the estimated density with respect to the true density. In this context simpler models can be preferred to $\mathcal{M}_{(K_0,\, S_0)}$, in which too many parameters can entail estimators which overfit the data. Actually there is no need to assume that P_0 belongs to one of the competing models $\mathcal{M}_{(K,S)}$.

BIC relies on a strong asymptotic assumption, and can thus require large sample

sizes to reach its asymptotic behavior ; practically **BIC** is known to overpenalize, and therefore selects too small models for small or medium values of n (see (**?**)). On the contrary Akaike's Information Criterion (**AIC**) is known to underpenalize, and selects too large models for large and medium values of n. We would like a criterion which gathers the virtues of both **AIC** and **BIC**, and performs well for different values of n.

In this article, we propose a non asymptotic penalized criterion based on the metric entropy theory of Massart (in particular (Massart, 2007)). It leads to a non asymptotic oracle inequality, which compares the risk of the selected estimator to the risk of the estimator associated with the unknown best model (see Theorem 1 below). There exists a large literature on model selection via penalization from a non asymptotic perspective. This literature is still in development with the appearance of sophisticated tools of probability such as concentration and deviations inequalities (see (Massart, 2007) and the references therein). In mixture models the non asymptotic approach is very recent, the first related work being (Maugis and Michel, 2009) for the Gaussian mixture model.

However, the obtained penalty function presents drawbacks : it depends on a multiplicative constant for which sharp upper bounds are not available, and it leads in practice to an overpenalization — even worse than **BIC**. Therefore our theoretical result mainly suggests the shape of the penalty function :

$$\mathbf{pen}_n(m) = \lambda D_m/n,$$

where D_m is the dimension of model m, and λ an unknown parameter depending on the sample size and the complexity of the collection of models under competition, which has to be calibrated. A calibration of λ with the so-called slope heuristics has been proposed in (Birgé and Massart, 2007) in such a case. We propose a modified version based on a sliding window of this calibration method. The resulting criterion does not require an ad-hoc choice of the penalty parameters and adapts automatically to the data. Although the full theoretical validation of slope heuristics is provided only in the Gaussian homoscedastic and heteroscedastic regression frameworks (Birgé and Massart, 2007; Arlot and Massart, 2009), they have been implemented in several other frameworks (see (**?**Lebarbier, 2002; Verzelen, 2009; Villers, 2007) for applications in density estimation, genomics, etc.). The simulations performed in Subsection 4.4.3 illustrate that our criterion behaves well with respect to more classical criteria as **BIC** and **AIC**, both to estimate the density, even when n is relatively small, and to retrieve the true model. It can be seen as a representative of the family of the General Information Criteria (see for instance (**?**) whose criterion is less intuitive but presents some analogy with the slope heuristics).

The paper is organized as follows. Section 4.2 is devoted to the presentation of the mixture models framework and to the model selection paradigm. In Section 4.3 we state and prove our main result, the oracle inequality. Section 4.4 is devoted to the practical aspect of our procedure which has been implemented in the stand alone software `MixMoGenD` (Mixture Model using Genotypic Data) (see (Toussile and Gassiat, 2009)). Results on simulated experiments are also presented : we compare our proposed criterion to classical **BIC** and **AIC**, in both points of view of the selection of the true model and of density estimation. Eventually, the Appendices contain several technical results used in the main analysis.

4.2 Model and methods

4.2.1 Framework

We suppose we deal with independent and identically distributed (iid) realizations of a multivariate random vector $X = (X^l)_{1 \leq l \leq L}$. We consider two main settings :

1. Each X^l is a multinomial variable taking values in $\{1, \ldots, A_l\}$.
2. Each X^l consists in a (non ordered) set $\{X^{l,1}, X^{l,2}\}$ of two (that may be equal) qualitative variables taking their values in the same set $\{1, \ldots, A_l\}$.

All along this article, these two settings will be referred to as Case 1 and Case 2. In both cases, the numbers A_l of allowed states are supposed to be known, and to verify $A_l \geq 2$.

The first case is a usual latent class model with various applications (psychomatrics, marketing, credit scoring, genomics, etc.), while the last one is more specific to genotypic data. In this context $X = (X^l)_{1 \leq l \leq L}$ represents the genotype of an individual at L loci of its DNA. Case 1 corresponds to haploid organisms, with a single representative of each chromosome ; at any locus l a single allele X^l is measured. Case 2 corresponds to diploid organisms, with two representatives of each chromosome ; at any locus l, two alleles $X^{l,1}$ and $X^{l,2}$ are observed together.

We consider a model-based clustering, which means that the sample is a finite mixture of an unknown number K of populations (clusters), each being characterized by a set of frequencies of the states. Let denote by Z the (unobserved) population an individual comes from. Variable Z takes its values in the set $\{1, \ldots, K\}$ of the labels of the different clusters. Its distribution is given by the vector $\pi = (\pi_k)_{1 \leq k \leq K}$, where $\pi_k = P(Z = k)$. Conditionally to Z, the variables $X^1, \ldots, X^L$ are supposed to be independent. In Case 2, the states $X^{l,1}$ and $X^{l,2}$ for the l^{th} variable are also supposed to be independent conditionally to Z. The preceding two assumptions are what biologists respectively call *Linkage Equilibrium* (LE) and *Hardy-Weinberg Equilibrium* (HWE). According to these assumptions, the probability distribution of a genotype $x = \left(x^l\right)_{1 \leq l \leq L}$ in a population k is given in the following equations

$$\begin{aligned} P\left(x \middle| \, Z = k\right) &= \prod_{l=1}^{L} P\left(x^l | Z = k\right) \\ \text{Case 1} : P\left(x^l | Z = k\right) &= \alpha_{k,l,x^l} \\ \text{Case 2} : P\left(x^l | Z = k\right) &= \left(2 - \mathbb{1}_{x^{l,1} = x^{l,2}}\right) \alpha_{k,l,x^{l,1}} \alpha_{k,l,x^{l,2}} \end{aligned} \tag{4.1}$$

where $\alpha_{k,l,j}$ is the probability of state j associated to variable X^l in population k. The mixing proportions π_k and the probabilities $\alpha_{k,l,j}$ will be treated as parameters.

In the context of genomics, Hardy-Weinberg and linkage equilibria are based on several simplifying assumptions that can seem unrealistic ; however they have still proven to be useful in describing many population genetic attributes and serve as a base model in the development of more realistic models of microevolution. Further, the choice of estimators derived from the maximum likelihood estimator (MLE) responds to the wish of

biologists to group the sample into clusters minimizing the Hardy-Weinberg and linkage desequilibria, and this brings some robustness to our modeling (see (Latch et al., 2006) and references therein).

Going deeper, the oracle approach emphasizes that we should often prefer simplified and misspecified models. This introduces a modeling bias in order to get more robust estimators and classifiers, and at the end we get a smaller estimation error. This legitimizes also the following simplification.

It may happen that the structure of interest is contained in only a subset S of the L available variables, the others been useless or even harmful to detect a reasonable clustering into statistically different populations. For the variables in S, the frequencies of the states in at least two populations are different : we will call them clustering variables. For the other variables, the states are supposed to be equally distributed across the clusters. This approximation is theoretically justified by the oracle heuristics, which is able to take advantage of the misspecification ; the simulations performed in (Toussile and Gassiat, 2009) illustrate its benefits.
We denote by $\beta_{l,j}$ the frequency of state j associated to variable X^l in the whole population :

$$\beta_{l,j} = \alpha_{1,l,j} = \cdots = \alpha_{k,l,j} \cdots = \alpha_{K,l,j} \text{ for any } l \notin S \text{ and } 1 \leq j \leq A_l.$$

Obviously, $S = \emptyset$ if $K = 1$, otherwise S belongs to $\mathcal{P}^*(L)$, the set of all non empty subsets of $\{1, \ldots, L\}$.

Summarizing all these assumptions, we can write down the likelihood of an observation $x = \left(x^l\right)_{1 \leq l \leq L}$:

$$\begin{aligned}
\text{Case 1} : P_{(K,S,\theta)}(x) &= \left[\sum_{k=1}^{K} \pi_k \prod_{l \in S} \alpha_{k,l,x^l}\right] \times \prod_{l \notin S} \beta_{l,x^l} \\
\text{Case 2} : P_{(K,S,\theta)}(x) &= \left[\sum_{k=1}^{K} \pi_k \prod_{l \in S} \left(2 - \mathbb{1}_{x^{l,1}=x^{l,2}}\right) \alpha_{k,l,x^{l,1}} \times \alpha_{k,l,x^{l,2}}\right] \\
&\quad \times \prod_{l \notin S} \left(2 - \mathbb{1}_{x^{l,1}=x^{l,2}}\right) \beta_{l,x^{l,1}} \beta_{l,x^{l,2}}
\end{aligned} \tag{4.2}$$

where $\theta = (\pi, \alpha, \beta)$ is a multidimensional parameter, with

$$\begin{aligned}
\alpha &= \left(\alpha_{k,l,j}\right)_{1 \leq k \leq K;\ l \in S;\ 1 \leq j \leq A_l} \\
\beta &= \left(\beta_{l,j}\right)_{l \notin S;\ 1 \leq j \leq A_l}.
\end{aligned}$$

For a given K and S, $\theta = \theta_{(K,S)}$ ranges in the set

$$\Theta_{(K,S)} = \mathbb{S}_{K-1} \times \left[\prod_{l \in S} \mathbb{S}_{A_l - 1}\right]^K \times \prod_{l \notin S} \mathbb{S}_{A_l - 1}, \tag{4.3}$$

where $\mathbb{S}_{r-1} = \left\{p = (p_1,\ p_2, \ldots,\ p_r) \in [0,\ 1]^r : \sum_{j=1}^{r} p_j = 1\right\}$ is the $(r-1)$-dimensional simplex.

Then we consider the collection of all parametric models

$$\mathcal{M}_{(K,S)} = \left\{ P_{(K,S,\theta)} : \theta \in \Theta_{(K,S)} \right\} \tag{4.4}$$

with $(K,S) \in \mathbb{M} := \{(1,\emptyset)\} \cup (\mathbb{N}\backslash\{0,1\}) \times \mathcal{P}^*(L)$. To alleviate notations, we will often use the single index $m \in \mathbb{M}$ instead of (K,S).

Each model $\mathcal{M}_{(K,S)}$ corresponds to a particular structure situation with K clusters and a subset S of clustering variables. Inferring K and S becomes a model selection problem in a density estimation framework. It also leads to a data clustering, via the estimation $\widehat{\theta}$ of the parameter $\theta_{(K,S)}$ and the prediction of the class z of an observation x by the MAP method :

$$\widehat{z} = \underset{1\leq k\leq K}{\arg\max}\, P_{(K,S,\widehat{\theta})}\left(Z = k | X = x\right).$$

4.2.2 Model selection via penalization

A common method to solve model selection problems consists in the minimization of a penalized maximum likelihood criterion. In each model $\mathcal{M}_{(K,S)}$, consider the maximum likelihood estimator (MLE) $\widehat{P}_{(K,S)} = P_{(K,S,\widehat{\theta})}$, which minimizes the log-likelihood contrast

$$\gamma_n\left(P\right) = -\frac{1}{n}\sum_{i=1}^{n} \ln P\left(X_i\right) \tag{4.5}$$

where X_i describes the individual i in the sample. Then a data driven selected model $\mathcal{M}_{\left(\widehat{K}_n,\widehat{S}_n\right)}$ is chosen, where $\left(\widehat{K}_n, \widehat{S}_n\right)$ minimizes a penalized maximum likelihood criterion of the form

$$\mathbf{crit}(K,S) = \gamma_n\big(\widehat{P}_{(K,S)}\big) + \mathbf{pen}_n(K,S),$$

where $\mathbf{pen}_n : \mathbb{M} \to \mathbb{R}_+$ is the penalty function. Eventually the selected estimator is $\widehat{P}_{(\widehat{K}_n,\ \widehat{S}_n)}$.

The penalty function is designed to avoid overfit problems. Classical penalties, such as the ones used in **AIC** and **BIC** criteria, are based on the dimension of the model. In the following, we will refer to the number of free parameters

$$D_{(K,S)} = K - 1 + K\sum_{l\in S}\left(A_l - 1\right) + \sum_{l\notin S}\left(A_l - 1\right) \tag{4.6}$$

as the dimension of the model $\mathcal{M}_{(K,S)}$. The penalty functions of **AIC** and **BIC** are respectively defined by

$$\begin{aligned} \mathbf{pen_{AIC}}\left(m\right) &= \frac{1}{n}\cdot D_m; \\ \mathbf{pen_{BIC}}\left(m\right) &= \frac{\ln n}{2n}\cdot D_m. \end{aligned} \tag{4.7}$$

Our work is centered on the MLE estimator $\widehat{P}_{(K,\,S)}$, but this last one presents a drawback. For the sake of density estimation, we would like to use the Kullback-Leibler divergence **KL** as a risk function to measure the quality of an estimator. Unfortunately, when an state is not present in the sample, the MLE estimator assigns to it a zero probability. As a consequence, the Kullback risk $\mathbf{E}_{P_0}\left[\mathbf{KL}\left(P_0,\ \widehat{P}_{(K,\,S)}\right)\right]$ is infinite.

The Hellinger distance offers an alternative to the Kullback-Leibler divergence. Let us consider two probability distribution P and Q, admitting respectively s and t as density functions with respect to a common σ-finite measure μ. We call Hellinger distance between P and Q the quantity $\mathbf{h}(P,Q)$ defined by

$$\mathbf{h}(P,Q)^2 = \int \left(\sqrt{s(x)} - \sqrt{t(x)}\right)^2 d\mu(x). \tag{4.8}$$

Let $(K^*,\ S^*)$ be a minimizer in $(K,\ S)$ of the Hellinger risk of the MLE estimator

$$R_{(K,\,S)} = \mathbf{E}_{P_0}\left[\mathbf{h}^2\left(P_0,\ \widehat{P}_{(K,\,S)}\right)\right]. \tag{4.9}$$

The density $\widehat{P}_{(K^*,\,S^*)}$ is called oracle for the Hellinger risk. It is not an estimator, since it depends on the true density P_0. However it can be used as a benchmark to quantify the quality of our model selection procedure : in the simulation performed in paragraph 4.4.3, we compare the Hellinger risk of the selected estimator $\widehat{P}_{(\widehat{K}_n,\,\widehat{S}_n)}$ to the oracle risk.

4.3 New criteria and non asymptotic risk bounds

4.3.1 Main result

Our main theorem provides an oracle inequality for both Case 1 and Case 2. It links the Hellinger risk of the selected estimator to the Kullback-Leibler divergence **KL** between the true density and each model in the models collection. Unlike **KL** which is not a metric, the Hellinger distance **h** permits to take advantage of the metric properties (metric entropy) of the models.

Theorem 1. *We consider the collection $\mathbb{M}$ of models defined above, and a corresponding collection of ρ-MLEs $\left(\widehat{P}_{(K,S)}\right)_{(K,S)\in\mathbb{M}}$, which means that for every $(K,S)\in\mathbb{M}$*

$$\gamma_n\left(\widehat{P}_{(K,S)}\right) \leq \inf_{Q\in\mathcal{M}_{(K,S)}} \gamma_n(Q) + \rho.$$

Let $A_{\max} = \sup_{1\leq l\leq L} A_l$, and let ξ be defined by $\xi = \dfrac{4\sqrt{A_{\max}}\sqrt{L}}{2^{L+1}-1}$ in Case 1 and $\xi = \dfrac{4\sqrt{A_{\max}}\sqrt{L}}{2(1+3\sqrt{2})^L - 1}$ in Case 2. Assume that $\xi < 1$ or $n > \xi^2 K$.

There exists absolute constants κ and C such that whenever

$$\mathbf{pen}_n(K,S) \geq \kappa\left(5 + \sqrt{\max\left(\frac{1}{2}\ln n + \frac{1}{2}\ln L, \frac{\ln 2}{2} + \ln L\right)}\right)^2 \frac{D_{(K,S)}}{n} \tag{4.10}$$

for every $(K,S) \in \mathbb{M}$*, then the model* $\mathcal{M}_{\left(\widehat{K}_n,\widehat{S}_n\right)}$ *where* $\left(\widehat{K}_n,\widehat{S}_n\right)$ *minimizes*

$$\mathbf{crit}(K,S) = \gamma_n\left(\widehat{P}_{(K,S)}\right) + \mathbf{pen}_n(K,S)$$

over $\mathbb{M}$ *exists and moreover, whatever the underlying probability* P_0*,*

$$\mathbf{E}_{P_0}\left[\mathbf{h}^2\left(P_0, \widehat{P}_{\left(\widehat{K}_n,\widehat{S}_n\right)}\right)\right] \leq C\left(\inf_{(K,S)\in\mathbb{M}}\left(\mathbf{KL}\left(P_0,\mathcal{M}_{(K,S)}\right) + \mathbf{pen}_n(K,S)\right) + \rho + \frac{(3/4)^L}{n}\right)$$

where, for every $(K,S) \in \mathbb{M}$*,* $\mathbf{KL}\left(P_0,\mathcal{M}_{(K,S)}\right) = \inf_{Q\in\mathcal{M}_{(K,S)}} \mathbf{KL}(P_0,Q)$*.*

The condition $\xi < 1$ is used in the proof to avoid more complicated calculations. In practice, ξ is very likely to be smaller than 1 for L not too small.

Note that as soon as $n \geq 2L$, (4.10) is simplified in the following way

$$\mathbf{pen}_n(K,S) \geq \kappa\left(5 + \sqrt{\frac{1}{2}\ln n + \frac{1}{2}\ln L}\right)^2 \frac{D_{(K,S)}}{n}.$$

The leading term for large n is $\kappa\frac{\ln n}{2}\frac{D_{(K,S)}}{n}$, which is a multiple of the penalty function of **BIC**. As a consequence, we can apply Theorem 2 from (Toussile and Gassiat, 2009) : when the underlying distribution P_0 belongs to one of the competing models, the smallest model (K_0,S_0) containing P_0 is selected with probability tending to 1 as n goes to infinity.

Such a penalty is not surprising in our context : it is in fact very similar to the one obtained in (Maugis and Michel, 2009) in a Gaussian mixture framework.

Sharp estimates of κ are not available. Theorem 1 is too conservative in practice, and leads to an over-penalized criterion which is outperformed by smaller penalties. So it is mainly used to suggest the shape of the penalty function

$$\mathbf{pen}_n(K,S) = \lambda\frac{D_{(K,S)}}{n} \tag{4.11}$$

where λ is a parameter to be chosen depending on n and the collection $\mathbb{M}$ — but not on (K,S). Slope heuristics (Birgé and Massart, 2007; Arlot and Massart, 2009) can be used in practice to calibrate λ : this is done in Section 4.4, where we use change-point detection (Lebarbier, 2002) in relation to slope heuristics.

Since $\mathbf{h}^2$ is upper bounded by 2, the non-asymptotic feature of Theorem 1 is interesting when n is large enough with respect to $D_{(K,S)}$. However, even with small values of n, the simulations performed in Subsection 4.4.3 show that the penalized criterion calibrated using the slope heuristics keep good behaviors.

4.3.2 A general tool for model selection

Theorem 1 is obtained from (Massart, 2007, Theorem 7.11). This last result deals with model selection problems by proposing penalty functions related to geometrical properties of the models, namely metric entropy with bracketing for Hellinger distance.

The framework here is the following. We consider some measurable space $(A, \mathcal{A})$, and μ a σ-finite positive measure on A. A collection of models $(\mathcal{M}_m)_{m\in\mathbb{M}}$ is given, where each model $\mathcal{M}_m$ is a set of probability density functions s with respect to μ. The following relation permits us to extend the definition of $\mathbf{h}$ to positive functions s or t whose integral is finite but not necessary 1. Denoting $\sqrt{s}$ the function defined by $\sqrt{s}(x) = \sqrt{s(x)}$, and by $\|\cdot\|_2$ the usual norm in $\mathbb{L}^2(\mu)$, then

$$\mathbf{h}(s,t) = \|\sqrt{s} - \sqrt{t}\|_2.$$

Let us now recall the definition of metric entropy with bracketing. Consider some collection F of measurable functions on A, and d one of the following metrics on F : $\mathbf{h}$, $\|\cdot\|_1$, or $\|\cdot\|_2$. A bracket $[l,u]$ is the collection of all measurable functions f such that $l \leq f \leq u$. Its d-diameter is the distance $d(u,l)$. Then, for every positive number ε, we denote by $N_{[.]}(\varepsilon, F, d)$ the minimal number of brackets with d-diameter not larger than ε which are needed to cover F. The d-entropy with bracketing of F is defined as the logarithm of $N_{[.]}(\varepsilon, F, d)$, and is denoted by $H_{[.]}(\varepsilon, F, d)$.

We assume that for each model $\mathcal{M}_m$ the square entropy with bracketing $\sqrt{H_{[.]}(\varepsilon, \mathcal{M}_m, \mathbf{h})}$ is integrable at 0. Let us consider some function ϕ_m on $\mathbf{R}_+$ with the following properties

(I). ϕ_m is nondecreasing, $x \mapsto \phi_m(x)/x$ is non-increasing on $(0, +\infty)$ and for every $\sigma \in \mathbf{R}_+$ and every $u \in \mathcal{M}_m$

$$\int_0^\sigma \sqrt{H_{[.]}\left(x, S_m(u,\sigma), \mathbf{h}\right)}\mathbf{d}x \leq \phi_m(\sigma),$$

where $S_m(u,\sigma) = \left\{t \in \mathcal{M}_m : \|\sqrt{t} - \sqrt{u}\|_2 \leq \sigma\right\}$.

(I) is verified in particular with $\phi_m(\sigma) = \int_0^\sigma \sqrt{H_{[.]}\left(x, \mathcal{M}_m, \mathbf{h}\right)}\mathbf{d}x$.

In order to avoid measurability problems, we suppose that for each $m \in \mathbb{M}$, the following separability condition is verified for $\mathcal{M}_m$:

(M). There exists some countable subset $\mathcal{M}'_m$ of $\mathcal{M}_m$ and a set $A' \subset A$ with $\mu(A') = \mu(A)$ such that for every $t \in \mathcal{M}_m$, there exists some sequence $(t_k)_{k\geq 1}$ of elements of $\mathcal{M}'_m$ such that for every $x \in A'$, $\ln(t_k(x))$ tends to $\ln(t(x))$ as k tends to infinity.

Theorem 2. *Let $X_1, \ldots, X_n$ be iid random variables with unknown density s with respect to some positive measure μ. Let $\{\mathcal{M}_m\}_{m\in\mathbb{M}}$ be some at most countable collection of models, each fulfilling (M). We consider a corresponding collection of ρ-MLEs $(\widehat{s}_m)_{m\in\mathbb{M}}$. Let $\{x_m\}_{m\in\mathbb{M}}$ be some family of nonnegative numbers such that*

$$\sum_{m\in\mathbb{M}} e^{-x_m} = \Sigma < \infty,$$

and for every $m \in \mathbb{M}$ considering ϕ_m with property (i) define σ_m as the unique positive solution of the equation

$$\phi_m(\sigma) = \sqrt{n}\sigma^2. \tag{4.12}$$

Let $\mathbf{pen}_n : \mathbb{M} \to \mathbf{R}_+$ and consider the penalized log-likelihood criterion

$$\mathbf{crit}(m) = \gamma_n\left(\widehat{s}_m\right) + \mathbf{pen}_n(m).$$

Then, there exists some absolute constants κ and C such that whenever

$$\mathbf{pen}_n(m) \geq \kappa\left(\sigma_m^2 + \frac{x_m}{n}\right) \text{ for every } m \in \mathbb{M},$$

some random variable $\widehat{m}$ minimizing **crit** *over $\mathbb{M}$ exists and moreover, whatever the density s*

$$\mathbf{E}_s\left[\mathbf{h}^2\left(s, \widehat{s}_{\widehat{m}}\right)\right] \leq C\left(\inf_{m\in\mathbb{M}}\left(\mathbf{KL}\left(s, \mathcal{M}_m\right) + \mathbf{pen}_n(m)\right) + \rho + \frac{\Sigma}{n}\right).$$

In Theorem 2, σ_m^2 has the role of a variance term of $\widehat{s}_m$, while the weights x_m take into account the number of models m having the same dimension.

4.3.3 Proof of Theorem 1

In order to apply Theorem 2, we need to compute the metric entropy with bracketing of each model $\mathcal{M}_{(K,S)}$. This is done in the following result, which is proved in Appendix 4.A.

Proposition 1 (Bracketing entropy of a model)**.** *Let $\eta : \mathbf{R}_+ \to \mathbf{R}_+$ be the increasing convex function defined by*

$$\begin{aligned}
&\textit{Case 1} : \eta(\varepsilon) = (1+\varepsilon)^{L+1} - 1,\\
&\textit{Case 2} : \eta(\varepsilon) = (1+\varepsilon)\left(1 + \sqrt{2}\,\varepsilon\,(2+\varepsilon)\right)^{L} - 1.
\end{aligned}$$

For any choice of K and S, $\mathcal{M}_{(K,S)}$ fulfills (M). For any $\varepsilon \in (0,1)$,

$$H_{[\cdot]}\left(\eta(\varepsilon), \mathcal{M}_{(K,S)}, \mathbf{h}\right) \leq D_{(K,S)} \ln\left(\frac{1}{\varepsilon}\right) + C_{(K,S)},$$

where

$$\begin{aligned}
C_{(K,S)} = \frac{1}{2}\Bigg(&\ln(2\pi e) D_{(K,S)} + \ln(4\pi e)\left(\mathbb{1}_{K\geq 2} + L + (K-1)|S|\right)\\
&+ \mathbb{1}_{K\geq 2}\ln(K+1) + \sum_{l=1}^{L}\ln(A_l+1) + (K-1)\sum_{l\in S}\ln(A_l+1).\Bigg)
\end{aligned} \tag{4.13}$$

$C_{(K,S)}$ is a technical quantity measuring the complexity of a model $\mathcal{M}_{(K,S)}$.

In the next step we establish an expression for ϕ_m. All following results are proved in Appendix 4.B.

Proposition 2. *For any choice of $m = (K, S)$, the function ϕ_m defined on $(0, \eta(1)]$ by*

$$\phi_m(\sigma) = \left(2\sqrt{\ln 2}\sqrt{D_{(K,S)}} + \sqrt{C_{(K,S)} - D_{(K,S)} \ln \eta^{-1}(\sigma)}\right) \sigma$$

fulfills (I).

We do not define ϕ_m for σ bigger than $\eta(1)$, to avoid more complicated expressions. This is why a condition on ξ appears in the following lemma :

Lemma 1. *Let $A_{\max} = \sup_{1 \le l \le L} A_l$, $\xi = \dfrac{4\sqrt{A_{\max}}\sqrt{L}}{2^{L+1} - 1}$ in Case 1, and $\xi = \dfrac{4\sqrt{A_{\max}}\sqrt{L}}{2(1 + 3\sqrt{2})^L - 1}$ in Case 2. Then, for all $n \ge 1$ if $\xi < 1$, and for $n > \xi^2 K$ otherwise, the solution σ_m of (4.12) verifies $\sigma_m < \eta(1)$.*

From Proposition 2 we can deduce an upper bound for σ_m, with a similar reasoning to (Maugis and Michel, 2009). First, $\sigma_m \le \eta(1)$ entails $\eta^{-1}(\sigma_m) \le 1$, and we obtain the lower bound $\sigma_m \ge \widetilde{\sigma}_m$, where

$$\widetilde{\sigma}_m = \frac{1}{\sqrt{n}}\left(2\sqrt{\ln 2}\sqrt{D_m} + \sqrt{C_m}\right). \tag{4.14}$$

This can be used to get an upper bound

$$\sigma_m \le \frac{1}{\sqrt{n}}\left(2\sqrt{\ln 2}\sqrt{D_m} + \sqrt{C_m - D_m \ln \eta^{-1}(\widetilde{\sigma}_m)}\right). \tag{4.15}$$

Let us now choose the weights x_m. If we take something bigger than $n\sigma_m^2$, this will change the shape of the penalty in Theorem 2. We define

$$x_m = (\ln 2) D_m.$$

The following Lemma shows that this choice is suitable.

Lemma 2. *For any model $\mathcal{M}_m$, with $m \in \mathbb{M}$ as above, let us set $x_m = (\ln 2) D_m$. Then*

$$\sum_{m \in \mathbb{M}} e^{-x_m} \le (3/4)^L.$$

To express the penalty function we have to lower bound $\eta^{-1}(\widetilde{\sigma}_m)$. This is done in the following Lemma.

Lemma 3. *Using the preceding notations,*

$$\sigma_m^2 + \frac{x_m}{n} \le \frac{D_{(K,S)}}{n}\left(5 + \sqrt{\max\left(\frac{1}{2}\ln n + \frac{1}{2}\ln L, \frac{\ln 2}{2} + \ln L\right)}\right)^2.$$

This ends the proof of Theorem 1.

4.4 In practice

In real datasets the number A_l of distinct allele states at each locus l is not necessarily known. The number $\widehat{A}_l$ of observed alleles can be used instead. In fact, the MLE estimator select a density with null weight on non-observed alleles. Then, in each model $\mathcal{M}_{(K,S)}$, an approximated MLE estimator can be computed thanks to the Expectation-Maximization (EM) algorithm (Dempster et al., 1977).

The other two points that have to be addressed before reaching the final estimator $\widehat{P}_{(\widehat{K}_n,\ \widehat{S}_n)}$ are the choice of the penalty function, and the sub-collection of models among which to select the optimal model. These two points are discussed in Subsections 4.4.1 and 4.4.2. Then simulations are presented in Subsection 4.4.3.

4.4.1 Slope heuristics and Dimension jump

Theorem 1 suggests to take a penalty function of the shape (4.11), defined modulo a multiplicative constant λ which has to be calibrated. Slope heuristics, as presented in Birgé and Massart (2007) and Arlot and Massart (2009), provide a practical method to find an optimal penalty

$$\mathbf{pen}_{\text{opt}}(m) = \lambda_{\text{opt}} D_m.$$

These heuristics are based on the conjecture that there exists a minimal penalty

$$\mathbf{pen}_{\text{min}}(m) = \lambda_{\text{min}} D_m$$

required for the model selection procedure to work : when the penalty is smaller that $\mathbf{pen}_{\text{min}}$, the selected model is one of the most complex models, and the risk of the selected estimator is large. On the contrary, when the penalty is larger than $\mathbf{pen}_{\text{min}}$, the complexity of the selected model is much smaller. Then the optimal penalty is close to twice the minimal penalty :

$$\mathbf{pen}_{\text{opt}}\,(m) \approx 2\lambda_{\text{min}} D_m.$$

The name "slope heuristics" comes from λ_{min} being the slope of the linear regression $\gamma_n\left(\widehat{P}_m\right) \sim D_m$ for a certain sub-collection of the most competing models m. For example, on Figure 4.1(a) below, a slope is visible for the models containing the true one $\mathcal{M}_{(K_0,\ S_0)}$. But the true model is unknown in real situation. A solution is to plot the function $D_{(K,\ S)} \mapsto -\gamma_n\left(\widehat{P}_{(K,\ S)}\right)$ and choose a large enough threshold D_0 such that that function has a linear behavior on the set of dimensions greater than D_0. But in our context, exploring large models is very expensive in computing time and may leads to suboptimal approximation of the maximum likelihood estimator by the EM algorithm.

Instead of estimating λ_{min} by linear regression, another method is the detection of the biggest jump on the selected dimension with respect to the candidate values of λ. In practice, suppose we have at hand a reasonable grid $\lambda_1 < \ldots < \lambda_r$ of candidate estimates of λ_{min}, and a sub-collection $\mathcal{C}_{ex}$ of most competitive models. Each λ_i leads to a selected

model $\widehat{m}_i$ with dimension $D_{\widehat{m}_i}$. If you plot $D_{\widehat{m}_i}$ as a function of λ_i, λ_{min} is expected to lie at the locus of the biggest jump. However, Figure 4.1(b) illustrates an important point : in that example the biggest jump is at $\lambda \approx 5.1$, but the optimal value of $\lambda_{\min}$ is around 0.9, which corresponds to several successive jumps. We propose an improved version of the dimension jump method of Arlot and Massart (2009), based on a sliding window : we consider at a time all jumps in a window of $h \geq 1$ following intervals in the grid. Algorithm 3 below describes the procedure. An example of the application of this variant of dimension jump method is given in Figure 4.1(c).

Algorithm 3 Calibration of Penalty $\left(\mathcal{C}_{ex}, (\lambda_i)_{i=1,\ldots,r}, h\right)$

for $i = 1$ to n_λ **do**

$\widehat{m}_i \leftarrow \arg\min_{m \in \mathcal{C}_{ex}} \left\{ \mathbb{P}_n \left(-\ln \widehat{P}_m \right) + \lambda_i D_m \right\}$

end for

$i_{jump} \leftarrow \min \arg \max_{i \in \{h+1,\ldots,r\}} \left\{ D_{\widehat{m}_{i-h}} - D_{\widehat{m}_i} \right\}$

$i_{init} \leftarrow \max \left\{ j \in [i_{jump} - h, i_{jump} - 1], D_{\widehat{m}_j} - D_{\widehat{m}_{i_{jump}}} = D_{\widehat{m}_{i_{jump}-h}} - D_{\widehat{m}_{i_{jump}}} \right\}$

$\widehat{\lambda}_{\min} \leftarrow \dfrac{\lambda_{i_{init}} + \lambda_{i_{jump}}}{2}$

return $\widehat{\lambda}_{\min}$

4.4.2 Sub-collection of models for calibration of the penalty

For a given maximum value $K_{\max}$ of the number of clusters, the number of models under competition is equal to $1 + (K_{\max} - 1) * \left(2^L - 1\right)$. Since this number is huge in most situations, it is very painful to consider all competing models for calibration of the constant λ. On the other hand, we need enough models to ensure that there is a clear jump in the sequence of selected dimension. We consider the modified backward-stepwise algorithm proposed in Toussile and Gassiat (2009), which enables to gather the most competitive models among all possible S for a given number K of clusters and a given penalty function $\mathbf{pen}_n$. It gives also the possibility to add a complementary exploration step based on a similarly modified forward strategy. We will refer to this algorithm as *explorer* $(K,\ \mathbf{pen}_n)$.

Since we do not know the final penalty during the exploration step, we consider a reasonable grid

$$\frac{1}{2n} = \lambda_1 < \ldots < \lambda_r = \frac{\ln n}{n}$$

containing both penalty functions associated to **AIC** and **BIC**. To each value λ_i of the grid is associated a penalty function $\mathbf{pen}_{\lambda_i}$. We launch *explorer* $\left(K,\ \mathbf{pen}_{\lambda_i}\right)$ for all values of K in $\{1, \ldots,\ K_{\max}\}$ and for all values of λ_i of the above grid, and we gather the explored models in $\mathcal{C}_{ex}$. This sub-collection seemly contains the most competitive models and it is then used to calibrate λ.

4.4.3 Numerical experiments

Our proposed procedure with a data-driven calibration of the penalty function has been implemented in the software `MixMoGenD` (Mixture Model for Genotypic Data), which already proposed a selection procedure based on asymptotic criteria **BIC** and **AIC** (Toussile and Gassiat, 2009). Here, we conduct numerical experiments on simulated datasets for performances assessment of the new non asymptotic criterion with respect to **BIC** and **AIC**. The penalty functions of these last criteria are respectively defined by

$$\mathbf{pen_{BIC}}(m) = \frac{\ln n}{2n} \cdot D_m;$$
$$\mathbf{pen_{AIC}}(m) = \frac{1}{n} \cdot D_m.$$

We present two experiments. The first one considers the consistency of the selected model : we study how the procedure retrieves the main features of the true model as the number of individuals in the datasets increases. In the second one, we are rather interested in the density estimation : we compare the risk of the selected estimator to the oracle risk.

Consistency performances In this experiment we consider a setting with $L = 10$ loci of 10 alleles each. We chose a parameter with $K_0 = 5$ populations of equal probability. The allelic frequencies have been chosen such that the genetic differentiation between the populations is decreasing with the locus number. In the first 6 loci, the populations are more separated. In the following 2 loci, the populations are poorly differentiated. In the last 2 loci, the alleles follow the same uniform distribution in all populations. The whole parameter is available at `http://www.math.u-psud.fr/~toussile/`.

We considered different values n of the sample size in $[50,\ 900]$ and for each of them, 10 datasets have been simulated. The results are summarized in Figure 4.2 and Table 4.1. The left panel gives the proportion of selecting the subset $\widehat{S}_n$ of clustering variables containing the first 6 variables, which are the most genetically differentiated variables. The right panel gives the proportion of selected models with $\widehat{K}_n = K_0$.

In this experiment, the **AIC** seems to be the best criterion for variable selection ; however the different between **AIC** and the new criterion is not significant. It also appears that **AIC** estimates the number of clusters better than the other criteria for small sample sizes (around $n = 100$ and $n = 200$), but it overestimates this number from $n = 500$ (see Table 4.1). On the contrary, the new criterion perfectly estimates the number of clusters for sample sizes ≥ 300. **BIC** performs poorly for both variables selection and classification on datasets with small sizes. As expected, the data-driven calibration of the penalty function improves globally the performances of the selection procedure, and it gives thus an answer to the question "Which penalty for which sample size ?".

It may happen that the results obtained on small sample sizes change a little from one run to another. In fact, the EM algorithm can miss the global maximum on such sample

		$\widehat{K}_n$							
n	**crit**	1	2	3	4	5	6	7	8
50	Cte*dim	0	10	0	0	0	0	0	0
	AIC	0	10	0	0	0	0	0	0
	BIC	10	0	0	0	0	0	0	0
100	Cte*dim	0	2	8	0	0	0	0	0
	AIC	0	0	9	0	0	0	0	0
	BIC	10	0	0	0	0	0	0	0
200	Cte*dim	0	0	3	2	5	0	0	0
	AIC	0	0	0	1	9	0	0	0
	BIC	10	0	0	0	0	0	0	0
300	Cte*dim	0	0	0	0	10	0	0	0
	AIC	0	0	0	0	10	0	0	0
	BIC	0	9	1	0	0	0	0	0
400	Cte*dim	0	0	0	0	10	0	0	0
	AIC	0	0	0	0	10	0	0	0
	BIC	0	9	1	0	0	0	0	0
500	Cte*dim	0	0	0	0	10	0	0	0
	AIC	0	0	0	0	8	1	1	0
	BIC	0	5	4	1	0	0	0	0
600	Cte*dim	0	0	0	0	10	0	0	0
	AIC	0	0	0	0	7	3	0	0
	BIC	0	0	0	0	5	5	0	0
700	Cte*dim	0	0	0	0	10	0	0	0
	AIC	0	0	0	0	6	2	2	0
	BIC	0	0	0	0	10	0	0	0
800	Cte*dim	0	0	0	0	10	0	0	0
	AIC	0	0	0	0	9	1	0	0
	BIC	0	0	0	0	10	0	0	0

TABLE 4.1 – Selection of the number of populations using different criteria : AIC, BIC and Cte*dim for criterion with penalty function on the form $\mathbf{pen}(K,\ S) = \lambda \cdot d_{K,S}$.

sizes, in particular in models of higher dimension. In our experiments, it is probably the case with some datasets of size $n \leq 300$, when the number of free parameters in the simulated model is ≥ 310.

Oracle performances of the estimator Since the new criterion is designed for density estimation, it is interesting to compare the associated estimator to the oracle for Hellinger risk. Recall that the oracle is the estimator associated to the model indexed by the minimizer $(K^*,\ S^*)$ of the risk $\mathbf{E}\left[\mathbf{h}^2\left(P_0,\ \widehat{P}_{(K,\ S)}\right)\right]$ over the collection of models $\mathcal{C}$.

In this experiment, we consider simulated datasets with reduced variability in order to reduce the computation time. The parameter underlying the data admits $L = 6$ loci, 3 alleles for each locus, and $K_0 = 3$ populations with equal proportions. The allelic frequencies have been chosen in such a way that the genetic differentiation between the population is significant on the first 3 loci, very small on the 4^{th} and 5^{th} loci, while the alleles of the 6^{th} locus follow the uniform distribution in all populations. Thus the true model is defined by $K_0 = 3$ and $S_0 = \{1,\ 2,\ 3,\ 4,\ 5\}$. The whole parameter is available at http://www.math.u-psud.fr/~toussile/.

We estimated the oracle using a Monte Carlo procedure on 100 simulated datasets of size 500 each, and got $\widehat{K^*} = 3$ and $\widehat{S^*} = \{1,\ 2,\ 3,\ 4\}$. The results we obtained are summarized in Figure 4.3 and Table 4.2.

	AIC	**BIC**
AIC	-	$< 5.40e-05$
Cte*Dim	$< 2.02e-05$	$< 2.20e-16$

TABLE 4.2 – The p-values of pairwise student tests comparing the means of the $\mathbf{h}^2\left(P_0,\ \widehat{P}_{(\widehat{K}_n,\ \widehat{S}_n)}\right)$. The alternative hypothesis is that the mean of the Hellinger distance associated to the criterion in the first column is less than the one associated to the criterion in the first line.

The worst behavior comes from **BIC** and it is not a surprise for two main reasons. First **BIC** is designed to find the true model which is different to the oracle in our experiments. Second, it is based on asymptotic approximation and then may requires large samples. In contrary, compared to **AIC** and **BIC**, the new criterion with data-driven calibration of the penalty function is significantly the best in the sense of Hellinger risk and the capacity of selecting the oracle. Recall that both **AIC** and the new criterion are designed to find the oracle (see Table 4.2). But like **BIC**, **AIC** is based on asymptotic approximations. So the advantage of the new criterion over **AIC** is probably that it is designed in a non asymptotic perspective.

4.5 Conclusion

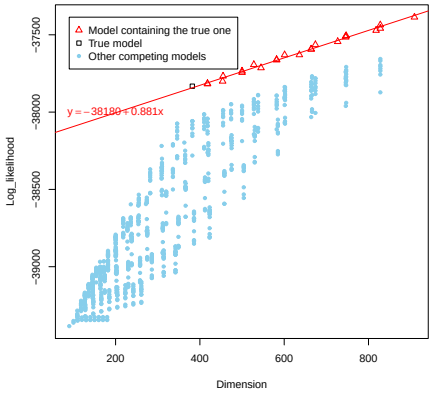
In this paper, we have considered a model selection via penalization, which performs simultaneously a variables selection and a detection of the number of populations, in the specific framework of multivariate multinomial mixture. This leads to a clustering in a second time. Our main result provides an oracle inequality, under the condition of some lower bound on the penalty function. The weakness of such a result is that the associated penalized criterion is not directly usable. Nevertheless, it suggests a shape of the penalty function which is of the form $\mathbf{pen}_n(m) = \lambda D_m/n$, where $\lambda = \lambda\,(n,\ \mathbb{M})$ is a parameter which depends on the data and the collection of the competing models. In practice λ is calibrated via the slope heuristics.

In the simulated experiments we conducted, the new criterion with penalty calibration shows good behaviors for density estimation as well as for the selection of the true model. It also performs well both when the number of individuals is large and when it is reasonnably small. This gives an answer to the question "Which criterion for with sample size ?"

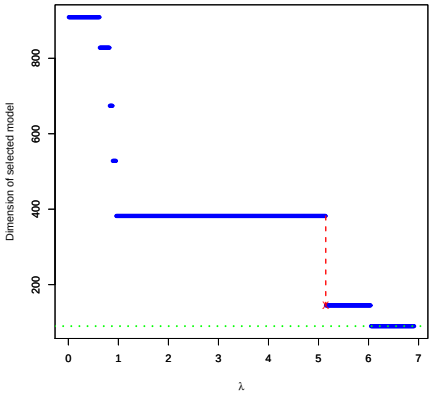
In the modeling we considered, the model dimension grows rapidly. In real experiments the number of individuals can be small, so other modeling with reduced dimension may be needed. We currently work on models which cluster the populations differently for each variable, as well as models which allocate the same probability to several states.

Acknowledgment

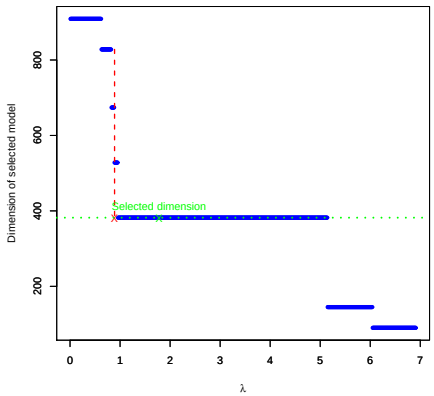
The authors gratefully acknowledge the comments and advice of Elisabeth Gassiat, Pascal Massart, and Gilles Celeux. Many thanks also to Nathalie Akakpo, Nicolas Verzelen, and Cathy Maugis, for the useful discussion we had. s the number of individuals can be small, so other mo



(a) $\widehat{\lambda}_{\min} \approx 0.88$ by linear regression on sub-models



(b) $\widehat{\lambda}_{\min} \approx 5.10$. The grid is too thin and the biggest jump corresponds to a wrong value.



(c) $\widehat{\lambda}_{\min} \approx 0.90$ by dimension jump detection combined with sliding window of size $h = 14$.

FIGURE 4.1 – Two ways to compute the slope, on a simulated sample of 1000 individuals, with 8 clustering loci among 10, and 5 populations. Models have been explored via the modified backward-stepwise described in subsection 4.4.2, the number K of clusters varying from 1 to 10. The size of the sliding window is 0.15.

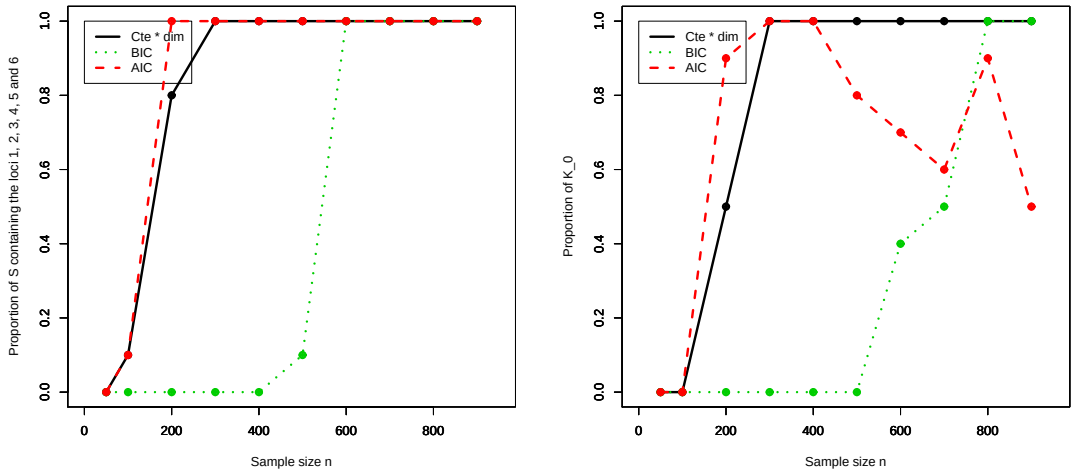


FIGURE 4.2 – The figure in the left panel gives the proportion of selected models with $\widehat{S}_n \supseteq \{1, \ldots, 6\}$, and the one in the right gives the proportion of selected models with $\widehat{K}_n = K_0$, versus the sample size.

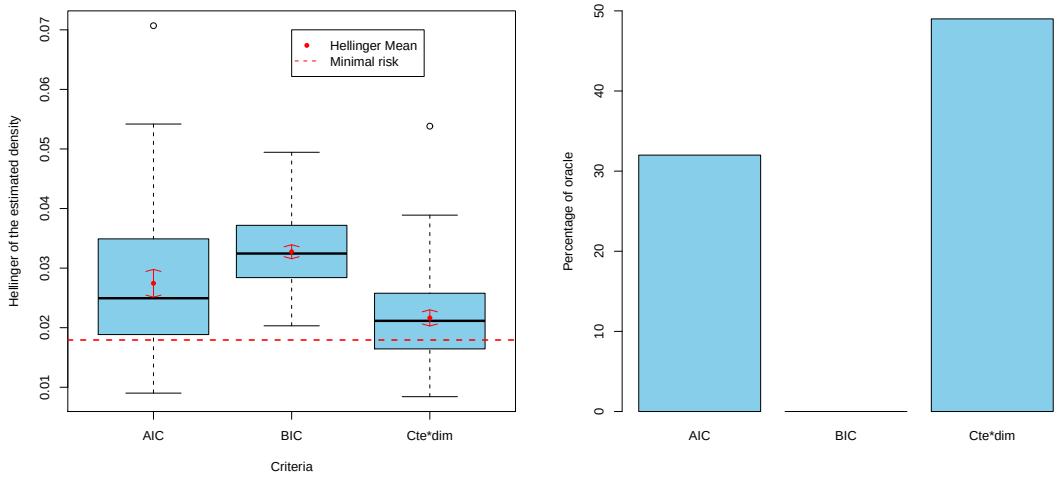


FIGURE 4.3 – The left panel gives the box-plots, means and their 95% confident intervals, for $\mathbf{h}^2\left(P_0,\ \widehat{P}_{(\widehat{K}_n,\ \widehat{S}_n)}\right)$; the right panel gives the percentages of selection of the estimated oracle $\left(\widehat{K^*},\ \widehat{S^*}\right)$; three criteria have been used : **AIC**, **BIC**, and **Cte*Dim** which denotes the new criterion with data-driven calibration of the penalty function.

Appendices

4.A Metric entropy with bracketing

We first state several results about the entropy with bracketing, which will be used to prove Proposition 1. They are mainly adapted from (Genoveve and Wasserman, 2000), but several are improved or writen here in a more general form. These lemmas can be seen as a toolbox to calculate the metric entropy with bracketing of complex models from the metric entropy of simpler elements.

We consider a measurable space $(A, \mathcal{A})$, and μ a σ-finite positive measure on A. We consider a model $\mathcal{M}$, which is a set of probability density functions with respect to μ. All functions considered in the following will be positive functions in $\mathbb{L}^1(\mu)$.

Lemma 4. *Let $\varepsilon > 0$. Let $[l, u]$ be a bracket in $\mathbb{L}^1(\mu)$, with $\mathbf{h}$-diameter less than ε, and containing s, a probability density function with respect to μ. Then*

$$\int l \, \mathbf{d}\mu \leq 1 \leq \int u \, \mathbf{d}\mu \leq (1+\varepsilon)^2.$$

Démonstration. First two inequalities are immediate, from $l \leq s \leq u$. For the last one, we use triangle inequality in $\mathbb{L}^2(\mu)$, and the definition of $\mathbf{h}$:

$$\begin{aligned}\int u \, \mathbf{d}\mu &= \int \left(\sqrt{l} + \left(\sqrt{u} - \sqrt{l}\right)\right)^2 \mathbf{d}\mu \\ &\leq \left(\sqrt{\int l \, \mathbf{d}\mu} + \mathbf{h}(u, l)\right)^2 \\ &\leq (1+\varepsilon)^2.\end{aligned}$$

□

Lemma 5 (Bracketing entropy of product densities)**.** *Let $n \geq 2$, and consider a collection $(A_i, \mathcal{A}_i, \mu_i)_{1 \leq i \leq n}$ of measured space. For any $1 \leq i \leq n$, let $\mathcal{M}_i$ be a collection of probability density functions on A_i fulfilling (M). Consider the product model*

$$\mathcal{M} = \left\{ s = \otimes_{i=1}^n s_i ; \forall 1 \leq i \leq n, s_i \in \mathcal{M}_i \right\}.$$

$\mathcal{M}$ contains density functions on $A = \prod_{i=1}^n A_i$ with respect to $\mu = \otimes_{i=1}^n \mu_i$.

$\mathcal{M}$ fulfills (M) and, for any sequence of positive numbers $(\delta_i)_{1 \leq i \leq n}$, if $\varepsilon \geq \prod_{i=1}^n (1 + \delta_i) - 1$ then

$$H_{[\cdot]}(\varepsilon, \mathcal{M}, \mathbf{h}) \leq \sum_{i=1}^n H_{[\cdot]}(\delta_i, \mathcal{M}_i, \mathbf{h}).$$

Démonstration. Let us consider some $s = \otimes_{i=1}^n s_i$ in $\mathcal{M}$. For $1 \leq i \leq n$, let $\mathcal{M}_i'$, A_i' and a sequence $(t_{i,k})_{k \geq 1}$ be such as needed for $\mathcal{M}_i$ to verify (M). Then, with the choice $t_k = \otimes_{i=1}^n t_{i,k}$ and $A' = \prod_{i=1}^n A_i'$, (M) is true for $\mathcal{M}$ too.

Let $\delta > 0$. For any $1 \leq i \leq n$, let $[l_i, u_i]$ a bracket containing s_i, with $\mathbf{h}$-diameter less than δ_i. Let us set $l = \otimes_{i=1}^n l_i$, and $u = \otimes_{i=1}^n u_i$. Then s belongs to bracket $[l, u]$. We can compute its $\mathbf{h}$-diameter :

$$\begin{aligned}
\mathbf{h}(l,u) &= \sqrt{\int_A \left(\sum_{j=1}^n \left(\prod_{i=1}^{j-1} \sqrt{l_i} \prod_{i=j}^n \sqrt{u_i} - \prod_{i=1}^{j} \sqrt{l_i} \prod_{i=j+1}^n \sqrt{u_i} \right) \right)^2 \mathbf{d}\mu} \\
&\leq \sum_{j=1}^n \prod_{i=1}^{j-1} \sqrt{\int_{A_i} l_i \, \mathbf{d}\mu_i} \prod_{i=j+1}^n \sqrt{\int_{A_i} u_i \, \mathbf{d}\mu_i} \;\; \mathbf{h}(l_j, u_j) \\
&\leq \sum_{j=1}^n \delta_j \prod_{i=j+1}^n (1+\delta_i) = \prod_{j=1}^n (1+\delta_j) - 1
\end{aligned}$$

thanks to triangle inequality and Lemma 4 (empty products equal 1).

Let $\varepsilon \geq \prod_{i=1}^n (1+\delta_i) - 1$. For any $1 \leq i \leq n$ consider a minimal covering of $\mathcal{M}_i$ with brackets of $\mathbf{h}$-diameter less than δ_i. With the previous process we can build a covering of $\mathcal{M}$ with brackets of $\mathbf{h}$-diameter less than ε. So the minimal cardinality of such a covering verifies

$$N_{[\cdot]}(\varepsilon, \mathcal{M}, \mathbf{h}) \leq \prod_{i=1}^n N_{[\cdot]}(\delta_i, \mathcal{M}_i, \mathbf{h}) .$$

□

Lemma 6 (Bracketing entropy of mixture densities)**.** *Let $n \geq 2$, and for any $1 \leq i \leq n$, let $\mathcal{M}_i$ be a set of probability density functions, all on the same measured space $(A, \mathcal{A}, \mu)$ and fulfilling (M). Let us consider the set of all mixture densities*

$$\mathcal{M} = \left\{ \sum_{i=1}^n \pi_i s_i : \pi = (\pi_i)_{1 \leq i \leq n} \in \mathbb{S}_{n-1}; \forall 1 \leq i \leq n, s_i \in \mathcal{M}_i \right\} .$$

Then $\mathcal{M}$ fulfills (M), and for any $\delta > 0$, $\eta > 0$, and $\varepsilon \geq \delta + \eta + \delta\eta$,

$$H_{[\cdot]}(\varepsilon, \mathcal{M}, \mathbf{h}) \leq H_{[\cdot]}(\delta, \mathbb{S}_{n-1}, \mathbf{h}) + \sum_{i=1}^n H_{[\cdot]}(\eta, \mathcal{M}_i, \mathbf{h}) .$$

Démonstration. First, let us note that $\mathbb{S}_{n-1}$ is separable for its usual topology. Then, checking that $\mathcal{M}$ fulfills (M) is easy, and we do not explicit it.

We do not develop either the proof of the last relation, because it is exactly the same as in (Genoveve and Wasserman, 2000, proof of Theorem 2). Let us just say that at the end we get, using our Lemma 4 instead of (Genoveve and Wasserman, 2000, Lemma 3),

$$\begin{aligned}
\mathbf{h}^2(l,u) &\leq \eta^2 (1+\delta)^2 + \delta^2 + 2\eta\,\delta\,(1+\delta) \\
&\leq \varepsilon^2 .
\end{aligned}$$

□

Next result is just Lemma 2 from (Genoveve and Wasserman, 2000) :

Lemma 7 (Bracketing entropy of the simplex)**.** *Let $n \geq 2$ be an integer. Let μ be the counting measure on $\{1, \ldots, n\}$. We identify any probability on $\{1, \ldots, n\}$ with its density $s \in \mathbb{S}_{n-1}$ with respect to μ. Then, if $0 < \delta \leq 1$,*

$$H_{[\cdot]}\left(\delta, \mathbb{S}_{n-1}, \mathbf{h}\right) \leq (n-1) \ln\left(\frac{1}{\delta}\right) + \frac{\ln 2 + \ln(n+1) + n \ln(2\pi e)}{2}.$$

To deal with Case 2, we also need the metric entropy of the collection of all Hardy-Weinberg genotype distributions for a given variable.

Lemma 8 (Bracketing entropy of Hardy-Weinberg genotype distributions)**.** *Suppose that, for some variable l, there exist $A_l \geq 2$ different states. Let Ω_l be the collection of all genotype distributions following Hardy-Weinberg model (4.1). Then Ω_l fulfills (M), and for any $\delta > 0$ and $\varepsilon \geq \sqrt{2}\,\delta\,(2+\delta)$,*

$$H_{[\cdot]}\left(\varepsilon, \Omega_l, \mathbf{h}\right) \leq H_{[\cdot]}\left(\delta, \mathbb{S}_{A_l - 1}, \mathbf{h}\right).$$

Démonstration. (4.1) permits to associate a parameter $\alpha = (\alpha_1, \ldots, \alpha_{A_l}) \in \mathbb{S}_{A_l-1}$ to any density in Ω_l. More generally, for any $\alpha \in [0,1]^{A_l}$, we define a function

$$d_\alpha(x) = \left(2 - \mathbb{1}_{x_1 = x_2}\right) \alpha_{x_1} \alpha_{x_2}$$

on the set of all genotypes $x = \{x^1, x^2\}$ on A_l states. Consider some $\delta > 0$ and $d_\alpha \in \Omega_l$. Let $[l, u]$ be some bracket containing α, with $\mathbf{h}$-diameter less than δ. Then d_α belongs to the bracket $[d_l, d_u]$. Let us calculate its diameter.

$$\begin{aligned}
\mathbf{h}^2(d_l, d_u) &= \sum_{a=1}^{A_l} (u_a - l_a)^2 + \sum_{1 \leq a < b \leq A_l} \left(\sqrt{2u_a u_b} - \sqrt{2 l_a l_b}\right)^2 \\
&\leq 2 \sum_{a=1}^{A_l} \sum_{b=1}^{A_l} \left(\sqrt{u_a u_b} - \sqrt{u_a l_b} + \sqrt{u_a l_b} - \sqrt{l_a l_b}\right)^2 \\
&\leq 2 \left(\sqrt{\sum_{a=1}^{A_l} u_a \sum_{b=1}^{A_l} \left(\sqrt{u_b} - \sqrt{l_b}\right)^2} + \sqrt{\sum_{a=1}^{A_l} \left(\sqrt{u_a} - \sqrt{l_a}\right)^2 \sum_{b=1}^{A_l} l_b} \right)^2 \\
&\leq 2 \left((1+\delta)\,\delta + \delta\right)^2
\end{aligned}$$

using Lemma 4. So $\mathbf{h}(d_l, d_u) \leq \sqrt{2}\,\delta\,(2+\delta)$.

Let $(\alpha^{(k)})_{k \geq 1}$ a sequence of elements of $\mathbb{S}_{A_l - 1} \cap \mathbb{Q}^{A_l}$, which tends to α for the usual topology as k tends to infinity. Then, for any genotype $x = \{x^1, x^2\}$, $\ln d_{\alpha^{(k)}}(x)$ tends to $\ln d_\alpha(x)$. Therefore Ω_l fulfills (M). □

Proof of Proposition 1. We build the proof for Case 2. For Case 1 everything is similar, with a simplification : we directly have $\mathbb{S}_{A_l - 1}$ instead of Ω_l.

Using (4.2) we see that a probability $P_{(K,S)}(\,\cdot\,|\theta)$ is the product of a mixture density corresponding to the variables in S, and a product density in $\bigotimes_{l\notin S}\Omega_l$ for the other variables. Let us call $\mathcal{M}$ the collection of all mixtures of K densities in $\bigotimes_{l\in S}\Omega_l$.

We first deal with the non clustering variables. Using Lemma 5 and Lemma 8, $\bigotimes_{l\notin S}\Omega_l$ fulfills (M). For any $\varepsilon\in(0,1)$,

$$H_{[\cdot]}\left((1+2\sqrt{2}\varepsilon+\sqrt{2}\varepsilon^2)^{L-|S|}-1,\bigotimes_{l\notin S}\Omega_l,\mathbf{h}\right)\leq\sum_{l\notin S}H_{[\cdot]}\left(2\sqrt{2}\varepsilon+\sqrt{2}\varepsilon^2,\Omega_l,\mathbf{h}\right)$$
$$\leq\sum_{l\notin S}H_{[\cdot]}\left(\varepsilon,\mathbb{S}_{A_l-1},\mathbf{h}\right).$$

On the same way

$$H_{[\cdot]}\left((1+2\sqrt{2}\varepsilon+\sqrt{2}\varepsilon^2)^{|S|}-1,\bigotimes_{l\in S}\Omega_l,\mathbf{h}\right)\leq\sum_{l\in S}H_{[\cdot]}\left(\varepsilon,\mathbb{S}_{A_l-1},\mathbf{h}\right).$$

We can apply Lemma 6, and get that $\mathcal{M}$ fulfills (M) and

$$H_{[\cdot]}\left((1+2\sqrt{2}\varepsilon+\sqrt{2}\varepsilon^2)^{|S|}(1+\varepsilon)-1,\mathcal{M},\mathbf{h}\right)$$
$$\leq\mathbb{1}_{K\geq2}H_{[\cdot]}\left(\varepsilon,\mathbb{S}_{K-1},\mathbf{h}\right)+K\sum_{l\in S}H_{[\cdot]}\left(\varepsilon,\mathbb{S}_{A_l-1},\mathbf{h}\right).$$

Lemma 5 again, applied to $\mathcal{M}$ and $\bigotimes_{l\notin S}\Omega_l$, gives that $\mathcal{M}_{(K,S)}$ fulfills (M), and for any $\varepsilon\in(0,1)$,

$$H_{[\cdot]}\left(\eta(\varepsilon),\mathcal{M}_{(K,S)},\mathbf{h}\right)$$
$$\leq\mathbb{1}_{K\geq2}H_{[\cdot]}\left(\varepsilon,\mathbb{S}_{K-1},\mathbf{h}\right)+K\sum_{l\in S}H_{[\cdot]}\left(\varepsilon,\mathbb{S}_{A_l-1},\mathbf{h}\right)+\sum_{l\notin S}H_{[\cdot]}\left(\varepsilon,\mathbb{S}_{A_l-1},\mathbf{h}\right).$$

At this point, it only remains to use Lemma 7 and to compute the constants. □

4.B Establishing the penalty

First, we need to establish some properties of function η.

Lemma 9 (Properties of function η). *We consider the function η defined in Proposition 1, from $\mathbf{R}_+$ into $\mathbf{R}_+$. η is nonnegative, increasing and convex. $\eta(0)=0$, and $\eta'(0)=L+1$ in Case 1 while $\eta'(0)=2\sqrt{2}\,L+1$ in Case 2.*

Démonstration. The proof in Case 1 is immediate, so we develop only Case 2.

Setting $u(x)=1+2\sqrt{2}\,x+\sqrt{2}\,x^2$, we can write $\eta(x)=(1+x)\,u(x)^L-1$. Then, calculus gives

$$\eta'(x)=(2L+1)\,u(x)^L+2L\,(\sqrt{2}-1)\,u(x)^{L-1}.$$

Since u is positive on $(0,+\infty)$, η is increasing. But $\eta(0)=0$, so η is nonnegative on $\mathbf{R}_+$. We also have $\eta'(0)=2\sqrt{2}\,L+1$. Next,

$$\eta''(x)=2\sqrt{2}\,(1+x)\,\left((2L^2+L)\,u(x)^{L-1}+2L\,(L-1)\,(\sqrt{2}-1)\,u(x)^{L-2}\right)$$

which is positive on $\mathbf{R}_+$. □

Proof of Proposition 2. Let $0<\sigma\leq\eta(1)$, and $\delta=\eta^{-1}(\sigma)$. Then, for any $u\in\mathcal{M}_m$,

$$\begin{aligned}
&\int_0^\sigma \sqrt{H_{[\cdot]}\left(x,\mathcal{M}_m(u,\sigma),\mathbf{h}\right)}\mathbf{d}x\\
&\quad\leq\sum_{j=1}^{\infty}\int_{\eta(2^{-j}\delta)}^{\eta(2^{-j+1}\delta)}\sqrt{H_{[\cdot]}\left(x,\mathcal{M}_m,\mathbf{h}\right)}\mathbf{d}x\\
&\quad\leq\sum_{j=1}^{\infty}\left(\eta\left(2^{-j+1}\delta\right)-\eta\left(2^{-j}\delta\right)\right)\sqrt{C_m-D_m\ln\delta+D_m j\ln 2}\\
&\quad\leq\eta(\delta)\sqrt{C_m-D_m\ln\delta}\\
&\qquad+\sqrt{D_m\ln 2}\sum_{j=1}^{\infty}\sqrt{j}\,\left(\eta\left(2^{-j+1}\delta\right)-\eta\left(2^{-j}\delta\right)\right).
\end{aligned}$$

We deal with the last term of this sum in the following way :

$$\begin{aligned}
\sum_{j=1}^{\infty}\sqrt{j}\,\left(\eta\left(2^{-j+1}\delta\right)-\eta\left(2^{-j}\delta\right)\right)&\leq\sum_{j=1}^{\infty}j\,\left(\eta\left(2^{-j+1}\delta\right)-\eta\left(2^{-j}\delta\right)\right)\\
&=\sum_{k=1}^{\infty}\eta\left(2^{-k+1}\delta\right)\\
&\leq\sum_{k=1}^{\infty}2^{-k+1}\eta(\delta)=2\sigma.
\end{aligned}$$

So

$$\int_0^\sigma \sqrt{H_{[\cdot]}\left(x,\mathcal{M}_m(u,\sigma),\mathbf{h}\right)}\mathbf{d}x\leq\phi_m(\sigma).$$

Since η is increasing, $\phi_m(x)/x$ is decreasing. To check that ϕ_m is nondecreasing, it is enough to prove that function $f(x)=x\sqrt{b-\ln\eta^{-1}(x)}$ is nondecreasing on $(0,\eta(1)]$, where $b=\frac{C_m}{D_m}$. From (4.13), we get $C_m>\frac{\ln(2\pi e)}{2}D_m>D_m$, so $b>1$. Calculus gives

$$f'(x)=\sqrt{b-\ln\eta^{-1}(x)}-\frac{x}{2\eta^{-1}(x)\,\eta'\left(\eta^{-1}(x)\right)\sqrt{b-\ln\eta^{-1}(x)}}.$$

Let $y\in(0,1]$. η is convex on $(0,1]$, and that entails $\frac{\eta(y)}{y\,\eta'(y)}\leq 1$. Thus

$$\sqrt{b-\ln y}\,f'\left(\eta(y)\right)\geq b-\ln y-1/2>0.$$

□

Proof of Lemma 1. Since $\phi_m(x)/x$ is non-increasing, for any $\sigma > 0$ such that $\sqrt{n}\,\sigma^2 > \phi_m(\sigma)$, $\sigma > \sigma_m$. So, we look for situations such that $\sqrt{n} > \dfrac{\phi_m(\eta(1))}{\eta^2(1)}$.

For all $1 \leq l \leq L$, $A_l \geq 2$. Since $\frac{1}{2}\ln(1+x) \leq x-1$ for $x \geq 2$, we get the following bounds

$$\frac{1+\ln(2\pi)}{2} D_m \leq C_m \leq \left(2 + \ln(2\pi) + \frac{\ln 2}{2}\right) D_m. \tag{4.16}$$

Therefore

$$\frac{\phi_m(\eta(1))}{\eta^2(1)} < \frac{4\sqrt{D_m}}{\eta(1)}$$

On the other hand, we have

$$D_m \leq K\,L\,A_{\max}.$$

So, since $\phi_m(x)/x^2$ is decreasing, $\sigma_m < \eta(1)$ as soon as $n > \xi^2 K$. This is true when $\xi < 1$, since $K \leq n$: the number of clusters is not bigger than the number of individuals. □

Proof of Lemma 2. We define $\delta = 1/2$, from which $e^{-x_m} = \delta^{D_m}$. If we consider the collection $\mathbb{M}$, we can discern two cases : $K = 1$ and $S = \emptyset$, or $K \geq 2$ and $S \neq \emptyset$. So, using (4.6),

$$\begin{aligned}
\sum_{m\in\mathbb{M}} e^{-x_m} &= \delta^{\sum_{l=1}^{L}(A_l-1)}\left(1 + \sum_{S\neq\emptyset}\sum_{K\geq 2}\left(\delta^{1+\sum_{l\in S}(A_l-1)}\right)^{K-1}\right) \\
&= \delta^{\sum_{l=1}^{L}(A_l-1)}\left(1 + \sum_{S\neq\emptyset}\frac{\delta^{1+\sum_{l\in S}(A_l-1)}}{1-\delta^{1+\sum_{l\in S}(A_l-1)}}\right) \\
&\leq \delta^{L}\left(1 + \frac{\delta}{1-\delta}\sum_{S\neq\emptyset}\delta^{|S|}\right) \\
&= \delta^{L}(1+\delta)^{L}.
\end{aligned}$$

□

Proof of Lemma 3. η^{-1} in concave and nondecreasing, $\eta(0) = 0$, so for any $0 \leq x \leq \eta(1)$,

$$\eta^{-1}(x) \geq \frac{\eta^{-1}(2)}{2}\,\min(x, 2).$$

On the other hand (4.14) and (4.16) entail

$$\widetilde{\sigma}_m \geq C_1\sqrt{\frac{D_m}{n}} \geq C_1\sqrt{\frac{L}{n}} \tag{4.17}$$

where $C_1 = 2\sqrt{\ln 2} + \sqrt{\frac{1+\ln(2\pi)}{2}} > 2\sqrt{2}$. Therefore

$$-\ln\eta^{-1}\left(\widetilde{\sigma}_m\right) \leq -\ln\left(\frac{\eta^{-1}(2)}{2}\right) - \ln 2 + \max\left(0, \frac{1}{2}\left(\ln n - \ln L - \ln 2\right)\right).$$

Consider Case 1. Since η is a convex function and $\eta'(0) = L + 1$,

$$\eta^{-1}(2) \leq \frac{2}{L+1}.$$

Now,

$$\eta\left(\frac{2}{L+1}\right) = \left(1 + \frac{2}{L+1}\right)^{L+1} - 1 \leq e^2 - 1.$$

Then

$$\frac{\eta^{-1}(2)}{2} \geq \frac{2/(L+1)}{\eta\left(2/(L+1)\right)} \geq \frac{2}{(e^2-1)(L+1)}.$$

Therefore

$$-\ln\left(\frac{\eta^{-1}(2)}{2}\right) \leq \ln(e^2-1) - \ln 2 + \ln L + \ln(3/2)$$

and

$$-\ln \eta^{-1}\left(\widetilde{\sigma}_m\right) \leq \ln(e^2-1) - \frac{7}{2}\ln 2 + \ln 3 + \max\left(\frac{1}{2}\ln n + \frac{1}{2}\ln L, \frac{\ln 2}{2} + \ln L\right).$$

Using now (4.15), we get

$$\begin{aligned}
\sigma_m^2 + \frac{x_m}{n} &\leq \frac{D_m}{n}\left(\frac{1}{2} + \left(2\sqrt{\ln 2} + \sqrt{2 + \ln(2\pi) + \frac{\ln 2}{2} - \ln \eta^{-1}\left(\widetilde{\sigma}_m\right)}\right)^2\right) \\
&\leq \frac{D_m}{n}\left(\frac{1}{\sqrt{2}} + 2\sqrt{\ln 2} + \sqrt{2 + \ln(2\pi) - 3\ln 2 + \ln 3 + \ln(e^2-1)}\right. \\
&\qquad \left. + \sqrt{\max\left(\frac{\ln n + \ln L}{2}, \frac{\ln 2}{2} + \ln L\right)}\right)^2 \\
&\leq \frac{D_m}{n}\left(5 + \sqrt{\max\left(\frac{\ln n + \ln L}{2}, \frac{\ln 2}{2} + \ln L\right)}\right)^2.
\end{aligned}$$

Next, consider Case 2, and follow the same method. Then

$$\eta^{-1}(2) \leq \frac{1}{\sqrt{2}\,L}$$

and

$$\eta\left(\frac{1}{\sqrt{2}\,L}\right) \leq 2\exp\left(2 + \frac{1}{\sqrt{2}}\right).$$

This leads to

$$-\ln \eta^{-1}(x) \leq 2 + \frac{1}{\sqrt{2}} + \frac{3\ln 2}{2} + \ln L - \ln \min(x, 2)$$

and

$$-\ln \eta^{-1}\left(\widetilde{\sigma}_m\right) \leq 2 + \frac{1}{\sqrt{2}} + \max\left(\frac{1}{2}\ln n + \frac{1}{2}\ln L, \frac{\ln 2}{2} + \ln L\right).$$

Now we obtain

$$\begin{aligned}\sigma_m^2 + \frac{x_m}{n} &\leq \frac{D_m}{n}\left(\frac{1}{\sqrt{2}} + 2\sqrt{\ln 2} + \sqrt{4 + \ln(2\pi) + \frac{\sqrt{2} + \ln 2}{2}}\right. \\ &\qquad \left. + \sqrt{\max\left(\frac{\ln n + \ln L}{2}, \frac{\ln 2}{2} + \ln L\right)}\right)^2 \\ &\leq \frac{D_m}{n}\left(5 + \sqrt{\max\left(\frac{\ln n + \ln L}{2}, \frac{\ln 2}{2} + \ln L\right)}\right)^2.\end{aligned}$$

□

Chapitre 5

MixMoGenD, a software for both loci selection and clustering on genotypic data

Abstract

`MixMoGenD` (Mixture Model for Genotypic Data) is a stand alone computer package implementing a loci selection procedure in model-based clustering using multilocus genetic data. It is implemented using $C++$ language with object-oriented programming and dynamic memory allocation. Windows and Linux versions are available free of charge on http://www.math.u-psud.fr/~toussile. `MixMoGenD` is concerned with the problem of grouping diploid individuals into genetically homogeneous clusters on the basis of their genotype at a certain number of loci. It may happen that some loci are just noise for classification into statistically different populations. `MixMoGenD` simultaneously solves loci selection and classification problems in a model selection procedure. The competing models are compared using penalized likelihood criteria : the Bayesian Information Criterion (BIC), the Akaike Information Criterion (AIC) and a criterion with data-driven calibration of the penalty function via the so called "slope heuristics". To avoid an exhaustive search of the optimum model, the selection procedure of `MixMoGenD` is based on backward and forward stepwise strategies, which enables a better search of the optimum model among the most competitive models with all possible cardinalities of S.

5.1 Introduction

`MixMoGenD` (Mixture Model for Genotypic Data) is a stand alone program that implements model-based clustering using multilocus genotypic data. Compared to other existing programs, its classification process is combined with a loci selection procedure. `MixMoGenD` 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 instance, the genetic stratification may be due to differences in allelic frequencies in (unknown) ancestral populations. Also, there may be preferences in the choice of partners of reproduction in such a way that the target population is structured in units of reproduction. It is then reasonable to assume that clusters are characterized by Hardy-Weinberg Equilibrium (HWE), Linkage Equilibrium (LE) and a set of allelic frequencies. Such assumptions are common in unsupervised classification on genotypic data as evidenced practically all existing programs : cite for instance `Structure` proposed by Pritchard et al. (2000), `BAPS` (Bayesian Application of Population Structured) proposed by Corander et al. (2004), the `R` package `Geneland` proposed by Guillot et al. (2005), `Admixture` by Alexander et al. (2009) and `Fastruct` proposed by Chen et al. (2006). HWE and LE define a genetic equilibrium almost impossible to achieve. But such a genetic equilibrium is an ideal state that provides a baseline to measure genetic changes. Population structure is known to create Hardy-Weinberg Disequilibrium (HWD) and Linkage Disequilibrium (LD). So discovering the genetic stratification in a target population is an important task in population genetics. For instance, such a stratification has long been recognized as a confounding factor in genetic association studies. The statistical modeling commonly used for this purpose is model-based clustering. For example, this model has allowed Rosenberg et al. (2001) to highlight gene flow between some of the eight Jews populations in the Middle East.

`MixMoGenD` deals with genotypic data from diploid individuals at a certain number L of autosomic loci. We assume that the n-sample of genotypes we have at hand come from a population structured into a certain (unknown) number K of populations. On the other hand, the number L of genotyped loci may be very large due to the explosion of genomic projects. But, the population structure of interest may be contained in only a subset S of the L genotyped loci, the others (loci in S^c with $S \cup S^c = \{1, \ldots, L\}$) 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 sample in the "best" way. So our problem is twofold : we are looking for a relevant subset of variables that produce an unobserved classification.

`MixMoGenD` solves simultaneously the loci selection problem and the classification problem (via estimation of the number K of populations) in a model selection procedure. A specific collection of competing models is defined in such a way that each model corresponds to a particular classification. The selection procedure is based on penalized maximum likelihood criteria. Since we are in unsupervised classification settings, the population of origin of each individual of the sample is unobserved. The maximum likelihood is estimated using the Expectation-Maximization (EM) algorithm. We consider classical asymptotic criteria such as the Bayesian Information Criterion (BIC) (Schwarz, 1978)

and Akaike Information Criterion (AIC) (Akaike, 1973). We also consider a family of penalized criteria associated to penalty functions on the shape $\lambda d_{K,S}$, where $\lambda = \lambda(\mathcal{C},\ n)$ depends of the data and the collection $\mathcal{C}$ of competing models, and $d_{K,S}$ is the number of free parameters in the model defined by (K, S) (Toussile and Bontemps, 2010). In `MixMoGenD` , λ is calibrated using a data-driven procedure based on "slope heuristics" (Arlot and Massart, 2009, and references therein). We found on simulated data that the data calibration of the penalty term gives an answer to the question "Which penalty function for which sample size ?".

`MixMoGenD` is implemented using $C++$ programming language with object-oriented programming. The memory is dynamically allocated so that the memory capacity of the user's computer is the only limit of the size of data sets.

This document describes how `MixMoGenD` proceeds and how to use the software. Section 5.2 presents the data format. How to open a session and the main menus of `MixMoGenD` are given in Section 5.3. During the analysis of a data set, `MixMoGenD` produces several output files described in Section 5.4. The competing models and model selection via penalization are presented in Section 5.5.

5.2 Data format

The entire dataset must be arranged as a matrix with n rows and L or $L+1$ columns in a single file : n is the number of individuals sampled and L the number of loci (the same for all individuals). The i-th row contents the genotypes of individual i at the L loci separated by tabulation or space. The eventual $(L+1)$-th column indicates the a priori population of origin of the individuals in the sample. Data of this column are treated as character strings. The a priori classification may be based on visible characters such as geographical location, linguistic, cultural, physical characters. The only purpose of this classification is to enable `MixMoGenD` to produce the confusion matrix after the assignment of individuals of the sample to clusters.

Table 5.1 gives an example of a dataset. It shows some useful features of the input file :

- Line i-th gives the genotypes of individual i at the $L = 7$ loci.
- The a priori group of origin of individual i (if it exists) is a string of characters or an integer in the last column of line i.
- The observed alleles are labeled from 1 to 9, or 01 to 99, or 001 to 999, or ... if needed. In 3-digits and 4-digits formats, homozygous for allele 80 are noted 080080 and 00800080 respectively, not 8080 as in 2-digits format. Different digits coding of alleles can be intermixed among loci, but not within loci.
- `MixMoGenD` input files are ASCII text files with the extension `.txt`. The user must make sure that there is no blank line at the end of the data file.

080107	1001	101105	01060109	102104	108109	102105	5
101102	0208	101108	01070108	106106	104110	105101	3
107106	0406	104106	01040107	108101	107107	107107	4
107108	0710	105102	01070110	102101	101105	103109	2
101080	0702	108104	01040107	110101	107101	107106	1
080107	1009	108108	01040090	107101	109107	102108	5

(a)

080107	1001	101105	01060109	102104	108109	102105
101102	0208	101108	01070108	106106	104110	105101
107106	0406	104106	01040107	108101	107107	107107
107108	0710	105102	01070110	102101	101105	103109
101080	0702	108104	01040107	110101	107101	107106
080107	1009	108108	01040090	107101	109107	102108

(b)

TABLE 5.1 – Example of data files : one row per individual genotyped at $L = 7$ loci and one column per locus. (a) and (b) represent the same genotype data set with and without an a priori classification respectively in the last column

5.3 Open a session

`MixMoGenD` is very simple to use. Just read on the screen and choose what you want by entering the corresponding integer displayed in the menu. The software does not need to be installed. For both Linux and Windows Operating Systems (OS), just copy the executable file in the same directory as the dataset file to be analyzed and proceed as follows.

- Under Linux OS, open a console window in the directory where the executable file named `MixMoGenD` and the data file to be analyzed have been copied and execute this `./MixMoGenD`. If this does not work, go to the properties of the executable file by right-clicking on it, and enable its execution as a program.
- Under Windows OS, just double click on the executable file named `MixMoGenD.exe`.

Then `MixMoGenD` will display some informations about the program and display the following main menu :

```
-----------------------------------------------------------------
 > No current dataset

                          MixMoGenD 2.0

 Data file ..................................................   1

 Analyse many datasets with a single set of parameters ........   2

 Quit MixMoGenD .............................................   0

 > Choose a value in [0, 2]:
```

At this step, you have three possible choices : select 1 for the analysis of a single dataset, 2 for the analysis of many datasets with the same set of parameters of the selection process, or 0 to quit the program. Let give more details.

1. For the analysis of a single dataset, you have to answer the following questions. The default answer (if it exists) is given in capital letter.

   – `> Data file name 'file.txt' :`

     Give the name of the file containing you data (see Section 5.2 for the format of the data). The name of the data file must be on the form **yourdata.txt**.

   – `> Is there a prior classification (Y/n)?`

     The prior classification, if it exists, is used only for observed allelic frequencies and for the confusion matrix after classification by Maximum A Posteriori (MAP) rule.

   – `> Is the prior classification in the last column (Y/n)?`

     The prior classification must be in the first or in the last column. Answering "n" at this question means that the prior classification is in the first column.

   – `> Is the first line gives the names of variables (y/N)?`

     The first line of your file may contain the names of columns or not.

   – `> Create basic description file(s) (y/N)?`

     The basic description file gives the observed allelic and genotypic frequencies (see Section 5.4). If a prior classification is present in the data file, then the basic description file also contains these statistics for each prior class.

2. Selecting option 2 in the previous menu allows to analyze many datasets using the same set of parameters of the selection process. Before choosing this option, you must have create a file name `datasets.txt` in the same directory as the datasets to be analyzed. The file `datasets.txt` must have two or four columns. Here is an example

   ```
   yourdata1.txt 1 1 0
   yourdata2.txt 1 1 1
   yourdata3.txt 1 0 0
   ```

   – Column 1 gives the names of the datasets ;
   – Column 2 indicates if the prior classification exists (1) or not (0) ;
   – Column 3 indicates if the prior classification is in the last (1) or the first (0) column ;

– Column 4 indicates if the first line in the data file contents the names of the columns (1) or not (0).

In the case this file contains only two columns, the default values are used for columns 3 and 4. The default values are 1 and 0 respectively.

If the datasets are successfully loaded, `MixMoGenD` will display the size n and the number L of loci of each dataset, and the following menu will appear. You can now choose what to do with the data.

```
                    MixMoGenD 2.0

 Analyse many datasets with a single set of parameters

 Return to the main menu (keep only one dataset) .........   1

 Display data ............................................   2
 Observed alleles and their observed frequencies ..........   3
 Observed genotypes and their observed frequencies ........   4

 Estimates of the allelic frequencies for a given K and S .   5

 Selection procedure:
     Selection of K for a given S ........................   6
     Selection of S for a given K ........................   7
     Selection of both K and S ...........................   8

     Perform Model Selection from file(s) ................   9

 File conversions: .......................................  10

 Analyse many datasets with a single set of parameters ....  11

 Quit MixMoGenD ..........................................   0

 > Choose a value in [1, 10]:
```

In this menu and the following, K denotes the number of clusters, and S the relevant subset of loci that cluster the sample in the best way. The other loci gathered in S^c are just noise or are irrelevant for classification purposes. It is clear that for any locus in S, there exists at least 2 populations with different sets of allelic frequencies. We assume that the alleles of the loci in S^c are identically distributed across the populations. The options in the menu correspond to the following tasks :

1. Change the current data file.
2. Display the datasets that have been loaded.

3. Display the observed alleles and their frequencies.
4. Display the observed genotypes and their frequencies.
5. Perform the EM algorithm for an approximation of the the maximum likelihood estimate of the vector of parameters for a given number K of populations, and a subset S of clustering loci (see Appendix 5.A for the EM equations). The parameters are gathered in a multidimensional vector $\theta_{K,S} := (\pi,\ \alpha,\ \beta)$, where
   - $\pi := (\pi_k)_{1\leq k\leq K}$, is the vector of mixing proportions (π_k is the proportion of population k in the overall population).
   - $\alpha := (\alpha_{k,l,j})_{1\leq k\leq K;1\leq l\leq L;1\leq j\leq A_l}$ is a 3-dimensional table given the allelic frequencies. More precisely, $\alpha_{k,l,j}$ is the frequency of allele j at locus l in population k. Thus, α is associated to loci in the subset S of selected loci.
   - $\beta := (\beta_{l,j})_{1\leq l\leq L;1\leq j\leq A_l}$ is a 2-dimensional table given the allelic frequencies in the overall population. β contains the allelic frequencies of the loci $l \notin S$.
6. Perform the selection of the number K of populations for a given subset S of clustering loci, and a maximum number $K_{\max}$ of populations.
7. Select the relevant subset S of loci that gives the "best" classification of the sample into K clusters.
8. Select both the optimum number K of clusters and the optimum subset S of clustering loci for a given maximum number $K_{\max}$ of clusters.
9. Also select both K and S among competing models gathered by the `Explorer` algorithm (see Subsection 5.5.2) in previous analysis. These competing models are in the file `ExploredModels_yourdata.txt` (see Section 5.4).
10. Convert a dataset from `MixMoGenD` format to `Genepop` or `Structure` format (Raymond and Rousset, 1995; Pritchard et al., 2000).
11. Analyze many datasets with the same set of parameters of the selection process.

In options 6, 7, 8 and 9 of the previous menu, the selection process is based on penalized maximum likelihood criteria. We consider Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) and a family of criteria associated to penalty functions of the shape $\mathbf{pen}(K,\ S) = \lambda d_{K,S}$, where $\lambda = \lambda(n,\ \mathcal{C})$ is a multiplicative term assumed to depend on the sample size n and the complexity of the collection $\mathcal{C}$ of the competing models (see Subsection 5.5.1), and $d_{K,S}$ the number of free parameters associated to $(K,\ S)$. The multiplicative term λ is data-driven calibrated thanks to "slope heuristics".

In options 5, 6, 7 , 8 and 9 of the previous menu, the maximum likelihood is estimated via the Expectation and Maximization (EM) algorithm (see Appendix 5.A for the EM equations). This algorithm is known to converge slowly in some situations and its solution can highly depends on its starting position. Consequently it can produce sub-optimal maximum likelihood estimates. To act against this high dependency of EM on its initial position, CEM (Classification EM) (Celeux and Diebolt, 1991) and SEM (Stochastic EM) (Celeux and Govaert, 1992) have been proposed. We decide for the strategy of short runs of EM (at least $NBER\_ITER\_SMALL\_EM = 20$ iterations) from at least

$NBER\_SMALL\_EM = 50$ random positions followed by a long run from the solution maximizing the observed log-likelihood (Biernacki et al., 2001). For this purpose, the user has to provide some EM parameters :

- $EPSI \in [1e-20;\ 1e-8]$: this is the threshold used to stop long runs of the EM algorithm. Expectation and Maximization steps are repeated until $\frac{\mathcal{L}_n\left(\widehat{\theta}_{K,S}^{(j+1)};\ \mathbf{x}\right) - \mathcal{L}_n\left(\widehat{\theta}_{K,S}^{(j)};\ \mathbf{x}\right)}{|\mathcal{L}_n\left(\widehat{\theta}_{K,S}^{(j)}\right)|}$ is smaller than $EPSI$ or until the maximum number 10 000 of iterations is reached, where $\mathcal{L}_n\left(\widehat{\theta}_{K,S}^{(j)};\ \mathbf{x}\right)$ is the observed maximum log-likelihood at iteration j.
- $NBER\_ITER\_SMALL\_EM \in [20,\ 100]$: the number of iterations for each small EM.
- $NBER\_SMALL\_EM \in [50,\ 500]$: the number of short runs of EM.

5.4 Output files

During the analysis of a dataset, `MixMoGenD` produces several output files summarizing the analysis process and the results. The output files are saved in the same folder as the datasets being analyzed. Denote by `yourdata.txt` the data file to be analyzed. The content of each output file is described as follows.

- `Descript_yourdata.txt` : this file contains the observed allelic and genotypic frequencies. If a prior classification is specified in the dataset, this output file also contains these observed statistics for each prior class.
- `ExploredModels_yourdata.txt` : this file contains the models gathered by our "`Explorer`" algorithm (see Section 5.5). It is created in options 6, 7 and 8 of the previous menu.
- `Parameters_X_yourdata.txt` : this file contains the maximum likelihood estimates of the mixing proportions and the allelic frequencies of the selected model : with
  - `X`=`KS` in option 5 ;
  - `X`=`K_AIC`, `K_BIC` or `K_Cte_Dim` according on the choice of the penalty function in option 6 ;
  - `X`=`S_AIC`, `S_BIC` or `S_Cte_Dim` according on the choice of the penalty function in option 7 ;
  - `X`=`KS_AIC`, `KS_BIC` or `KS_Cte_Dim` according on the choice of the penalty function in options 8 et 9.

  `Cte_Dim` stands for the criterion associated to a penalty function of the form $\lambda d_{K,S}$. In case a prior classification exists in the dataset, this file also contains the confusion matrix obtained after MAP assignment of individuals to clusters.
- `MAP_X_yourdata.txt` : this file contains data and Maximum A Priori (MAP) classification. It is produced in options 5, 6, 7, 8 and 9 of the previous menu with `X` defined as above.

  In fact, in these options, `MixMoGenD` computes the maximum likelihood estimates of the mixing proportions and the allelic frequencies in the selected model. These estimates yields the MAP rule which produces the assignment of each individual

of the sample to the cluster with hight probability. These data can be useful for additional analysis of the population structure produced by `MixMoGenD` .

5.5 Models and methods

We deal with a sample of diploid individuals without any prior information on the population they come from. We have to face the problem of unsupervised classification. Several unsupervised classification methods exist and roughly fall into two categories. The first one gathers the so-called distance-based methods which are based on similarity and/or dissimilarity distances. As example of these methods, we can cite hierarchical classification (Fowlkes et al., 1988), and also the K-means algorithm (Brusco and Cradit, 2001). The main drawback of these methods is the interpretation of the obtained clusters. Indeed, the resulting classification depends on the choice of the distance or the dissimilarity measure. The second category gathers model-based methods which consist of using a model for each cluster and attempting to optimize the fit between the data and the resulting mixture of probability distributions. In practice, each cluster is represented by a parametric distribution like a Gaussian distribution in case of continuous variables or the multinomial distribution in case of discrete nominal variables. The entire data set is therefore modeled as a mixture of these distributions. Thus, each cluster is characterized by a set of parameters. One advantage of model-based clustering is to provide a rigorous framework to assess the number of mixture components and the role of variables in the clustering process (Keribin, 2000; Maugis et al., 2009).

The other fold of our problem is the variable selection. It is a recent research topic in clustering (Kohavi and John, 1997; Guyon and Elisseeff, 2003) compared with regression and supervised classification. In clustering, variable selection methods roughly fall into three approaches. The first approach is to weight the variables in clustering process or to reduce the dimension (see for instance Friedman and Meulman (2004) and McLachlan et al. (2002)). The specificity of this approach is an implicit variable selection. In contrast, the two other approaches, called ”filter” and ”wrapper”, select explicitly the relevant variables. The "filter" method consists of independently select the variables before clustering analysis (see for instance Dash et al. (2002); Jouve and Nicoloyannis (2005)). Independence between the selection and classification processes is the main weakness of this approach. In contrast, the so-called "wrapper" methods combine the two steps, allowing more interpretability of the selected variables (Maugis et al., 2009).

In `MixMoGenD` , we considered a “wrapper“ approach : we simultaneously solve the variable selection and classification problems via a model selection procedure. The maximum likelihood estimator associated to the selected model is then used for classification by Maximum A Posteriori (MAP) rule.

5.5.1 Competing models

Multilocus genotype data of diploid individuals can be considered as realizations of a random vector $\mathbf{X} = \left(X^l\right)_{1\leq l\leq L}$, where :

- L is the number of genotyped loci ;
- Each component X^l, called locus, is a pair of unordered nominal variables $\left\{X^{l,1},\ X^{l,2}\right\}$, called alleles, that take values in the same set of states.

Denote by $1,\ldots,\ A_l$, the distinct allele states at locus X^l (or equivalently locus l), with A_l being their number. We assume that $L \geq 2$ and $A_l \geq 2$ for any $l \in \{1,\ldots,\ L\}$. A locus with only one allele is not useful for clustering purposes. Let Z be the unobserved population of origin of an individual.

As the most existing model-based clustering methods using such data, `MixMoGenD` aims to group individuals into clusters of random mating individuals so that the *Hardy-Weinberg* (HWD) and linkage disequilibria (LD) are minimized across the sample (Latch et al., 2006, and references therein). Clusters are thus characterized by a set of allelic frequencies, *Hardy-Weinberg* equilibrium (HWE) and complete linkage equilibrium (LE). Under HWE, the probability of the genotype $x_i^l = \left\{x_i^{l,1},\ x_i^{l,2}\right\}$ of individual i at locus l is given by

$$P_{\alpha_{l,\cdot}}\left(x_i^l\right) = \left(2 - \mathbb{1}_{\left[x_i^{l,1}=x_i^{l,2}\right]}\right)\alpha_{l,x_i^{l,1}} \times \alpha_{l,x_i^{l,2}}, \tag{5.1}$$

for a given vector $\alpha_{l,\cdot} := (\alpha_{l,j})_{1\leq j\leq A_l}$ of allelic frequencies at locus l. Complete linkage equilibrium in each cluster means that within population, genotypes at different loci are independent random variables. 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 micro-evolution.

Now consider a given number K of clusters and a relevant subset S of loci for classification. Under HWE and LE, we consider the following form of probability distribution :

$$P_{(K,S,\theta)}\left(\mathbf{x}_i\right) = \left[\sum_{k=1}^{K}\pi_k \prod_{l\in S} P_{\alpha_{k,l,\cdot}}\left(x_i^l|\ Z=k\right)\right] \times \prod_{l\in S^c} P_{\beta_{l,\cdot}}\left(x_i^l\right) \tag{5.2}$$

in which

- $\mathbf{x}_i = \left(x_i^l\right)_{1\leq l\leq L}$ gives the genotypes of individual i at the L considered loci ;
- $\pi = (\pi_k)_{1\leq k\leq K}$ is the vector of mixing proportions ($\pi_k = P\left(Z=k\right)$) ;
- $\alpha_{k,l,\cdot} := (\alpha_{k,l,j})_{1\leq j\leq A_l}$ is the vector of the allelic frequencies at the loci l in S, in population k ;
- $\beta_{l,\cdot} := (\beta_{l,j})_{1\leq j\leq A_l}$ is the vector of allelic frequencies of the loci $l \in S^c$ in the overall population.

K, S and $\theta = (\pi,\ \alpha,\ \beta)$ are treated as parameters, but K and S are treated in a particular way. Indeed, for a given $(K,\ S)$, the multidimensional parameter θ belongs to

$$\Theta_{K,S} = \mathbb{S}_{K-1} \times \prod_{l\in S}\left[\mathbb{S}_{A_l-1}\right]^K \times \prod_{l\notin S}\mathbb{S}_{A_l-1}, \tag{5.3}$$

where $\mathbb{S}_r$ is the r-simplex

$$\mathbb{S}_r := \left\{ p = (p_1, \ldots, p_{r+1}) \in [0,\ 1]^{r+1} : \sum_{j=1}^{r+1} p_j = 1 \right\}.$$

Each couple $(K,\ S)$ defines thus a model

$$\mathcal{M}_{K,S} = \left\{ P_{(K,S,\theta)} :\ \theta \in \Theta_{K,S} \right\} \tag{5.4}$$

of probability distributions. Since each model $\mathcal{M}_{K,S}$ corresponds to a particular situation with K populations discriminated by loci in S, a choice of a model automatically leads to loci selection and classification. Indeed, once a model $\mathcal{M}_{\widehat{K}_n,\widehat{S}_n}$ is selected, one can derive a classification $\widehat{\mathbf{z}}_i = (\widehat{z}_{i,k})_{1\leq k\leq K}$ of each individual i in the sample by the MAP rule using the associated maximum likelihood estimator $\widehat{\theta}_{K,S}$:

$$\widehat{z}_{i,k} = \begin{cases} 1 & \text{if } P_{\widehat{\theta}_{K,S}}\left(Z_i = k |\ \mathbf{x}\right) > P_{\widehat{\theta}_{K,S}}\left(Z_i = h |\ \mathbf{x}\right) \text{ for all } h \neq k \\ 0 & \text{else.} \end{cases}$$

In our settings of incomplete data, the Expectation and Maximization (EM) algorithm (Dempster et al., 1977) is widely used to compute an approximation of the maximum likelihood estimate $\widehat{\theta}_{K,S}$ (see Appendix 5.A).

5.5.2 Principle of model selection via penalization

Model selection principle

Recall that L is the number of the genotyped loci. Denote by $\mathcal{P}^*(L)$ the set of all non-empty subsets S of the genotyped loci $\{1, \ldots,\ L\}$. Now we have at hand a collection

$$\mathcal{C} = \left\{ \mathcal{M}_{K,S} :\ (K,\ S) \in \{1\} \times \{\emptyset\} \bigcup \{\mathbb{N} \setminus \{0,\ 1\}\} \times \mathcal{P}^*(L) \right\} \tag{5.5}$$

of competing models. An old idea in model selection consists of selecting a model $\left(\widehat{K}_n,\ \widehat{S}_n\right)$ minimizing a penalized criterion

$$\mathbf{crit}\,(K,S) = \gamma_n\left(\widehat{P}_{K,S}\right) + \mathbf{pen}\,(K,S), \tag{5.6}$$

in which γ_n is an empirical contrast measuring the fit of the model to the data, $\widehat{P}_{K,S}$ a minimizer of γ_n, and $\mathbf{pen} :\ \mathcal{C} \longmapsto \mathbb{R}_+$ a penalty function that enables to make a trade off between the fit of the selected model to the data and its complexity. Since we are in discrete settings, the log-likelihood contrast is considered :

$$\gamma_n(P) = -\frac{1}{n}\sum_{i=1}^{n} \ln P(X_i), \text{ for any probability distribution } P.$$

Thus, the minimizer $\widehat{P}_{K,S}$ of the empirical contrast γ_n is the maximum likelihood estimator.

On the other hand, it is reasonable that the penalty function depends on the dimension of the competing models. We define the dimension of a model $\mathcal{M}_{(K,\ S)}$ as the number of free parameters :

$$d_{K,S} = K - 1 + K \sum_{l \in S} (A_l - 1) + \sum_{l \notin S} (A_l - 1) . \tag{5.7}$$

Maximum likelihood criteria implemented in MixMoGenD

The Bayesian Information Criterion (BIC) (Schwarz, 1978) and the Akaike Information Criterion (AIC) (Akaike, 1981; Burnham, 2004; Wang and Liu, 2006) are the must used asymptotic penalized maximum likelihood criteria. They can be defined by

$$\begin{aligned} BIC\,(K,S) &= -\frac{1}{n}\sum_{i=1}^{n} \ln\left\{\widehat{P}_{K,S}\,(X_i)\right\} + \frac{\ln n}{2n} d_{K,S} \\ AIC\,(K,S) &= -\frac{1}{n}\sum_{i=1}^{n} \ln\left\{\widehat{P}_{K,S}\,(X_i)\right\} + \frac{1}{n} d_{K,S}. \end{aligned} \tag{5.8}$$

Despite that both two are penalized maximum likelihood criteria, their objective models may be different. The BIC aims at choosing the "true" model $(K_0,\ S_0)$ assumed to belong to the considered collection $\mathcal{C}$ of competing models (Burnham and Anderson, 2002). It is designed to be consistent for the dimension (Keribin, 2000; Toussile and Gassiat, 2009). In a different approach, the AIC aims at choosing the model $\mathcal{M}_{(K^*,\ S^*)}$ associated to the minimal risk called "oracle". This approach does not mind if the true distribution P_0 of the observations belongs to one of the competing models or not.

We also consider a family of penalized maximum likelihood criteria with data-oriented penalty functions of the shape

$$\mathbf{pen}_{\lambda}\,(K,\ S) = \lambda d_{K,S}, \tag{5.9}$$

where $\lambda = \lambda\,(\mathcal{C},\ n)$ is a multiplicative data-dependent parameter. It typically depends on the collection $\mathcal{C}$ of competing models and on the sample size n. Such a family of penalized criteria is also known as General Information Criteria (GIC) (Bai et al., 1999). The shape of penalty function (5.9) is not new in model selection for density estimation. See for instance Maugis and Michel (2008) where such a shape of penalty function is used in a Gaussian context.

In `MixMoGenD` , the data-driven calibration of the multiplicative parameter λ is based on the so called "slope heuristics" proposed by Birgé and Massart (2007). These heuristics are based on the idea that the selection procedure requires a minimal penalty to work. In practice, the estimation of the minimal penalty $\mathbf{pen}_{\lambda_{\min}}$ by a linear regression as described in Maugis and Michel (2008) requires to explore a certain number of huge models. This

can be very costly in computing time. We consider the "dimension jump" version of these heuristics. The minimal penalty is such that the selected model associated to the smaller penalty is huge, and the one associated to the bigger penalty is reasonably small. In our context, the minimal penalty is associated with a value $\lambda_{\min}$ of the multiplicative term λ. The optimal penalty is then twice the minimal penalty

$$\mathbf{pen}_{opt}(K,\ S) = 2\lambda_{\min} d_{K,S}. \tag{5.10}$$

The estimation procedure of $\lambda_{\min}$ in `MixMoGenD` is as follows. First, we gather the most competitive models in $\mathcal{C}_{ex}$ using the "Explorer" algorithm described in Subsection 5.5.2. Now, consider the grid

$$\mathbf{grid}_\lambda = \left(\frac{1}{2n} = \lambda_1 < \ldots < \lambda_r = \frac{\ln n}{n}\right)$$

of candidate estimates of $\lambda_{\min}$. The associated family of penalized criteria contains both BIC and AIC. Each λ_i in the grid $\mathbf{grid}_\lambda$ leads to a selected model $\widehat{m}(\lambda_i)$ with dimension $d_{\widehat{m}(\lambda_i)}$. If you plot $d_{\widehat{m}(\lambda_i)}$ as a function of λ_i, λ_{min} is expected to lie at the position of the biggest jump. However the grid may be too thin, in which case the biggest jump may be at a wrong position like in Figure 5.1(b). We consider a slightly modified version of dimension jump method by introducing a sliding window (see Algorithm 4). Figure 5.1(a) gives the estimate of $\lambda_{\min}$ using linear regression on the sub-models containing the true one. In the example given in Figure 5.1(c), the sliding window of size 0.15 candidate values. The estimate obtained is approximately the same as the one obtained in Figure 5.1(a) by linear regression.

Algorithm 4 Calibration of Penalty$\left(\mathcal{C}_{ex}, (\lambda_i)_{i=1,\ldots,r}, h\right)$

for $i = 1$ to n_λ **do**
$\widehat{m}_i \leftarrow \arg\min_{m\in\mathcal{C}_{ex}} \left\{\mathbb{P}_n\left(-\ln \widehat{P}_m\right) + \lambda_i d_m\right\}$
end for
$i_{jump} \leftarrow \min\arg\max_{i\in\{h+1,\ldots,r\}} \left\{d_{\widehat{m}(\lambda_{i-h})} - d_{\widehat{m}(\lambda_i)}\right\}$
$i_{init} \leftarrow \max\left\{j \in [i_{jump}-h, i_{jump}-1], d_{\widehat{m}(\lambda_j)} - d_{\widehat{m}(\lambda_{i_{jump}})} = d_{\widehat{m}(\lambda_{i_{jump}-h})} - d_{\widehat{m}(\lambda_{i_{jump}})}\right\}$
$\widehat{\lambda}_{\min} \leftarrow \dfrac{\alpha_{i_{init}} + \alpha_{i_{jump}}}{2}$
return $\widehat{\lambda}_{\min}$

"Explorer" algorithm

For a given maximum value $K_{\max}$ of the number of populations, the number of models under competition is equal to $1+(K_{\max}-1)*\left(2^L-1\right)$. Since this number is huge in most situations, it is very painful to consider all competing models for both penalty calibration

and the search of the optimum model. On the other hand, we need enough models to ensure that there is a clear jump in the sequence of selected dimensions associated to $\mathbf{grid}_\lambda$. We propose in `MixMoGenD` the strategy described in the Algorithm 5 below. This algorithm is designed to gather the most competitive models among all possible cardinalities of the set S of relevant loci for classification. We refer to this algorithm as *Explorer* $(K,\ \mathbf{grid}_\lambda)$.

Algorithm 5 Explorer$(K,\ \mathbf{grid}_\lambda)$

for $\lambda \in \mathbf{grid}_\lambda$ **do**
  Backward-Stepwise$\left(K,\ \mathbf{crit}_{\mathbf{pen}_\lambda}\right)$ (Algorithm 6 below)
  Forward-Stepwise$\left(K,\ \mathbf{crit}_{\mathbf{pen}_\lambda}\right)$ (Algorithm 7 below)
end for

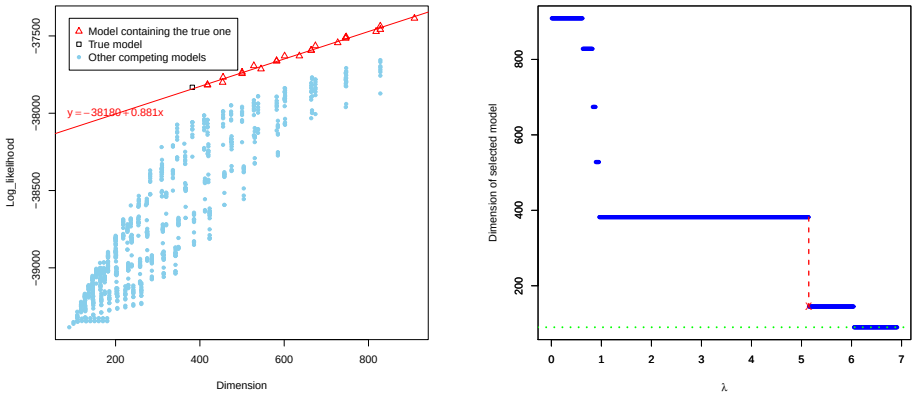
The optimum model is then selected among the models visited by the `Explorer` algorithm.

Algorithm 6 Backward-Stepwise(**crit**, K)

$S \leftarrow \{1, \ldots,\ L\}$, $c_{ex} \leftarrow 0$, $c_{in} \leftarrow 0$;
repeat
  EXCLUSION $(K,\ S)$;
  $c_{ex} \leftarrow \arg\min_{l \in S} \mathbf{crit}\,(K, S \setminus \{l\})$;
  if $\mathbf{crit}\,(K,\ S) - \mathbf{crit}\,(K,\ S \setminus \{c_{ex}\}) \geq 0$ or $c_{in} = 0$ **then**
    $S \leftarrow S \setminus \{c_{ex}\}$
  else
    $c_{ex} \leftarrow 0$;
  end if
  INCLUSION $(K,\ S)$
  $c_{in} \leftarrow \arg\min_{l \notin S} \mathbf{crit}\,(K,\ S \cup \{l\})$;
  if $\Big(\mathbf{crit}\,(K,\ S \cup \{c_{in}\}) - \mathbf{crit}\,(K,\ S) < 0$ and $S \cup \{c_{in}\}$ has never been the current set in an EXCLUSION step$\Big)$ **then**
    $S \leftarrow S \cup \{c_{in}\}$
  else
    $c_{in} \leftarrow 0$;
  end if
until $|S| = 1$

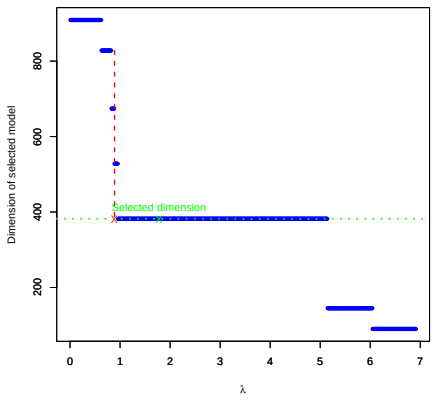
Algorithm 7 Forward-Stepwise(**crit**, K)

$S \leftarrow \emptyset,\ c_{ex} \leftarrow 0,\ c_{in} \leftarrow 0$;
repeat
  INCLUSION $(K,\ S)$;
  $c_{in} \leftarrow \arg\min_{l \notin S}$ **crit** $(K, S \cup \{l\})$;
  if crit $(K,\ S)$ − **crit** $(K,\ S \cup \{c_{in}\}) > 0$ or $c_{ex} = 0$ **then**
    $S \leftarrow S \cup \{c_{in}\}$
  else
    $c_{in} \leftarrow 0$;
  end if
  EXCLUSION $(K,\ S)$
  $c_{ex} \leftarrow \arg\min_{l \in S}$ **crit** $(K,\ S \setminus \{l\})$;
  if (**crit** $(K,\ S \setminus \{c_{ex}\})$ − **crit** $(K,\ S) \leq 0$ and $S \setminus \{c_{ex}\}$ has never been the current set in an INCLUSION step) **then**
    $S \leftarrow S \setminus \{c_{ex}\}$
  else
    $c_{ex} \leftarrow 0$;
  end if
until $|S^c| = 1$



(a) $\widehat{\lambda}_{\min} \approx 0.88$ by linear regression on sub-models

(b) $\widehat{\lambda}_{\min} \approx 5.10$. The grid is too thin and the biggest jump corresponds to a wrong value.



(c) $\widehat{\lambda}_{\min} \approx 0.90$ by dimension jump detection combined with sliding window of size $h = 14$.

FIGURE 5.1 – Example of calibration of the multiplicative term λ of the penalty function.

Appendices

5.A EM equations

We have considered a model selection procedure based on penalized maximum likelihood criterion to solve our two-fold problem of loci selection and classification. Recall that we are in unsupervised classification settings. More precisely, the population of origin of the sample we deal with is missing. For the estimation of the maximum likelihood in such a situation, Expectation and Maximization (EM) algorithm is widely used. It consists of iteratively maximizing the conditional expectation of the log-likelihood of the complete data, given the observations $\mathbf{x}$ and a current parameter $\theta^{(r)}$. Recall that we deal with a n-sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ that we denote by $\mathbf{x}$. For a given density $P_{K,S,\theta}$ in a model $\mathcal{M}_{K,S}$, the log-likelihood is given by

$$L_n(\theta;\ \mathbf{x}) = \sum_{i=1}^{n} \ln \left\{ \sum_{k=1}^{K} \pi_k P_{k,\alpha}\left(\mathbf{x}_i^S\right) \right\} + \sum_{i=1}^{n} \ln \left\{ P_\beta\left(\mathbf{x}_i^{S^c}\right) \right\},$$

where

- $\mathbf{x}_i^A = \left(x_i^l\right)_{l \in A}$ for a give subset A of the set $\{1, \ldots, L\}$ the considered loci ;
- π_k is the probability that an individual come from population k ;
- $P_{k,\alpha}(\cdot)$ is the density of the random vector $\mathbf{x}_i^S$ in population k, with α the allelic frequencies of the loci in S, in different populations ;
- $P_\beta(\cdot)$ is the density of the random vector $\mathbf{x}_i^{S^c}$, with β the allelic frequencies of loci not in S, in the overall population ;
- $\theta = (\pi,\ \alpha,\ \beta)$.

Let $\gamma = (\pi,\ \alpha)$. The maximum likelihood estimator (MLE) $\widehat{\theta}_{MLE}$ of θ is obviously given by $\widehat{\theta}_{MLE} = \left(\widehat{\gamma}_{MLE},\ \widehat{\beta}_{MLE}\right)$. Since we are in multinomial settings, $\widehat{\beta}_{MLE}$ is given by the observed allelic frequencies of the loci not in S. Thus the EM algorithm concerns only $\gamma = (\pi,\ \alpha)$.

The conditional expectation is given as follows,

$$Q\left(\gamma |\ \gamma^{(r)},\ \mathbf{x}^S\right) = \mathbf{E}_Z\left[\ln\left(P_\gamma\left(\mathbf{x}^S,\ Z\right)\right) |\ \mathbf{x},\ \gamma^{(r)}\right], \tag{5.11}$$

where the completed log-likelihood is given by

$$\ln\left(P_\gamma\left(\mathbf{x}^S,\ \mathbf{z}\right)\right) = \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{i,k} \ln\left(P_{k,\alpha}\left(\mathbf{x}_i^S\right)\right).$$

In our settings, the principle of the EM algorithm is to iteratively replace $Z_{i,k}$ by its conditional expectation for a given set $\mathbf{x}$ of observations and a current parameter $\gamma^{(r)}$. This expectation is the posterior assignment probability of individual i in population k. The algorithm starts with an initial parameter $\gamma^{(0)}$, and alternates between the two following steps. At the r^{th} iteration,

- **E step :** This step is to calculate $Q\left(\gamma |\ \gamma^{(r)},\ \mathbf{x}^S\right)$, which is to express the conditional

probabilities $\tau_{ik}^{(r)}$ that individual i come from population k :

$$\tau_{ik}^{(r)} = \frac{\pi_k^{(r)} \prod_{l \in S} \left(2 - \mathbb{1}_{[x_i^{l,1} = x_i^{l,2}]}\right) \alpha^{(r)}_{k,l,x_i^{l,1}} \alpha^{(r)}_{k,l,x_i^{l,2}}}{\sum_{h=1}^{K} \pi_h^{(r)} \prod_{l \in S} \left(2 - \mathbb{1}_{[x_i^{l,1} = x_i^{l,2}]}\right) \alpha^{(r)}_{h,l,x_i^{l,1}} \alpha^{(r)}_{h,l,x_i^{l,2}}}, \tag{5.12}$$

where $\alpha^{(r)}_{k,l,\,.}$ is the vector of allelic frequencies at locus l in population k, at iteration r.

– **M step :** This step consists of updating the parameters by estimating the parameter $\gamma^{(r+1)}$ that maximizes $Q\left(\gamma |\ \gamma^{(r)},\ \mathbf{x}^S\right)$. The update formula for the parameters can be derived using the standard method of the EM algorithm

$$\pi_k^{(r+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{(r)} \tag{5.13}$$

and

$$\alpha^{(r+1)}_{k,l,j} = \frac{\sum_{i=1}^{n} \tau_{ik}^{(r)} \left(\mathbb{1}_{[x_{i,1}^l = j]} + \mathbb{1}_{[x_{i,2}^l = j]}\right)}{2 \sum_{i=1}^{n} \tau_{ik}^{(r)}}. \tag{5.14}$$

The growth of $Q\left(\gamma,\ \gamma^{(r)},\ \mathbf{x}^S\right)$ at each iteration of the algorithm implies that of the observed log-likelihood $L_n\left(\gamma;\ \mathbf{x}^S\right)$, since

$$Q\left(\gamma |\ \gamma^{(r)},\ \mathbf{x}^S\right) = L_n\left(\gamma;\ \mathbf{x}^S\right) + H\left(\gamma;\ \gamma^{(r)}\right),$$

where

$$H\left(\gamma;\ \gamma^{(r)},\ \mathbf{x}^S\right) := \mathbf{E}\left[\ln\left(P_\gamma\left(Z |\ \mathbf{x}^S\right)\right) \mid \mathbf{x}^S,\ \gamma^{(r)}\right]$$

satisfies $H\left(\gamma;\ \gamma^{(r)},\ \mathbf{x}^S\right) \leq H\left(\gamma^{(r)};\ \gamma^{(r)},\ \mathbf{x}^S\right)$, from the Jensen inequality.

Under certain regularity conditions, EM algorithm is known to converge slowly in some situations and its solution can highly depend on its starting position and consequently produces sub-optimal maximum likelihood estimates. To act against this high dependency of EM on its initial position, CEM (Classification EM) (Celeux and Govaert, 1992) and SEM (Stochastic EM) (Celeux and Diebolt, 1991) have been proposed. We recommend here the strategy of short runs of EM from random positions followed by a long run of EM from the solution maximizing the observed log-likelihood (Biernacki et al., 2001).

Chapitre 6

Gametocytes infectiousness to mosquitoes : variable selection using random forests, and zero inflated models

This chapter presents a work in collaboration with Isabelle Morlais and Robin Genuer.

Abstract

Malaria control strategies aiming at reducing disease transmission intensity may impact both oocyst intensity and infection prevalence in the mosquito vector. Thus far, mathematical models failed to identify a clear relationship between *Plasmodium falciparum* gametocytes and their infectiousness to mosquitoes. Natural isolates of gametocytes are genetically diverse and biologically complex. Infectiousness to mosquitoes relies on multiple parameters such as density, sex-ratio, maturity, host immune factors and parasite genotypes. In this article, we investigated how density and genetic diversity of gametocytes impact on the success of transmission in the mosquito vector. We analyzed data for which the number of covariates plus attendant interactions is at least of order of the sample size, precluding usage of classical models such as general linear models. We then considered the variable importance from random forests to address the problem of selecting the most influent variables. The selected covariates were assessed in the zero inflated negative binomial model which accommodates both over-dispersion and the sources of non infected mosquitoes. We found that the most important covariates related to infection prevalence and parasite intensity are gametocyte density and multiplicity of infection.

6.1 Introduction

Malaria still represents a major health problem in more than one hundred tropical countries. The disease is caused by the parasite *Plasmodium* and its transmission occurs through the bite of an infective *Anopheles* female mosquito. In the last decades, insecticide and drug resistance has seriously hampered its control and alternative measures are urgently needed. Because *Plasmodium* transmission relies on the success of its development within the mosquito vector, called the sporogonic development, new strategies to fight malaria aim at controlling *Plasmodium* during the mosquito life cycle. Within the mosquito vector, malaria parasites undergo several life-stages and their successful development from one transition stage to an other will determine the outcome of infection. When ingested with the blood meal, male and female gametocytes fuse to form a zygote that differentiates into a mobile ookinete. The ookinete then traverses the midgut epithelium and encysts as an oocyst along the basal lamina. The oocyst, after several days of maturation, will release large number of sporozoites into the hemocoel. Sporozoites that will reach salivary glands will then be transmitted to a new host at a subsequent blood meal. *Plasmodium* parasites encounter severe losses during these successive phases and factors controlling parasite densities are not yet completely understood. Blood digestion processes and mosquito immune responses account for parasite decrease, but also the complex interplay between vector and parasite genotypes (Vaughan, 2007; Jaramillo-Gutierrez et al., 2009).

Transmission of *Plasmodium falciparum* sexual stages, the gametocytes, to the mosquito mainly depends on their maturity and density in the human host at the time of the mosquito bite. Even if it has been demonstrated that high gametocyte densities do not guarantee high mosquito infection, a greater infection of mosquitoes is generally observed with higher gametocyte densities (Hogh et al., 1998; Drakeley et al., 1999; Targett et al., 2001; Boudin et al., 2004; Paul et al., 2007; Nwakanma et al., 2008). Gametocyte densities vary greatly between human hosts, due to host acquired immunity, genetic factors of the parasite strain and other environmental parameters (blood quality, fever, anemia, anti-malarial drug uptake). In malaria endemic areas, human hosts are typically infected with multiple genotypes of parasites (Day et al., 1992; Babiker et al., 1999; Anderson et al., 2000; Nwakanma et al., 2008) and within-host competition of parasite genotypes is likely to drive transmission success. Indeed, from experiments using *Plasmodium* animal models, it has been shown that different genotypes of parasites in mixed infections have distinct ability to transmit, the more virulent strain having a competitive advantage (de Roode et al., 2005; Bell et al., 2006; Wargo et al., 2007). If different models have been proposed to correlate the gametocyte density to the transmission success of wild isolates of *Plasmodium falciparum* (Pichon et al., 2000; Boudin et al., 2005; Paul et al., 2007), to date no study related the outcome of infection to parasite complexity within the gametocyte population. Understanding relationships between co-infecting genotypes and how they influence the disease transmission is however of great importance as these might help to predict the spread of resistant strains of parasites and guide strategies for malaria control.

In this paper, we investigate how density and genetic diversity of gametocytes impact on infectiousness to mosquitoes. We analyze mosquito infection data consisted of oocyst counts with corresponding gametocyte data : densities and genotypes at 7 microsatellite loci. Data were obtained from experiments of membrane feeding of a local colony of *Anopheles gambiae* mosquitoes on blood from volunteers naturally infected by *Plasmodium falciparum* isolates from Cameroon. Gametocyte genotypes are occurrences of several unordered categorical variables, each having numerous levels. Therefore the number of variables plus attendant interactions is at least of order of the sample size. We considered as response variables : the intensity of infection as measured by the mean of oocyst counts in infected mosquitoes, and the infection prevalence defined by the proportion of mosquitoes that became infected. The high number of variables in our data set will obviously lead to over-fitting of many familiar regression techniques such as general linear model (GLM). In addition, we deal with unordered categorical variables with several levels and potentially accompanying interactions. Therefore, following Segal et al. (2001), we use regression trees techniques.

We address the problem of selecting the most influent variables related to the response variable by applying a variable selection procedure, which comes from Genuer et al. (2010), and is based on variable importance from random forests (Breiman, 2001). The resulting method is completely non-parametric and thus can be used on data with a large number of variables of various types. Moreover, it solves the two following constraints about variable selection : 1) to find all variables highly related to the response variable ; and 2) to find a small number of variables sufficient for a good prediction of the response variable. The selected variables are then assessed in a modeling for oocyst count which takes into account the complexity of the experiment we deal with. The key point of our modeling is the introduction of a new unobserved variable that enables to distinguish two possible sources of non infected mosquitoes. Indeed, the heterogeneity in the quantity and quality of gametocytes in blood-meal (Vaughan, 2007), and natural variation in mosquito susceptibility (Riehle et al., 2006) are well known phenomena. We then suggest here that mosquitoes with no oocyst can be non infected either because they did not ingest enough gametocytes with the blood-meal, or because they were refractory to the ingested parasites. We fitted a model, called Zero-Inflated (ZI) model, which is a two components mixture model combining a point mass at zero with a proper count model. Since we deal with count data, the typical candidate models were Zero-Inflated Poisson (ZIP) and Zero-Inflated Negative Binomial (ZINB) ; ZINB having a slight advantage because it captures over-dispersion which is likely to appear in such data.

The rest of the paper is organized as follows. Section 6.2 presents the data to be analyzed in Subsection 6.2.1, the principle of variable selection based on variable importance from random forests in Subsection 6.2.2, and the modeling of oocyst count in Subsection 6.2.3. Section 6.3 is devoted to the application of these methods on our data. Finally a discussion is given in Section 6.4.

6.2 Material and methods

6.2.1 Data collection and description

The data we considered consist of parasite densities and genotypes at 7 microsatellite loci for gametocyte isolates of *Plasmodium falciparum* on one hand, and oocyst counts 7 days post feeding for each engorged females on the other hand. *Plasmodium falciparum* gametocyte carriers were identified among asymptomatic children aged from 5 to 11 in primary schools of the locality of Mfou, a small town located 30 km apart from Yaounde, the Cameroon capital city. Volunteers were enrolled upon signature of an informed consent form by their parents or legal guardian. The protocol was approved by the National Ethics Committee of Cameroon. Gametocyte densities were expressed as the number of parasites seen against 1 000 leukocytes in a fresh thick blood smear, assuming a standard concentration of 8 000 leukocytes per μl (see Table 6.1 for summary of log-transformed gametocyte densities). Venous blood (2 to 3 mL) was taken from consenting gametocyte carriers, centrifuged and the serum replaced by a non-immune AB serum. This procedure avoids the introduction of human transmission blocking factors in the experiment. 3 to 5 old females of a laboratory strain of *Anopheles gambiae* mosquito were used for the membrane feeding assays placed in cups of approximately 60-80 mosquitoes. Females were allowed to feed for 20 minutes through a Parafilm membrane on glass feeders maintained at $37°C$ and fully engorged females were kept in insectar until dissections 7 days post-infection. Midguts were removed, stained in a 0.4% Mercurochrome solution and the number of developed oocysts counted by light microscopy ($X20$ *lens*). A total of 7 364 mosquitoes (see Table 6.1) were dissected, giving a mean of 39 females per experiment.

Gametocytes were separated from 1 mL of serum free blood using MACS® columns as previously described (Ribaut et al., 2008). DNA extractions from purified gametocytes were performed with DNAzol® and 20 ng of gametocyte DNA were subjected to whole-genome amplification (WGA) using the GenomiPhi V2 DNA Amplification Kit to generate sufficient amounts of DNA for microsatellite genotyping. Genetic polymorphism was assessed at 7 microsatellite loci as previously described (Annan et al., 2007). Their chromosome location and GenBank accession number are as follows : POLYa (chr. 4, G37809), TA60 (chr. 13, G38876), ARA2 (chr. 11, G37848), Pfg377 (chr. 12, G37851), PfPK2 (chr. 12, G37852), TA87 (chr. 6, G38838), and TA109 (chr.6, G38842). Alleles were analyzed using `GeneMapper`® software. Multiple alleles were scored when minor peaks were at least 20% of the height of the predominant allele. The number of observed alleles per locus is 21, 9, 10, 5, 15, 10 and 17 respectively (see Figure 6.1).

TABLE 6.1 – Summary of the numbers of mosquitoes per isolate (N) and log-transformed of gametocyte densities (**log_gameto**).

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
N	11.000	29.000	38.000	39.380	47.000	79.000
log_gameto	1.816	3.156	3.832	3.973	4.612	7.742

Feedings for which the number of dissected mosquitoes was below 20 were not considered. Then 110 experiments were included in the analysis.

6.2.2 Variable selection procedure

The selection procedure we considered is based on variable importances (VI) from random forests (RF). The principle of RF is to aggregate regression or classification trees built on several bootstrap samples drawn from the learning set (more details are given in Appendix 6.A). It is shown to exhibit very good performance for lots of diverse applied situations (Breiman, 2001). Moreover, it computes a variable importance index, defined in Appendix 6.A. Roughly, this index is a measure of the degradation of forest predictions when values of a variable are permuted.

RF variable importance is the key point of the selection procedure (see Genuer et al. (2010) for more backgrounds on RF variable importance). This procedure presents two main benefits. First the method is completely non-parametric and can be applied on data with lots of variables of various types. Second, it achieves two main variable selection objectives : (1) to magnify all the variables related to the response variable, even with high redundancy, for interpretation purpose ; (2) to find a parsimonious set of variables sufficient for prediction of the outcome variable.

Let us now describe the procedure, which comes from Genuer et al. (2010), with the following algorithm. The R package `randomForest` (Liaw and Wiener, 2002; R Development team, 2009) was used in all computations.

To both illustrate and give details about this procedure, we apply it on a simulated dataset with $n = 200$ observations described by 25 continuous variables and 25 binary variables. We assume standard normal distribution $\mathcal{N}(0,\ 1)$ for all continuous variables and binomial distribution $\mathcal{B}(0.5)$ for all binary variables. We consider the following linear model

$$Y = \sum_{j=1}^{25} \beta_{c_j} X_{c_j} + \sum_{j=1}^{25} \beta_{b_j} X_{b_j}$$

in which only 8 over a total of $p = 50$ variables are related to the outcome, the others being just noise. The set of significant variables is composed by the first 4 continuous variables $\left(X_{c_j}\right)_{1\leq j\leq 4}$ and the first 4 binary ones $\left(X_{b_j}\right)_{1\leq j\leq 4}$. Their associated coefficients are given by

$$(\beta_{c_j})_{1\leq j\leq 25} = (\beta_{b_j})_{1\leq j\leq 25} = (4, 4, 2, 2, 0, \ldots, 0).$$

We also assume a 0.9 correlation between X_{c_1} and X_{c_2}, X_{c_3} and X_{c_4}, X_{b_1} and X_{b_2}, and X_{b_3} and X_{b_4}.

The selection process uses a certain number $nfor$ of random forests. In addition of this number, the user has also to provide the number $ntree$ of trees in each random forest, and the number $mtry$ of variables among which to select the best split at each node. The default parameters in the **R** package `randomForest` we used are $mtry = p/3$, $ntree = 500$. In our example, we choose the following parameters : $mtry = p/3$, and

we choose $nfor = 50$ and $ntree = 1000$ to increase the VI stability. The results are summarized in Figure 6.2.

Let us detail the main stages of the procedure together with, in italics, the results obtained on simulated data. In the following, out of bag (OOB) error refers to an estimation of the prediction error (which is defined in Appendix 6.A and is close to a cross-validation estimate).

- **Elimination step**
  First the variables are sorted in descending order according to VI (averaged from the $nfor$ runs).
  The result is drawn on the top left graph. The 8 variables of interest arrive in the first 8 positions of the ranking.
  Keeping this order in mind, the corresponding standard deviations of VI are plotted. A threshold for importance is computed using this graph. More precisely, the threshold is set as the minimum prediction value given by a Classification And Regression Tree (CART) model fitting this curve (for details about CART, see Breiman et al. (1984)). Then only variables with an averaged VI exceeding this level are kept. This rule is, in general, conservative and leads to retain more variables than necessary, in order to make a careful choice later.
  The standard deviations of VI can be found in the top right graph. We can see that true variables standard deviation is large compared to the noisy variables one, which is very close to zero. The threshold leads to retain $p_{elim} = 14$ variables. Note that the threshold value is based on VI standard deviations (top right panel of Figure 6.2) while the effective thresholding is performed on VI mean (top left panel of Figure 6.2).
- **Interpretation step**
  Then, OOB error rates (averaged on $nfor$ runs and using default parameters) of the nested random forests models are computed; starting from the one with only the most important variable, and ending with the one involving all important variables kept previously. The set of variables leading to the smallest OOB error is selected.
  Note that in the bottom left graph the error decreases and reaches its minimum when the first $p_{interp} = 9$ variables are included in the model. This set of selected variables for interpretation contains the 8 true variables plus one noisy one. Note that the associated error is closed to the one of the model with the 6 first variables (see bottom left panel of Figure 6.2) suggesting that a smaller model should be preferred for prediction purposes.
- **Prediction step**
  Finally a sequential variable introduction with testing is performed : a variable is added only if the error gain exceeds a data-driven threshold. The rationale is that the error decrease must be significantly greater than the average variation obtained by adding noisy variables.
  The bottom right graph shows the result of this step, the final model for prediction purpose involves 6 out of the 8 true variables. It is of interest that each of the two true variables non-selected is correlated to one selected variable. The threshold is set to twice the mean of the absolute values of the first order differentiated

OOB errors between the model with $p_{interp} = 9$ variables (the model we selected for interpretation, see the bottom left graph) and the one with all the $p_{elim} = 14$ variables :

$$ave_{jump} = \frac{1}{p_{elim} - p_{interp}} \sum_{j=p_{interp}}^{p_{elim}-1} |\, errOOB(j+1) - errOOB(j) \,|$$

where $errOOB(j)$ is the OOB error of the RF built using the j most important variables.

Since the number of variables after the variable elimination step is small (14), we tried some variants more computationally expensive, in order to validate the two last steps of the algorithm. Instead of the interpretation step, we launch a forward procedure. The principle is, at each time, to seek the best variable (in terms of OOB error rate, averaged on $nfor$ runs and using default parameters) to add in the current variable set. The set of variables leading to the smallest OOB error is then selected.

For our example, it leads, as the interpretation step, to retain the 8 true variables plus one noisy variable (this last noisy variable being different from the one selected by interpretation step). We remark however that the initial ranking according to VI is quite changed with this procedure.

To validate the prediction step, we tried an exhaustive procedure, i.e. we compute the OOB error rate (averaged on $nfor$ runs and using default parameters) for all models formed with the variables selected by the forward procedure. The set of variables leading to the smallest OOB error is then selected.

This procedure selects all 9 variables selected previously.

This validates the interpretation and the prediction step of our algorithm, since the variables sets in these variants are close to ours. In addition the errors reached by the two procedures are comparable. However this comparison was done on the easy simulated dataset we considered in this section.

6.2.3 Modeling oocyst count with Zero-Inflated models

The key point of our modeling is to consider that there are two possible sources of non-infected mosquitoes. First, some mosquitoes may not ingest enough parasites with sufficient sex-ratio to ensure fertilization. The reason is seemingly the high heterogeneity in the number of gametocytes in blood-meals (Pichon et al., 2000). Second, some other mosquitoes may not be genetically susceptible to the parasites ingested (Riehle et al., 2006). We introduce a new variable U materializing this situation of non-infected mosquitoes : for mosquito j fed with blood coming from gametocytes carrier i,

$$U_{i,j} = \begin{cases} 1 & \text{if enough parasites are present in its blood-meal} \\ 0 & \text{otherwise.} \end{cases}$$

$U_{i,j}$ is an unobserved variable in our experiment. We assume that for a given i, $U_{i,1}, \ldots, U_{i,n_i}$ are independent and identically distributed. Here n_i is the number of mosquitoes associated to gametocytes carrier i. For any gametocytes carrier i, denote by

$$\pi_i := P\left(U_{i,j} = 0\right)$$

the probability that mosquito j does not ingest enough gametocytes in its blood-meal. Let $Y_{i,j}$ be the number of oocysts developed in mosquito j associated to gametocytes carrier i. The probability distribution of $Y_{i,j}$ is given by

$$P\left(Y_{i,j} = y_{i,j}\right) = \pi_i \mathbb{1}_{(y_{i,j}=0)} + \left(1 - \pi_i\right) P\left(Y_{i,j} = y_{i,j} | \; U_{i,j} = 1\right), \tag{6.1}$$

where $P\left(Y_{i,j} = y_{i,j} | \; U_{i,j} = 1\right)$ is a suitable count probability distribution.

Consequently, for any gametocytes carrier i, the zero class is a mixture of two components with π_i and $1 - \pi_i$ as the mixture proportions. The resulting model of probability distribution is known as a zero-inflated count model. Such a model is a two components mixture model combining a point mass at zero with a count distribution such as Poisson, geometric or negative binomial (see Zeileis and Jackman (2008) and references therein). Thus there are two sources of zeros : zeros may come from point mass or from count component. In our framework, the zeros coming from the point mass are assumed to represent mosquitoes which did not ingest enough gametocytes to produce an infection.

Let $\lambda_i := \mathbf{E}\left(Y_{i,j} | \; U_{i,j} = 1\right)$ be the conditional mean of the count component. In the regression setting, both the mean λ_i and the excess zero proportion π_i are related to covariates vectors $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p})$ and $\mathbf{z}_i = (z_{i,1}, \ldots, z_{i,q})$, respectively. The components of these covariates are typically the observations of the previously selected variables. They contain gametocyte density and / or their genetic profile. We consider canonical link functions **log** and **logit** for the mean of count component and the point mass component respectively. The corresponding regression equations are

$$\left\{ \begin{array}{rcl} \lambda_i & = & \exp\left(\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_p x_{i,p}\right) \\ \pi_i & = & \dfrac{\exp\left(\gamma_0 + \gamma_1 z_{i,1} + \ldots + \gamma_p z_{i,q}\right)}{1 + \exp\left(\gamma_0 + \gamma_1 z_{i,1} + \ldots + \gamma_p z_{i,q}\right)}, \end{array} \right.$$

where $\beta := (\beta_0, \ldots, \beta_p)$ and $\gamma := (\gamma_0, \ldots, \gamma_q)$ are the parameters to be estimated. Note that different sets of regressors can be specified for the zero inflated component and count component. In the simplest case, only an intercept is used for modeling the unobserved state (zero vs. count).

Typical candidate of zero-inflated models for count data are zero inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) (see Xiang et al. (2007) and references therein). ZINB and ZIP specifications are given in Appendix 6.B. For the estimation of the parameters of these models, we used the package named `pscl` (Zeileis and Jackman, 2008) in **R** statistical software (R Development team, 2009).

6.3 Application on the real data

6.3.1 Variable selection

Here, the results are given following the main stages of the selection procedure given in Subsection 6.2.2. The details are given once, in the case where the response variable is the infection prevalence of mosquitoes measured by proportion of infected mosquitoes. We will just give the selected variables at each stage in the other case where the response variable is the mean number of oocysts per infected mosquitoes. In these results, the binary variables associated to the observed alleles are coded as `locus_allele`. For example, $Pfg377\_093$ is allele 093 at locus $Pfg377$. In addition to the log-transformed of gametocytes density (log_gameto), we also consider the multiplicity of infection (MOI) defined as the maximum number of observed alleles across the considered microsatellite loci.

Here are the main stages of the procedure.

- **Elimination Step**
  - The top left panel in Figure 6.3 gives the **VI** mean of all the 88 variables sorted in decreasing order.
  - The top right panel of Figure 6.3 plots the standard deviations of VI and the fitted CART model. The threshold $\min_{CART}$ represented by the horizontal dashed line leads to retain $p_{elim} = 36$ variables over 88.
- **Interpretation Step**
  This step is illustrated in the bottom left panel of Figure 6.3 in which the minimum **OOB** error rate is reached with $p_{interp} = 11$ variables for interpretation :

$$S_{interp} = \{log\_gameto,\ Pfg377\_093,\ PfPK2\_180,\ MOI,\ Pfg377\_102,\ PfPK2\_183,\ Pfg377\_099,\ TA60\_071,\ PfPK2\_169,\ PfPK2\_166,\ POLYa\_135\}.$$

- **Prediction Step**
  The bottom right panel in Figure 6.3 shows the behavior of the OBB error of the nested models corresponding to the selected variables for prediction :

$$S_{pred} = \{log\_gameto,\ Pfg377\_093,\ MOI,\ PfPK2\_183\}.$$

The 4 selected variables in S_{pred} lead to the OOB error of 0.074. We also launch the variant based on forward and exhaustive search of the selection procedure. Finally it retains a set of 9 variables containing S_{pred}. The associated OOB error is 0.062 which is not far from 0.074. So we prefer a model with variables in S_{pred} which is more parsimonious.

The same procedure was applied when the outcome variable is the infection intensity as measured by the mean number of oocysts in infected mosquitoes. Figure 6.4 gives the behavior of **VI** and the OOB error at each stage of the selection procedure. 25 variables were selected by thresholding the **VI** in the first stage, the 2 most important being log_gameto and MOI. Even if only log_gameto is selected in the interpretation and prediction stages, we also keep MOI. Indeed, as can be seen in the bottom left graph

of Figure 6.4, the model with these two variables is still competitive compared with the model built with log_gameto only.

6.3.2 Zero-Inflated models fitting oocyst count

Zero-Inflated negative binomial (ZINB) and Poisson (ZIP) were fitted to the data in two situations : (i) using only log-transformed of the gametocyte density as covariate, (ii) using the set of variables selected for prediction for both infection prevalence and infection intensity (see Subsection 6.3.1). The estimates of the parameters of ZINB and ZIP models are given in Table 6.2 and 6.3.

In situation (i), it is of interest how the zero counts are captured by the two models : they perfectly predict the observed number of non infected mosquitoes (see the left panel of Figure 6.5). Also, the mean oocyst estimates from both two models are similar (see the right panel of Figure 6.5). But according to the $\mathcal{X}^2$ goodness-of-fit test ($\mathcal{X}^2 = 48.162$, $df = 45$, $p.value \geq 0.3461$ for ZINB against $\mathcal{X}^2 = 2964.606$, $df = 46$, $p.value = 0$ for ZIP model), ZINB model is more adapted to our data. Over-dispersion is probably the main reason : there are more mosquitoes with no or few oocysts than the ones with high oocyst loads. ZIP model underestimates the number of mosquitoes with lower oocyst loads (see the left panel of Figure 6.5). We then consider the ZINB model in the rest of the analysis.

In situation (ii), since the data are over-dispersed, only ZINB is considered. The selected variables in the prediction step of our variable selection process using the infection prevalence as response variable are used in point mass component, and the ones using the infection intensity as response variable are used in the count component. Recall that the infection prevalence is measured by the proportion of mosquitoes that became infected, and the infection intensity by the mean number of oocysts in infected mosquitoes. We found that allele PfPK2_183 is the only variable not significant ($Z = -0.8329$, $p.value \geq 0.40$). In contrary, gametocyte density log_gameto, gametocyte genetic complexity MOI and allele 093 of locus $Pfg377$ significantly influence the oocyst load in mosquitoes. The significance of the gametocyte density confirms the result obtained by ZINB model in situation (i). The significance of the effect of MOI in both zero and count components is very interesting. The significance is more important in the zero component (t-test $Z = -4.5711$, $p.value < 4.9e-06$) than in the count one (t-test $Z = -2.1058$, $p.value < 3.5e-02$) ; also note that the correlation is negative in both two components ($\widehat{\beta}_{MOI} = -0.0333$ and $\widehat{\gamma}_{MOI} = -0.1499$ in count and zero components respectively). So mono infected gametocyte isolates increase the probability that a mosquito do not ingest enough parasites to ensure the transmission success of *Plasmodium* through its vector mosquito. Hence, low values of MOI tend to decrease the infection prevalence. In contrary, a lower genetic diversity of gametocytes in an isolate increases the mean number of oocysts in infected mosquitoes. Also note that the presence of allele 093 of the genetic marker Pfg377 increases the proportion of non-infected mosquitoes ($\widehat{\gamma}_{Pfg377\_093} = 1.2242$, $SE = 0.1204$; t-test $Z = 10.177$, $p-value < 2.7e-24$).

TABLE 6.2 – Maximum likelihood estimates of the parameters of ZINB and ZIP models with data from 110 gametocyte carriers using only log_gameto as the variable. Significant codes : 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.'; 0.1 ' '. $\mathcal{X}^2$ Goodness-of-fit test : $\mathcal{X}^2 = 47.0992$, $df = 45$, $p.value \geq 0.3866$ for ZINB against $\mathcal{X}^2 = 2834.848$, $df = 46$, $p.value = 0$ for ZIP model

		Estimate	Std. Error	z value	Pr(>|z|)	
ZINB						
Count	(Intercept)	-1.3021	0.1163	-11.1985	4.1E-29	***
	log_gameto	0.8402	0.0257	32.6835	2.7E-234	***
	Log(theta)	-0.5693	0.0557	-10.2235	1.6E-24	***
Zero	(Intercept)	0.0029	0.2405	0.0119	9.9E-01	
	log_gameto	-0.2618	0.0531	-4.9294	8.2E-07	***
ZIP						
Count	(Intercept)	-0.7941	0.0199	-40.0016	0.0E+00	***
	log_gameto	0.7717	0.0036	213.9455	0.0E+00	***
zero	(Intercept)	1.4508	0.1284	11.2996	1.3E-29	***
	log_gameto	-0.4383	0.0316	-13.8930	7.0E-44	***

6.4 Discussion

Plasmodium development within its vector mosquito follows complex biological processes and factors controlling parasite dynamics are not well understood. In the rodent malaria parasite *Plasmodium berghei*, it has been previously shown that the efficiency of parasite transmission from one developmental stage to another followed density-dependent models and the best fitted mathematical model differed from one developmental transition to the other one (Sinden et al., 2007). For natural populations of *Plasmodium falciparum*, the human malaria parasite, modeling becomes more challenging because of unknown genetic factors and uncontrolled environmental parameters. Nonetheless, Paul et al. (2007) found a sigmoid relationship between *Plasmodium falciparum* gametocyte density and mosquito transmission and the authors argued that parasite aggregation within mosquitoes represents an adaptive mechanism for transmission efficiency. The great variability in *Plasmodium falciparum* oocyst numbers observed in natural *Anopheles gambiae* populations suggests that parasite transmission is the result of complex interactions between vectors and parasites, which rely on both genetic and environmental factors. Understanding factors that determine transmission intensity and then parasite population structure is of crucial importance in predicting the impact of current malaria control strategies.

In this study, we analyzed patterns of mosquito infection from experiments performed with field isolates of *Plasmodium falciparum* from Cameroon, an area of high malaria endemicity. The infection prevalence for each parasite isolate was scored using two variables, the mean number of oocysts developed within the mosquito midgut 7 days post-infection and the number of mosquitoes carrying at least one oocyst. Gametocyte isolates were

TABLE 6.3 – Maximum likelihood estimates of the parameters of ZINB and ZIP models with the data from 110 gametocytes carriers, using $S_{pred} = \{log\_gameto,\ Pfg377\_093,\ MOI,\ PfPK2\_183\}$ as variables. Significant codes : 0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.'; 0.1 ' '.

		Estimate	Std. Error	z value	Pr(>|z|)	
ZINB						
Count	(Intercept)	-0.9985	0.1436	-6.9539	3.6E-12	***
	log_gameto	0.8009	0.0261	30.6432	3.3E-206	***
	MOI	-0.0333	0.0158	-2.1058	3.5E-02	*
	Log(theta)	-0.5210	0.0500	-10.4296	1.8E-25	***
Zero	(Intercept)	0.9651	0.2679	3.6030	3.1E-04	***
	log_gameto	-0.3769	0.0534	-7.0615	1.6E-12	***
	Pfg377_093	1.2242	0.1204	10.1717	2.7E-24	***
	MOI	-0.1499	0.0328	-4.5711	4.9E-06	***
	PfPK2_183	-4.5225	5.4301	-0.8329	4.0E-01	

genetically characterized at seven microsatellite loci, allowing estimation of the number of co-infecting parasite clones, the MOI, and of the genetic polymorphism, given by the number of alleles at each locus. In such a situation with potentially a high number of unordered categorical variables with numerous levels and accompanying interactions, many familiar statistical techniques such as GLM over-fit the data. Then we had to face the problem of selecting the most important variables related to the outcome variables. We have addressed this issue with a selection procedure based on variable importance from random forests. The procedure has two main benefits. First, it is completely non-parametric and thus can be used on data with lots of variables of various types. Second, it answers the two distinct objectives about variable selection : (1) to find all variables related to the outcome variable and (2) to find a small number of variables sufficient for a good prediction of the outcome variable.

Recall that we are in a critical situation with the number of variables of the order of the sample size ($n = 110$). The application of the variable selection procedure on our data revealed that only 4 among the 88 variables we considered suffice to predict the infectiousness of *Plasmodium falciparum* to *Anopheles gambiae* in our experimental settings. The procedure indicates that the log-transformed of gametocyte density is the most influent variable for both infection prevalence and infection intensity. But whereas gametocyte density was positively correlated with mean oocyst burden, it was negatively correlated with mosquito infection prevalence. This contrasting feature of transmission parameters probably reflects that Plasmodium parasites have developed complex and diverse strategies to ensure their transmission through the mosquito vector. The fact that higher oocyst counts are found for higher gametocyte densities conforms to previous observations showing that infectiousness generally increases with gametocytemia. Interestingly, Paul et al. (2007) described upper gametocyte densities at which mosquito infection rates level off, which is consistent with our results. In their models, mosquitoes

with no oocyst were treated as non infected without further consideration about the putative factors responsible of the non infected status. However, a mosquito population fed on the same gametocyte carrier results in individuals carrying high number of parasites while others do not have any. Failure to infection of a mosquito can result from various factors such the heterogeneity of gametocyte environment (Vaughan, 2007) and natural variation in mosquito susceptibility in the other hand (Riehle et al., 2006). We have described in this article an approach based on that the non-infected mosquitoes represent two distinct populations : one genetically refractory vector population and another population for which the no-oocyst status results from other biological or interacting factors. Further study to quantify the gametocyte uptake in mosquitoes fed on a single carrier would help to determine the individual variation of gametocyte density between blood-meals, and thus the real part of mosquitoes that are refractory and those that did not develop any oocyst because of other environmental factors. Nonetheless, our model perfectly predicts the number of non infected mosquitoes. Our fitting models revealed that over-dispersion of oocysts affects mosquito infection intensity. In addition, a higher over-dispersion of oocysts is observed for mosquitoes fed on blood with high gametocyte density (over 90 gametocytes/μl). The over-dispersed distribution of oocysts has often been explained as the result of the aggregation of gametocytes in the capillary blood at the time of the mosquito bite (Pichon et al., 2000). In this study, mosquitoes were membrane fed and membrane feeding is thought to suppress gametocyte over dispersion (Vaughan, 2007). Nonetheless, the fact that the maximum aggregation is found for high gametocyte densities is indicative of aggregation of sexual stages ; aggregation may occur within the mosquito midgut after parasite intake and genetic factors from the parasites may play a role in parasite recognition. This speculation is consistent with the hypothesis of adaptive aggregation, where gamete aggregation would favor fertilization and then increase infection intensity (Paul et al., 2007; Pichon et al., 2000).However, this increased oocyst burden coincided with a lower infection prevalence, possibly indicating that other factors operate in limiting mating (see below).

In malaria endemic areas, intensive use of treatments for malaria has led to the emergence of drug-resistant parasites. Despite their low efficacy, malaria therapies such as chloroquine (CQ) and sulphadoxine-pyrimethamine (SP) are still widely used in sub Saharan Africa. It has been shown that, upon treatment, drug-resistant parasites have a selective advantage, leading to higher transmission by the vector (Hallett et al., 2004, 2006). Our samples originated from an area with high drug pressure and volunteers carrying single parasite genotype may have received an early anti malarial treatment that cured them from drug-sensitive genotypes, thus allowing an optimal growth and transmission of a resistant genotype. However, children who received a malaria treatment in the one month period preceding the gametocyte carriage detection were not included in the study and genotyping of pfcrt-K76T mutation in a subset of our gametocyte samples identified single infections both as CQ resistant or sensitive parasite strains. This result indicates that other factors contribute to the better transmission capacity of the mono-infected *Plasmodium falciparum* isolates.

We found that the Multiplicity Of Infection is negatively correlated to the response variable in both zero and count components. This indicates that the genetic complexity

of gametocyte populations modulates the mosquito infection outcomes in an opposite manner : while gametocyte isolates containing a single clone of *Plasmodium falciparum* resulted in a higher mean number of oocysts in infected mosquitoes, gametocyte isolates with multiple genotypes gave rise to a higher infection prevalence. These results may suggest that malaria parasites use kin discrimination to adapt strategies allowing optimal parasite transmission.

Our results showed that the genetic complexity of gametocyte isolates affects the mosquito infection intensity. Mosquito infections with isolates of lower complexity resulted in higher oocyst counts. This may reflects a higher virulence of genotypes in these infections, where the gametocyte genotypes in the mono-infected isolates could have suppressed their competitors in a prior step of the infection, within the human host. Nonetheless, the lower infection prevalence in mono clonal infections indicates that the higher number of oocysts arises at the cost of a reduced ability to infect the mosquito vector population. This could result from blood quality/quantity such as agglutinating antibodies or anaemia. It was shown that mixed infections resulted in increased anaemia, a possible adaptive response for sex ratio adjustment (Taylor and Read, 1998; Paul et al., 2004). Sex allocation theory predicts that sex ratio becomes less female-biased as clone number increases (Read et al., 1992; Paul et al., 2002; Reece et al., 2008; Schall, 2009). Then, if parasite aggregation is an adaptive trait to promote gamete fertilization, by contrast the highly female biased sex ratio in mono infected isolates will affect infection prevalence because male availability will constitute a limiting factor for mating.

Our results may have important implications for the genetic structuring of *Plasmodium falciparum* populations. For *Plasmodium falciparum*, fertilization of gametes can occur between genetically-identical gametes (inbreeding) or between different gametes (outbreeding). Levels of inbreeding differ from one malaria area to another but they roughly correlate with the disease endemicity (Anderson et al., 2000). In areas of high malaria endemicity, inbreeding levels are generally more reduced, mostly because parasite genetic diversity is high and multiple infections predominant. However, population genetics studies, after genotyping of oocysts from wild mosquitoes collected in intense malaria transmission areas, gave rise to conflicting results and the extent of inbreeding in natural settings remains controversial (Razakandrainibe et al., 2005; Annan et al., 2007; Mzilahowa et al., 2007). The higher fitness of inbred parasites, as suggested in this study and others (Hastings and Wedgwood-Oppenheim, 1997; Razakandrainibe et al., 2005), could explain the departs from panmixia frequently found in areas of high malaria transmission.

Finally, our results comfort the idea that malaria parasites are able to discriminate the genetic complexity of their infections and to adjust accordingly adaptive traits implicated in transmission (aggregation, sex ratio). Deciphering specific processes involved in parasite recognition and competition within the mosquito vector would help for our understanding of within host behaviour of malaria parasites. This may have important implications for future malaria interventions strategies.

FIGURE 6.1 – Alleles detected for the 7 microsatellite loci and their frequencies in *Plasmodium falciparum* gametocyte carriers.

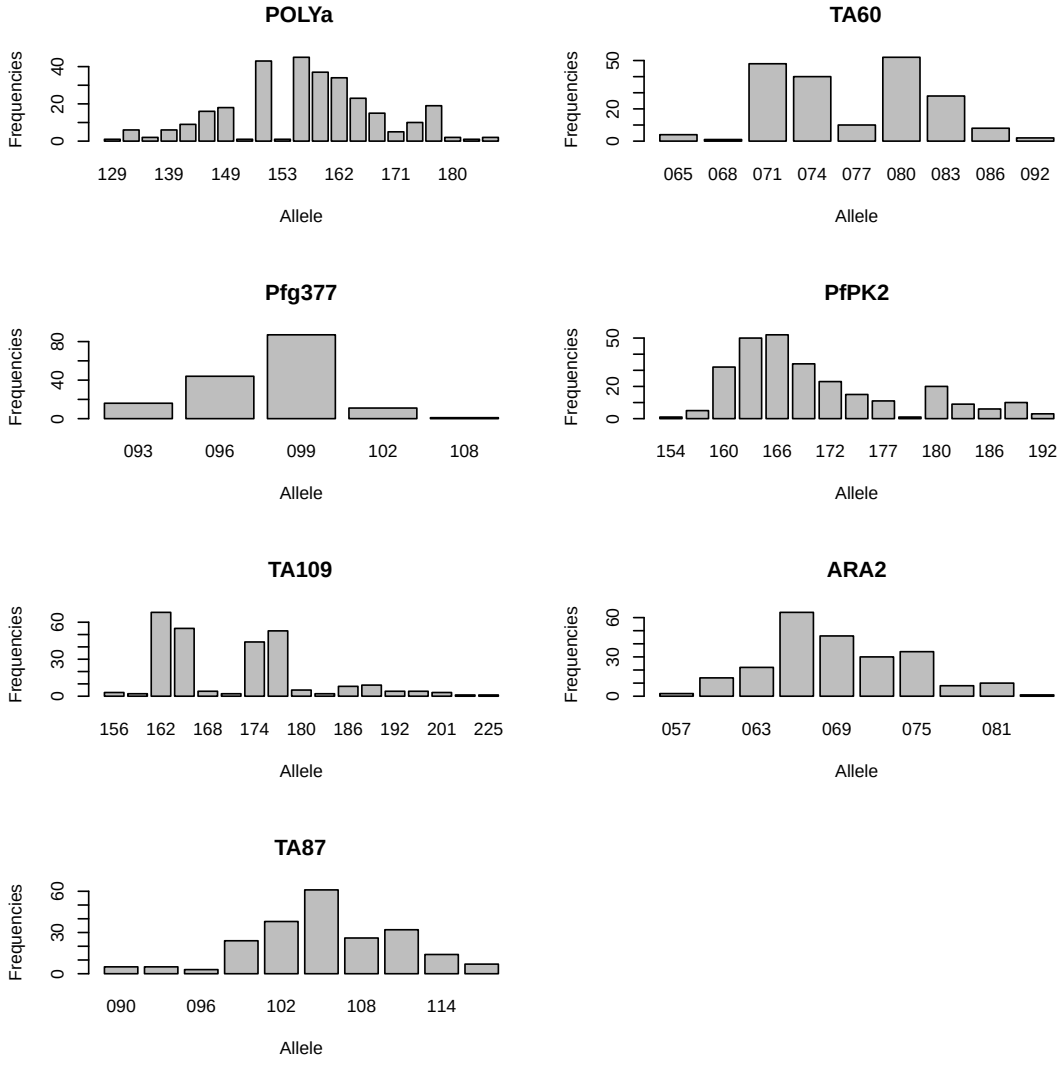


FIGURE 6.2 – Variable selection procedures for interpretation and prediction for simulated data

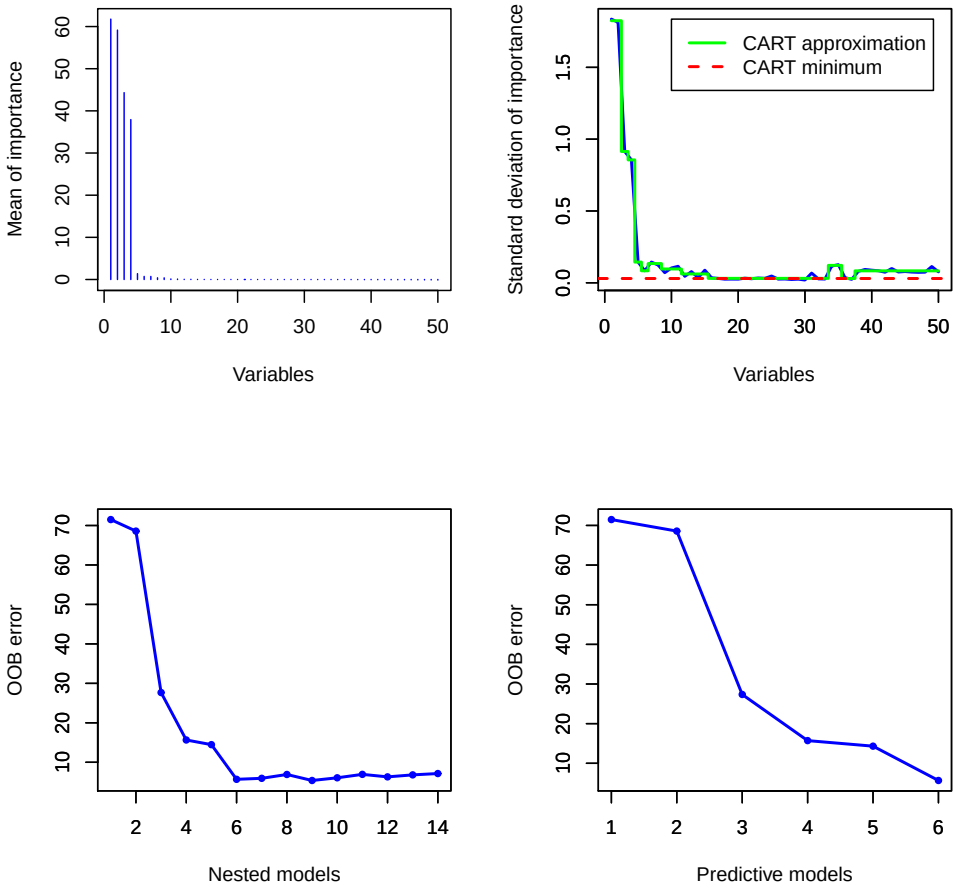


FIGURE 6.3 – Variable selection for interpretation and prediction. The response variable is the infection prevalence measured by the proportion of infected mosquitoes.

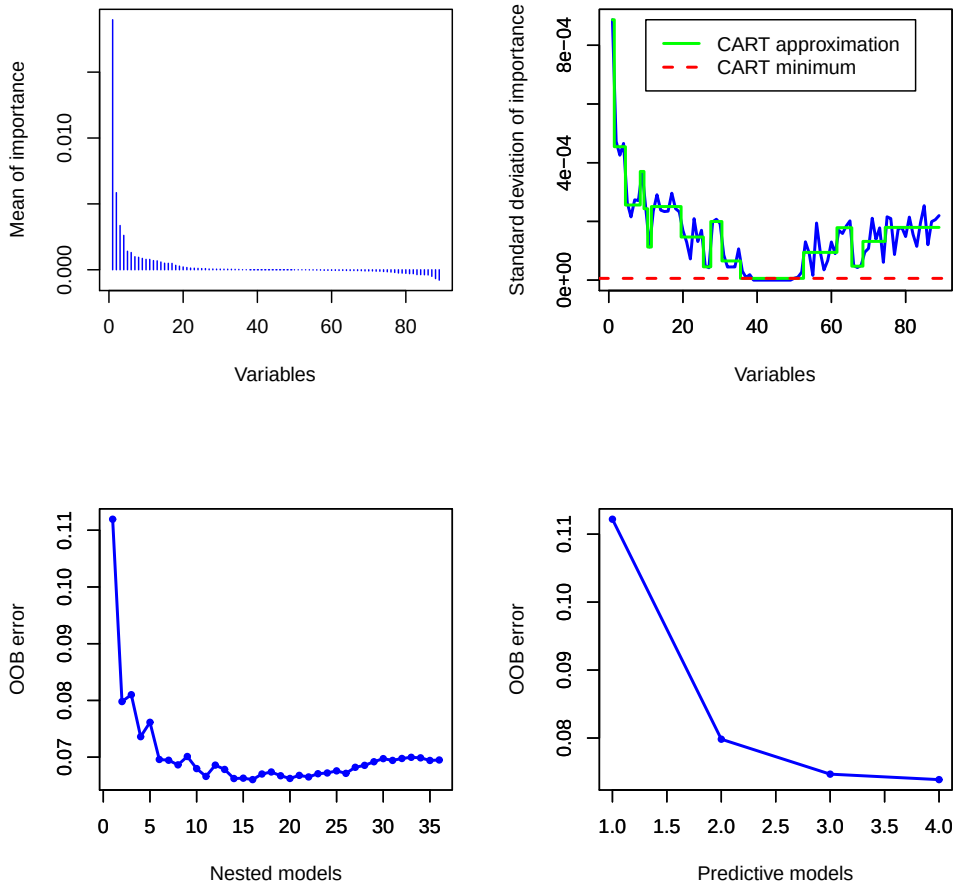


FIGURE 6.4 – Variable selection for interpretation and prediction. The response variable is the mean number in infected mosquitoes.

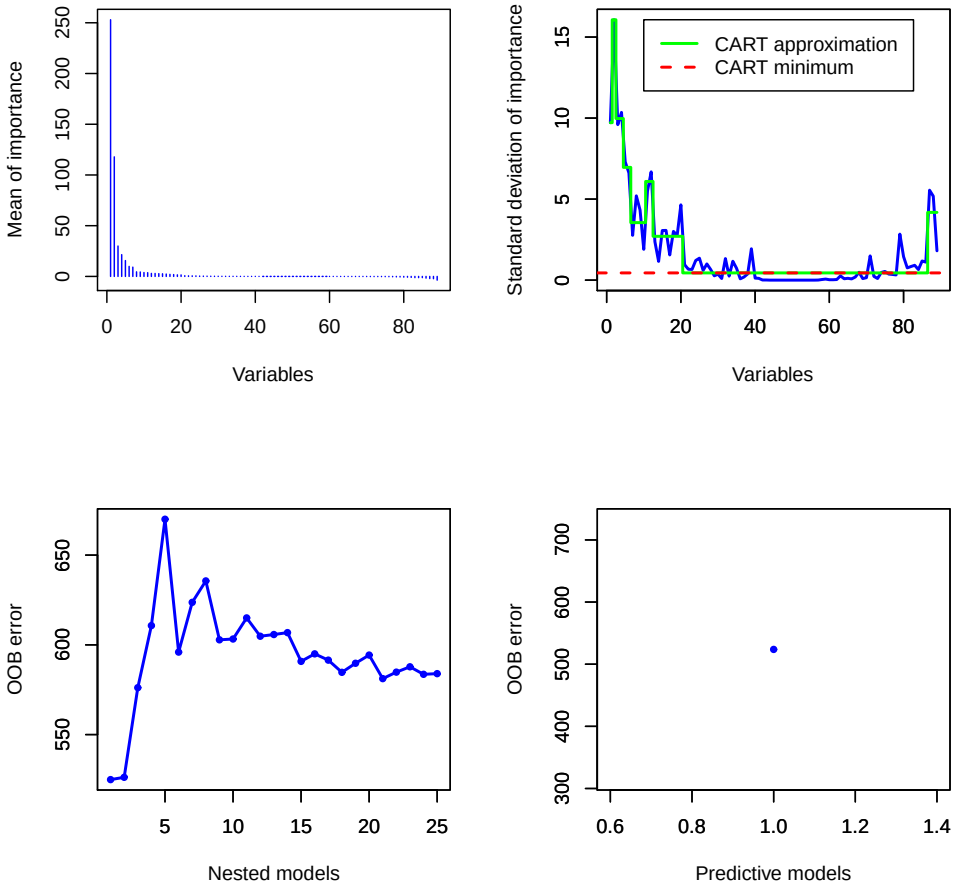
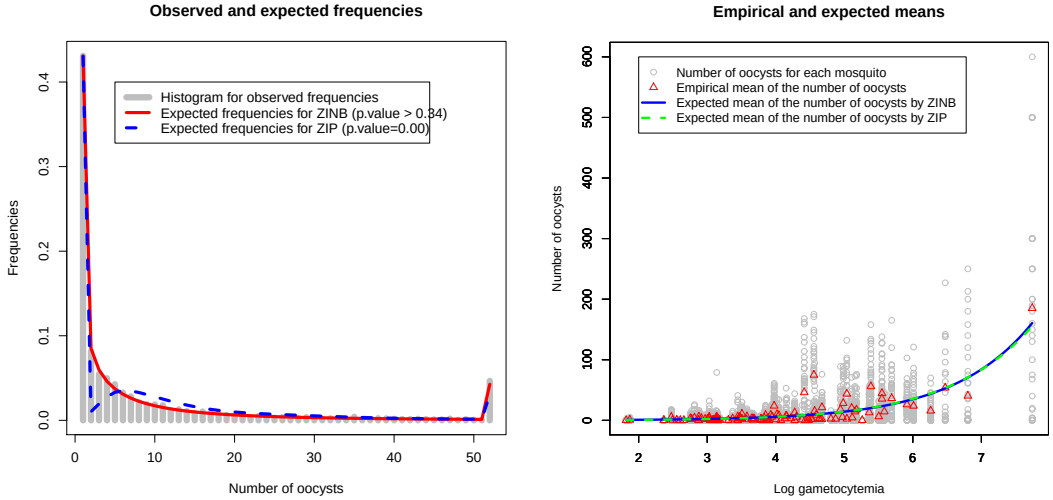


FIGURE 6.5 – The right panel gives observed and predicted frequencies from ZINB and ZIP models, and the right one the empirical and predicted mean number of oocysts versus log-gametocytemia.



Appendices

6.A Random Forests

RF estimator

The principle of random forests is to aggregate a given number *ntree* of binary decision trees built on several bootstrap samples drawn from the learning set. The bootstrap samples are obtained by uniformly drawing n samples among the learning set with repetition. The decision trees are fully developed binary trees and the split rule is the following.

First, the whole dataset (also called the root of the tree) is split into two subsets of data (called two children nodes). To do that, one randomly chooses a given number *mtry* of variables, and computes all the splits only for the previously selected variables. A split is of the form $\{X^i \leq s\} \cup \{X^i > s\}$, which means that data with the i-th variable value less than the threshold s go to the left child node and the others to the right one. Finally the selected split is the one minimizing the variance children nodes.

Then, one restraints to one child node, randomly chooses another set of *mtry* variables and calculates the best split. And so on, until each node is a terminal node, i.e. it comprises less than 5 observations.

A new data item X, starting in the root of the tree, goes down the tree following the splits and falls in a terminal node. Then the tree predicts for X, $\bar{Y}$ the mean of response of data in this terminal node. To finally get the RF predictor, one aggregates all the tree predictors by averaging their predictions.

RF error estimate : the OOB error

Inside the variable selection procedure, we use an estimation of the prediction error directly computed by the RF algorithm. This is the Out Of Bag (OOB) error and is calculated as follows. Fix one data in the learning sample, and consider all the bootstrap samples which do not contain this data (i.e. for which the data is "out of bag"). Now perform an aggregation only among trees built on these bootstrap samples. After doing this for all data, compare to the true response and get an estimation of the prediction error (which is a kind of cross-validated error estimate).

RF variable importance

Let us now detail the computation of the RF variable importance for the first variable X^1. For each tree, one has a bootstrap sample associated with an OOB sample. Predict the OOB data with the tree predictor. Now, randomly permute the values of the first variable of the OOB observations, predict these modified OOB data with the tree predictor. The variable importance of X^1 is defined as the mean increase of prediction errors after permutation. The more the error increases, the more important the variable is (note that it can be slightly negative, typically for irrelevant variables).

6.B $\mathcal{ZIP}$ and $\mathcal{ZINB}$ specifications

These two models are defined by equation (6.1) with the count model given by :

– $\mathcal{ZIP}$:

$$\left\{\begin{array}{rcl} P\left(Y_{i,j}=y_{i,j}|\ U_{i,j}=1\right) & = & \exp\left(-\lambda_i\right)\frac{\lambda_i^{y_{i,j}}}{y_{i,j}!} \\ \lambda_i & := & \mathbf{E}\left(Y_{i,j}|\ U_{i,j}=1\right) \\ & = & \mathbf{var}(Y_{i,j}|\ U_{i,j}=1) \end{array}\right.$$

– $\mathcal{ZINB}$:

$$\left\{\begin{array}{rcl} P\left(Y_{i,j}=y_{i,j}|\ U_{i,j}=1\right) & = & \frac{\Gamma\left(y_{i,j}+\theta\right)}{\Gamma\left(\theta\right)\cdot y_{i,j}!}\frac{\lambda_i^{y_{i,j}}\cdot\theta^{\theta}}{\left(\lambda_i+\theta\right)^{y_{i,j}+\theta}} \\ \lambda_i & := & \mathbf{E}\left(Y_{i,j}|\ U_{i,j}=1\right) \\ \mathbf{var}(Y_{i,j}|\ U_{i,j}=1) & = & \lambda_i+\frac{1}{\theta}\lambda_i^2 \end{array}\right.$$

where $\Gamma\left(t\right)=\int_0^{\infty}x^{t-1}\mathbf{e}^{-x}\mathbf{d}x$, and θ is the over-dispersion parameter. The expectation and the variance of $Y_{i,j}$ are given by :

$$\begin{array}{rcl} \mu\left(x\right) & := & \mathbf{E}\left(Y_{i,j}|\right)=\Big(1-\pi_i\Big)\lambda_i \\ \mathbf{var}(Y_{i,j}) & = & \left\{\begin{array}{ll} \Big(1-\pi_i\Big)\Big(\lambda_i+\pi_i\lambda_i^2\Big) & ZIP \\ \Big(1-\pi_i\Big)\Big(\lambda_i+\left(\frac{1}{\theta}+\pi_i\right)\lambda_i^2\Big) & ZINB. \end{array}\right. \end{array}$$

Chapitre 7

Conclusion et perspectives

La sélection de variable suscite un intérêt croissant aussi bien en régression qu'en classification supervisée et non supervisée, comme en témoigne une littérature abondante à ce sujet (voir Massart (2007) et les références citées). En effet, on s'intéresse de plus en plus aux données décrites par un nombre de variables sans cesse croissant au regard de la taille des échantillons. En principe plus on dispose d'informations sur chaque individu de l'échantillon d'intérêt, meilleures devraient être les performances des méthodes statistiques. Certaines variables peuvent cependant ajouter du bruit à ces méthodes. Il devient alors important de chercher le sous-ensemble de variables pertinent pour répondre à la question posée. Dans cette thèse, nous avons considéré deux problèmes tous liés à la sélection de variable. Contrairement aux méthodes implicites de sélection qui procèdent par pondération des variables, les méthodes de sélection de variable que nous avons choisies donnent explicitement les variables sélectionnées, permettant ainsi une meilleure interprétation des résultats.

Classification non supervisée

Le premier problème de sélection de variable auquel nous nous sommes intéressés concerne la classification non supervisée par mélange fini dans un contexte de données discrètes nominales. Les modèles de mélange fini constituent un cadre intuitif et rigoureux pour la classification non supervisée, particulièrement en ce qui concerne les données génétiques multilocus. D'abord, l'idée que chaque classe recherchée est caractérisée par un jeu de paramètres et les équilibres de Hardy-Weinberg et de liaison est compatible avec la notion de population vue comme unité de reproduction ou comme ensemble d'individus partageant la même structure génétique. La quasi-totalité des méthodes de classification non supervisée sur les données génétiques multilocus sont fondées sur cette idée là (Pritchard et al., 2000; Guillot et al., 2005; Corander et al., 2008; Alexander et al., 2009). L'autre avantage des modèles de mélange fini est qu'ils permettent d'évaluer de façon rigoureuse le nombre de composants de la population et le rôle des variables dans le processus de classification. Pour les mélanges gaussiens, citons par exemple les travaux

de Keribin (2000) sur la consistance du critère BIC, et ceux de Baudry (2009) sur les performances des critères analogues à ICL (Integrated Completed Likelihood). Le résultat de consistance que nous avons obtenu au Chapitre 3 est nouveau, en tous cas en ce qui concerne le problème double de sélection de variable et de classification sur variables qualitatives nominales.

Par ailleurs, un critère pénalisé non-asymptotique et une inégalité oracle associée sont proposés dans le Chapitre 4. L'approche non asymptotique est très récente en ce qui concerne la sélection de modèle dans le cadre des mélanges finis comme en témoignent les premiers travaux en la matière dûs à MAUGIS (2009) dans un contexte gaussien. Notre critère non-asymptotique et celui de MAUGIS sont construits sur la base des mêmes outils, à savoir le théorème général de sélection de modèle pour l'estimation de densité dû à Massart (2007). L'application de ce théorème dans notre cas spécifique a nécessité le contrôle des entropies à crochets des modèles de mélange fini de lois multinomiales dont les paramètres sont obtenus à partir du modèle de Hardy-Weinberg. L'autre particularité de notre critère non asymptotique est d'être aussi consistant.

Sur le plan pratique, nous proposons un logiciel autonome nommé `MixMoGenD` (Mixture Model for Gentypic Data) qui implémente les procédures décrites dans les Chapitres 2, 3 et 4. Ce logiciel est implémenté dans une approche orienté objet avec une allocation dynamique de la mémoire qui fait en sorte que seules la capacité mémoire et la puissance de calcul de l'ordinateur de l'utilisateur soient les seules limites à la taille des données. La procédure de sélection est basée sur les critères du maximum de vraisemblance pénalisé BIC, AIC et un critère dont la fonction de pénalité est sous la forme $\mathbf{pen}(m) = \lambda \cdot d_m$ où d_m est la dimension du modèle m et λ un nombre données-dépendant à calibrer. Cette forme de pénalité n'est pas nouvelle. Elle a été suggérée pour des problèmes divers. C'est par exemple le cas pour le problème de détection de rupture traité par Lebarbier (2002), et pour celui de la classification non supervisée par MAUGIS (2009) dans le cadre Gaussien. Dans notre cas précis, cette forme de la pénalité est en quelque sorte justifiée par le résultat théorique obtenu dans le Chapitre 4. La calibration de λ est réalisée dans une procédure automatique basée sur l'"heuristique de la pente" proposée par Birgé and Massart (2007). Nous avons montré empiriquement que la procédure de sélection qui en résulte est globalement meilleure que BIC et AIC par rapport à la taille de l'échantillon, permettant ainsi de répondre en partie à la question "quel critère pour quelle taille de l'échantillon ?"

Concernant la sélection de modèle de mélange fini pour résoudre un problème de classification, il n'y a pas encore de consensus clair sur le modèle objectif et ce que les gens s'attendent à être une classe. Par exemple, en utilisant les critères BIC et ICL, Baudry (2009) montre sur des simulations qu'il est difficile de choisir entre un bon ajustement (qui est nécessaire pour avoir un cadre rigoureux d'étude) et un nombre pertinent de classes. Néanmoins, les résultats obtenus des données simulées montrent dans notre contexte précis de données génétiques multilocus que la sélection de modèle pour l'estimation de densité permet de bien estimer le nombre de populations et le sous-ensemble pertinent pour la classification, même dans des situations où le niveau de différentiation génétique est dans un intervalle réputé critique pour la classification non supervisée.

Les modèles considérés dans cette thèse ne prennent pas en compte les situations complexes où ce n'est pas le même sous-ensemble de variables qui discrimine toutes les populations entre-elles. Il est vrai que dans certains cas, comme l'exemple de données réelles traité dans le Chapitre 4, on peut surmonter la difficulté en faisant plusieurs analyses. Il est toutefois envisageable de concevoir des modèles prenant ces situations en comptent, c'est l'objet d'un travail en cours en collaboration avec Dominique Bontemps. L'autre question qui reste en suspens est liée à la dimension des modèles candidats. Les écarts de dimension sont très grands d'un modèle à l'autre, et on se retrouve très vite avec des dimensions plus grandes que la taille de l'échantillon. Nous souhaitons continuer ce travail dans le sens de la réduction de la dimension des modèles candidats.

Transmission de *Plasmodium* à travers le moustique

Le deuxième problème auquel nous nous sommes intéressés dans cette thèse concerne la transmission de *Plasmodium* à travers son vecteur moustique. Comprendre les interactions génétiques entre ce parasite et son vecteur peut ouvrir la voie à de nouvelles stratégies de lutte contre sa transmission. Notre travail dans cette partie a consisté à proposer des outils statistiques pour analyser les données d'expériences réalisées à cet effet. Le premier problème que posent ces données est le nombre élevé de covariables (diverses) au regard de la taille de l'échantillon. Nous avons considéré une procédure de sélection de variable basée sur le calcul de l'importance des variables des forêts aléatoires. La procédure de sélection résultante est non paramétrique et convient aux cas où les covariables sont diverses (quantitatives ou qualitatives). Cette procédure de sélection a permis de mettre en évidence des facteurs intervenant dans le succès du cycle sporogonique de *Plasmodium*. Le résultat le plus pertinent est le rôle de la diversité génétique du parasite sur sa transmission. Cette diversité jouerait un rôle clé dans le phénomène d'agrégation du parasite aussi bien chez l'hôte vertébré que chez le moustique. La mise en évidence de ce résultat a été rendu possible grâce à une modélisation prenant en compte les deux sources possibles de moustiques non infectés à l'issue de l'expérience. En effet, dans le repas de sang, la densité parasitaire ne serait pas uniforme, diminuant ainsi la probabilité qu'un moustique ingère suffisamment de parasites lors de son repas de sang. Par ailleurs, ce même phénomène d'agrégation augmenterait les chances du succès du cycle sporogonique en favorisant la fertilisation.

Des études antérieures ont montré que le système immunitaire du moustique est capable de réagir à l'infection au *Plasmodium*. On imagine donc que le succès du cycle sporogonique dépend aussi de facteurs multiples liés au moustique. Nous envisageons de proposer des modèles statistiques pour la compatibilité génétique parasite / vecteur en général.

Bibliographie

Akaike, H. (1973). 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.

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, 16(1) :3–14.

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, 19(9) :1655–64.

Allman, E. S., Matias, C., and Rhodes, J. A. (2008). Identifiability of latent class models with many observed variables. arxiv.

Anderson, T. J., Haubold, B., Williams, J. T., Estrada-Franco, J. G., Richardson, L., Mollinedo, R., Bockarie, M., Mokili, J., Mharakurwa, S., French, N., Whitworth, J., Velez, I. D., Brockman, A. H., Nosten, F., Ferreira, M. U., and Day, K. P. (2000). Microsatellite markers reveal a spectrum of population structures in the malaria parasite plasmodium falciparum. *Mol Biol Evol*, 17(10) :1467–82.

Annan, Z., Durand, P., Ayala, F. J., Arnathau, C., Awono-Ambene, P., Simard, F., Razakandrainibe, F. G., Koella, J. C., Fontenille, D., and Renaud, F. (2007). 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.

Arlot, S. and Massart, P. (2008). Slope heuristics for heteroscedastic regression on a random disign. *Submited*.

Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10 :245–279.

Azais, J.-M., Gassiat, E., and Mercadier, C. (2009). The likelihood ratio test for general mixture models with possibly structural parameter. *ESAIM P&S, to appear*.

Babiker, H. A., Ranford-Cartwright, L. C., and Walliker, D. (1999). Genetic structure and dynamics of plasmodium falciparum infections in the kilombero region of tanzania. *Trans R Soc Trop Med Hyg*, 93 Suppl 1 :11–4.

Bai, Z., Rao, C., and Wu, Y. (1999). Model selection with data-oriented penalty. *Journal of Statistical Planning and Inference*, 77(1) :103–118.

Banfield, J. and Raftery, A. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3) :803–821.

Baudry, J.-P. (2009). *Sélection de Modèle pour la Classification Non Supervisée. Choix du Nombre de Classes.* PhD thesis, UNIVERSITÉ PARIS-SUD XI.

Bell, A. S., de Roode, J. C., Sim, D., and Read, A. F. (2006). Within-host competition in genetically diverse malaria infections : parasite virulence and competitive success. *Evolution*, 60 :1358–71.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7) :719–725.

Biernacki, C., Celeux, G., and Govaert, G. (2001). Strategies for getting highest likehood in mixture models. Technical Report 4255, INRIA.

Birgé, L. and Massart, P. (2007). Minimal penalties for gaussian model selection. *Probability Theory and Related Fields*, 138(1-2) :33–73.

Boudin, C., Diop, A., Gaye, A., Gadiaga, L., Gouagna, C., Safeukui, I., and Bonnet, S. (2005). Plasmodium falciparum transmission blocking immunity in three areas with perennial or seasonal endemicity and different levels of transmission. *Am J Trop Med Hyg*, 73(6) :1090–5.

Boudin, C., Van Der Kolk, M., Tchuinkam, T., Gouagna, C., Bonnet, S., Safeukui, I., Mulder, B., Meunier, J. Y., and Verhave, J. P. (2004). Plasmodium falciparum transmission blocking immunity under conditions of low and high endemicity in cameroon. *Parasite Immunol*, 26 :105–10.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1) :5.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees.* Statistics/Probability Series. Wadsworth Publishing Company, Belmont, California, U.S.A.

Brusco, M. and Cradit, J. (2001). A variable-selection heuristic for K-means clustering. *Psychometrika*, 66(2) :249–270.

Burnham, K. and Anderson, D. (2002). *Model Selection and Multimodel Inference : A Practical Information-theoretic Approach.* Springer.

Burnham, K. P. (2004). Multimodel inference : Understanding aic and bic in model selection. *Sociological Methods & Research*, 33(2) :261.

Celeux, G. and Diebolt, J. (1991). The EM and SEM algorithms for mixtures : Statistical ans numerical aspects. *Cahiers du CERO*, 32 :135–151.

Celeux, G. and Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational statistics and data analysis*, 14(3) :315–332.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5) :781–793.

Chambaz, A., Garivier, A., and Gassiat, E. (2008). A MDL approach to HMM with Poisson and Gaussian emissions. Application to order identification. *To appear JSPI*.

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

Corander, J., Marttinen, P., Sirén, J., and Tang, J. (2008). Enhanced bayesian modelling in baps software for learning genetic structures of populations. *BMC Bioinformatics*, 9 :539.

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

Dash, M., Choi, K., Scheuermann, P., and Liu, H. (2002). Feature selection for clustering - a filter solution. In *ICDM*, pages 115–122. IEEE Computer Society.

Dawkins, R. (1999). *The extended phenotype : The long reach of the gene*. Oxford University Press, USA.

Dawson, K. J. and Belkhir, K. (2001). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet Res*, 78(1) :59–77.

Day, K. P., Koella, J. C., Nee, S., Gupta, S., and Read, A. F. (1992). Population genetics and dynamics of plasmodium falciparum : an ecological view. *Parasitology*, 104 Suppl :S35–52.

de Roode, J. C., Helinski, M. E., Anwar, M. A., and Read, A. F. (2005). Dynamics of multiple infection and within-host competition in genetically diverse malaria infections. *Am Nat*, 166 :531–42.

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

Development, C. T. R. (2009). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Dimopoulos, G., Richman, A., Müller, H., and Kafatos, F. (1997). Molecular immune responses of the mosquito Anopheles gambiae to bacteria and malaria parasites. *Proceedings of the National Academy of Sciences of the United States of America*, 94(21) :11508.

Drakeley, C. J., Secka, I., Correa, S., Greenwood, B. M., and Targett, G. A. (1999). Host haematological factors influencing the transmission of plasmodium falciparum gametocytes to anopheles gambiae s.s. mosquitoes. *Trop Med Int Health*, 4 :131–8.

Everitt, B. and Hand, D. (1981). Finite mixture distributions.

Fowlkes, E., Gnanadesikan, R., and Kettenring, J. (1988). Variable selection in clustering. *Journal of classification*, 5(2) :205–228.

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

Friedman, J. and Meulman, J. (2004). Clustering objects on subsets of attributes. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 66(4) :815–849.

Gassiat, E. (2002). 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.

Genoveve, C. R. and Wasserman, L. (2000). Rates of convergence for the gaussian mixture sieve. *Ann. Statist.*, 28(4) :1105–1127.

Genuer, R., Poggi, J. M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*. To appear. doi :10.1016/j.patrec.2010.03.014.

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

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3 :1157–1182.

Hallett, R. L., Dunyo, S., Ord, R., Jawara, M., Pinder, M., Randall, A., Alloueche, A., Walraven, G., Targett, G. A., Alexander, N., and Sutherland, C. J. (2006). Chloroquine/sulphadoxine-pyrimethamine for gambian children with malaria : transmission to mosquitoes of multidrug-resistant plasmodium falciparum. *PLoS Clin Trials*, 1 :e15.

Hallett, R. L., Sutherland, C. J., Alexander, N., Ord, R., Jawara, M., Drakeley, C. J., Pinder, M., Walraven, G., Targett, G. A., and Alloueche, A. (2004). Combination therapy counteracts the enhanced transmission of drug-resistant malaria parasites to mosquitoes. *Antimicrob Agents Chemother*, 48 :3940–3.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning : data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2) :83–85.

Hastie, T., Tibshirani, R., and Walther, G. (2001). Estimating the number of clusters in a dataset via the gap statistic. *Journal of the Royal Statistical Society, B*, 63 :411–423.

Hastings, I. M. and Wedgwood-Oppenheim, B. (1997). Sex, strains and virulence. *Parasitol Today*, 13 :375–83.

Hogh, B., Gamage-Mendis, A., Butcher, G. A., Thompson, R., Begtrup, K., Mendis, C., Enosse, S. M., Dgedge, M., Barreto, J., Eling, W., and Sinden, R. E. (1998). The differing impact of chloroquine and pyrimethamine/sulfadoxine upon the infectivity of malaria species to the mosquito vector. *Am J Trop Med Hyg*, 58 :176–82.

Jaramillo-Gutierrez, G., Rodrigues, J., Ndikuyeze, G., Povelones, M., Molina-Cruz, A., and Barillas-Mury, C. (2009). Mosquito immune responses and compatibility between plasmodium parasites and anopheline mosquitoes. *BMC Microbiol*, 9 :154.

Jouve, P.-E. and Nicoloyannis, N. (2005). A filter feature selection method for clustering. In Hacid, M.-S., Murray, N. V., Ras, Z. W., and Tsumoto, S., editors, *ISMIS*, volume 3488 of *Lecture Notes in Computer Science*, pages 583–593. Springer.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā Ser. A*, 62(1) :49–66.

Kohavi, R. and John, G. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2) :273–324.

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

Lebarbier, É. (2002). *Quelques approches pour la détection de rupture à horizon fini*. PhD thesis, Univ. Paris-Sud 11, F-91405 Orsay.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R news*, 2(3) :18–22.

Marcotorchino, F. and Michaud, P. (1982). Agregation de similarites en classification automatique. *Revue de Statistique Appliquée*, 30(2) :21–44.

Massart, P. (2007). *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

MAUGIS, C. (2009). *Sélection de variables pour la classiiňĄcation e non supervisée par mélanges gaussiens. Application l'étude de données transcriptomes.* PhD thesis, UNIVERSITE PARIS-SUD 11.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009). Variable selection for clustering with gaussian mixture models. *Biometrics*.

Maugis, C. and Michel, B. (2008). Slope heuristics for variable selection and clustering via Gaussian mixtures. *Research Report*, 6550.

Maugis, C. and Michel, B. (2009). A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM : P&S. To appear*.

McLachlan, G. and Basford, K. (1988). Mixture models. Inference and applications to clustering.

McLachlan, G., Bean, R., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18(3) :413.

McLachlan, G. and Krishnan, T. (1997). *The EM algorithm and extensions.* Wiley New York.

McLachlan, G. and Peel, D. (2000). *Finite mixture models.* Wiley-Interscience.

Miller, L. H. and Greenwood, B. (2002). Malaria–a shadow over Africa. *Science*, 298(5591) :121–2.

Mzilahowa, T., McCall, P. J., and Hastings, I. M. (2007). population structure and genetics of the malaria agent p. falciparum. *PLoS One*, 2 :e613.

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, 8(4) :343–366.

Nwakanma, D., Kheir, A., Sowa, M., Dunyo, S., Jawara, M., Pinder, M., Milligan, P., Walliker, D., and Babiker, H. A. (2008). High gametocyte complexity and mosquito infectivity of plasmodium falciparum in the gambia. *Int J Parasitol*, 38(2) :219–27.

Osta, M. A., Christophides, G. K., and Kafatos, F. C. (2004). Effects of mosquito genes on Plasmodium development. *Science*, 303(5666) :2030–2.

Paul, R. E., Lafond, T., Muller-Graf, C. D., Nithiuthai, S., Brey, P. T., and Koella, J. C. (2004). Experimental evaluation of the relationship between lethal or non-lethal virulence and transmission success in malaria parasite infections. *BMC Evol Biol*, 4 :30.

Paul, R. E. L., Bonnet, S., Boudin, C., Tchuinkam, T., and Robert, V. (2007). Aggregation in malaria parasites places limits on mosquito infection rates. *Infect Genet Evol*, 7(5) :577–86.

Paul, R. E. L., Brey, P. T., and Robert, V. (2002). Plasmodium sex determination and transmission to mosquitoes. *Trends in Parasitology*, 18 :32–38.

Pearson, K. (1894). Contributions to the theory of mathematical evolution. *Phil. Trans. Roy. Soc. London A*, 186 :71–110.

Pichon, G., Awono-Ambene, H. P., and Robert, V. (2000). High heterogeneity in the number of plasmodium falciparum gametocytes in the bloodmeal of mosquitoes fed on the same host. *Parasitology*, 121 (Pt 2) :115–20.

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

Raftery, A. (1995). Bayesian model selection in social research (with discussion by Andrew Gelman & Donald B. Rubin, and Robert M. Hauser, and a rejoinder). *Sociological Methodology*, pages 111–196.

Raftery, A. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473) :168–178.

Raymond, M. and Rousset, F. (1995). Genepop version 1.2 : population genetics software for exact tests and ecumenicism. *J. Hered.*, 86 :248–249.

Razakandrainibe, F. G., Durand, P., Koella, J. C., Meeüs, T. D., Rousset, F., Ayala, F. J., and Renaud, F. (2005). "clonal" population structure of the malaria agent Plasmodium falciparum in high-infection regions. *Proc Natl Acad Sci U S A*, 102(48) :17388–93.

Read, A. F., Narara, A., Nee, S., Keymer, A. E., and Day, K. P. (1992). Gametocyte sex ratios as indirect measures of outcrossing rates in malaria. *Parasitology*, 104 (Pt 3) :387–95.

Reece, S. E., Drew, D. R., and Gardner, A. (2008). Sex ratio adjustment and kin discrimination in malaria parasites. *Nature*, 453 :609–14.

Ribaut, C., Berry, A., Chevalley, S., Reybier, K., Morlais, I., Parzy, D., Nepveu, F., Benoit-Vical, F., and Valentin, A. (2008). Concentration and purification by magnetic separation of the erythrocytic stages of all human plasmodium species. *Malar J*, 7 :45.

Riehle, M. M., Markianos, K., Niaré, O., Xu, J., Li, J., Touré, A. M., Podiougou, B., Oduol, F., Diawara, S., Diallo, M., Coulibaly, B., Ouatara, A., Kruglyak, L., Traoré, S. F., and Vernick, K. D. (2006). Natural malaria infection in anopheles gambiae is regulated by a single genomic control region. *Science*, 312(5773) :577–9.

Rosenberg, N. A., Woolf, E., Pritchard, J. K., Schaap, T., Gefel, D., Shpirer, I., Lavi, U., Bonne-Tamir, B., Hillel, J., and Feldman, M. W. (2001). Distinctive genetic signatures in the libyan jews. *Proc Natl Acad Sci U S A*, 98(3) :858–63.

Schall, J. J. (2009). Do malaria parasites follow the algebra of sex ratio theory ? *Trends Parasitol*, 25 :120–3.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Segal, M. R., Cummings, M. P., and Hubbard, A. E. (2001). Relating amino acid sequence to phenotype : analysis of peptide-binding data. *Biometrics*, 57(2) :632–42.

Sinden, R. E., Dawes, E. J., Alavi, Y., Waldock, J., Finney, O., Mendoza, J., Butcher, G. A., Andrews, L., Hill, A. V., Gilbert, S. C., and Basanez, M. G. (2007). Progression of plasmodium berghei through anopheles stephensi is density-dependent. *PLoS Pathog*, 3 :e195.

Targett, G., Drakeley, C., Jawara, M., von Seidlein, L., Coleman, R., Deen, J., Pinder, M., Doherty, T., Sutherland, C., Walraven, G., and Milligan, P. (2001). Artesunate reduces but does not prevent posttreatment transmission of plasmodium falciparum to anopheles gambiae. *J Infect Dis*, 183 :1254–9.

Taylor, L. H. and Read, A. F. (1998). Determinants of transmission success of individual clones from mixed-clone infections of the rodent malaria parasite, plasmodium chabaudi. *Int J Parasitol*, 28 :719–25.

Titterington, D., Smith, A., and Makov, U. (1985). *Statistical analysis of finite mixture distributions*. John Wiley & Sons.

Toussile, W. and Bontemps, D. (2010). A new penalized criterion for variable selection and clustering using genotypic data. *Submitted, arxiv*.

Toussile, W. and Gassiat, E. (2009). Variable selection in model-based clustering using multilocus genotype data. *Advances in Data Analysis and Classification*, 3(2) :109–134.

Vaart., A. W. V. D. (1998). *Asymptotic Statistic*. Cambridge Series in Statistical and Probabilistic Mathematics.

Vaughan, J. A. (2007). Population dynamics of Plasmodium sporogony. *Trends Parasitol*, 23(2) :63–70.

Verzelen, N. (2009). *Adaptative estimation to regular Gaussian Markov random fields*. PhD thesis, Université Paris-Sud 11.

Villers, F. (2007). *Tests et selection de modèles pour l'analyse de données protéomiques et transcriptomiques*. PhD thesis, Université Paris-Sud 11.

Wang, Y. and Liu, Q. (2006). Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of stock–recruitment relationships. *Fisheries Research*, 77(2) :220–225.

Wargo, A. R., de Roode, J. C., Huijben, S., Drew, D. R., and Read, A. F. (2007). Transmission stage investment of malaria parasites in response to in-host competition. *Proc Biol Sci*, 274 :2629–38.

Xiang, L., Lee, A., Yau, K., and McLachlan, G. (2007). A score test for overdispersion in zero-inflated poisson mixed regression model. *Statistics in medicine*, 26(7) :1608–1622.

Zeileis, A. and Jackman, C. K. S. (2008). Regression Models for Count Data in **R**. *Journal of Statistical Software*, 27.